

Metrology and Pharmaceutical Calculations

Roger L Schnaare, PhD
Shelly J Prince, PhD



One of the first technical operations that the student of pharmacy must learn is the manipulation of balances, weights, and measures of volume. This entails a study of the various systems of weights and measures, their relationships, and a mastery of the mathematics involved. This chapter considers the fundamental principles of metrology underlying the testing, manufacturing, and compounding of pharmaceutical preparations:

Weights and Measures—An accumulation of facts concerning the various systems, with tables of conversion factors and practical equivalents. The relationships among the various systems of weights and measures are clarified.

Weighing and Measuring—A discussion of the various types of balances, particularly prescription balances and methods of using, testing, and protecting them; also of various devices and methods for measuring large or small volumes of fluids.

Density and Specific Gravity—A consideration of the mass/volume ratio of a substance (density), and the ratio of the weight (mass) of one substance to the weight (mass) of another substance taken as the standard (specific gravity).

Pharmaceutical Calculations—A review of basic mathematical principles and their use in solving pharmaceutical problems.

WEIGHTS AND MEASURES

Weight is a measure of the gravitational force acting on a body; weight is directly proportional to the body's mass. The latter, being a constant based on inertia, never varies, whereas weight varies slightly with latitude, altitude, temperature, and pressure. The effect of these factors usually is not considered unless very precise weighing and large quantities are involved.

Measure is the determination of the volume or extent of a body. Temperature and pressure have a pronounced effect, especially on gases or liquids. These factors, therefore, are considered when making precise measurements.

All standard weights and measures in the US are derived from or based on the United States National Prototype Standards of the Meter and the Kilogram. The standards are made of platinum-iridium, and are in the custody of the National Institute of Standards and Technology (NIST) in Washington, DC.

History

A brief outline of the origin of the many systems of weights and measures may help clarify the essential distinctions between them. The sense of the weight of a body cannot be conveyed intelligibly to the mind unless a means of comparison is chosen. As weight is the measure of the gravitational force of a body, this force is expressed in terms of standards of resistance, which exactly balance the body and keep it in equilibrium when used with a mechanical device constructed for this specific purpose. Such standards are termed *weights* and the mechanical devices are called *balances* or *scales*.

The standards that have been chosen by various nations are arbitrary, and instances are common where different standards are in use at the same time in the same country. Many of the ancient standards clearly are referable to variable parts of the human body, such as nail, foot, span, pace, cubit (length of

the forearm), and fathom or faethm (stretch of the arms). In the history of metrology three periods may be traced:

1. The *Ancient* period, during which the old classical standards originated, terminated with the decline of the Roman Empire. The unit of distance used by all nations for maritime measurements, the *nautical* or *meridian* mile (1/60 of a degree of the earth's equatorial circumference) is exactly equal to 1000 Egyptian fathoms or 4000 Egyptian cubits. These Egyptian measurements, which have persisted for more than 4000 years, were based on astronomical or meridian measurements that were recorded imperishably in the great Pyramid at Ghizeh, whose perimeter is exactly 500 of these fathoms, or 1/2 nautical mile.
2. The *Medieval* period extended to the 16th century. During this period the old standards were lost, but their names were preserved, and European nations adopted various independent standards.
3. The *Modern* period extends from the 16th century to the present. Since the 17th century, the efforts of most enlightened nations have been directed toward scientific accuracy and simplicity, and during the present century toward international uniformity.

Historical metrology, also referred to as *documentary metrology*, is concerned with the study of monuments and records of ancient periods. *Inductive metrology* is concerned with the accumulation of data concerning the measurement of large numbers of objects that have been referred to as standards but which have no exact measure except by statutory regulation.

THE ENGLISH SYSTEMS—In Great Britain, in 1266, the 51st Act of the reign of Henry III declared

“that by the consent of the whole realm of England the measure of the King was made—that is to say, that an English silver penny called the sterling, round and without clipping, shall weigh *thirty-two grains of wheat*, well dried and gathered out of the middle of the ear; and twenty pence (pennyweights) do make an ounce and twelve ounces a pound, and eight pounds do make a gallon of wine, and eight wine gallons do make a bushel, which is the eighth of a quarter.”

The 16-ounce pound (*avoirdupois pound*), undoubtedly of Roman origin, was introduced at the time of the first civilization of the British island. However, according to Gray, the word *haberdepois* was first used in English laws in 1303. A statute of Edward I (1304 AD) states “that every *pound* of money or of *medicines* is of *twenty shillings weight*, but the *pound* of all other things is *twenty-five shillings weight*. The *ounce of medicines* consists of *twenty pence*, and the *pound* contains *twelve ounces* [the Troy Pound], but in other things the *pound* contains *fifteen ounces*, in both cases the *ounce* weighing *twenty pence*.”

These laws unfold the theory of the ancient weights and measures of Great Britain, and reveal the standards (ie, a natural object, grains of wheat). A difference existed then between the Troy and the avoirdupois pound, but the weights now in use are $\frac{1}{16}$ heavier than those of Edward I, due to the change subsequently made in the value of the coin by the sovereign. In addition, the true pennyweight standard was lost, and, in the next revision of the weights and measures, the present troy and avoirdupois standards were adopted.

The *troy weight* is of still earlier origin. The great fairs of the 8th and 9th centuries were held at several French cities, including Troyes, the gathering place of traders from all countries. Coins frequently were mutilated, so they were sold by weight, and the standard weight of Troyes for selling coin was adopted for precious metals and medicines in all parts of Europe. The troy ounce and the avoirdupois ounce originally were intended to have the same weight, but after the revision it was found that the avoirdupois ounce was lighter by $42\frac{1}{2}$ gr (grains) than the troy ounce. The subsequent adoption of troy weight by the London College of Physicians in 1618, on the recommendation of Sir Theodore Turquet de la Mayerne who compiled their first pharmacopoeia, has entailed upon all apothecaries who are governed by British customs to this day the very great inconvenience of buying and selling medicines by one system of weights (the *avoirdupois*) and compounding them by another (the *apothecary* or *troy*).

In the next century efforts were made toward reforming the standards, and in 1736 the Royal Society began the work that ended in the preparation, by Mr. Bird under the direction of the House of Commons, of the standard *yard* and standard *pound* troy in 1760. Copies of these were prepared and no intentional deviation has been made since.

The growing popularity of the French metric system—and the desirability of securing a standard that could be recovered easily in case of loss or destruction, and that should be commensurable with a simple unit—prompted steps in England to secure these advantages in 1816. The labors of English scientists led to the adoption of the *imperial* measures and standards, which were legalized January 1, 1826; imperial standards, are now in use in Great Britain, thus introducing another element of confusion into an already complicated subject. In this system the *yard* is equivalent to 36 inches, and its length was determined by comparison with a pendulum beating seconds of mean time, in a vacuum, at the temperature of 62°F at the level of the sea in the latitude of London, a length that was found to be 39.1393 inches. The *pound troy* (containing 5760 gr) was determined by comparison with a given measure of distilled water under specified conditions. Thus, a cubic inch of distilled water was weighed with brass weights in air at 62°F, the barometer at 30 inches, and it weighed 252.458 gr. The standard for measures of capacity in Great Britain (either dry or liquid) is the *imperial gallon*, which contains 10 lb avoirdupois of distilled water weighed in air at 62°F, the barometer standing at 30 inches. The *bushel* contains eight such gallons.

Washington, in his first annual message to Congress, January 1790, recommended the establishment of uniformity in currency, weights, and measures. Action was taken with reference to the currency and recommendations were made by Jefferson, then the Secretary of State, for the adoption of either the currently used English systems or a decimal system. However,

nothing was accomplished until 1819 to 1820, when efforts again were made in the US to secure uniformity in the standards that were in use by the several states. Finally, after a lengthy investigation, on June 14, 1836, the Secretary of the Treasury was directed by Congress to furnish each state in the Union with a complete set of the revised standards, and thus the *troy pound* (5760 gr), the *avoirdupois pound* (7000 gr), and the *yard* (36 inches) are all identical with the British standards. However, the US *gallon* is quite different; the old wine gallon of 231 inch³—containing 58,372.2 gr of distilled water at its maximum density, weighed in air at 62°F, the barometer standing at 30 inches—was retained. The bushel contained 77.274 lb of water under the same conditions, thus making the dry quart about 16% greater in volume than the liquid quart.

In 1864 the use of the metric measures was legalized in Great Britain, but was not made compulsory, and in 1866 the US followed the same course. By the US law of July 28, 1866, all lengths, areas, and cubic measures are derived from the international meter equivalent to 39.37 inches. Since 1893 the US Office of Standard Weights and Measures has been authorized to derive the yard from the meter: one yard equals 3600/3937 m, and the customary weights are referred to the kilogram by an Executive order approved April 5, 1893. Capacities were to be based on the equivalent; dm³ equals one liter, the decimeter being equal to 3.937 inches. The gallon still remains at 231 inch³ and the bushel contains 2150.42 inch³. This makes the liquid quart equal to 0.946 liter and the dry quart equal to 1.1013 liter, whereas the imperial quart is 1.1359 liter. The customary weights are derived from the international kilogram, based on the value that one avoirdupois lb equals 453.5924277 g and that 5760/7000 avoirdupois lb equals one troy lb.

Avoirdupois weight is used in general in the US for commercial purposes, including the buying and selling of drugs on a large scale and occasionally on prescription orders.

THE METRIC SYSTEM—The idea of adopting a scientific standard for the basis of metrology that could be reverified accurately was suggested by a number of individuals after the Renaissance. Jean Picard, the 17th-century French astronomer, proposed to take as a unit the length of a pendulum beating one sec of time at sea level, at latitude 45°.

In 1783, the English inventor James Watt first suggested the application of decimal notation, and the commensurability of weight, length, and volume. The French National Assembly in 1790 appointed a committee to decide the preferability of the pendulum standard or a terrestrial measure of some kind as a basis for the new system. The committee reported in 1791 in favor of the latter, and commissions were appointed to measure an arc of meridian and to perfect the details of the commensurability of the units and of nomenclature. However, certain inaccuracies were inherent in the early standards, so they do not bear to each other the intended exact relationships. The present accepted standards are defined in publications of the National Institute of Standards and Technology (NIST).

In its original conception, the meter was the fundamental unit of the metric system, and all units of length and capacity were to be derived directly from the meter, which was intended to be equal to one ten-millionth of the earth's quadrant. Furthermore, it originally was planned that the unit of mass, the kilogram, should be identical with the mass of a cubic decimeter of water at its maximum density. At present, however, the units of length and mass are defined independently of these conceptions.

For all practical purposes, calibration of length standards in industry and scientific laboratories is accomplished by comparison with the material standard of length: the distance between two engraved lines on a platinum-iridium bar, *The International Prototype Meter*, which is kept at the International Bureau of Weights and Measures.

The *kilogram* is defined independently as the mass of a definite platinum-iridium standard, the *International Prototype Kilogram*, which also is kept at the International Bureau of Weights and Measures. The *liter* is defined as the volume of a

kilogram of water, at standard atmospheric pressure, and at the temperature of its maximum density, approximately 4°C. The *meter* is thus the fundamental unit on which are based all metric standards and measurements of length and area and of volumes derived from linear measurements.

Of basic scientific interest is that on October 14, 1960, the 11th General Conference on Weights and Measures, meeting in Paris, adopted a new international definition for the standard of length: the meter is now defined as the length equal to 1,650,763.73 wavelengths of the orange-red light of the krypton-86 isotope. This standard will be used in actual measurements only when extreme accuracy is needed.

The kilogram is the fundamental unit on which are based all metric standards of mass. The liter is a secondary or derived unit of capacity or volume. The liter is larger by about 27 ppm (parts per million) than the cube of the tenth of the meter (the cubic decimeter): one liter = 1.000027 dm³.

The conversion tables in this publication that involve the relative length of the yard and meter are based upon the relation: one m = 39.37 inch, contained in the act of Congress of 1866. From this relation it follows that one inch = 25.40005 mm (nearly).

In recent years engineering and industrial interests the world over have urged the adoption of the simpler relation, one inch = 25.4 mm exactly, which differs from the preceding value by only five ppm. This simpler relation has not as yet been adopted officially by either Great Britain or the US but is in wide industrial use.

In the US, the abbreviation *cc* (for cubic centimeter) still persists in general use and is taken as synonymous for the more correct milliliter. The US Pharmacopeia (USP) IX and National Formulary (NF) IV adopted the term *milliliter* with its abbreviated form *mil*, but it proved so unpopular in practice that the following pharmacopeial convention directed the return to the older term cubic centimeter (*cc*). However, in 1955, USP XV and NF X once again adopted the term milliliter with the abbreviation mL.

National jealousies and the natural antipathy to changing established customs interfered greatly with the adoption of the metric system during the early part of the 19th century. At present the metric system is in use in every major country of the world. In the US and Great Britain it is legalized for reference to and definition of other standards, and it is in exclusive use by nearly all scientists and by increasing segments of industry and the public. In the US the metric system was legalized in 1866, but not made compulsory; in the same year the international prototype meter and kilogram were adopted as fundamental standards. The US silver coinage was based upon the metric system, the half dollar being exactly 12.5 g and the quarter and the dime being of the proportionate weights.

As corporations became more international, the need for a universal standard increased. Since 1875 there has been established and maintained an International Bureau of Weights and Measures, with headquarters at Paris. This Bureau is managed by an international committee that enjoys universal representation. One objective of the committee is to make and provide prototypes of the meter and kilogram for the subscribing nations; approximately 40 such copies have been prepared.

The US prototype standards of both the meter and the kilogram mass, constructed of a platinum-iridium alloy, were brought from Paris in 1890 and are now in the custody of the NIST in Washington, DC. They have been reproduced and distributed by our own government to the various states having bureaus needing such replicas. The original US prototype meter was taken back to Paris in 1957 for reverification and was found to have altered only 3 parts in 100,000,000 after 67 years of use. Thus, there was no demonstrable change within the limits of experimental error.

Adoption of the krypton-86 wavelength of light definition for the meter gives the different countries the means to check their prototype meter bars without returning them to Paris at periodic intervals for comparison with the international meter bar.

Orthography and Reading

ORTHOGRAPHY—There are two methods of orthography of the metric units in use. In the original French, the units are spelled *metre*, *litre*, and *gramme*; in the method proposed by the American Metric Bureau, the units are spelled *meter*, *liter*, and *gram*. For three decades after the original adoption of the metric system, the USP and NF adopted *meter* and *liter*, but used the French *gramme*. Now these official compendia use the spelling *gram*.

READING—Some difficulty usually is experienced by those unfamiliar with the metric system in reading the quantities. In the linear measures in pharmacy, centimeters and millimeters are used almost exclusively; thus, 0.05 m would not be read five hundredths of a meter, but rather five centimeters (5 cm); if the millimeter column contains a unit, as in 0.055 m, it is read 55 millimeters (55 mm) in preference to fifty-five thousandths of a meter.

Fractions of a millimeter must be read decimally, as 0.0555 m, fifty-five and five-tenths millimeter (55.5 mm). In measures of capacity, cubic centimeters (cc) or milliliters (mL) are used exclusively for quantities of less than a liter. The terms half-liter, quarter-liter, 100 milliliters, and one milliliter are denoted by 500 mL, 250 mL, 100 mL, and one mL.

In weight, when the quantity is relatively large and in commercial transactions, the *kilogram* is abbreviated to *kilo*. When less than a *kilogram* and not less than a *gram*, the quantity is read with the gram for the unit. Thus, 2000 g would be read either as 2000 grams or as 2 kilos, and 543 g would be read 543 grams; 2543 g is sometimes read 2 kilos and 543 grams, although 2543 grams usually is preferred.

For quantities below the *gram*, decigram and centigram usually are not used, but rather *milligram* has been regarded as the most convenient unit. With the increase in the use of extremely small doses of very potent drugs and the wide application of more delicate analytical procedures, the term *microgram* (mcg, µg, or γ), for thousandths of a milligram, is used frequently to designate quantities up to 999 µg (less than 1.000 mg).

Both the metric and English systems of weights and measures are in use in the US. Even though the metric system nearly has replaced the English system, the pharmacist must have a practical knowledge of both.

WEIGHTS

The Metric System

The USP of 1890 adopted the metric system of weights and measures to the exclusion of all others except for equivalent dosage statements, and the British Pharmacopoeia of 1914 did likewise. In 1944 the Council on Pharmacy and Chemistry of the American Medical Association adopted the metric system exclusively. The advantages of the metric or decimal system, and its simplicity, brevity, and adaptability to everyday needs are now conceded universally.

FRACTIONAL AND MULTIPLE PREFIXES—In many experimental procedures, including some in the pharmaceutical sciences, very small (and occasionally very large) quantities of weight, length, volume, time, or radioactivity are measured. To avoid the use of numbers with many zeros in such cases, the NIST recognizes prefixes to be used to express fractions or multiples of the International System of Units (SI), which was established in 1960 by the General Conference on Weights and Measures (see the foregoing discussion). The recognized prefixes, which in use are adjoined to an appropriate unit (as, for example, in such quantities as nanogram, picomole, microcurie, microsecond, or megavolt) are defined in Table 11-1.

Table 11-2 lists some metric weights. The prefixes, which indicate multiples, are of Greek derivation: deka, 10; hecto, 100; kilo, 1000. Fractions of the units are expressed by Latin prefixes: deci, 1/10; centi, 1/100; milli, 1/1000.

Table 11-1. Prefixes for Fractions and Multiples of SI Units

FRACTION	PREFIX	SYMBOL	MULTIPLE	PREFIX	SYMBOL
10^{-1}	deci	d	10	deka	da
10^{-2}	centi	c	10^2	hecto	h
10^{-3}	milli	m	10^3	kilo	k
10^{-6}	micro	μ	10^6	mega	M
10^{-9}	nano	n	10^9	giga	G
10^{-12}	pica	p	10^{12}	tera	T
10^{-15}	femto	f	10^{15}	peta	P
10^{-18}	atto	a	10^{18}	exa	E

Only a few of the most convenient denominations are employed in practical work. Whole numbers from one to 1000 usually are expressed in terms of grams, while the kilogram is used as the unit for larger quantities. Quantities between one milligram and one gram usually are referred to in terms of milligrams; microgram (μg or mcg) is used in quantitative analysis, biological studies, and for minute dosage statements.

The English Systems

In the US, both the avoirdupois and apothecary systems of weight measurement sometimes are used in handling medicines. It must be emphasized *that pharmacists may buy their drugs by avoirdupois weight*. These two systems differ:

1 pound avoirdupois = 7000 gr and is abbreviated lb.
 1 pound apothecary = 5760 gr and is abbreviated ℔.
 1 ounce avoirdupois = 437.5 gr and is abbreviated oz.
 1 ounce apothecary = 480 gr and is abbreviated ʒ.

The *grain* avoirdupois is exactly the same as the *grain* apothecary. The apothecary pound is therefore 1240 gr *lighter* than the avoirdupois pound, and the apothecary ounce is therefore 42.5 gr *heavier* than the avoirdupois ounce.

The abbreviations of the denominations of apothecary weight are represented by the signs ounce, ʒ; dram, ʒ; scruple, ʒ; and grain, gr. These long have been in use but possibly may be mistaken for one another in rapid or careless writing. The abbreviations or signs of avoirdupois weight differ from those of apothecary weight, and care should be used not to confound them; they are lb (sometimes written #), pound: oz, ounce: gr, grain. Tables 11-3, 11-4, and 11-5 show three English systems of weight.

Jewelers evaluate precious stones with troy weight, which is very similar to apothecary weight. The apothecary and troy grain, ounce, and pound are identical, but the ounces are subdivided differently. The *carat*, used by jewelers, is equal to 3.168 troy grains or four carat grains. When used to express the fineness of gold, one carat signifies 1/24 part. A 14-carat ring is 14/24 pure gold.

As indicated in the footnote to Table 11-6, a number of special metric system units are used in various pharmacopeial and nonofficial descriptions, tests, and assays of drugs and other substances to express linear measurements of very small dimension. These units and their symbols or abbreviations are listed in Table 11-7, together with their equivalents in terms of the other metric units and the inch.

Table 11-2. Metric Weight

1 microgram	μg	=	0.000001	g
1 milligram	mg	=	0.001	g
1 centigram	cg	=	0.01	g
1 decigram	dg	=	0.1	g
1 gram	g	=	1	g
1 dekagram	dag	=	10	g
1 hectogram	hg	=	100	g
1 kilogram	kg	=	1000	g

Note: The abbreviation μg or mcg is used for microgram in pharmacy, rather than gamma (γ) as in biology.

Table 11-3. Avoirdupois Weight

POUNDS	OUNCES	GRAINS
1 =	16 =	7000
	1 =	437.5

Note: 2000 lb = 1 ton, and 2240 lb = 1 long ton.

Table 11-4. Apothecary Weight

POUNDS	OUNCES	DRAMS	SCRUPLES	GRAINS
1 =	12 =	96 =	288 =	5760
	1 =	8 =	24 =	480
		1 =	3 =	60
			1 =	20

Table 11-5. Troy Weight

POUNDS	OUNCES	PENNYWEIGHTS	GRAINS
1 =	12 =	240 =	5760
	1 =	20 =	480
		1 =	24

Table 11-6. Metric Linear Measure

1 nanometer	(nm)	=	0.000000001 m (0.001 μm : 10^{-9} m: 10 \AA)
1 micrometer	(μm)	=	0.000001 m (0.001 mm: 10^{-6} m: 10,000 \AA)
1 millimeter	(mm)	=	0.001 m
1 centimeter	(cm)	=	0.01 m
1 decimeter	(dm)	=	0.1 m
1 meter	(m)	=	1 m
1 dekameter	(dam)	=	10 m
1 hectometer	(hm)	=	100 m
1 kilometer	(km)	=	1000 m

Although the meter (m) is observed to be the initial unit, it is seldom necessary to use it in pharmaceutical practice, and the same holds true for a number of the above measures. The micrometer (μm), millimeter (mm), and centimeter (cm) are employed in the description of many official drugs. Measurements pertaining to spectrometric and colorimetric tests and assays of many official drugs are recorded in micrometers (μm) or reciprocal centimeters (cm^{-1}) for infrared and in nanometers (nm) for ultraviolet and visible wavelengths of light, respectively.

Table 11-7. Equivalent Linear Measurements

UNIT	INCHES	MM	μM	NM	\AA
1 inch	1	25.4	25,400	2.54×10^7	2.54×10^8
1 mm (millimeter)	0.0394	1	1000	10^6	10^7
1 μm (micrometer)	3.94×10^{-5}	10^{-3}	1	1000	10,000
1 nm (nanometer)	3.94×10^{-8}	10^{-6}	10^{-3}	1	10
1 \AA (angstrom unit)	3.94×10^{-9}	10^{-7}	10^{-4}	0.1	1

MEASURES

Systems

Two systems of linear measure are used in the US: English and metric. Two systems of liquid measure are used: apothecary (also called the wine measure or US liquid measure) and metric. The units of the English system of linear measure (inch, foot, yard, and mile) are well-known, and needn't be described here. The units of the metric systems of linear and liquid measure, and of the apothecary (wine, US liquid) system of liquid measure, with their respective equivalents, are given in Tables 11-7, 11-8, and 11-9.

Pharmacists who fill Canadian or British prescriptions should also be familiar with the substantially different British imperial liquid measure system; the units, with their equivalents, are given in Table 11-10.

The following facts concerning the US system of liquid measure (see Table 11-9) should be noted:

1. The apothecary fluidounce (f℥) of distilled water weighs 455 gr at 25°C.
2. The apothecary pint contains 16 f℥.
3. The US gallon contains 128 f℥ or 231 inch³. One gallon of distilled water weighs 8.337 avoirdupois lb at 62°F. The US pint therefore weighs 1.04 avoirdupois lb and the pound of distilled water measures only 0.96 pt. *One pound does not measure one pt.*

The following facts concerning the imperial system (see Table 11-10) should be noted:

1. The imperial fluidounce of distilled water weighs 437.5 gr at 15.6°C (60°F). It therefore weighs one avoirdupois oz.
2. The imperial pint contains 20 f℥.
3. The imperial gallon contains 160 f℥. One gal of distilled water weighs 10 avoirdupois lb; 16 f℥ in this system therefore weighs one avoirdupois lb.

From the above, one can deduce the following:

1. The US fluidounce and minim are larger than the imperial fluidounce and minim (℥). One US minim or fluidounce equals 1.04 imperial minims or fluidounces.
2. The imperial pint and gallon are much larger than the US pint and gallon.

It is, therefore, inaccurate to use measuring devices calibrated in the US system in measuring quantities directed in English prescriptions when the imperial measure is intended. Conversely, devices calibrated in the imperial system should not be used to measure quantities directed in US prescriptions when the US measure is intended. For example, Canadian pharmacists using American graduated cylinders should calculate percentage solutions on the basis of 454.6 gr of distilled water to the fluidounce. This is one more argument in favor of adoption internationally by all pharmacists of the metric system of weights and measures.

Table 11-8. Metric Liquid Measure

1 microliter (μL)	=	0.000001 L
1 milliliter (mL)	=	0.001 L
1 centiliter (cL)	=	0.01 L
1 deciliter (dL)	=	0.1 L
1 liter (L)	=	1 L
1 dekaliter (daL)	=	10 L
1 hectoliter (hL)	=	100 L
1 kiloliter (kL)	=	1000 L

Note: The standard of capacity is the *liter*, which is the volume of one kg of distilled water at its maximum density (approx 4°C). Microliters (μL) are used to measure volumes of solutions used in chromatographic procedures for the separation and quantitative determination of some official drugs.

Table 11-9. Apothecary or Wine Measure (US)

GALLON	PINTS	FLUIDOUNCES	FLUIDRAMS	MINIMS
1	8	128	1024	61,440
	1	16	128	7,680
		1	8	480
			1	60

THE RELATIONSHIPS OF WEIGHTS AND MEASURES

When the systems of weights and measures in use in the US are examined, the lack of close relation between the different units is appreciated at once. Nevertheless, if the following points are used carefully, many pharmaceutical problems will be greatly simplified.

1. Pharmacists may weigh themselves, buy merchandise, sell over the counter, and calculate postage, etc., using avoirdupois weight, which contains 437.5 gr in one oz.
2. Pharmacists may compound formulas by apothecary weight, which contains 480 gr in one ℥.
3. One apothecary fluidounce of water weighs 455 gr at 25°C. Since 480 ℥ weigh 455 gr, one m weighs 455/480 = 0.95 gr.
1 ℥ does *not* weigh one gr.
1 f℥ does *not* weigh one ℥.

Practical Equivalents

Tables of weights and measures and a table of practical equivalents should be kept in a conspicuous and convenient place in the prescription department, and the following equivalents, which are given with practical accuracy, should be committed to memory. Other equivalents may be calculated from these.

Linear Measure

1 meter = **39.4 inches**
1 inch = 2.54 cm = **25.4 mm**
1 micrometer = **1/1000 mm** = 10⁻⁶ m = 1/25,400 inch

Liquid Measure

1 milliliter = **16.2 m**
1 fluidounce = **29.6 mL**
1 pint = **473 mL**
1 gallon = **3790 mL**

Weight

1 kilogram = **2.20 lb avoirdupois**
1 pound avoirdupois = **454 g**
1 ounce avoirdupois = **28.4 g**
1 ounce apothecary = **31.1 g**
1 pound apothecary = **373 g**
1 gram = **15.4 gr**
1 grain = **64.8 mg**

The USP *Table of Metric Doses with Approximate Apothecary Equivalents* is reproduced in the Appendix, along with information concerning its permissible uses.

Approximate Measures

In apportioning doses for a patient, the practitioner usually is compelled to order the liquid medicine to be administered in

Table 11-10. Imperial Measure (British)

GALLON	PINTS	FLUIDOUNCES	FLUIDRAMS	MINIMS
1	8	160	1280	76,800
	1	20	160	9,600
		1	8	480
			1	60

certain quantities that have been established by custom, and estimated as:

HOUSEHOLD MEASUREMENT	APOTHECARY NOTATION	METRIC VOLUME
1 tumblerful	f ʒ viii	240 mL
1 teacupful	f ʒ iv	120 mL
1 wineglassful	f ʒ ii	60 mL
2 tablespoonfuls	f ʒ i	30 mL
1 tablespoonful	f ʒ iii or ʒ s̄s	15 mL
1 dessertspoonful	f ʒ ii	8 mL
1 teaspoonful	f ʒ i	5 mL
½ teaspoonful	f ʒ s̄s	2.5 mL

Note: one drop is often considered to be one minim, but this is incorrect, as drops are variable.

In almost all cases, careful tests have found that modern teacups, tablespoons, dessertspoons, and teaspoons to average 25% greater capacity than the theoretical quantities just given. The physician and the pharmacist therefore should recommend the use of accurately graduated medicine droppers, teaspoons, and calibrated measuring devices, which may be procured at a small cost (Fig 11-1).

Approximate Dose Equivalents

For many years the apothecaries' system of weights and measures was used widely by physicians and pharmacists when considering the doses of medicinal substances, and it was customary

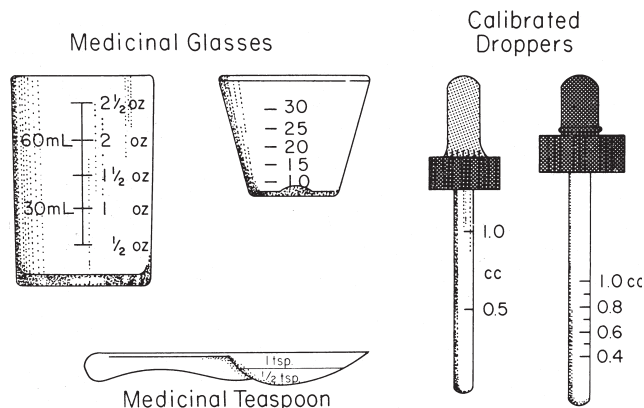


Figure 11-1.

to translate these apothecary doses into relatively exact amounts when the metric equivalents were mentioned. Today, however, doses are established primarily in the metric system without considering the relation of these metric figures to the corresponding quantities in any other system of weights and measures.

It should be emphasized that exact alternative formulas in the *avoirdupois* system of weights and measures are not obtained by using approximate equivalents but, for the purpose of compounding, should be calculated with the use of practical equivalents.

WEIGHING AND MEASURING

Having studied the several systems of weights and measures, students may now learn to apply their knowledge to the *weighing* and *measuring* of pharmaceuticals. The former process requires the use of the *balance*, or, for manufacturing purposes, *scales*, and the latter process requires the use of the *measure*, the *graduate*, and the *pipet*. The successful performance of many of the operations in pharmacy depends on a thorough knowledge of the principles of the balance and a correct understanding of its care and use; because weighing is nearly always the preliminary step in any compounding, it will be discussed first.

There is a relativity of accuracy in weighing (or measuring) that must not be overlooked, as illustrated by the following graded list: coal, salt, sugar, epsom salt, penicillin G, morphine, digoxin, vitamin B₁₂, and radium. One of the most important things for the pharmacist to learn is the degree of tolerance or error permissible in weighing or measuring any particular ingredient. Obviously, the final item on the list, radium, must be measured with much greater precision and accuracy than coal, the first item.

The empiric weighing and measuring methods of the kitchen, embodied in such concepts as a handful, a pinch, or "sweeten to suit your taste," have no place in pharmacy. Accurate work can be accomplished only by means of suitable apparatus.

WEIGHING

In pharmacy, weighing usually refers to ascertaining a definite weight of material to be used in compounding a prescription or manufacturing a dosage form.

The *balance* may be defined as an instrument for determining the relative weights of substances. It should be *selected correctly* for the specific task at hand, *used skillfully*, *protected from damage*, and *checked periodically*, if accurate results are to be obtained. Of even greater importance is its *construction*. Standards for balances are given by the NIST.¹

Construction of the Balance

For systematic consideration pharmaceutical balances may be classified as follows: single-beam (equal-arm or unequal arm), compound lever, torsion and electronic.

SINGLE-BEAM EQUAL-ARM BALANCES—The principle on which single-beam equal-arm balances (or scales) operate is clearly evident in the construction of the classical two-pan analytical balance. This type has a metallic lever or beam, divided into two equal arms at the center by a knife-edge, on which it is supported. At exactly equal distances from this point of support, and situated in the same plane, are placed the end knife-edges; these suspend the pans, which carry the substances to be weighed. A properly constructed balance of this type should meet the following requirements:

1. When the beam is in a horizontal position, the center of gravity should be slightly below the point of support, or central knife-edge, and perpendicular to it.
The relative sensitivity of the balance depends on the fulfillment of this principle, which may be illustrated roughly by forcing a pin through the center of a circular piece of pasteboard. If the edge of the pasteboard is touched slightly, it does not oscillate at all, but rather revolves around the center to a degree corresponding to the impulse given it. In this position it illustrates neutral equilibrium. If the pin is removed and inserted at a very short distance above the center, and the edge of the pasteboard touched as before, it will oscillate slowly, corresponding to a very sensitive beam, the point of support being slightly above the center of gravity as in the balance. If the pin is removed again and inserted far above the center, and the same impulse imparted to the edge, it will oscillate quickly, illustrating stable equilibrium characteristic of a beam that comes to rest quickly and is not particularly sensitive. Unstable equilibrium may be illustrated by balancing the disc so that the point of support is below the center. The slightest touch then causes it to reverse its position completely and finally come to rest with the center of gravity below the point of support.
2. The end knife-edges must be exactly equal distances from the central knife-edge; they all must be in the same plane and the edges absolutely parallel to each other.

It is very apparent that the conditions of a good prescription balance cannot be satisfied if there is inequality in the length of the arms of the beam. The distance from the central knife-edge to the one on the left must be exactly the same as the distance from the central knife-edge to the one on the right, otherwise unequal weights would be required to establish equilibrium. If the central knife-edge is placed either above or below a line drawn so that it connects the end knife-edges, the loading of the pans either will cause the beam to cease oscillating or diminish the sensitivity in proportion to the load. If the knife-edges are not parallel, the weight of a body will not be constant upon every part of the pan, but will be greater if placed near the edge on one side, and correspondingly less at a point directly opposite.

3. *The beam should be inflexible, but as light in weight as possible, and the knife-edges in fine balances should bear upon agate plates.* The rigidity of the beam is necessary because any serious deflection caused by a loading of the pans would lower the end knife-edges and thus accuracy in weighing would be impossible. The beam should not be heavier than necessary because the sensitiveness of the balance thereby would be lessened; to diminish friction, which constantly increases with the age and use of a balance, the bearings of the knife-edges should be agate plates, which are polished flat pieces of the very hard mineral called agate.

A single-beam equal-arm balance with the rider beam graduated to 28 g in increments of 0.2 g, and to 1 oz in increments of 0.01 oz, is shown in Figure 11-2.

UNEQUAL-ARM BALANCES—The unequal-arm balance is the type is preferred for laboratory work when large amounts are to be weighed (Fig 11-3). The lever principle on which these scales are constructed is based on the law of physics that at equilibrium the force applied at one end of the lever multiplied by the length of the arm (distance from the fulcrum to the point where the force is applied) must be equal to the product of the force acting at the opposite end of the lever and the length of the other arm. The inequality in the length of the arms of this beam permits the convenient use of movable weights upon the graduated longer arm of the beam, thus dispensing with the use of small weights, which are liable to be lost. This scale is of great advantage in laboratory or manufacturing work because it is particularly adapted for weighing liquids; a sliding tare is set on one beam for the weight of the container, and other sliding weights can be adjusted to the weight of liquid desired. These are available with the beams graduated either in the avoirdupois or metric system.

COMPOUND-LEVER BALANCES—The principle of the compound lever was first applied in the construction of balances by Robervahl of Paris, in about 1660 AD. It was skillfully adapted for both prescription balances and the general counter and platform scales. The principal objection to this type of scale, when compared with single-beam balances, consists in the multiplicity of points of contact and suspension, thus necessarily in-

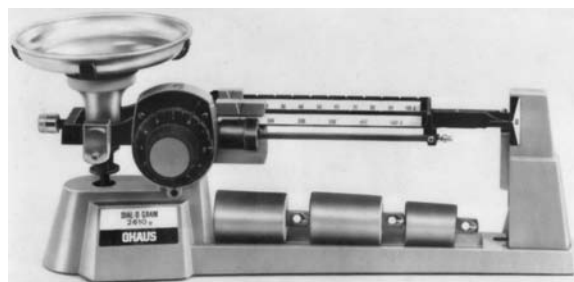


Figure 11-3. Manufacturing laboratory scale and weights (courtesy, Ohaus).

creasing friction and the liability to disarrangement; however, their general convenience has made them popular.

TORSION BALANCES—A simple illustration of the principle of torsion is afforded by tying a stout piece of cord to a firm support and inserting a lead pencil in the middle of the cord between the strands, at right angles to it. If the free end of the cord is stretched tightly, resistance is offered to any effort to turn the lead pencil over; if the pencil is released, it at once flies back to its original position. *Torsion* is the term applied to this method of twisting. The principle of supporting the beam of a balance on a tightly stretched wire, with the view of doing away with knife-edges and diminishing friction, occupied the attention of inventors for years.

In 1882 Prof. Roeder and Dr. Springer contrived an ingenious torsion balance that gave promise of valuable results. Two illustrations of this original balance were shown on page 54 of the first edition of *Remington's Practice of Pharmacy* in 1885. Improvements have increased its efficiency greatly. The most important difficulty in applying the principle of torsion resistance was overcome by placing a weight just above the center of gravity (Fig 11-4). Torsional resistance tends to keep the beam in a horizontal position, while the elevation of a weight above the center of gravity, by its tendency to produce unstable equilibrium, exercises an opposite effect—the beam is inclined to be top heavy and, therefore, to tip on either side. If now the weight is made adjustable by mounting it upon a perpendicular screw so that it can be raised or lowered, it is possible to arrange these opposite forces so that one exactly neutralizes the other. In this manner sensitivity is obtained.

The torsion principle has been applied to prescription balances, as well as analytical balances and scales designed to carry heavier loads. In the torsion prescription balance

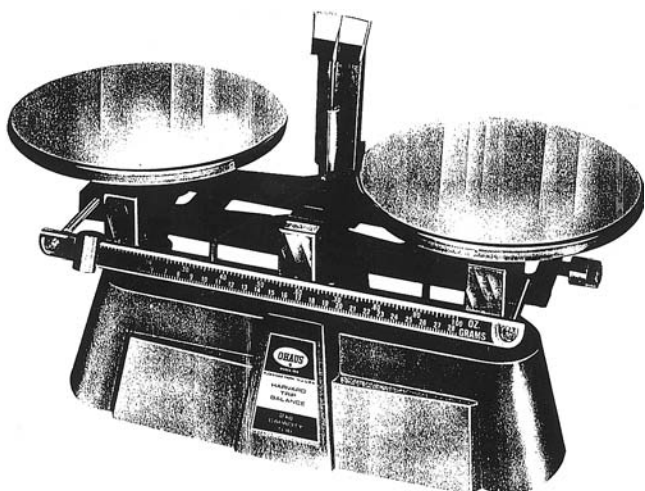


Figure 11-2. Single-beam equal-arm balance (courtesy, Ohaus).

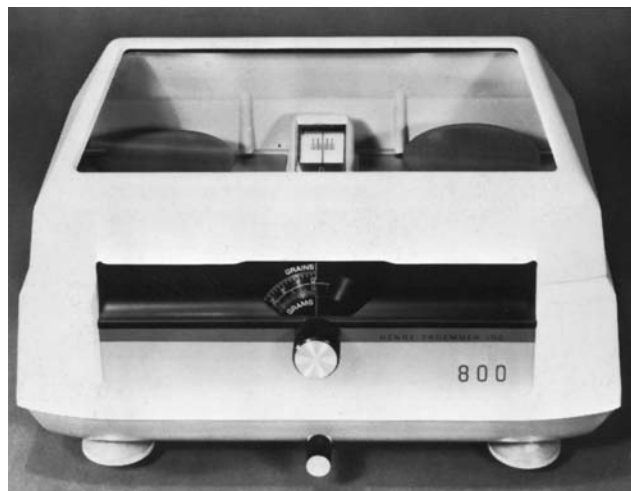


Figure 11-4. Troemner/800 prescription balance (courtesy, Troemner).

two beams are used, supported on three frames, each of the latter having a flattened metallic band stretched tightly over its edge.

The torsion balance, which has a rider beam graduated upon the upper edge from 1/8 to 15 gr and on its lower edge from 0.01–1.0 g, furnishes a very convenient means of weighing small quantities without having to use small weights. Most modern balances have a direct-reading dial instead of a rider beam, with the metric scale on the upper scale and the apothecary scale on the lower.

The prescription balance may be placed upon a base containing a drawer that can be used for holding weights or powder papers.

ELECTRONIC BALANCES—Electronic balances are single pan balances with digital or direct-reading features (Fig 11-5). Taring a weighing paper, weigh boat, or beakers is done automatically with the push of a button or lever without the need of external balancing weights. These balances are much more sensitive than the traditional prescription balance, are easier and quicker to use, but are usually more expensive than a torsion balance.

Prescription Balances

The most common type of prescription balance uses the taut-wire frame or torsion principle (see Fig 11-4). Such balances, manufactured to meet the requirements of the NIST Class III balances, have a maximum maintenance sensitivity of six mg with no load and with full load (ie, addition of the 6 mg weight to one pan causes the indicator or the rest point to be shifted not less than one division on the index plate). The Class III balance is used to weigh quantities up to 60 g, depending on the stated capacity and subject to the physical limit of the amount of the material that can be placed on the pan. Electronic balances typically have a sensitivity less than 10 mg (easily meeting standards for a Class III balance) and can weigh small quantities of drug more accurately than a torsion balance. All prescription departments must have a Class III balance.

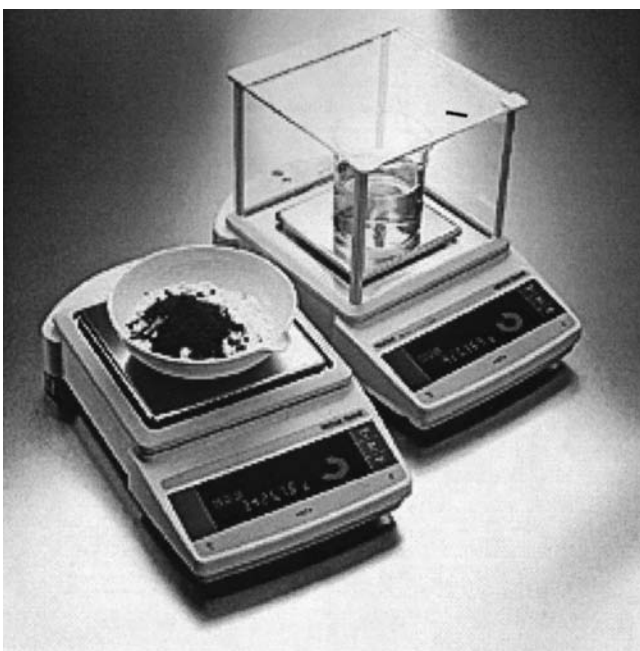


Figure 11-5. Electronic single pan balance.

REQUIREMENTS—A prescription balance should meet the following general requirements:

1. It should be constructed so as to support its full capacity without developing undue stresses, and should not be thrown out of adjustment by repeated weighings of the capacity load. (The capacity of the balance will be seen on the metal plate attached to it.) If the capacity is not stated, it is assumed to be at least 15 g (1/2 oz). The Class III balances usually have a capacity of 60 g (2 oz).
2. The removable pans of a torsion prescription balance should be of equal weight. If the pans show any difference in weight, they should be adjusted by leveling the balance or using small pieces of paper. Pans with any appreciable corrosion or wear should be refinished or replaced.
3. A prescription balance should have a leveling device, usually leveling feet or screws, so that the balance can be adjusted to a level position. A balance that does not have these is not entitled to be designated as a prescription balance.
4. The balance that has a rider or graduated dial should have, at the end of the graduation, a stop that halts the rider or dial at the zero reading. The reading edge of the rider should be parallel to the graduations on the beam.
5. The indicator points, when there are two on the balance, should be sharp, and their ends should not be separated by more than one mm (0.04 inch) when the scale is in balance. The distance from the face of the index plate to the indicator pointer or pointers should be small (1 mm or less) to protect the operator against making errors resulting from parallax, because it is unlikely that the eye of the operator will be exactly in line with the indicator and the division on the index plate. The indicating elements as well as the lever system of the balance should be protected against drafts. The balance should have a lid that allows a weighing to be made when the lid is closed.
6. A torsion prescription balance must have a mechanical means for arresting the oscillation of the mechanism.

TESTING—Certain tests may be used to satisfy the user regarding the construction and character of a torsion balance when its origin, history, or condition is in doubt. Additional tests are carried out by the NIST, manufacturers, and local and state testing agencies.

A Class III torsion prescription balance meets the following basic tests. Use a set of *test weights* and keep the rider or graduated dial at zero unless directed to change its position.

1. **Sensitivity Requirement:** Level the balance, determine the rest point, place a 10-mg weight on one of the empty pans, and again determine the rest point. Repeat the operation with a 10-mg weight in the center of each pan. The rest point is shifted not less than one division of the index plate each time the 10-mg weight is added. The sensitivity requirement for an electronic balance is supplied by the manufacturer.
2. **Arm Ratio Test:** This test is designed to check the equality of length of both arms of the balance. Determine the rest point of the balance with no weight on the pans. Place 30 g of test weights in the center of each pan and determine the rest point. If the second rest point is not the same as the first, place a 20-mg weight on the lighter side; the rest point should move back to the original place on the index plate scale or farther.
3. **Shift Tests:** These tests are designed to check the arm and lever components of the balance.
 - a. Determine the rest point of the indicator without any weights on the pans.
 - b. Place one of the 10-g weights in the center of the left pan, and place the other 10-g weight successively toward the right, left, front, and back side of the right pan, noting the rest point in each case. If in any case the rest point differs from the rest point determined in (a), add the 10-mg weight to the lighter side; this should cause the rest point to shift back to the rest point determined in (a) or farther.
 - c. Place a 10-g weight in the center of the right pan, and place a 10-g weight successively toward the right, left, front, and back sides of the left pan, noting the rest point in each case. If in any case the rest point is different from that obtained with no weights on the pans, this difference should be overcome by addition of the 10-mg weight to the lighter side.

A balance that does not measure up to these tests *must* be corrected.
4. **Rider- and Graduated-Dial Tests**—Determine the rest point for the balance with no weight on the pans. Now place on the left

pan the 500-mg test weight and move the rider to the 500-mg point on the beam. Now determine the rest point. If it is different from the zero rest point, add a 10-mg weight to the lighter side. This should bring the rest point back to its original position or farther. Repeat this test, using the 1-g test weight and moving the rider or graduated dial to the 1-g division. If the rest point is different it should be brought back at least to the zero rest point position by the addition of 10 mg to the lighter pan. If the balance does not meet this test, the graduated beam or the rider must be corrected. For balances equipped with a dial scale, the dial must be corrected.

PROTECTION—The necessity for protecting the delicate mechanism of a balance is overlooked frequently, notwithstanding the possibility of having a precision apparatus irretrievably ruined by lack of care in using or cleaning it or in protecting it while at rest. The position chosen for the balance or scales should be on a level and firm counter, desk, or table, where it will be subjected to little risk of damage from dampness, dust, or corrosive vapors and where the knife-edges will not be liable to become dulled by jarring or other vibrations.

In the analytical class of balances, protection is afforded by enclosing them in glass cases having sash doors in the front, sides, or back. They are protected against damage from vibration by a lever for elevating or locking the beam, so that the knife-edges are not in contact with any surface when not in use. To prevent damage from jarring while the balance is in use, from a weight falling on the pan, or other accident, the finest balances are provided with pan supports, which break the fall and serve the additional purpose of quickly arresting the beam, thus saving time while weighing.

In using a prescription balance, neither the weights nor the substance that is to be weighed should be placed on the balance pans while the beam is free to oscillate. The desired weight should be placed upon one pan (usually the one on the right-hand side) and an amount of the substance to be weighed, approximately the desired weight, upon the opposite pan. The beam should be released by means of the lever, and if the substance is in excess, the beam should be locked and a small portion removed and the beam again released and the oscillations observed. This procedure should be repeated until the correct amount is obtained. In case of a deficiency of the substance to be weighed, the reverse procedure is followed until the correct amount is obtained. With practice this can be done very deftly and very quickly and the sensitivity of the balance retained for years.

Substances that react with metals, such as iodine, and those that are adhesive, such as the extracts, should not be weighed directly upon the pans, but rather upon counterpoised watch crystals, or upon glazed paper, care being taken to balance the papers before weighing the substance. In cleaning the balances, great care should be exercised; polishing powders should be used sparingly, as a portion is very apt to find its way into crevices and elude detection until an attempt is made to adjust the balances, when the increased weight of one of the sides of the beam leads to its discovery. Frequent cleaning with soft leather generally is sufficient to keep a balance in good order, but once neglect makes it necessary to use more active measures, some simple polishing powder for the metal work, soap-suds for the nickel plate, and simple brushing for the lacquered brass are all that is necessary.

As the pans are subjected to more wear and tear than any other part of the balance, it is economical to use *solid* rather than *plated* pans because constant friction wears off the plating and the additional cost for replating soon absorbs the difference in price. Equipped in this way, and with agate bearings, a prescription balance is durable and really inexpensive because it will remain fully equal to the most exacting demands for a long time.

Weights Used in Pharmacy

The weights used by the pharmacist are very important, and care in their selection and examination is necessary. False economy must be avoided, as the use of cheap, inaccurate weights ultimately leads to serious consequences. Official inspectors have

found pharmacies using prescription weights that were so worn that the characters on their faces had disappeared; also, weights have been found with bits of hardened extract and dirt almost entirely obscuring their characters. An unused set of standard weights should be kept on hand so that at least once a year the weights in daily use can be tested and adjusted or rejected if necessary. The standard weights should be used also when the balance is tested. The set should contain the following weights in a well-fitted box with forceps: one 50-g, two 20-g, one 10-g, one 5-g, two 2-g, one 1-g, one 500-mg, two 200-mg, one 100-mg, one 50-mg, two 20-mg, and one 10-mg, all adjusted to NIST tolerances for analytical or Class P weights.

METRIC WEIGHTS—For weighing larger quantities, japanned iron metric weights are available. They are preferably hexagonal, to distinguish them from the round avoirdupois weights. Sets of brass weights, usually in the range of 10 g to 1000 g, fitted into holes of appropriate size in a block of plastic (*block weights*), are especially convenient for many weighing operations. For prescription compounding, accurate sets of weights ranging from 10 mg to 50 g are available.

For analytical purposes, metric weights are used exclusively; usually, the highest weight is 100 g, the lowest 1 mg. The weights from one g upward are of finely lacquered brass or of nonmagnetic stainless steel or rhodium-plated bronze. The smaller weights are made of squares of platinum or aluminum foil, with one edge turned up to permit them to be handled easily with the forceps. Fractions of a milligram are weighed by means of the rider on the graduated beam of the balance.

In analytical work and in using the Class III balance in prescription work, the weights should never be handled with the fingers but always with the forceps, which accompany an accurate set of weights. In the more expensive sets of weights the forceps are tipped with bone, ivory, or plastic to prevent the wearing away of the weights during handling. With proper care the accuracy of a fine set of weights may be maintained for years.

COMMON AVOIRDUPOIS WEIGHTS—Avoirdupois weights usually are made of iron, and they are flat and circular and japanned to prevent rusting. These weights form a pyramidal pile, and range from $\frac{1}{2}$ oz to 4 lb; if found to be incorrect, they may be adjusted by adding to or diminishing the amount of lead that is hammered into a depression in the base of each weight. They sometimes are made of brass in this form, and sometimes of zinc (the latter, however, are brittle and unserviceable). For general use in the pharmacy, the cylindrical weights, known technically as block weights, are preferable. The advantages of block weights are that the gaps left by missing weights are readily noticeable, and the greater part of the surface of the weight is protected from the action of corrosive vapors when the weights are not in use.

APOTHECARY WEIGHTS—Apothecary weights may be obtained either as *block weights* or in the less-desirable *flat* forms. The round, flat, brass *dram* weights, which have the denomination stamped on their faces in raised characters, still are used but should be replaced. With flat weights, the denomination is often only faintly stamped on the face and thus is liable to be obliterated by constant use or by corrosive contact.

Undoubtedly, the best grain weights are the aluminum wire weights. The wire weights are less susceptible to corrosive action than are the brass weights. Also, the wire weights are more easily and quickly distinguished from one another than are other weight forms, so there is less likelihood of dangerous mistakes: the number of sides in the wire weights at once gives the denomination (Fig 11-6).

Aluminum grain weights, which are cut out of aluminum plates, are also less liable to be corroded. They usually can be more accurately adjusted than brass weights. The corners of the aluminum weights are clipped, and each weight usually is pressed into a curved form so that it may be picked up easily (Fig 11-7).

The need for apothecary weights in modern practice is decreasing. Apothecary weights can easily be converted to the corresponding metric weights, which are easier to use and less



Figure 11-6. Metric and Apothecary weight set (courtesy, Troemner).

prone to error. In addition, electronic balances do not use external weights. Typically, an electronic balance can display weights in several systems as the discretion of the operator.

Minimum Weighable Quantity

All of the balances described must be used within a degree of error that can be tolerated in prescription compounding and in pharmaceutical manufacturing. The USP allows a maximum error of 5% in a single weighing operation. Since the sensitivity requirement of a balance represents the absolute error in using that balance, the percent error will depend on the amount of drug weighed and will increase as the amount of drug decreases.

The Minimum Weighable Quantity (MWQ) with no more than 5% error can be calculated for any balance knowing the sensitivity requirement (SR) (ie, the absolute error) from the following:

$$\text{MWQ} = \text{SR} \frac{100\%}{5\%}$$

Examples

1. Calculate the MWQ with no more than a 5% error for a balance with a sensitivity requirement of 10 mg.

$$\text{MWQ} = 10 \text{ mg} \frac{100\%}{5\%} = 200 \text{ mg}$$

2. Calculate the MWQ with no more than a 3% error for a balance with a sensitivity requirement of 2.5 mg.

$$\text{MWQ} = 2.5 \text{ mg} \frac{100\%}{3\%} = 83.3 \text{ mg}$$

MEASURING

In pharmacy, *measuring* usually refers to the exact determination of a definite volume of liquid. Many types of apparatus are

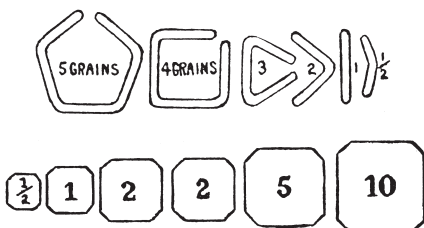


Figure 11-7. Aluminum wire and aluminum grain weights.

used in this operation, depending on the kind and quantity of liquid to be measured and the degree of accuracy required. (The NIST has requirements for graduates.¹)

Large Quantities

Glass measures are preferred for measuring liquids. Although glass measures are subject to breakage, they can indicate volume more accurately because of the transparency of glass.

THE MENISCUS—When an aqueous or alcoholic liquid is poured into a graduate, surface forces cause its surface to become concave—the portion in contact with the vessel is drawn upward. This phenomenon is known as the formation of a *meniscus* (Fig 11-8), and in determining the volume of a liquid the reading must be made at the bottom of this meniscus. This regulation has been established by the NIST, and all glass measuring vessels are graduated on this basis. Liquids with large contact angles, such as mercury, form an *inverted meniscus*, and the reading then is made at the top of the curved surface.

PROCEDURE—Pharmaceutical manufacturers package liquid preparations in glass or plastic containers equipped with a plastic screw-cap. These containers serve as a stock bottle from which liquids may be poured directly into a graduate. The procedure for pouring liquid from screw-capped containers is as follows:

1. Remove the cap and place it on the counter while the transfer of liquid is made.
2. While holding the graduate in the left hand, grasp the original container with the label in such a position that any excess of liquid will not soil the label if it should run down the side of the bottle.
3. Raise the graduate and hold it so that the graduation point to be read is on a level with the eye, and measure the liquid. (The extension of the graduating mark into a circle that passes entirely around the graduate is an improvement that obviates the necessity of placing the graduate upon a level place, as the corresponding mark upon the opposite side may be seen through the glass and the graduate easily leveled even when held in the hand.)
4. Replace the cap, and return the bottle to the counter or shelf.
5. Pour the liquid into the bottle or mortar for dispensing or compounding.

METALLIC MEASURES—Metallic measures are nearly cylindrical in shape, but are slightly wider at the bottom. These are generally used for measuring liquids when the quantity is over a pint. A set usually consists of five (gallon, half-gallon, quart, pint, and half-pint) of these measures. Measures made of tinned iron, or of the enameled sheet iron called agaware, are greatly inferior to those made of *tinned copper* or *stainless steel*; tinned-iron measures soon become rusty, and particles of enameling can chip off, leaving the exposed iron to contaminate the measured liquids.

The initial cost of copper or stainless-steel measures is greater than tinned iron, but they are far more durable. Care

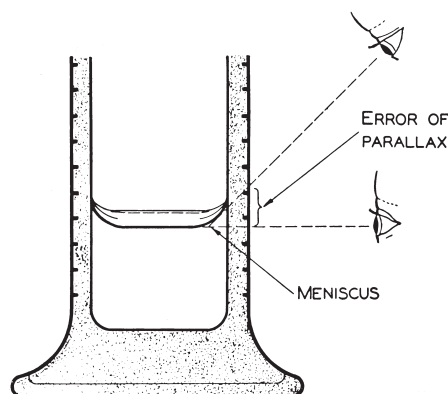


Figure 11-8. Error of measurement due to parallax.

must be taken to protect them from blows that will cause dents as these may be serious enough to detract from their accuracy. Cylindrical metric measures, usually made of monel metal or stainless steel and having a diameter just half their height, are available in various sizes. Such containers are relatively expensive, but their resistance to corrosion and wear is a tremendous advantage. Copper, of course, should not be used where it is likely to catalyze oxidation.

GRADUATED GLASS MEASURES—Graduated glass measures nearly always are used for quantities of 500 mL or one pt or less. There are of two forms, *conical* and *cylindrical* (Fig 11-9, 11-10). The conical graduate is suitable for some measurements because of the greater ease with which it can be handled, but cylindrical measures are more accurate because of their uniform and smaller average diameter. In a graduated cylinder, the error in volume caused by a deviation of ± 1 mm in reading the meniscus remains constant along the height of the uniform column; the same deviation causes a progressively larger error in a conical graduate because the diameter, and thus the volume of the 1-mm column, increases along its vertical axis. It is safe to assume that practically all good-grade modern graduates comply with the NIST requirements for internal diameters at stated volumes.

A study has indicated that, to improve accuracy, the lower portions of graduates should not be used, and therefore should not be marked.² A composite tabulation (Table 11-11) shows the calculated and the assigned blank portions of graduates. The elimination of the lower markings on graduates was suggested, and in 1955 the NIST specifications for graduates used this principle.¹ The NIST Handbook states, "A graduate shall have an initial interval that is not subdivided, equal to not less than one-fifth and not more than one-fourth of the capacity of the graduate." For accurate measurement of volumes less than 1.5 mL, a graduated pipet or a graduated dropper could be used.

EFFECT OF LIQUID AND CONTAINER—It is difficult to measure accurately when pouring from a completely filled bottle because of the uneven flow of the liquid. After the first portion of the liquid is removed, the shape of the bottle does not influence the ease of pouring to any appreciable extent unless the neck is extremely narrow.

Viscous liquids pour slowly, but their accurate measurement is not difficult. Experiments showed that when glycerin is poured into a graduate without letting the liquid run down the inside surface, the precision of measurement can be very high. Naturally, the chance of hitting the inner surface is greater with smaller than with larger graduates. The increase in possible deviation then is caused by the slow movement of the viscous liquid to the desired mark.

Viscous liquids introduce another factor: drainage time. Graduates are calibrated to contain or deliver indicated volumes within specified limits. Aqueous, alcoholic, and hydroalcoholic liquids can be drained from a graduate in 30 seconds so completely that the delivered and contained volumes are fairly close. When 25 mL of glycerin was measured in the same cleaned and dried cylinders, the received volume measured 23.7 mL af-

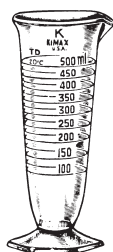


Figure 11-9. Glass conical graduate (courtesy, Kimble Glass).



Figure 11-10. Glass cylindrical graduate (courtesy, Kimble Glass).

ter the same time period. Silicone-treated glassware, which now is used frequently, drains completely in a few seconds.

The viscosity factor might be altered when another liquid is to be mixed with the glycerin by measuring and mixing both liquids in a suitable graduate.

Small Quantities

For measuring smaller quantities of liquids, graduated glass tubes of small diameter should be used. The narrower bore permits greater distances between the graduations on the apparatus, thus allowing greater accuracy in making the reading. For example, with a buret the pharmaceutical chemist can estimate volumes to the nearest 1/100 mL.

Pipets and similar apparatus are more accurate and convenient than very small graduates. The graduations on very small graduates are necessarily in the very small, lowest portion of a comparatively tall measure. To measure 1 mL of a volatile oil in a graduate, the surface that the oil must traverse when this measure is inverted is so great that probably 20% of the oil will be left adhering to the measure. In liquid preparations in which the smaller liquid is miscible with the larger quantity of diluting liquid, the graduate may be rinsed and this loss recovered, but inconveniences are largely overcome and greater accuracy secured by using a pipet.

In administering small quantities of liquids, the very convenient *drop* is almost always used. It should be emphasized that one *drop* is not equivalent to 1 m^3 and that 60 *drops* are not equivalent to one *f 3*. This impression doubtlessly arose because 60 ordinary drops of *water* are about equal to one *f 3*, but the volume of a drop of fluid depends on many factors, including density, temperature, viscosity, surface tension, and the size and nature of the orifice from which it is dropped. Thick, viscous liquids, such as the mucilages and the syrups, necessarily produce large drops because the drop adheres to the surface of the glass as long as its weight does not overcome its power of adhesion, whereas chloroform, a mobile liquid that has very little adhesion to the dropping surface, produces very small drops. The greater the surface tension, the larger the

Table 11-11. Unmarked (Unreliable) Portions of Graduates

CAPACITY OF GRADUATE (mL)	CALCULATED BLANKS (1951)		NIST BLANKS (mL)
	2.5% ^a ALLOWED (mL)	5% ^a ALLOWED (mL)	
5	3.0	1.5	1
10	4.4	2.2	2
25	11.8	5.9	5
50	15.8	7.9	10
100	20.9	10.5	20
250	36.3	18.2	50
500	66.5	33.2	100
1000	—	—	200

^a Calculations by Goldstein and Mattocks² based on deviation of ± 1 mm from graduation mark and allowable errors of 2.5% and 5%.

drop, and the greater the extent of surface to which the drop adheres, the larger, proportionally, the drop.

A *normal or standard drop measure* was recommended by the Brussels Conference of 1902 for international adoption. This dropper is recognized in the USP.

MEDICINE DROPPER³

The Pharmacopeial medicine dropper consists of a tube made of glass or other suitable transparent material that generally is fitted with a collapsible bulb and, while varying in capacity, is constricted at the delivery end to a round opening having an external diameter of about 3 mm. The dropper, when held vertically, delivers water in drops each of which weighs between 45 mg and 55 mg.

When drops are specified on a prescription, the usual custom has been to employ an *eyedropper*, but now the standard dropper should be supplied. When accuracy is required, it is particularly important to use a specially calibrated dropper for administering potent medicines. The volume error incurred in measuring any

liquid by means of a calibrated dropper should not exceed 15%, under normal conditions.³

TEASPOON

For household purposes, an American Standard Teaspoon has been established by the American National Standards Institute as containing 4.93 ± 0.24 mL. In view of the almost universal practice of employing teaspoons ordinarily available in the household for the administration of medicine, the teaspoon may be regarded as representing 5 mL and is so accepted by the USP.

It must be kept in mind that the actual volume delivered by a teaspoon of any given liquid is related to the latter's viscosity and surface tension, among other influencing factors.

THE HUMAN FACTOR—The *human factor of carefulness* is of paramount importance in every pharmaceutical operation in which accuracy is essential. Accurate measurement of liquids requires accurate equipment, careful manipulation, good vision, and a steady hand.

DENSITY AND SPECIFIC GRAVITY

Several terms are used to express the mass (weight) of equal volumes of different substances.

Absolute density is the ratio of the mass of an object, determined in or referred to a vacuum, at a specified temperature, to the volume of the object at the same temperature. This relationship is expressed mathematically as:

$$\text{Absolute Density} = \frac{\text{Mass in grams (in a vacuum)}}{\text{Volume in millimeters}}$$

Apparent density differs from absolute density only in that the mass of the object is determined in air; the mass is influenced by the difference in the buoyant effect of air on the object being weighed, and on the standard masses (weights) used for comparison. If the object and masses are made of the same material, or have the same density, there will be no difference in the buoyant effect, and the apparent density will be identical with the absolute density.

Relative density is an expression sometimes employed to indicate the mass of 1 mL (not cc, which is very slightly different) of a standard substance, such as water, at a specified temperature, relative to water at 4°C taken as unity. Thus, at 4°C the relative density of water is 1.0000, whereas its absolute density at the same temperature is 0.999973. Water attains its maximum absolute density of 0.999973 at 3.98°C. To convert a relative density of water to absolute density, the former should be multiplied by 0.999973.

Specific gravity may be defined as the ratio of the mass of a substance to the mass of an equal volume of another substance taken as the standard. For gases, the standard may be hydrogen or air; for liquids and solids, it is water.

From what has been stated, it is obvious that in a determination of specific gravity there will be, in general, a difference in the result if the masses (weights) are determined in air or in vacuum. If the masses are determined in, or referred to, a vacuum, the result is a *true specific gravity* (sometimes called *absolute specific gravity*); if the masses are determined in air, the calculated result is an *apparent specific gravity*. The difference between these specific gravities is, as a rule, very small.

A very important variable in specific gravity determinations is temperature, and this is doubly important because both the temperature of the substance under examination and the temperature of the standard may be different. The temperatures are commonly shown as a ratio, with the temperature of the water always being indicated in the denominator. The common practice with regard to the determination of specific gravity is that defined by the USP: "Unless otherwise stated, the specific

gravity basis is 25°/25°, ie, the ratio of the weight of a substance in air at 25° to that of an equal volume of water at the same temperature."

But it is not always convenient, or desirable, to determine the weight of both the substance and the water at 25°, or even to determine the weight of the substance at the same temperature as that at which the water is weighed. Thus, the substance may be weighed at 25° and compared with the weight of an equal volume of water at 4°, in which case the specific gravity is reported as being on a 25°/4° basis. In the case of theobroma oil, which is solid at 25°, the specific gravity is determined on a 100°/25° basis; for alcohol, it is determined on a 15.56°/15.56° basis because many years ago the US government adopted 60°F (15.56°C) as the temperature at which alcoholometric measurements are to be made for government control of alcoholic liquids.

It is apparent that a completely informative statement of specific gravity must indicate the temperature of the substance under examination, as well as that of the equal volume of water. Furthermore, it should be stated whether the determinations of mass (weight) were made on an *in-vacuum* or *in-air* basis; the latter case, the material of construction of the weights also should be indicated (as the buoyant effect of air on weights depends on their volume).

Calculations

The principle underlying the determination of the specific gravity of either a liquid or a solid is the same: to find the ratio of the mass (weight) of the substance to that of an equal volume of water. This may be expressed by a simple relationship:

$$\text{Specific gravity} = \frac{W_s}{W_w}$$

where W_s is the weight of the substance, and W_w the weight of an equal volume of water.

DENSITY

Density is defined as the mass of a substance per unit volume. It has the units of mass over volume. *Specific gravity* is the ratio of the weight of a substance in air to that of an equal volume of water. In the metric system both density and specific gravity may be numerically equal, although the density figure has units. In the English system, density and specific gravity are

not numerically equal; for example, the density of water is 62.4 lb/ft³ and the specific gravity is 1. This shows the convenience of the metric system. The equations for calculating density, weight, and volume are

$$\text{Density} = \frac{\text{Weight}}{\text{Volume}}$$

$$\text{Weight} = \text{Density} \times \text{Volume}$$

$$\text{Volume} = \frac{\text{Weight}}{\text{Density}}$$

Given any two variables, the third one can be calculated.

Examples

1. A pharmacist weighs out 2 kg of glycerin (density, 1.25 g/mL). What is the volume of the glycerin?

$$\text{Volume} = \frac{2000 \text{ g}}{1.25 \text{ g/mL}} = 1600 \text{ mL}$$

2. What is the weight of 60 mL of oil whose density is 0.9624 g/mL?

$$\text{Weight} = 60 \text{ mL} \times 0.9624 \text{ g/mL} = 57.7 \text{ g}$$

3. Calculate the weight of 30 mL of sulfuric acid (density, 1.8 g/mL).

$$\text{Weight} = 30 \text{ mL} \times 1.8 \text{ g/mL} = 54 \text{ g}$$

4. If a prescription order requires 25 g of concentrated hydrochloric acid (density, 1.18 g/mL), what volume should the pharmacist measure?

$$\text{Volume} = \frac{25 \text{ g}}{1.18 \text{ g/mL}} = 21.2 \text{ mL}$$

Problems (Answers on page 125)

1. What is the weight in grams of one L of alcohol (density, 0.816 g/mL)?
2. What is the volume (mL) of one lb (avoirdupois) of glycerin (density, 1.25 g/mL)?
3. What is the volume (mL) of 65 g of an acid whose density is 1.2 g/mL?

PHARMACEUTICAL CALCULATIONS

Pharmaceutical dispensing and compounding calculations use simple arithmetic. The errors that may arise often are due to carelessness, as in improper placing of decimal points, incorrect conversion from one system of measurement to another, or uncertainty over the system of measurement to be used. Before proceeding with any calculation, it is imperative that the problem presented (in a prescription, chart order, formula, etc) be read carefully, that the information given and required be identified, and that the procedure to be used in the calculation be selected.

Before students read this part of the chapter and attempt to solve the problems, the information in the preceding part of this chapter must be understood thoroughly. Often, several steps are necessary to solve problems. Shortcuts should not be taken unless one is certain they are proper. Many problems can be solved by more than one procedure, such as by ratio and proportion or by dimensional analysis. If students find a procedure that is more logical to them and gives the correct answer, it should be used. Thus, the solutions to sample problems used here generally should be considered suggestions, rather than the only way to solve a given type of problem.

Mathematical Principles

A few mathematical principles (eg, common decimal fractions, exponents, powers and roots, significant figures, and logarithms) will be reviewed, as these are areas where students often become careless or have forgotten skills. Following this, various types of practical pharmaceutical problems that the pharmacist may be required to solve are discussed and solutions are given. Where practical, rules for solving these problems are given. No attempt is made to elaborate on any mathematical theory.

The problems generally consist of determining the quantity or quantities of material(s) required to compound prescriptions properly and make products used to aid the compounding of prescriptions. The materials used to compound prescription orders may be pure or mixtures of substances in varying strengths. The strengths of mixtures may be denoted in different ways. Conversions may be necessary between systems of varying strengths or between different measuring systems. At the end of each section, sample problems are given for the student to solve, the answers to which appear on page 125.

Because of the decreasing importance of the apothecary system, the metric system is emphasized here. Chemicals and preparations most likely will be purchased using the avoirdupois or metric systems. Prescription orders are filled in the system indicated on the order, usually the apothecary or metric systems.

The student should become familiar with the terminology used in writing prescription orders, such as Latin words and abbreviations used in giving directions to the pharmacist and patient. The prescriber occasionally may use Roman numerals instead of Arabic numerals, so students must be familiar with these (even if the practice is declining).

SIGNIFICANT FIGURES

Weighing and measuring can be carried out with only a certain maximum degree of accuracy; the result always is approximate due to the many sources of error such as temperature, limitations of the instruments employed, personal factors, and so on. Pharmacists must achieve the greatest accuracy possible with their equipment, but it would be erroneous to claim that they have weighed 1 mg of a solid on a Class III prescription balance, which has a sensitivity requirement of 10 mg, or that they have measured 76.32 mL of a liquid in a 100-mL graduate, which can be read only to 1 mL. When quantities are written, the numbers should contain only those digits that are *significant* within the precision of the instrument.

Significant figures are digits that have practical meaning. In some instances zeros are significant; in other instances they merely indicate the order of magnitude of the other digits by locating the decimal point. For example, in the measurement 473 mL all the digits are significant, but in the measurement 4730 mL the zero may or may not be significant. In the weight 0.0316 g the zeros are not significant but only locate the decimal point. In any result the last significant figure is only approximate, but all preceding figures are accurate.

When 473 mL is recorded, it is understood that the measurement had been made within ± 0.5 mL or somewhere between 472.5 and 473.5 mL. The student should stop to consider the full implications of this specifically that the measurement is subject to a maximum error of:

$$\frac{0.5}{473} \times 100 = (\text{approx}) 0.1\% \text{ or } 1 \text{ part in } 1000$$

A zero in a quantity such as 473.0 mL is a significant figure and implies that the measurement has been made within the limits 472.95 mL and 473.05 mL or with a possible error of:

$$\frac{0.05}{473} \times 100 = (\text{approx}) 0.01\% \text{ or } 1 \text{ part in } 10,000$$

Thus, 473 is correct to the nearest mL, and 473.0 is correct to the nearest 0.1 mL.

Rules

1. When adding or subtracting, retain in the sum or remainder no more decimal places than the least number entering into the calculations. For example,

11.5 g	11.50 g
2.65 g	2.65 g
3.49 g	3.49 g
17.64 g	17.64 g
Answer: 17.6 g	Answer: 17.64 g

In the first column 11.5 g was weighed to 0.1 g or with an accuracy of ± 0.05 g. Although the other two weighings were made with an accuracy of ± 0.005 g, the sum can be expressed properly only to one decimal place.

In the second column 11.50 g was weighed to the nearest 0.01 g or with an accuracy of ± 0.005 g. Since all weighings were made with this degree of accuracy, the sum may be stated as in the example, 17.64 g.

Retain all figures possible until all the calculations are completed and then retain only the significant figures for the answer. Additions or subtractions involving both large and small quantities, each expressed with maximum significance, are often useless. For example, if one were to add 1.2 and 0.041 g, the physical sum would be 1.2 g, regardless of the fact that the two numbers add numerically to 1.241. To express the physical sum as 1.241 g would convey an erroneous degree of accuracy with which the quantity was known.

2. When multiplying or dividing, retain in the answer no more significant figures than the least number entering into the calculation. The meaning of this rule may be illustrated by the use of equivalents during conversions from one measuring system to another. Table 11-12 gives different equivalent values and the number of significant figures to which the answer is correct. Always use an equivalent that will give the desired degree of accuracy. Repeated multiplication of an approximation increases the error progressively; therefore, retain all figures during calculations and drop insignificant figures as the final step.

FRACTIONS

Common Fractions

An example of a common fraction is $\frac{3}{8}$. It is read as “three-eighths” and indicates three parts divided by eight parts of the same thing. The units with both numbers must be the same. Pharmacists measure $\frac{3}{8}$ of a fluidounce into a graduate, they measure 3 fluidrams, out of 8 fluidrams (a fluidounce contains 8 fluidrams).

The following principles should be applied when using common fractions:

1. The value of a fraction is not altered by multiplying or dividing both numerator and denominator by the same number.

Table 11-12.

WEIGHT(g)		EQUIVALENT WEIGHT (GR/G)		EQUIVALENT WEIGHT (GR)	SIGNIFICANT FIGURES
4.522	×	15.432	=	69.78	4
4.522	×	15.43	=	69.77	4
4.522	×	15.4	=	69.6	3
4.522	×	15	=	68	2

2. Multiplying the numerator or dividing the denominator by a number, multiplies the fraction by that number.
3. Dividing the numerator or multiplying the denominator by a number divides the fraction by that number.
4. To add or subtract fractions, form fractions with the *lowest common denominator*, perform the arithmetical operation, and reduce to the lowest common denominator.
5. To multiply fractions, multiply all numbers above the line to form the new numerator and multiply all numbers below the line to form the new denominator. Cancel if possible to simplify and reduce to the lowest common denominator.
6. To divide by a fraction, multiply by the reciprocal of the fraction.

Decimal Fractions

Fractions with the power of 10 as the denominator are known as *decimal fractions* and are written by omitting the denominator and inserting a decimal point in the numerator as many places from the last number on the right as there are ciphers of 10 in the denominator.

The following principles should be applied when using decimal fractions:

1. When adding or subtracting decimals, align the decimal points under each other.
2. When multiplying decimals, proceed as with whole numbers, then place the decimal point in the product as many places from the first number on the right as the sum of the decimal places in the multiplier and the multiplicand.
3. When dividing by a decimal fraction, move the decimal point to the right, in both divisor and dividend, as many places as it is to the left in the divisor to form a whole number in the divisor; proceed as with whole numbers. The decimal point in the quotient should be placed immediately above the decimal point in the dividend.
4. When converting a common fraction into a decimal fraction, divide the numerator by the denominator and place the decimal point in the correct place.
5. When converting a decimal fraction into a common fraction, place the entire number, as the numerator, over the power of 10 containing the same number of ciphers of 10 as there are decimal places. Cancel, if possible, to simplify.

EXPONENTS, POWERS, AND ROOTS

In the expression $2^4 = 16$, the following names are given to the terms: 16 is called the *power* of the *base* 2 and 4 is the *exponent* of the power. If the exponent is 1, it usually is omitted. The following laws should be recalled:

1. The product of two or more powers of the same base is equal to that base with an exponent equal to the sum of the exponents of the powers; eg, $2^5 \times 2^3 = 2^8$.
2. The quotient of two powers of the same base is equal to that base with an exponent equal to the exponent of the dividend minus the exponent of the divisor; eg, $2^8 \div 2^3 = 2^5$.
3. The power of a power is found by multiplying the exponents; eg, $(2^8)^3 = 2^{24}$.
4. The power of a product equals the product of the powers of the factors; eg, $(2 \times 3 \times 4)^2 = 2^2 \times 3^2 \times 4^2$.
5. The power of a fraction equals the power of the numerator divided by the power of the denominator; eg,

$$\left(\frac{2}{3}\right)^2 = \frac{2^2}{3^2}$$

The root of a power is found by dividing the exponent of the power by the index of the root; eg,

$$\sqrt[3]{3^6} = 3^{\frac{6}{3}} = 3^2$$

Any number other than 0 with an exponent 0 equals 1; eg, $2^0 = 1$. A number with a negative exponent equals one divided by the number with a positive exponent equal in numerical value to the negative exponent; for example,

$$2^{-4} = \frac{1}{2^4}$$

Logarithms

Logarithms (logs) were invented to facilitate the solution of involved and lengthy problems. Many calculations that are difficult by ordinary arithmetical processes are performed rapidly and easily with the aid of logs; the advent of modern calculators and computer spreadsheet programs has made this use of logs obsolete. Logs still appear, however, in many chemical and pharmacokinetic equations.

The log of a number is the exponent of the power to which a given base must be raised in order to equal that number.

$$Y = a^x$$

$$\log_a Y = x$$

John Napier, of Scotland, who discovered logs over three centuries ago, used the Natural Log Number, 2.71828+, as the base. Henry Briggs, using Napier's discovery a few years later, introduced 10 as the base, which is the most convenient for practical purposes. Napier's system is called natural logs, and Briggs' system is called common logs. In this latter system the natural numbers are regarded as powers of the base 10 and the corresponding exponents are the logs; eg,

$$6 = 10^{0.7782}$$

$$\log_{10} 6 = 0.7782$$

For natural logs,

$$6 = e^{1.792}$$

$$\ln_e 6 = 1.792$$

LAWS AND RULES

The following laws, governing the use of logs, are based on the laws of exponents, and hence hold for any log system.

1. The log of a product equals the *sum* of the log of the component numbers; for example, for 25×2 :

$$\log(25 \times 2) = \log 25 + \log 2 = 1.3979 + 0.3010 = 1.6989$$

2. The log of a quotient equals the log of the numerator minus the log of the denominator; for example, for $25 \div 2$:

$$\log(25/2) = \log 25 - \log 2 = \log 10^{1.3979} - \log 10^{0.3010}$$

$$= 1.3979 - 0.3010 = 1.0969$$

3. The log of a power of a number equals the log of the number multiplied by the exponent of the power; for example, for $(25)^{12}$:

$$\log(25)^{12} = 12 \log 25 = 12 \times 1.3979 = 16.7748$$

4. The log of a root of a number equals the log of the number divided by the index of the root; for example, for $\sqrt{25}$

$$\log \sqrt{25} = \log 25^{1/2} = \frac{\log 25}{2} = \frac{1.3979}{2} = 0.6990$$

5. The log of a negative power of a number equals the reciprocal of the number multiplied by the exponent of the power; for example, $(5)^{-2}$:

$$\log(5)^{-2} = -2 \log 5 = -2 \times 0.6990 = -1.398$$

The Log of a Number

The logarithm of a number can be easily obtained from a calculator or computer spreadsheet program.

1. Find the logarithm of 273.

$$\log 273 = 2.4362$$

$$\ln 273 = 5.6095$$

2. Find the logarithm of 0.08206.

$$\log 0.08206 = -1.08587$$

$$\ln 0.08206 = -2.5003$$

The Antilog of a Number

To find the number corresponding to a given log (or antilog), the reverse procedure of that discussed above is employed (ie, the appropriate numerical base is raised to the exponent expressed by the logarithm).

1. Find the number corresponding to the antilog 3.8357.

$$\log X = 3.8357$$

$$X = 10^{3.8357} = 6850$$

2. Find the number corresponding to the natural log 0.4351.

$$\ln X = 0.4351$$

$$X = e^{0.4351} = 2.71828^{0.4351} = 1.5451$$

3. Using the Henderson-Hasselbalch equation for an acidic substance, find the ratio of ionized to un-ionized drug at a pH of 3.0. The pK_a of the drug is 7.4.

$$pH = pK_a + \log \frac{[\text{Salt}]}{[\text{Acid}]}$$

$$\log \frac{[\text{Salt}]}{[\text{Acid}]} = pH - pK_a$$

$$\log \frac{[\text{Salt}]}{[\text{Acid}]} = 3.0 - 7.4 = -4.4$$

$$\frac{[\text{Salt}]}{[\text{Acid}]} = 10^{-4.4} = 3.98 \times 10^{-5}$$

Pharmaceutical Problems

The student who knows algebra, has studied the previous sections of this chapter, and recognizes the Roman numerals and Latin abbreviations used on prescription orders (for directions to the pharmacist and patient by the prescriber) should have sufficient knowledge to solve the routine problems encountered in a pharmacy. The various symbols and abbreviations and their meanings must be well understood. Explanation of practical problems, representative of those faced in practice, are presented below. Practice problems follow each section and the answers to these problems are found at the end of this chapter (page 125).

To solve each problem properly, the following procedure is suggested:

1. Analyze the problem carefully so that all data are clearly fixed in the mind; determine what is given and what is asked.
2. Select the most direct method of solving the problem. Not all problems can be solved properly in one step. Look up doses, equivalents, and abbreviations when you are not sure.
3. Prove or check the result.

Many problems encountered in pharmacy still utilize the apothecary and avoirdupois systems; however, solving these problems in contemporary practice is based on converting these systems into metric units prior to solving the problem mathematically. This approach will be followed in this text. Methods for the mathematical manipulation of apothecary and avoirdupois units and direct problem solving in these systems can be found in previous editions of *Remington*.

ADDITION

Review weighing and measuring systems discussed earlier in this chapter.

Rules

1. Add like quantities. Using the metric system, if the quantities are not alike, change them to a common unit.
2. When adding decimals, keep the decimal points directly under each other.
3. When adding fractions, reduce to the lowest common denominator (LCD), add the resulting numerators, and reduce the fraction, if possible, by canceling.

Examples

1. Add 3 kg, 33 g, and 433 mg.
Convert to a common unit. The gram is convenient because it is the unit of weight.

$$3 \text{ Kg} \times \frac{1000 \text{ g}}{\text{Kg}} = 3000 \text{ g}$$

$$33 \text{ g} = 33 \text{ g}$$

$$433 \text{ mg} \times \frac{1 \text{ g}}{1000 \text{ mg}} = 0.433 \text{ g}$$

Answer = 3033.433 g

Problems

1. Add 25 mg, 25 g, 210 mg, 2 kg, 1.75 g, 215 mg, 454 g, and 30 mg.
2. The following quantities of a drug were removed from a container: 31 g, 225 g, 855.6 g, and 45.4 g. What is the total weight removed from the container?

SUBTRACTION

Rules

1. Subtract only like quantities. If the quantities are not alike, change to a common unit.
2. Treat common and decimal fractions as indicated in the section on addition.

Examples

1. Subtract 285 mL from 1 L. Convert to a common unit.

$$\begin{array}{r} 1000 \text{ mL} \\ - 285 \text{ mL} \\ \hline 715 \text{ mL} \end{array}$$

Answer: 715 mL

Problems

1. How much is left in a 5 L container after the removal of 895 mL?
2. A pharmacist buys 5 g of a potent drug and at different times dispenses 0.2 g, 0.85 g, 90 mg, and 150 mg on prescription orders. How much of the drug remains?

MULTIPLICATION

Rules

1. The product has the same denomination as the multiplicand.
2. If the multiplicand is composed of different denominations in the metric system, form a common unit before multiplying and reduce the product to measurable units.
3. Multiply fractions and decimals as in any arithmetic problem, and reduce fractional quantities to measurable or weighable units.

Examples

1. What will be the total weight of the ingredients in a prescription order for 25 units, each unit containing 0.4 g of Solid F, 0.01 g of

Solid G, and 5 mg of Solid H? First, convert to a common unit such as grams.

$$0.4 \text{ g} + 0.01 \text{ g} + 0.005 \text{ g} = 0.415 \text{ g total weight of one unit}$$

$$0.415 \text{ g/unit} \times 25 \text{ units} = 10.375 \text{ g total weight of all units}$$

2. Multiply 22.4 mL by 2.65.

$$\begin{array}{r} 22.4 \text{ mL} \\ \times 2.65 \\ \hline 59.36 \text{ mL} \end{array}$$

Problems

1. Multiply 48.5 mL by 3.24.
2. A certain preparation is to contain 0.0325 g of a chemical in each mL of solution. How much must be weighed out to make 5 L of the solution?
3. How much cod liver oil is necessary to make 2500 capsules, each containing 0.33 mL?
4. How many mg are used to make 1500 units, each of which contains 250 μg of a drug?

DIVISION

Rules

1. The quotient always has the same denomination as the dividend.
2. If the dividend is composed of different denominations, form a common unit in the metric system before dividing and reduce the quotient to weighable or measurable quantities.
3. Treat fractions, and decimals as explained in the multiplication section.

Examples

1. Divide 3 L by 25.

$$\frac{3\text{L}}{25} = 0.120 \text{ L or } 120 \text{ mL}$$

Problems

1. How many 65 mg capsules can be made from 50 g of a drug?
3. The dose of a drug is 0.1 mg. How many doses are contained in 15 mg of the drug?
5. How many 325 mg capsules of a drug can be filled from a 454 g amount?

CONVERSION

As long as the student knows the interrelationships of the various units within the different weighing and measuring systems (eg, 20 gr = 1 \mathfrak{D} , 3 \mathfrak{D} = 1 \mathfrak{z} ; 1000 mg = 1 g), there are only three conversions necessary to memorize in order to convert between the apoth, avoird, and metric systems. These are

$$1 \text{ gr (avoird)} = 1 \text{ gr (apoth)}$$

$$1 \text{ gr} = 64.8 \text{ mg}$$

$$1 \text{ f}\mathfrak{z} = 29.6 \text{ mL}$$

Learn them!

With these three conversions the student is able to derive all other necessary conversions.

Apothecary Conversions

Various equalities within the apothecary system may be calculated.

1. The number of grains in a dram, grains in a pound, and so on may be calculated using the following steps.

$$\frac{20 \text{ gr}}{\mathfrak{D}} \times \frac{3\mathfrak{D}}{3} = \frac{60 \text{ gr}}{3}$$

$$\frac{60 \text{ gr}}{3} \times \frac{8\mathfrak{z}}{3} \times \frac{12\mathfrak{z}}{\mathfrak{lb}} = \frac{5760 \text{ gr}}{\mathfrak{lb}}$$

Cancel the units. If they do not cancel properly, something has been omitted.

2. Convert 1 ℥ (apoth) to weighable quantities in the avoirdupois system

$$1 \text{ gr (apoth)} = 1 \text{ gr (avoird)}.$$

Since 1 gr (apoth) = 1 gr (avoird), the number of grains in one system equals the number of grains in the other system; e.g., 480 gr (apoth) = 480 gr (avoird).

$$\frac{20 \text{ gr}}{\text{℥}} \times \frac{3 \text{ ℥}}{3} \times \frac{8 \text{ ℥}}{3} = \frac{480 \text{ gr}}{3 \text{ (apoth)}}$$

$$480 \text{ gr (apoth)} = 480 \text{ gr (avoird)}$$

$$437.5 \text{ gr} = 1 \text{ oz avoird}$$

$$\begin{array}{r} 480 \text{ gr} \\ - 437.5 \text{ gr} \\ \hline 42.5 \text{ gr} \end{array}$$

Answer: 1 ℥ (apoth) = 1 oz, 42.5 gr (avoird).

3. Conversions in the metric system are made in the same manner. Convert 1 g to mg.

$$1 \text{ g} \times \frac{1000 \text{ mg}}{\text{g}} = 1000 \text{ mg}$$

Convert 1 g to kg.

$$1 \text{ g} \times \frac{1 \text{ kg}}{1000 \text{ g}} = 0.001 \text{ kg}$$

The same procedure is valid for volume measurements in the metric system.

4. Conversions between the apothecary and metric weight systems can be based on the conversion factor; 15.4 gr = 1 g, which may be restated as 15.4 gr/g or 1 g/15.4 gr.

- a. How many mg equal 1 gr?

$$\frac{1 \text{ g}}{15.4 \text{ gr}} = 0.0648 \text{ g / gr} = 64.8 \text{ mg / gr} \text{ or } 64.8 \text{ mg} = 1 \text{ gr}$$

- b. How many grams are in 1 ℥?

$$\frac{1 \text{ g}}{15.4 \text{ gr}} \times \frac{480 \text{ gr}}{3} = \frac{311 \text{ g}}{3}$$

- c. How many grams are in 1 oz (avoird)? Remember: 1 gr (apoth) = 1 gr (avoird).

$$\frac{1.000 \text{ g}}{15.4 \text{ gr}} \times \frac{437.5 \text{ gr}}{\text{oz}} = \frac{28.4 \text{ gr}}{\text{oz}}$$

- d. Other weight conversions are then found in a similar manner.

5. Conversions between the apothecary and metric measuring systems can be based on the conversion factor; 1 f℥ = 29.6 mL, which may be restated as 1 f℥/29.6 mL or 29.6 mL/f℥.

- a. How many ℥ are in 1 mL?

$$\frac{480 \text{ ℥}}{\text{f℥}} \times \frac{1 \text{ f℥}}{29.6 \text{ mL}} = \frac{16.2 \text{ ℥}}{\text{mL}}$$

Rules

- The USP states that for prescription compounding one uses practical equivalents, defined as exact equivalents rounded to three (3) significant figures.
- To calculate quantities required in pharmaceutical formulas, the USP directs the use of practical equivalents.
- In converting doses the USP uses approximate equivalents. Use USP tables wherever possible.

Examples

1. Convert 1 pt, 4 f℥ into mL.
First, convert into f℥.

$$\frac{16 \text{ f℥}}{\text{pint}} \times 1 \text{ pint} + 4 \text{ f℥} = 20 \text{ f℥}$$

Second, convert f℥ to mL.

$$20 \text{ f℥} \times \frac{29.6 \text{ mL}}{\text{f℥}} = 592 \text{ mL}$$

Answer: 1 pt, 4 f℥ = 592 mL.

2. What is the weight of 1200 g in the apothecary system?

$$1200 \text{ g} \times \frac{15.4 \text{ gr}}{\text{g}} = 18,480 \text{ gr}$$

Or:

$$1200 \text{ g} \times \frac{1 \text{ lb}}{373 \text{ g}} = 3.22 \text{ lb}$$

3. Convert 1 pound (apoth) into grams.

$$\frac{1 \text{ g}}{15.4 \text{ gr}} \times \frac{480 \text{ gr}}{3} \times \frac{12 \text{ ℥}}{1 \text{ lb}} = 374 \text{ g}$$

4. Convert 25 gr to grams.

$$25 \text{ gr} \times \frac{1 \text{ g}}{15.4 \text{ gr}} = 1.62 \text{ g}$$

5. Convert 50 grams to grains.

$$50 \text{ g} \times \frac{15.4 \text{ gr}}{\text{g}} = 770 \text{ gr}$$

Problems

- Convert:
 - 6.50 grains into milligrams.
 - 3/10 grain into milligrams.
 - 3 1/2 apoth ounces into grams.
 - 2 ℥ into mg.
 - 3 1/2 avoird ounces into grams.
 - 1 lb avoird into grams.
- Convert:
 - 550 g into weighable quantities in the avoirdupois system.
 - 450 mg into grains.
 - 550 g into weighable quantities in the apoth system.
 - 100 μg into grains.
 - 1 kg into lb (avoird).
- Convert the following doses into metric weights:
 - 1/100 gr.
 - 1/320 gr.
 - 1/6 gr.
 - 5 gr.
 - 20 gr.
- Convert:
 - 200 m into mL.
 - 3 f℥ into mL.
 - 8 f℥ into mL.
 - 1 pt into mL.
 - 5 ℥ into mL.
 - 0.1 mg into gr.
 - 5 mg into gr.
- Answer the following questions.
 - How many gr are in 1 ℥?
 - How many drams are in 1 ℥?
 - How many grains are in 1 oz (avoird)?
 - How many gr are in 1/2 lb (apoth)?
 - Convert 250 gr to weighable quantities in the apothecary system.

HOUSEHOLD EQUIVALENTS

Common household equivalents are found on page 104. These are used to interpret the prescriber's instructions to the patient. The teaspoonful usually is indicated by the symbol f℥ or 5 mL, although 1 f℥ does not equal 5 mL. The problem of "the teaspoonful" has been discussed by Morrell and Ordway⁴ and by Madlon-Kay and Mosch⁵. For practical purposes, a

teaspoonful is equal to 5 mL, and 1 f3 in the directions to the patient on the prescription means 1 teaspoonful.

For purposes of solving most compounding and dispensing problems, the exact equivalents rounded to three significant places should be used.

DOSAGE CALCULATIONS

Over the past years various rules for calculating infants' and children's dosages have been devised. All of them give only approximate dosages because they erroneously assume that the child is a small adult; some of them are still used because as yet no absolute method of calculating an infant or child's dose has been found. Children are sometimes more susceptible than adults to certain drugs. Doses for infants and children, where they are known, may be found in the USP Drug Information, the *Pediatric Dosage Handbook* published by APhA, and textbooks on pediatrics.⁶⁻⁸ Doses should not be calculated when it is possible to obtain the actual infant or child's dose.

Rules for Approximate Doses for Infants and Children

1. *Young's Rule* (for children 2 years old and older)

$$\frac{\text{Age (years)}}{\text{Age (years)} + 12} \times \text{Adult dose} = \text{Child's dose (approx)}$$

2. *Clark's Rule*

$$\frac{\text{Weight (lb)}}{150} \times \text{Adult dose} = \text{Child's dose (approx)}$$

3. *Fried's Rule* (for infants up to 2 years old)

$$\frac{\text{Age (months)}}{150} \times \text{Adult dose} = \text{Child's dose (approx)}$$

4. *The Square Meter Surface Area Method* relates the surface area of individuals to dose. It is thought that this is a more realistic way of relating dosages.

$$\frac{\text{Body surface area of child}}{\text{Body surface area of adult}} \times \text{Adult dose} = \text{Child's dose (approx)}$$

The average body surface area for an adult has been given as 1.73 square meters (m²); hence,

$$\frac{\text{Body surface area of child (m}^2\text{)}}{1.73 \text{ m}^2} \times \text{Adult dose} = \text{Child's dose (approx)}$$

Calculating Doses for Individuals—of any age or size

Many drugs have doses stated as the amount of *drug/m² body surface area* and may be calculated as follows:

$$\frac{\text{Dose of drug}}{\text{m}^2 \text{ body surface area}} \times \text{Body surface area (m}^2\text{)} = \text{Dose}$$

Many physiological functions are proportional to body surface area, such as metabolic rate and kidney function.

Drug doses are often stated in *mg/kg body weight* and may be calculated as follows:

$$\frac{\text{Dose of drug}}{\text{kg body weight}} \times \text{Body weight (kg)} = \text{Dose}$$

This is the most common way of determining children's doses.

Drug doses also may be stated in *units*, as with vitamins A and D, penicillin, and hormones. This means that a certain quantity of biological activity of that drug is called 1 unit. When

the term unit is used in connection with a drug, the calculations involved are the same as those for more familiar weight or volume notations. The USP often standardizes the unit for such drugs, so the expression "USP Units" is used. This means the units are calculated based on a USP assay procedure and reference standard.

Examples

1. The adult dose of a drug is 325 mg. What is the dose for a 3-year-old child?

Use Young's Rule:

$$\text{Child's dose (approx)} = \frac{3}{3 + 12} \times 325 \text{ mg} = 65 \text{ mg}$$

2. What is the dose for a 40 lb child if the average adult dose of the medication is 10 mg?

Use Clark's Rule:

$$\text{Child's dose (approx)} = \frac{40}{150} \times 10 \text{ mg} = 2.67 \text{ mg}$$

3. What is the dose for an 8-month-old infant if the average adult dose of a drug is 250 mg?

Use Fried's Rule:

$$\text{Infant's dose (approx)} = \frac{8}{150} \times 250 \text{ mg} = 13.3 \text{ mg}$$

4. If the average adult dose of a drug is 50 mg, what is the dose for a child who has a body surface area equal to 0.57 m²?

$$\text{Child's dose (approx)} = \frac{0.57}{1.73} \times 50 \text{ mg} = 16.5 \text{ mg}$$

Problems

1. What is the dose of a drug for a 9-month-old infant if the average adult dose is 25 mg?
2. What is the dose of a drug for a 6-year-old child if the average adult dose is 98 mg?
3. What is the dose of a drug for a child who weighs 28 lb if the average adult dose is 100 mg?
4. What is the dose of a drug for an individual who has a 1.21 m² body surface area? The average adult dose is 400,000 units.
5. What is the dose of a medication for a child that weighs 66 lb if the dose is stated as 2.5 mg/kg body weight?
6. What is the dose of a drug for an average adult patient if the dose of the drug is 45 mg/m²?

PROBLEM-SOLVING METHODOLOGY

The problem-solving method illustrated in solving pharmaceutical problems is *dimensional analysis* (which is based on *ratio and proportion*). Dimensional analysis is widely used in many scientific disciplines and offers a consistent way to solve problems. Dimensional analysis also overcomes many difficulties students and pharmacy practitioners have in problem interpretation and provides a well-defined, consistent starting point in the solution of pharmaceutical problems.

DIMENSIONAL ANALYSIS

The basis for dimensional analysis is the formation of relationships between quantities, multiplication and canceling units until only the units of the desired answer remain.

As an example, if 100 g of a drug cost \$1.80, how much will 25 g cost?

Begin by collecting all of the information in the problem and identify all relationships with units and labels. In this problem, we know the following:

$$\frac{\$1.80}{100 \text{ g drug}}, 25 \text{ g drug}$$

Identify the units you want for the answer.

$$= \$$$

Identify a relationship from the problem that contains the unit(s) desired for the answer, forming the skeleton of the process.

$$\frac{\$1.80}{100 \text{ g drug}} \times ? = \$$$

Complete the process by using terms from the problem (or equivalents) necessary to cancel out units until only the unit(s) of the answer remain on the left side.

$$\frac{\$1.80}{100 \text{ g drug}} \times 25 \text{ g drug} = \$$$

Solve mathematically.

$$\text{Answer} = \$0.45$$

Dimensional analysis can be used to solve most pharmaceutical problems, regardless of complexity, using a consistent procedure:

1. Collect all the information and relationships in the problem complete with units and labels.
2. Identify the unit(s) and label of the answer.
3. Select a starting point corresponding to the unit(s) and label of the answer in the numerator.
4. Complete the process using relationships in the problem and known conversions to cancel units.
5. Solve the problem mathematically.

More complex problems use the same basic procedure; eg, if 100 g of a drug cost \$1.80, what would be the cost of the drug to prepare 4 f $\bar{3}$ of a solution containing 5 g of the drug per teaspoonful?

Step 1: Collect all information and relationships:

$$\frac{\$1.80}{100 \text{ g drug}}, \frac{5 \text{ g drug}}{1 \text{ tsp}}, 4 \text{ f}\bar{3}$$

Step 2:

$$= \$$$

Step 3:

$$\frac{\$1.80}{100 \text{ g drug}} \times ? = \$$$

Step 4:

$$\frac{\$1.80}{100 \text{ g drug}} \times \frac{5 \text{ g drug}}{1 \text{ tsp}} \times \frac{1 \text{ tsp}}{5 \text{ mL}} \times \frac{29.6 \text{ mL}}{1 \text{ f}\bar{3}} \times 4 \text{ f}\bar{3} = \$$$

(The 3rd and 4th terms are known definitions and equivalents needed to cancel units.)

Step 5:

$$\text{Answer} = \$0.53$$

With practice, steps 2 through 4 can be written in one operation.

Examples

1. Determine the amount of each ingredient contained in one dose of the following prescription.

℞ Solid A 300 mg
Solid B 150 mg
Solid C 200 mg

M ft capsules, D.T.D. No 12.

The directions to the pharmacist are to mix and make 12 capsules each containing in the three solids in the amounts indicated. Thus, the dose of each ingredient is as stated in the prescription.

2. How much of each ingredient is contained in one dose of the following prescription?

℞ Solid E 7.2 g
Solid F 0.24 g
Solid G 1.2 g

M div capsules, No 24.

In this prescription the prescriber requests that 24 capsules be made from the three ingredients. The amounts of the ingredients requested are considerable, and drugs usually do not have doses of 7.2 g or 1.2 g, so division of the amounts by the number of doses (24) is required. The pharmacist should check a textbook or compendium to confirm the average adult dose.

$$\text{Drug E: } \frac{7.2 \text{ g}}{24 \text{ capsules}} \times 1 \text{ capsule} = 0.300 \text{ g}$$

$$\text{Drug F: } \frac{0.24 \text{ g}}{24 \text{ capsules}} \times 1 \text{ capsule} = 0.010 \text{ g}$$

$$\text{Drug G: } \frac{1.2 \text{ g}}{24 \text{ capsules}} \times 1 \text{ capsule} = 0.050 \text{ g}$$

3. A prescription calls for 10 units of a drug to be taken 3 times a day. How much will the patient have taken after 7 days?

$$\frac{10 \text{ units}}{\text{dose}} \times \frac{3 \text{ doses}}{\text{day}} \times 7 \text{ days} = 210 \text{ units}$$

4. If 250 units of an antibiotic weigh 1 mg, how many units are in the 15 mg?

$$\frac{250 \text{ units}}{\text{mg}} \times 15 \text{ mg} = 3750 \text{ units}$$

5. If the dose of a drug is 0.5 mg/kg of body weight/24 hours, how many grams will a 33-lb infant receive per 24 hours and per week?

$$\frac{1 \text{ g}}{1000 \text{ mg}} \times \frac{0.5 \text{ mg}}{\text{kg} \times 24 \text{ hours}} \times \frac{1 \text{ kg}}{2.2 \text{ lb}} \times 33 \text{ lb} \times 24 \text{ hours} = 0.00750 \text{ g}$$

$$\frac{0.00750 \text{ g}}{\text{day}} \times \frac{7 \text{ days}}{\text{week}} \times 1 \text{ week} = 0.0525 \text{ g}$$

6. A patient is to receive 260 μg of a drug 4 times a day for 14 days. How many 1/250-gr tablets must be dispensed?

$$\frac{1 \text{ tablet}}{\frac{1}{250} \text{ gr}} \times \frac{1 \text{ gr}}{64.8 \text{ mg}} \times \frac{1 \text{ mg}}{1000 \mu\text{g}} \times \frac{260 \mu\text{g}}{\text{dose}} \times \frac{4 \text{ doses}}{\text{day}}$$

$$\times 14 \text{ days} = 56.2 \text{ tablets} = 57 \text{ tablets}$$

7. An antibiotic is available as an injection containing 10 mg antibiotic/mL. How many mL are needed for an infant weighing 8 kg, the dose being 1.4 mg/kg of body weight?

$$\frac{1 \text{ mL}}{10 \text{ mg}} \times \frac{1.4 \text{ mg}}{\text{kg}} \times 8 \text{ kg} = 1.12 \text{ mL}$$

8. A preparation for coughs contains 1.5 g of an expectorant per 100 mL. How many gr of the expectorant are there in a teaspoonful?

$$1 \text{ tsp} = 5 \text{ mL}$$

$$\frac{15.4 \text{ gr}}{1 \text{ g}} \times \frac{1.5 \text{ g}}{100 \text{ mL}} \times \frac{5 \text{ mL}}{1 \text{ tsp}} \times 1 \text{ tsp} = 1.16 \text{ gr}$$

Problems

1. Calculate the dose for each ingredient in the following prescription.

℞ Chemical J 10 mg
Chemical K 50 mg
Chemical L 300 mg
M ft capsules, D.T.D. No 14.

2. Calculate the dose of each ingredient in the following prescription.

℞ Drug Q 10.5 g
Drug R 6.3 g
M div 21 doses.

- An 8 f̄ prescription contains 6 f̄ of a tincture. If 1 teaspoonful 4 times a day is prescribed, how much tincture does the patient take per dose and how much is taken daily?
- How many 0.3-mL doses are contained in 15 mL of a solution?
- If 1 mg of a hormone equals 22.5 units, how many mg are required to obtain 1 unit?
- If a bottle contains 80 units of a drug/mL, how many mL must the patient take to get a 60-unit dose? If the bottle contains 10 mL total volume of the drug solution, how many days' supply will patients have if they use 60 units a day?
- A 10-mL ampule contains a 2.5% solution of a drug. How many mL are needed to give a dose of 150 mg?
- The dose of an antibiotic is 75 mg for a child. How much of a flavored suspension containing 125 mg antibiotic/5 mL must be given to the child per dose?
- How many mg of a drug are there in each teaspoonful of a syrup that contains 0.5% of the drug?

REDUCING AND ENLARGING FORMULAS

Determine the total weight or volume of ingredients and convert, if necessary, to the system of the quantities desired. The quantities in the original and new formulas will have the same ratio.

Examples

- The formula for a syrup is

Drug M	140 g
Sucrose	450 g
Purified Water qs	1000 mL

- Find the quantities required for 100 mL.

$$\text{Drug M: } \frac{140 \text{ g}}{1000 \text{ mL}} \times 100 \text{ mL} = 14.0 \text{ g}$$

$$\text{Sucrose: } \frac{450 \text{ g}}{1000 \text{ mL}} \times 100 \text{ mL} = 45.0 \text{ g}$$

Purified Water: to make 100 mL

- What quantities are required to compound 60 mL of the syrup?

$$\text{Drug M: } \frac{140 \text{ g}}{1000 \text{ mL}} \times 60 \text{ mL} = 8.40 \text{ g}$$

$$\text{Sucrose: } \frac{450 \text{ g}}{1000 \text{ mL}} \times 60 \text{ mL} = 27.0 \text{ g}$$

Purified Water: to make 60 mL

- Calculate the amounts needed for 100 g of antiseptic powder as follows:

℞	Solid A	2 g
	Solid B	1 g
	Solid C	7 g
	Solid D	25 g
	Solid E	115 g
		150 g

$$\text{Factor} = \frac{100 \text{ g}}{150 \text{ g}} = 0.667$$

$$\text{Solid A: } 2 \text{ g} \times 0.667 = 1.33 \text{ g}$$

$$\text{Solid B: } 1 \text{ g} \times 0.667 = 0.667 \text{ g}$$

$$\text{Solid C: } 7 \text{ g} \times 0.667 = 4.67 \text{ g}$$

$$\text{Solid D: } 25 \text{ g} \times 0.667 = 16.7 \text{ g}$$

$$\text{Solid E: } 115 \text{ g} \times 0.667 = 76.7 \text{ g}$$

- Prescriptions, where the instruction to the pharmacist calls for making a certain number of doses of an ingredient or mixture of

several ingredients, are a type of formula enlargement. The expression usually used is DTD, which means let such doses be given. Occasionally the prescriber will not use this expression, but inspection of amounts of the ingredients indicates that this is what is desired. For example,

℞	Solid H	50 mg
	Solid K	150 mg
	Liquid N	0.2 mL
	M ft capsules,	D.T.D. No 24.

The pharmacist checked the individual doses of the ingredients and found them to be slightly below the average adult dose, confirming that the prescriber wanted the quantities listed to be multiplied by 24.

$$\text{Solid H: } \frac{50 \text{ mg}}{\text{capsule}} \times 24 \text{ capsules} = 1200 \text{ mg or } 1.2 \text{ g}$$

$$\text{Solid K: } \frac{150 \text{ mg}}{\text{capsule}} \times 24 \text{ capsules} = 3600 \text{ mg or } 3.6 \text{ g}$$

$$\text{Liquid N: } \frac{0.2 \text{ mL}}{\text{capsule}} \times 24 \text{ capsules} = 4.8 \text{ mL}$$

Problems

- The formula for a liquid preparation is

Liquid C	35 mL
Solid B	9 g
Liquid R	2.5 mL
Liquid P	20 mL
Purified Water, sufficient to make	100 mL

Calculate the quantities of the ingredients to make 2.5 L.

- The formula for an ointment is

℞	Solid G	1
	Liquid D	30
	Solid M	3
	Ointment base, sufficient to make	100

Calculate quantities of the ingredients for 2 lb (apoth).

- How much of each of the three solids and how much purified water are needed to properly compound the following prescription order?

℞	Solid N	0.1 mg
	Solid Q	2.5 mg
	Solid R	150.0 mg
	Purified Water, qs	5 mL
	M ft solution,	D.T.D. No 48.

- How much of each ingredient is required to compound 90 mL of the following product?

Solid S	7.5 g
Solid T	25 g
Oil C	350 mL
Alcohol	250 mL
Purified Water, qs	1000 mL

PERCENTAGE

Percent, written as %, means per hundred. Fifteen percent is written 15% and means 15/100, 0.15, or 15 parts in a total of 100 parts. Percent is a type of ratio and has units of parts per 100 parts. Thus, 10% of 1500 tablets is 10/100 × 1500 tablets = 150 tablets.

To change percent to a fraction, the percent number becomes the numerator and 100 is the denominator. To change a fraction to percent, put the fraction in a form having 100 as its denominator; multiply by 100 so that the numerator becomes the percent.

$$\frac{1}{2} = \frac{50}{100}; \frac{50}{100} \times 100 = 50\%$$

$$\frac{1}{8} = \frac{12.5}{100}; \frac{12.5}{100} \times 100 = 12.5\%$$

Calculations involving percentages are encountered continually by pharmacists. They must be familiar not only with the arithmetical principles, but also with certain compendial inter-

pretations of the different type percentages involving solutions and mixtures.

The USP states

Percentage concentrations of solutions are expressed as follows:

Percent weight in weight—(w/w) expresses the number of g of a constituent in 100 g of product.

Percent weight in volume—(w/v) expresses the number of g of a constituent in 100 mL of product, and is used regardless of whether water or another liquid is the solvent.

Percent volume in volume—(v/v) expresses the number of mL of a constituent in 100 mL of product.

The term *percent* used without qualification means, for mixtures of solids, percent weight in weight; for solutions or suspensions of solids in liquids, percent weight in volume; for solutions of liquids in liquids, percent volume in volume; and for solutions of gases in liquids, percent weight in volume. For example, a one percent solution is prepared by dissolving one g of a solid or one mL of a liquid in sufficient of the solvent to make 100 mL of the solution.

Ratio Strength

Ratio strength is another manner of expressing concentration. Such phrases as “1 in 10” are understood to mean that one part of a substance is to be diluted with a diluent to make 10 parts of the finished product. For example, a 1:10 solution means 1 mL of a liquid or one g of a solid dissolved in sufficient solvent to make 10 mL of solution. Ratio strength can be converted to percent by:

$$\frac{1 \text{ g substance}}{10 \text{ mL solution}} \times 100 \text{ mL solution} = 10 \text{ g substance}$$

$$\frac{10 \text{ g substance}}{100 \text{ mL solution}} = 10\%$$

The expression “parts per thousand” (eg, 1:5000) always means parts weight in volume when dealing with solutions of solids in liquids and is similar to the above expression. A 1:5000 solution means 1 g of solute in sufficient solvent to make 5000 mL of solution. This can be converted to percent by

$$\frac{1 \text{ g substance}}{5000 \text{ mL solution}} \times 100 \text{ mL solution} = 0.02 \text{ g substance}$$

$$\frac{0.02 \text{ g substance}}{100 \text{ mL solution}} = 0.02\%$$

The expression “trituration” has two different meanings in pharmacy. One refers to the process of particle-size reduction, commonly by grinding or rubbing in a mortar with the aid of a pestle. The other meaning refers to a dilution of a potent powdered drug with a suitable powdered diluent in a definite proportion by weight. It is the second meaning that is used in this chapter.

When pharmacists refer to a “1 in 10 trituration” they mean a mixture of solids composed of 1 g of drug plus sufficient diluent (another solid) to make 10 g of mixture or *dilution*. In this case the “1 in 10 trituration” is actually a solid dilution of a drug with an inert solid. The strength of a trituration may also be stated as percent w/w . Thus, the term trituration has come to mean a solid dilution of a potent drug with a chemically and physiologically inert solid.

The meanings implied by the USP statements in the section on percentage are illustrated below with a few examples of the three types of percentages.

Weight-in-Volume Percentages

This is the type of percent problem most often encountered on prescriptions. The volume occupied by the solute and the volume of the solvent are *not* known because sufficient solvent is added to make a given or known final volume.

EXAMPLES

1. Prepare 1 f $\bar{3}$ of a 10% solution.
Since this is a solution of a solid in a liquid, this is a w/v solution.

$$\frac{10 \text{ g drug}}{100 \text{ mL soln}} \times \frac{29.6 \text{ mL}}{1 \text{ f}\bar{3}} \times 1 \text{ f}\bar{3} = 2.96 \text{ g drug}$$

2.96 g is dissolved in sufficient purified water to make 29.6 mL of solution.

2. How much of a drug is required to compound 4 f $\bar{3}$ of a 3% solution in alcohol?

$$\frac{3 \text{ g drug}}{100 \text{ mL soln}} \times \frac{29.6 \text{ mL}}{1 \text{ f}\bar{3}} \times 4 \text{ f}\bar{3} = 3.55 \text{ g drug}$$

3. How much 0.9% solution of sodium chloride can be made from $\frac{1}{2}$ $\bar{3}$ of NaCl?

$$\frac{100 \text{ mL soln}}{0.9 \text{ g NaCl}} \times \frac{31.1 \text{ g}}{1 \bar{3}} \times 0.5 \bar{3} = 1730 \text{ mL soln}$$

4. How many grams of a drug are required to make 120 mL of a 25% solution?

$$\frac{25 \text{ g drug}}{100 \text{ mL soln}} \times 120 \text{ mL} = 30 \text{ g drug}$$

5. How would you prepare 480 mL of a 1 in 750 solution of an antiseptic?

Remember: percent w/v is indicated.

1 in 750 means 1 g of the antiseptic dissolved in sufficient solvent to make 750 mL solution.

$$\frac{1 \text{ g drug}}{750 \text{ mL soln}} \times 480 \text{ mL} = 0.64 \text{ g drug}$$

Dissolve 0.64 g of antiseptic in sufficient solvent to make 480 mL solution.

6. How much of a substance is needed to prepare 1 L of a 1:10,000 solution?

The ratio 1:10,000 means 1 g of a substance in 10,000 mL of solution.

$$\frac{1 \text{ g substance}}{10,000 \text{ mL soln}} \times \frac{1000 \text{ mL}}{1 \text{ L}} \times 1 \text{ L} = 0.1 \text{ g substance}$$

7. How would you prepare 120 mL of 0.25% solution of neomycin sulfate? The source of neomycin sulfate is a solution which contains 1 g neomycin sulfate/10 mL.

$$\frac{10 \text{ mL stock soln}}{1 \text{ g drug}} \times \frac{0.25 \text{ g drug}}{100 \text{ mL soln}} \times 120 \text{ mL soln} = 3 \text{ mL stock soln}$$

Add sufficient purified water to 3 mL of stock solution to make 120 mL.

Problems

1. How would you make 3 f $\bar{3}$ of a 12.5% solution?
2. How many liters of a 4% solution can be made from 4 $\bar{3}$ of a solid?
3. How many liters of an 8% solution can be made from 500 g of a solid?
4. How many grams of a drug are needed to make 4 L of a 1 in 500 solution?

Weight-in-Weight Percentages

Density must be considered in some of these problems. If a weight-in-weight solution is requested on a prescription, both the solute and solvent must be weighed, or the solute and the solvent may be measured if their densities are taken into consideration in determining the volumes. Since the solutions are made to a given weight, a given volume is not always obtainable.

EXAMPLES

1. What weights of solute and solvent are required to make $2\frac{3}{4}$ of a 3% *w/w* solution of a drug in 90% alcohol?

$$\frac{3 \text{ g solute}}{100 \text{ g soln}} \times \frac{31.1 \text{ g soln}}{1\frac{3}{4} \text{ soln}} \times 2\frac{3}{4} \text{ soln} = 1.87 \text{ g solute}$$

$$\frac{31.1 \text{ g soln}}{1\frac{3}{4} \text{ soln}} \times 2\frac{3}{4} \text{ soln} = 62.2 \text{ g soln}$$

$$62.2 \text{ g soln} - 1.87 \text{ g solute} = 60.3 \text{ g solvent}$$

2. The solubility of boric acid is 1 g in 18 mL of water at 25°C. What is the percentage strength, *w/w*, of a saturated solution? 1 g of boric acid + 18 mL of water make a saturated solution, 18 mL of water weighs 18 g; hence, the weight of solution is 19 g. The amount of boric acid present is 1 g in 19 g of solution; therefore, the following relationship can be set up:

$$\frac{1 \text{ g drug}}{19 \text{ g soln}} \times 100 \text{ g soln} = 5.26 \text{ g drug}$$

$$\frac{5.26 \text{ g drug}}{100 \text{ g soln}} = 5.26\%$$

3. How many grams of a chemical are needed to prepare 200 g of a 10% *w/w* solution?
10% *w/w* means 10 g of solute in 100 g total solution. The following relationship may be set up:

$$\frac{10 \text{ g solute}}{100 \text{ g soln}} \times 200 \text{ g soln} = 20 \text{ g solute}$$

4. How would one make a 2% *w/w* solution of a drug in 240 mL of alcohol? The density of alcohol is 0.816 g/mL.
a. First, convert 240 mL to weight. Remember: alcohol is the solvent and it has a density different from that of water.

$$\frac{0.816 \text{ g alcohol}}{1 \text{ mL alcohol}} \times 240 \text{ mL alcohol} = 195.8 \text{ g (196 g) alcohol}$$

- b. 2% *w/w* means 2 g solute in 100 g solution. In this problem the final weight of solution is not known; 240 mL (196 g) of alcohol represents the solvent only. The solvent is 98% *w/w* of the total solution, so the following relationship may be set up:

$$\frac{2 \text{ g solute}}{98 \text{ g alcohol}} \times 196 \text{ g alcohol} = 4.00 \text{ g solute}$$

- c. Dissolve 4.00 g of the drug in 240 mL alcohol. The resulting solution will be 2% *w/w* and have a volume slightly larger than 240 mL because of the volume displacement of the drug.

5. How much of a 5% *w/w* solution can be made from 28.4 g of a chemical?

$$\frac{100 \text{ g soln}}{5 \text{ g chemical}} \times 28.4 \text{ g chemical} = 568 \text{ g soln}$$

6. How many mL of a 70% *w/w* solution having a density of 1.2 g/mL will be needed to prepare 600 mL of a 10% *w/v* solution?
a. Drug needed

$$\frac{10 \text{ g drug}}{100 \text{ mL soln (10\%)}} \times 600 \text{ mL soln (10\%)} = 60 \text{ g drug}$$

- b. Weight of 70% solution needed

$$\frac{100 \text{ g soln (70\%)}}{70 \text{ g drug}} \times 60 \text{ g drug} = 85.7 \text{ g soln (70\%)}$$

- c. Volume of 70% solution needed.

$$\frac{1 \text{ mL soln (70\%)}}{1.2 \text{ g soln (70\%)}} \times 85.7 \text{ g soln (70\%)} = 71.4 \text{ mL soln (70\%)}$$

Compounding problems involving solid preparations (such as mixtures of powder) and semisolid preparations (such as ointments, creams, and suppositories) are also percent *w/w*. The following is an example of this.

1. How much drug is required to make $2\frac{3}{4}$ of a 10% ointment?

$$\frac{10 \text{ g drug}}{100 \text{ g oint}} \times \frac{31.1 \text{ g oint}}{1\frac{3}{4} \text{ oint}} \times 2\frac{3}{4} \text{ oint} = 6.22 \text{ g drug}$$

The same procedure could be used for such mixtures as powders and suppository masses. Instead of using units in the various measuring systems, quantities can be indicated "by parts." The term "parts" then can mean any unit in any measuring system, as long as the units are kept constant.

2. How many grams of each of the following three ingredients are required to make 30 g of the product?

R	Solid A	0.5 part
	Powder B	3.0 parts
	Powder C, qs	30.0 parts

Since the product is a mixture of powders, percent *w/w* is indicated. In the above prescription order the total product is 30 parts because Powder C is used to "qs" or "make up to" 30 parts. Therefore, 0.5 g of Powder A and 3.0 g of Powder B are needed.

$$30 \text{ g total} - 0.5 \text{ g powder A} - 3.0 \text{ g powder B} = 26.5 \text{ g powder C}$$

3. How much of each of the following ingredients is needed to make 60 g of the ointment?

R	Solid D	3.0 parts
	Solid E	6.0 parts
	Ointment Base Q	30.0 parts
		39.0 parts total

$$\frac{3.0 \text{ g solid D}}{39.0 \text{ g oint}} \times 60 \text{ g oint} = 4.62 \text{ g solid D}$$

$$\frac{6.0 \text{ g solid E}}{39.0 \text{ g oint}} \times 60 \text{ g oint} = 9.23 \text{ g solid E}$$

$$\frac{30.0 \text{ g base Q}}{39.0 \text{ g oint}} \times 60 \text{ g oint} = 46.2 \text{ g base Q}$$

4. What is the percent strength of a salt solution obtained by diluting 100 g of a 5% solution to 200 g?

Assign the 5% solution as soln 1
Assign the final solution as soln 2

$$\frac{5 \text{ g salt}}{100 \text{ g soln 1}} \times \frac{100 \text{ g soln 1}}{200 \text{ g soln 2}} \times 100 \text{ g soln 2} = 2.5 \text{ salt}$$

$$\frac{2.5 \text{ g salt}}{100 \text{ g soln 2}} = 2.5\% \text{ w/w}$$

Problems

1. How much of the drug and solvent are needed to compound the following prescription?

R	Compound A	6% <i>w/w</i>
	Solvent, qs	$4\frac{3}{4}$

2. How many grams of solute are needed to prepare 240 g of a 12% *w/w* solution?
3. How many kg of a 20% *w/w* solution can be made from 1 kg of the solute?
4. How would you prepare, using 120 mL of glycerin (density, 1.25 g/mL), a solution that is 3% *w/w* with respect to a drug?

5. How much of each substance is needed to prepare a total of 24 g of the following suppository mass?
- | | |
|----------------------|--------|
| Compound K | 0.3 g |
| Solid H | 0.15 g |
| Suppository base, qs | 2.0 g |
6. How would one prepare 500 mL of a 15% *w/w* aqueous solution?
7. How much of each of the ingredients is required to make 1 kg of the following mixture?
- | | |
|----------|----------|
| Powder P | 1 part |
| Powder Q | 8 parts |
| Powder R | 12 parts |
| Powder S | 15 parts |
| Total | 36 parts |
8. How much of each ingredient is required to prepare the following ointment?
- | | | |
|---|--------------------------|------|
| ℞ | Coal Tar Solution | 10% |
| | Hydrophilic Ointment, qs | 30 g |

Volume-in-Volume Percentages

A direct calculation of percentage from the total volume is made. Volumes, unlike weights, may not be additive. However, this does not present a problem because the final solution is made up to the desired volume with the diluent.

Examples

1. How many minims of a liquid are needed to make 6 fʒ of a hand lotion containing 0.5% *v/v* of the liquid?

$$\frac{16.2 \text{ ℥ liq}}{1 \text{ mL liq}} \times \frac{0.5 \text{ mL liq}}{100 \text{ mL lotion}} \times \frac{29.6 \text{ mL lotion}}{1 \text{ fʒ lotion}} \times 6 \text{ fʒ lotion} = 14.4 \text{ ℥ liq}$$

Add sufficient lotion to 14.4 m of the liquid to make 6 fʒ of the product.

2. How much 90% alcohol is required to compound 500 mL of a 10% alcohol mixture?

$$\frac{100 \text{ mL (90\%)}}{90 \text{ mL alcohol}} \times \frac{10 \text{ mL alcohol}}{100 \text{ mL (10\%)}} \times 500 \text{ mL (10\%)} = 55.5 \text{ mL (90\%)}$$

Problems

- How many minims of a liquid are needed to make 4 fʒ of a 12.5% *v/v* solution?
- What volume of 50% *v/v* alcohol could be prepared from 1 L of 95% *v/v* alcohol?
- What is the percentage strength, weight in weight, of a liquid made by dissolving 16 g of a salt in 30 mL of water?
- How much drug will be required to prepare 1 fʒ of a 2.5% solution?
- What is the percentage, weight in weight, of sugar in a syrup made by dissolving 5 kg of sugar in 8 kg of water?
- How many grams of a drug are required to prepare 120 mL of a 12.5% aqueous solution?
- How much drug is needed to compound a liter of a 1:2500 aqueous solution?
- A solution contains 37% of active ingredient. How much of this solution is needed to prepare 480 mL of an aqueous solution containing 2.5% of the active ingredient?
- How much of a drug is required to make 2 qt of a 1:1200 solution?

STOCK SOLUTIONS

To facilitate the dispensing of certain soluble substances, the pharmacist frequently prepares or purchases solutions of high concentration. Portions of these concentrated solutions are diluted to give required solutions of lesser strength. These concentrated solutions are known as *stock solutions*. This procedure is satisfactory if the substances are stable in solution or if the solutions are to be used before they decompose.

In the case of potent substances, a properly prepared stock solution permits the pharmacist to obtain accurately a quantity of solid that might otherwise be difficult to weigh. In the case of frequently prescribed salt solutions, a stock solution readily provides the required amount of salt without the necessity of weighing and dissolving it every time.

Stock solutions may be of various concentrations depending on the requirements for use. The stock solutions should be labeled properly and fractional parts needed to make various strengths also may be listed as a further convenience.

There is a type of compounding and dispensing problem that involves the concept of stock solutions. This involves the patient diluting a dose from the prescription order to a given volume to obtain a solution of desired concentration.

For example, how many grams of a salt are required to make 90 mL of a stock solution, 5 mL of which makes a 1:3000 solution when diluted to 500 mL?

Assign the stock solution as Soln 1

Assign the final dilution as Soln 2

$$\frac{1 \text{ g salt}}{3000 \text{ mL soln 2}} \times \frac{500 \text{ mL soln 2}}{5 \text{ mL soln 1}} \times 90 \text{ mL soln 1} = 3.0 \text{ g salt}$$

Problems

- How much of a drug is needed to compound 120 mL of a prescription order such that when 1 teaspoonful of the solution is diluted to 1 qt, a 1:750 solution results?
- How many grams of a drug are needed to make 240 mL of a solution of such strength that when 5 mL is diluted to 2 qt, a 1:2500 solution results?
- An ampule of solution of an anti-inflammatory drug contains 4 mg of drug/mL. What volume of the solution is needed to prepare a liter of solution that contains 2 µg of the drug/mL?

PARTS PER MILLION

An expression that is occasionally used in compounding prescriptions is *parts per million* (ppm). This is another way of expressing concentration, particularly concentrations of very dilute preparations. A 1% solution may be expressed as 1 part/100; a 0.1% solution is 0.1 parts/100 or 1 part/1000. A one ppm solution contains 1 part of solute/1 million parts of solution; 5 ppm is 5 parts solute/1 million parts solution, and so on. Remember that the two parts must have the same units, except in the metric system where one g = one mL of water.

Sodium fluoride is a drug that may be prescribed by a dentist as a preventative for tooth decay in children. It is used only in very dilute solutions due to the drug's toxicity and because only minute quantities are needed. For example, how much sodium fluoride would be needed to prepare the following prescription?

℞ Sod Fluoride, qs
Purified water, qs 60 mL
Make soln such that when 1 fʒ is diluted to 1 glassful of water a 2 ppm soln results.
Sig: 1 fʒ in a glassful of water a day.

The mathematics to solve this compounding problem are easy once the steps for calculating the answer are outlined. This problem should be worked "backward."

- The amount of NaF needed is not known.
- One glassful of water has a volume of 240 mL. The concentration of NaF in 240 mL is 2 ppm.
- The NaF solution poured into the glass came from a teaspoonful dose (1 fʒ), which is equal to 5 mL.
- The 5-mL dose came from the prescription order bottle containing a NaF solution.

$$\frac{2 \text{ g NaF}}{1,000,000 \text{ mL dilution}} \times \frac{240 \text{ mL dilution}}{5 \text{ mL } \mathfrak{R}} \times 60 \text{ mL } \mathfrak{R} = 0.00576 \text{ g NaF}$$

The pharmacist would weigh out 5.76 mg (actually, one would weigh out a larger quantity and take an aliquot part) and qs to 60 mL.

Another variation of this problem is the prescriber requesting the concentration in terms of fluoride ion (F^-). In this case the atomic weight of F^- and molecular weight of NaF are used in the calculation. If the request called for 2 ppm fluoride, the initial calculations would be the same as above, and an additional step would be added at the end. The 5.76 mg would now represent the weight of fluoride ion needed. This must be converted to weight of NaF. The molecular weight of NaF is 42 and the atomic weight of fluorine is 19. The following proportion can be set up.

$$5.76 \text{ mg fluoride} \times \frac{42 \text{ mg NaF}}{19 \text{ mg fluoride}} = 12.7 \text{ mg NaF}$$

Problems

- How many mg of NaF are needed in the following prescription?

℞ Sodium Fluoride
Purified water, qs to 90 mL
M ft solution such that when 1 ℥ is diluted to 1 glassful of water a 3 ppm NaF soln results.

DILUTION AND CONCENTRATION

Stock solutions can be diluted to make a product that has a lower concentration; also mixtures of powders or semisolids (eg, ointments) can be diluted to give a product of lower concentration of the drug(s). The diluent is an inert solid or semisolid or base that does not contain any active ingredients.

Mixtures also may be concentrated by adding pure drug or mixing with a product containing a higher concentration of the drug. For example, how much of a diluent must be added to 50 g of a 10% ointment to make it a 5% ointment?

- How many grams of active ingredient are in 50 g of 10% ointment?

$$\frac{10 \text{ g drug}}{100 \text{ g oint (10\%)}} \times 50 \text{ g oint (10\%)} = 5 \text{ g drug}$$

- How many grams of a 5% ointment can be made from 5 g of active ingredient?

$$\frac{100 \text{ g oint (5\%)}}{5 \text{ g drug}} \times 5 \text{ g drug} = 100 \text{ g oint (5\%)}$$

- How many grams of base must be added to the 50 g of the original 10% ointment?

$$100 \text{ g oint (5\%)} - 50 \text{ g oint (10\%)} = 50 \text{ g base}$$

The term *trituration* was used previously to mean a dilute powder mixture of a drug. It is often necessary to dilute this mixture further to obtain the required amount of drug.

- How much of a 1 in 10 trituration of a potent drug contains 200 mg of the drug?

A 1 in 10 trituration means 1 g of drug in 10 g of mixture or 1 g of drug plus 9 g diluent. *Remember:* mixtures of solids are percent w/w .

$$\frac{10 \text{ g trituration}}{1 \text{ g drug}} \times \frac{1 \text{ g drug}}{1000 \text{ mg drug}} \times 200 \text{ mg drug} = 2 \text{ g trituration}$$

- How much diluent must be added to 10 g of a 1:100 trituration to make a mixture that contains 1 mg of drug in each 10 g of the final mixture?

- Determine the amount of drug in 10 g of trituration.

$$\frac{1 \text{ g drug}}{100 \text{ g trituration}} \times 10 \text{ g trituration} = 0.1 \text{ g drug}$$

- Determine the amount of mixture that can be made from 0.1 g (100 mg) of drug.

$$\frac{10 \text{ g mixture}}{1 \text{ mg drug}} \times \frac{1000 \text{ mg drug}}{1 \text{ g drug}} \times 0.1 \text{ g drug} = 1000 \text{ g mixture}$$

- Determine the amount of diluent needed.

$$1000 \text{ g mixture} - 10 \text{ g trituration} = 990 \text{ g diluent}$$

Problems

- The following prescription order was received in a pharmacy. If the only *R* cream available is a 10% concentration, how much of the 10% cream and how much diluent are required to compound the prescription?

℞ *R* Cream 3% . . . 30 g

- How many grams of a 1:100 trituration contain 100 μg of the active ingredient?
- How many grams of a 1:1000 dilution can be made from 1 g of a 1:25 trituration?

MIXING DIFFERENT STRENGTHS

Rules

- The sum of the products obtained by multiplying a series of quantities by their respective concentrations equals the product obtained by multiplying a concentration by the sum of the quantities. For example, the sum of the products—obtained by multiplying the individual weights or volumes of a series of preparations by the concentration of a given ingredient contained in each preparation—is equal to the product obtained by multiplying the total weight of the series of preparations by the percentage of the given ingredient resulting from a homogeneous mixture of the same series of preparations.
- When mixing products of varying strengths, the units and type of percent (w/w , w/v , v/v) must be kept constant.

Examples

- What is the percent of alcohol in a mixture made by mixing 5 L of 25%, 1 L of 50%, and 1 L of 95% alcohol?

- Determine the total amount of alcohol in the three solutions and the total amount of solution (1 L = 1000 mL). Assume additivity of volumes on mixing.

$$\frac{25 \text{ mL alcohol}}{100 \text{ mL (25\%)}} \times 5000 \text{ mL (25\%)} = 1250 \text{ mL alcohol}$$

$$\frac{50 \text{ mL alcohol}}{100 \text{ mL (50\%)}} \times 1000 \text{ mL (50\%)} = 500 \text{ mL alcohol}$$

$$\frac{95 \text{ mL alcohol}}{100 \text{ mL (95\%)}} \times 1000 \text{ mL (95\%)} = 950 \text{ mL alcohol}$$

- Determine the percent of alcohol in the mixture. There is a total of 2700 mL of alcohol in 7000 mL of total solution.

$$\frac{2700 \text{ mL alcohol}}{7000 \text{ mL mixture}} \times 100 \text{ mL mixture} = 38.6 \text{ mL alcohol}$$

$$\frac{38.6 \text{ mL alcohol}}{100 \text{ mL mixture}} = 38.6\%$$

- What is the strength of a mixture obtained by mixing 50 g of a 5%, 100 g of a 7.5% and 40 g of a 10% ointment?

$$\frac{5 \text{ g drug}}{100 \text{ g oint (5\%)}} \times 50 \text{ g oint (5\%)} = 2.5 \text{ g drug}$$

$$\frac{7.5 \text{ g drug}}{100 \text{ g oint (7.5\%)}} \times 100 \text{ g oint (7.5\%)} = 7.5 \text{ g drug}$$

$$\frac{10 \text{ g drug}}{100 \text{ g oint (10\%)}} \times 40 \text{ g oint (10\%)} = 4.0 \text{ g drug}$$

There is a total of 14.0 g of active ingredient in 190 g of total mixture.

$$\frac{14.0 \text{ g drug}}{190 \text{ g mixture}} \times 100 \text{ g mixture} = 7.37 \text{ g drug}$$

$$\frac{7.37 \text{ g drug}}{100 \text{ g mixture}} = 7.37\%$$

Problems

1. What percent of a drug is contained in a mixture of powder consisting of 0.5 kg, containing 0.038% of a drug, and 10 kg, containing 0.043% of a drug?
2. What is the strength of a mixture produced by combining the following lots of alcohol: 2 L of 95%, 2 L of 50%, and 7 L of 60%?
3. What is the percent of drug content in the following mixture: 2 kg of 3%, 300 g of 2.5%, and 500 g of 4.2% resin?

ALLIGATION ALTERNATE

Alligation is a rapid method of calculation that is useful to the pharmacist. The name is derived from the Latin *alligatio*, meaning the act of attaching, and it refers to lines drawn during calculation to bind quantities together. This method is used to find the proportions in which substances of different strengths or concentrations must be mixed to yield a mixture of desired strength or concentration. When the proportion is found, a calculation may be performed to find the exact amounts of the substances required.

Rules

1. Line up the concentrations of all the starting materials in a vertical column in order of concentration, traditionally from high to low. Pure drugs are defined as being 100%; solvents or vehicles are designated as 0%.
2. Place the concentration of the desired product in a second column such that it is bracketed by concentrations of starting materials. With two starting materials, the desired product simply falls between the two.
3. Cross subtract the two columns to give a parts formula that can be used to calculate specific amounts of each starting material.

Examples and Procedure

1. In what proportion must a preparation containing 10% of drug be mixed with one containing 15% of drug to produce a mixture of 12% drug strength?

Applying the above rules gives:

$$\begin{array}{r} 15\% \\ \swarrow \quad \searrow \\ 12\% \\ \swarrow \quad \searrow \\ 10\% \end{array} \quad \begin{array}{l} 2 \text{ parts of } 15\% \\ \hline 3 \text{ parts of } 10\% \\ \hline 5 \text{ parts of } 12\% \end{array}$$

The concentrations of the starting material are lined up in the first column in decreasing or increasing order and the desired percent or concentration is placed in the center column. The third column is obtained by cross-subtracting as indicated by the arrows and gives a parts formula for mixing the two starting materials. Thus, mixing 2 parts of 15% drug preparation with 3 parts of 10% drug preparation will produce 5 parts of a drug mixture of the desired 12% strength.

2. In what proportion must 30% alcohol and 95% alcohol be mixed to make 500 mL of 50% alcohol? Set up the problem in the following manner:

$$\begin{array}{r} 95\% \\ \swarrow \quad \searrow \\ 50\% \\ \swarrow \quad \searrow \\ 30\% \end{array} \quad \begin{array}{l} 20 \text{ parts of } 95\% \\ \hline 45 \text{ parts of } 30\% \\ \hline 65 \text{ parts of } 50\% \end{array}$$

In a total of 65 parts, 20 parts of 95% alcohol + 45 parts of 30% alcohol are needed. Since the total is proportional to 500 mL, the following can be calculated:

$$\frac{20 \text{ parts (mL) } 95\%}{65 \text{ parts (mL) } 50\%} \times 500 \text{ (mL) } 50\% = 154 \text{ mL } 95\%$$

$$\frac{45 \text{ parts (mL) } 30\%}{65 \text{ parts (mL) } 50\%} \times 500 \text{ mL } 50\% = 346 \text{ mL } 30\%$$

Since volumes are not additive, sufficient water may be needed to make 500 mL.

3. How many grams of an ointment containing 0.18% of active ingredient must be mixed with 50 grams of an ointment containing 0.14% of active ingredient to make a product containing 0.15% of active ingredient?

$$\begin{array}{r} 0.18\% \\ \swarrow \quad \searrow \\ 0.15\% \\ \swarrow \quad \searrow \\ 0.14\% \end{array} \quad \begin{array}{l} 0.01 \text{ parts of } 0.18\% \\ \hline 0.03 \text{ parts of } 0.14\% \\ \hline 0.04 \text{ parts of } 0.15\% \end{array}$$

$$\frac{0.01 \text{ parts (g) } 0.18\%}{0.03 \text{ parts (g) } 0.14\%} \times 50 \text{ g } 0.14\% = 16.6 \text{ g } 0.18\%$$

4. Occasionally, it is necessary for a pharmacist to increase the strength of a product. For example, a prescription calls for 50 g of a 10% ointment. The pharmacist only has a 5% ointment and the pure ingredient available. How much of the 5% ointment and the pure ingredient are needed to compound the prescription?

$$\begin{array}{r} 100\% \\ \swarrow \quad \searrow \\ 10\% \\ \swarrow \quad \searrow \\ 5\% \end{array} \quad \begin{array}{l} 5 \text{ parts of } 100\% \\ \hline 90 \text{ parts of } 5\% \\ \hline 95 \text{ parts of } 10\% \end{array}$$

$$\frac{5 \text{ parts (g) } 100\%}{95\% \text{ parts (g) } 10\%} \times 50 \text{ g } 10\% = 2.63 \text{ g } 100\%$$

$$\frac{90 \text{ parts (g) } 5\%}{95 \text{ parts (g) } 10\%} \times 50 \text{ g } 10\% = 47.4 \text{ g } 5\%$$

Problems

1. How much ointment containing 12% drug and how much ointment containing 16% drug must be used to make 1 kg of a product containing 12.5% drug?
2. In what proportion should 50% alcohol and purified water be mixed to make a 35% alcohol solution? (The purified water is 0% alcohol.)
Note: This problem may be solved by a method other than alligation as was shown above.
3. How many grams of 28% w/w ammonia water should be added to 500 g of 5% w/w ammonia water to produce a 10% w/w ammonia concentration?
4. How many mL of 20% dextrose in water and how many mL of 50% dextrose in water are needed to make 1 L of 35% dextrose in water?

PROOF SPIRIT

For tax purposes, the US government calculates the strength of pure or absolute alcohol (herein referred to as C_2H_5OH) by means of *proof degrees*. This means that 100 proof spirit contains 50% (by volume) or 42.49% (by weight) of C_2H_5OH , and its specific gravity is 0.93426 at 60°F. Thus, 2 proof degrees equals 1% (by volume) of C_2H_5OH . One proof gallon is one gal of 50% (by volume) of C_2H_5OH at 15.56°C (60°F). In other words, a proof gallon is a gallon that contains 1/2 gal of C_2H_5OH . A proof gallon is 100 proof.

The term *10 degrees under proof* (10° up) signifies that 100 volumes of the spirit contains 90 volumes of proof spirit plus 10

volumes of water, and *30 degrees over proof* (30° op) indicates that 100 volumes diluted with water yields 130 volumes of proof spirit. To prepare proof spirit, 50 volumes of C₂H₅OH are mixed with 53.71 volumes of water to allow for the contraction that occurs to yield 100 volumes of product.

The terms *proof strength*, *proof gallon*, and *proof spirit* are used so that the tax is levied only on the actual quantity of C₂H₅OH contained in any mixture. Therefore, it is sometimes necessary for the pharmacist to convert alcohol purchased to proof strength to compute tax refunds or convert proof strengths to percent for compounding purposes.

A quantity of solution that contains 1/2 gal of C₂H₅OH is said to contain one proof gal. Proof gallons may be calculated by the following two equations:

$$\text{Proof gal} = \frac{\text{gal} \times v/v \text{ strength}}{50\% v/v}$$

$$\text{Proof gal} = \frac{\text{gal} \times \text{proof strength}}{100 \text{ proof}}$$

The second equation is the same as the first because proof strength is always twice the % v/v strength. With these equations, given any two variables the third can be calculated.

Examples

1. What is the taxable alcohol in 1 pt of Alcohol USP?

$$1 \text{ pt} = \frac{1}{8} \text{ gal} \quad (8 \text{ pt} = 1 \text{ gal})$$

Alcohol USP is 95% v/v; therefore,

$$\begin{aligned} \text{proof gal} &= \frac{\text{gal} \times \% \text{ strength}}{50\%} = \frac{1/8 \text{ gal} \times 95\%}{50\%} \\ &= 0.2375 \text{ proof gal} \end{aligned}$$

2. How much Diluted Alcohol USP can be made from 1 qt of alcohol labeled 1/2 proof gallon?

Diluted Alcohol USP is 49% v/v; therefore,

$$\begin{aligned} \text{Proof gal} &= \frac{\text{gal} \times \% \text{ strength}}{50\%} \\ \text{gal} &= \frac{0.5 \text{ proof gal} \times 50\%}{49\%} = 0.510 \text{ gal} \end{aligned}$$

Problems

1. How many proof gallons are there in 1 qt of a preparation that is labeled 75% v/v alcohol?
2. How many proof gallons are there in a pint of an elixir that contains 14% alcohol?
3. How much Diluted Alcohol USP can be made from 1 gal of 190 proof alcohol?

SATURATED SOLUTIONS

Occasionally, it is necessary for a pharmacist to make saturated solutions. Solubility in the USP/NF is expressed as the number of milliliters of a solvent that will dissolve one g of a solid; for example, one g dissolves in 0.5 mL of water. In other words, if one g of a solid is dissolved in 0.5 mL of water, a saturated solution results. An example will illustrate this.

How much of a drug is needed to make 120 mL of a saturated solution if one g of the drug dissolves in 7.5 mL of water?

Calculate the amount of drug that can be dissolved in 120 mL water.

$$\frac{1 \text{ g drug}}{7.5 \text{ mL water}} \times 120 \text{ mL water} = 16 \text{ g drug}$$

When 16 g of the drug are dissolved in 120 mL of water, a saturated solution results that has a volume greater than 120 mL

because the solid will take up a certain volume. Only 120 mL would be dispensed.

What is the % w/w of the above solution?

$$120 \text{ g (mL) water} + 16 \text{ g drug} = 136 \text{ g solution}$$

$$\frac{16 \text{ g drug}}{136 \text{ g solution}} \times 100 \text{ g solution} = 11.8 \text{ g drug}$$

$$\frac{11.8 \text{ g drug}}{100 \text{ g solution}} = 11.8\% \text{ w/w}$$

Problems

1. What is the solubility of a chemical if a saturated aqueous solution is 0.5% w/w?
2. How many grams are needed to make 500 mL of a saturated solution if 1 g of the solute is soluble in 14 mL of solvent?

MILLIEQUIVALENTS

The quantities of electrolytes administered to patients are usually expressed by the term *milliequivalents* (mEq). The reason that weight units (mg, g) are not used is because the electrical activity of the ions, which in this instance is important, may be best expressed as mEq. (See Chapter 17 for additional discussion on electrolytic equilibria.)

A mEq is 1/1000 of an *equivalent* (Eq). An Eq is the weight of a substance that combines with or replaces one gram-atomic weight (g-at wt) of hydrogen. In pharmacy the terms equivalent and equivalent weight (Eq wt) have been used interchangeably. For problem solving it is convenient to identify the molar weight in terms of mg per mmol and the number of mEq per mmol as follows:

$$\text{Molecular weight} = \frac{\text{g}}{\text{mole}} = \frac{\text{mg}}{\text{mmol}}$$

$$\frac{\text{mEq}}{\text{mol}} = \text{valence}$$

For example, KCl has a molecular weight of 74.5; the above parameters would be 74.5 mg/mmol and one mEq/mmol.

Water of hydration contributes to the molecular weight (mol wt) of a compound but *not* to the valence, and the total mol wt is used to calculate mEq.

Examples

1. Calcium (Ca²⁺) has a gram-atomic weight of 40.08. Determine the number of mEq/mmol. As the valence of the calcium ion is 2, there are 2 mEq/mmol.
2. A solution (100 mL) that contains 409.5 mg of NaCl/100 mL has how many mEq of Na⁺ and Cl⁻? The molecular weight of NaCl is 58.5.

$$\begin{aligned} \text{There is } & \frac{1 \text{ mEq Cl}^-}{\text{mmol NaCl}} \text{ and } \frac{1 \text{ mEq Na}^+}{\text{mmol NaCl}} \\ & \frac{1 \text{ mEq Cl}^-}{\text{mmol NaCl}} \times \frac{1 \text{ mmol NaCl}}{58.5 \text{ mg NaCl}} \times \frac{409.5 \text{ mg NaCl}}{100} \text{ mL} \\ & \quad \times 100 \text{ mL} = 7.0 \text{ mEq Cl}^- \end{aligned}$$

Since NaCl is a 1:1 electrolyte, the solution contains 7.0 mEq of Cl⁻ and 7.0 mEq of Na⁺.

3. A prescription order calls for a 500 mL solution of potassium chloride to be made so that it will contain 400 mEq of K⁺. How many grams of KCl (mol wt: 74.5) are needed?

$$\begin{aligned} & \frac{1 \text{ mEq}}{\text{mmol}} \text{ and } \frac{74.5 \text{ mg}}{\text{mmol}} \\ & \frac{1 \text{ g KCl}}{1000 \text{ mg KCl}} \times \frac{74.5 \text{ mg KCl}}{\text{mmol KCl}} \times \frac{1 \text{ mmol KCl}}{\text{mEq K}^+} \\ & \quad \times 400 \text{ mEq K}^+ = 29.8 \text{ g KCl} \end{aligned}$$

4. How many mEq of K^+ are in a 250-mg tablet of potassium phenoxymethyl penicillin (mol wt: 388.5; valence: 1)?

$$\frac{1 \text{ mEq } K^+}{\text{mmol Pen}} \text{ and } \frac{388.5 \text{ mg Pen}}{\text{mmol Pen}}$$

$$\frac{1 \text{ mEq } K^+}{\text{mmol Pen}} \times \frac{1 \text{ mmol Pen}}{388.5 \text{ mg Pen}} \times \frac{250 \text{ mg Pen}}{\text{Tab}}$$

$$\times 1 \text{ Tab} = 0.644 \text{ mEq } K^+$$

5. How many mEq of Mg are there in 10 mL of a 50% Magnesium Sulfate Injection? The mol wt of $MgSO_4 \cdot 7H_2O$ is 246.

$$\frac{2 \text{ mEq } Mg^{2+}}{\text{mmol drug}} \text{ and } \frac{246 \text{ mg drug}}{\text{mmol drug}}$$

$$\frac{2 \text{ mEq } Mg^{2+}}{\text{mmol drug}} \times \frac{1 \text{ mmol drug}}{246 \text{ mg drug}} \times \frac{1000 \text{ mg drug}}{\text{g drug}} \times \frac{50 \text{ g drug}}{100 \text{ mL}}$$

$$\times 10 \text{ mL} = 40.7 \text{ mEq } Mg^{2+}$$

6. A vial of Sodium Chloride Injection contains 3 mEq/mL. What is the percentage strength of this solution? The mol wt of NaCl is 58.5.

$$\frac{1 \text{ mEq}}{\text{mmol}} \text{ and } \frac{58.5 \text{ mg}}{\text{mmol}}$$

$$\frac{1 \text{ g}}{1000 \text{ mg}} \times \frac{58.5 \text{ mg}}{\text{mmol}} \times \frac{1 \text{ mmol}}{1 \text{ mEq}} \times \frac{3 \text{ mEq}}{\text{mL}} \times 100 \text{ mL} = 17.6 \text{ g}$$

$$\frac{17.6 \text{ g}}{100 \text{ mL}} = 17.6\%$$

Problems

- What is the mEq wt of ferrous ion (Fe^{2+}) which has a atomic weight of 55.85 g?
- What is the mEq wt of sodium phosphate ($Na_2HPO_4 \cdot 7H_2O$)?
- How many mEq of Na^+ are in 60 mL of a 5% solution of sodium saccharin (mol wt: 241 g; valence: 1)?
- How many mEq of Ca^{2+} are there in a 600-mg calcium lactate pentahydrate (mol wt: 308.30 g) tablet?
- How many mEq of sodium are there in a 5 gr sodium bicarbonate tablet? The mol wt of $NaHCO_3$ is 84 and the valence is 1.
- How many mEq of Na are there in 500 mL of 1/2 normal saline solution? Normal saline solution contains 9 g NaCl/L; mol wt NaCl is 58.5.
- How much KCl is needed to make a pint of syrup that contains 10 mEq of K^+ in each tablespoonful? The mol wt of KCl is 74.5.

TEMPERATURE

Rules

The relationship of Centigrade (C) and Fahrenheit (F) degrees is:

$$9 (^{\circ}C) = 5 (^{\circ}F) - 160$$

Where $^{\circ}C$ is the number of degrees Centigrade, and $^{\circ}F$ is the number of degrees Fahrenheit.

Examples

1. Convert $77^{\circ}F$ into $^{\circ}C$.

$$9 (^{\circ}C) = 5 (77) - 160$$

$$^{\circ}C = \frac{385 - 160}{9} = 25^{\circ}C$$

2. Convert $10^{\circ}C$ into $^{\circ}F$.

$$9 (10) = 5 (^{\circ}F) - 160$$

$$^{\circ}F = \frac{90 + 160}{5} = 50^{\circ}F$$

Problems

Convert

- $30^{\circ}C$ into $^{\circ}F$
- $100^{\circ}C$ into $^{\circ}F$
- $37^{\circ}C$ into $^{\circ}F$
- $120^{\circ}F$ into $^{\circ}C$

REFERENCES

- Specifications, Tolerances, and Other Technical Requirements for Weighing and Measuring Devices*. NBS Handbook 44. Washington DC:US Department of Commerce, NBS, USGPO, 1989.
- Goldstein SW, Mattocks AM. *Professional Equilibrium and Compounding Accuracy* (pamphlet). Washington DC: APhA, 1967.
- USP XXVI, 2003.
- Morrell CA, Ordway EM. *Drug Std* 1954; 22:216.
- Madlon-Kay DF, Mosch FS. *J Family Pract* 2000; 49(8):741.
- Shirkey HC. Dosage (posology). In Shirkey HC, ed. *Pediatric Therapy*, 5th ed. St Louis: Mosby, 1975, p 19.
- Benitz WE, Tatro DS. *The Pediatric Drug Handbook*, 3rd ed. St Louis: Mosby, 1995.
- Nelson JD. *Pocketbook of Pediatric Antimicrobial Therapy*, 4th ed. Dallas: Jodane, 1981.

ANSWERS TO PROBLEMS

DENSITY

- 816 g
- 363 mL
- 54.2 mL

ADDITION

- 2480 g or 2.48 kg
- 1160 g or 1.16 kg

SUBTRACTION

- 4100 mL or 4.11 L
- 3.71 g

MULTIPLICATION

- 157 mL
- 163 g
- 825 mL
- 375 mg

DIVISION

- 769 capsules + 15 mg remainder
- 150 doses
- 1396 capsules + 300 mg remainder

CONVERSIONS

- 422 mg
 - 19.4 mg
 - 109 g
 - 7780 mg
 - 99.4 g
 - 454 g
- 1 lb, 3 oz, 173 gr
 - 6.94 gr
 - 1 lb, 5 3/8, 5 3/8, 26 gr
 - 0.00154 gr
 - 2.2 lb
- 0.648 mg
 - 0.203 mg

- c. 10.8 mg
 - d. 0.325 or 0.324 g
 - e. 1.299 or 1.296 g
4.
 - a. 12.3 mL
 - b. 11.1 mL
 - c. 237 mL
 - d. 473 mL
 - e. 0.309 mL
 - f. 0.00154 gr
 - g. 0.0772 gr
 5.
 - a. 480 gr
 - b. 8 $\bar{3}$
 - c. 437 $\frac{1}{2}$ gr
 - d. 2880 gr
 - e. 4 $\bar{3}$, 10 gr

DOSAGE CALCULATION

1. 1.5 mg
2. 32.7 mg
3. 18.7 mg
4. 280,000 units
5. 75 mg
6. 77.9 mg

PROBLEM-SOLVING METHODOLOGY

1. D.T.D. No. 14 means dispense 14 such doses. Assuming the doses have been checked, they are for chemicals J, K, and L (10 mg, 50 mg, and 300 mg, respectively).
2. Drug Q: 0.5 g
Drug R: 0.3 g
3. 0.469 mL/dose; 1.88 mL/day
4. 50 doses
5. 0.0444 mg
6. 0.75 mL contains 60 units; 13 $\frac{1}{3}$ -day supply.
7. 6 mL
8. 3 mL
9. 25 mg

REDUCING AND ENLARGING

1. Liquid C 875 mL
Solid B 225 g
Liquid R 62.5 mL
Liquid P 500 mL
2. Solid G 7.46 g
Liquid D 224 g
Solid M 22.4 g
Base 492 g
3. Solid N 4.8 mg
Solid Q 120 mg
Solid R 7.2 g
Add sufficient purified water to make 240 mL solution.
4. Solid S 0.675 g
Solid T 2.25 g
Oil C 31.5 mL
Alcohol 22.5 mL

PERCENTAGE

w/v Solutions

1. Dissolve 11.1 g in sufficient solvent to make 3 f $\bar{3}$.
2. 2.84 L
3. 6.25 L
4. 8 g

w/w Products

1. Compound A 7.46 g
Solvent 117 g
2. 28.8 g
3. 5 kg
4. Dissolve 4.64 g of drug in 120 mL (150 g) of glycerin.
5. Compound K 3.6 g
Solid H 1.8 g
Base 18.6 g
6. Dissolve 88.2 g of the solute in 500 mL of purified water. Dispense 500 mL

7. Powder P 27.8 g
Powder Q 222 g
Powder R 333 g
Powder S 416 g
8. 3 g of coal tar solution; 27 g of hydrophilic ointment

PERCENT

(*v/v*, *w/v*, and *w/w*)

1. 240 \bar{m}
2. 1900 mL
3. 34.8% *w/w*
4. 0.740 gr
5. 38.5% *w/w*
6. 15 g
7. 0.4 g
8. 32.4 mL of a 37% solution
9. 1.58 g

STOCK SOLUTIONS

1. 30.3 g
2. 36.3 g
3. 0.5 mL

PARTS PER MILLION

1. 13 mg

DILUTION AND CONCENTRATION

1. 9 g of 10% cream and 21 g of diluent (base)
2. 0.01 g
3. 40 g

MIXING PRODUCTS OF DIFFERENT STRENGTHS

1. 0.0428%
2. 64.5%
3. 3.16%

ALLIGATION ALTERNATE

1. 875 g of 12% ointment and 125 g of 16% ointment
2. 35 parts of 50% alcohol and 15 parts of purified water
3. 139 g of 28% ammonia water
4. 500 mL each of the 20% and 50% solutions are needed

PROOF SPIRIT

1. 0.375 proof gal
2. 0.035 proof gal
3. 1.94 gal

SATURATED SOLUTIONS

1. 1 g in 199 mL
2. 35.7 g of solute—dispense 500 mL

MILLIEQUIVALENTS

1. 27.9 mg/mEq
2. 134 mg/mEq
3. 12.5 mEq
4. 3.89 mEq
5. 3.86 mEq Na
6. 38.5 mEq Na
7. 23.5 g

TEMPERATURE

1.
 - a. 86°F
 - b. 212°F
 - c. 98.6°F
 - d. 48.9°C

Statistics

Sanford Bolton, PhD



Statistical methods are an integral part of the development, evaluation, and marketing of drug products. In this chapter, elementary definitions and some common statistical applications to problems of pharmaceutical interest will be presented and discussed.

Statistics is often thought of as a collection of numbers and averages, such as vital statistics, baseball statistics, or statistics derived from the census. Indeed, this is an important aspect of statistical thinking, and such collections of data and counting do play a role in pharmacy and medicine, such as in marketing or disease-incidence data. However, here more emphasis will be placed on the use of statistics in presenting, analyzing, and interpreting data that are often, but not necessarily always, derived from planned experiments.

OVERVIEW AND INTRODUCTION

Although the material in this chapter is elementary for the most part, those readers who have had little or no exposure to statistical methods may be overwhelmed by the large amount of information presented in a relatively small space. This introduction presents an overview of the chapter so that the student can get a feeling for what is contained here. Many illustrations are interspersed in the didactic discussion to show the applications in a practical way.

The first part of the chapter deals with *introductory definitions and methods*. An understanding of this material is essential if one wishes to use elementary techniques intelligently, or if one wishes to pursue more advanced topics. Definitions include statistical jargon, design of scientific experiments (both laboratory and clinical experiments), the concept of sampling (including methods of obtaining samples for experiments), and the concept and definition of probability distributions. These concepts lay the foundation for the understanding of practical applications of statistics to scientific research. Although not complete, an understanding of this introductory material should allow the student to feel confident about applying elementary methods to real data.

Some words of caution are necessary here. Real examples often have twists that are not obvious to the initiate, which make them different from simple textbook examples. At the beginning, students should try to seek advice from more experienced persons, preferably a statistician, to make sure that they are using the techniques in a proper manner.

For those with some background in statistics, the initial portion of the chapter should serve as a quick review and an introduction to the material that follows. The elementary definitions include the usual measures of central tendency and spread, such as the mean, median, standard deviation, variance, coefficient of variation, and range. The nature of *variation* and its ba-

sis for statistical thinking is discussed, as without variation, statistical reasoning would be unnecessary. Statistical approaches take the experimental variability (often referred to as *error*) into account during the analysis.

Statistical “proof” is different from mathematical proofs. In statistics, one is never sure of an answer or a decision. The decision has a given probability of being correct. Discrete and continuous variables are defined and discussed. Discrete variables include binomial measurements, which may have a “yes or no” outcome (eg, accept or reject). Continuous variables can have any number of outcomes and include typical measurements (eg, weight or assay).

Definitions of a population and a sample are presented; these are very important concepts in statistical reasoning. Definitions and examples of bias, precision, and accuracy are introduced. Examples are used to illustrate the fact that data may be precise but not accurate and vice versa.

The analysis of any data set depends on the *experimental design*, the detailed experimental procedure. A description of some common designs and the manner in which data may be collected are presented in this chapter. The integrity of data from any experiment is only as good as the design and the care that was taken to implement the design. Each experiment is different. Design and sampling considerations are different for questionnaire surveys, censuses (complete sampling), and laboratory or clinical experiments. Good experimental design should result in optimality, increased precision, and lack of bias. The *random selection* of objects to be included in an experiment and/or assigned to treatments is of vital importance in pharmaceutical and clinical research. In particular, controlled clinical studies should be designed as double-blind studies if at all possible. A *controlled study* is a designed study that includes a placebo or a positive control (eg, a known active drug).

Statistical inference and estimation are cornerstones of statistical applications in pharmaceutical research. Statistical inference results from the formulation and testing of a hypothesis, the *null hypothesis*. In this procedure, a hypothesis is formulated with regard to the true, but unknown, values of parameters of the data distribution that is investigated in an experiment. For example, the average potency of a commercial batch of tablets may be of interest, or the mean blood pressure reduction of a new drug compared to an effective marketed product may be assessed. The experimental outcome is observed and analyzed. Using statistical procedures that usually are based on the normal probability distribution, an inference based on probability is drawn as to whether the proposed hypothesis is true; eg, “Is the true average potency equal to 100 mg?” or “Are the two comparative drugs equally efficacious?”

Again, these inferences are not proofs. Two treatments may be declared to be equal, but only with a given degree of

assurance expressed in probability terms. For example, two treatments may be considered different, but there may be a 5% chance that this decision is in error; that means there is a 5% chance that the treatments are truly not different. These procedures are based on knowledge of the underlying probability distribution of the experimental outcome. In this chapter, some properties of the binomial and normal distributions are presented as a basis for the inference procedures.

When estimating a parameter, such as the mean, from sample data, computation of a *confidence interval* is a useful way of showing the precision of the estimate. For example, if an experiment shows that a generic drug is absorbed 90% relative to a reference drug, a confidence interval of 80 to 100% places limits on the true relative absorption. This statement suggests that the true relative absorption is *probably* between 80 and 100%. The concepts of a confidence interval and simple hypothesis testing are discussed following the presentation of the properties of the normal and binomial distributions.

The *t* test is a common and well-known test that is used to make statistical decisions. This test is used to determine significance when comparing average results from two groups or treatments (a two-sample *t* test), or when comparing an average result to some hypothetical value (a one-sample *t* test). In the latter case, an example would be the comparison of the average dissolution time to some given compendial standard value. The *null hypothesis* is the hypothesis that is tested, and the *alternative hypothesis* is the hypothesis accepted if the null hypothesis is rejected. The test is deemed significant if the null hypothesis is rejected at a given probability level, the *alpha error* or *level of significance*. Thus, the alpha error is the probability of mistakenly rejecting the null hypothesis, usually taken as 5%. This, and other important concepts relating to statistical inference are presented in more detail in another part of the chapter.

The *t* test is appropriate for normally distributed variables. For dichotomous experimental outcomes following the binomial distribution, other statistical methods may be used. With sufficiently large samples, a *chi-square* test may be appropriate to compare the proportion of responders in two groups. A discussion of these tests is included following the examples of use of the *t* test.

The *F* distribution is introduced as used in a test to compare the variances of two independent samples. The more common use of the *F* test is in *analysis of variance* (ANOVA). The comparison of two means using the *t* test is the most elementary of comparisons. In more complicated experiments where more than two groups are being compared, and where the experimental design is complicated and includes many factors, the *t* test cannot be used. In these cases ANOVA is indicated.

A good deal of this chapter is devoted to ANOVA applications. Briefly, ANOVA is a method of separating the variance due to factors imposed on the experiment. For example, in a crossover design, subjects are treated on two occasions, Period I and Period II. If the results in one period tend to be higher than in the other, but the treatment differences are not affected, the variance due to period differences may be substantial without affecting the treatment comparison. By separating the variance due to period differences from the variation in the experiment, the residual error that is used to test treatment differences is smaller. This results in a more sensitive experiment—differences are detected more easily. If period differences exist and are not taken into account, the variance becomes part of the residual error, which is inflated, resulting in a less sensitive test.

Because of the more complex structure of experimental designs that are analyzed using ANOVA, several problems arise that need special attention. Procedures based on multiple comparisons have been devised to compare means in a pairwise fashion when more than two means are compared and it is not obvious how to identify significant effects. Also, in complex designs, the choice of the proper error term for an effect is not always obvious; ie, different effects may not all have the same

denominator error term for the *F* test. Various designs common to pharmaceutical sciences are discussed, including crossover designs used in bioequivalence and some clinical studies, and repeated measure designs used in clinical studies.

If the assumptions underlying ANOVA are not met, eg, if distributions are highly skewed, *nonparametric* methods of analysis may be used. These analyses do not quite have the flexibility of the parametric ANOVA, but are generally almost as efficient in detecting treatment differences compared to ANOVA. Several nonparametric tests are discussed. For more details, and for a description of other nonparametric tests, see Siegal's *Nonparametric Statistics*.¹

A persistent problem in data analysis is the presence of *outliers*, one or more values that seem to be remote from the main body of data. If no obvious reason can be found to discard such data, the nature of the data, including the experimental technique and the history of such experiments, should be investigated carefully. If this investigation reveals no cause for the presence of the outlier(s), a statistical test may be applied to determine if the data can be discarded. If such procedures are applied, a report should include a description of what was done. Some people recommend performing the analysis with and without the outlier. In any event, before discarding an outlier, one should evaluate the consequences of this action.

The remainder of the chapter considers some specialized topics of interest to pharmaceutical scientists. An understanding of basic statistics is necessary to apply this material. Basic *Shewhart and Fraction Defective Control Charts* are discussed with examples. Often, Shewhart charts do not work for pharmaceutical processes where the material is heterogeneous (eg, solid dosage forms) or the manufacturing equipment is variable from batch to batch. In these cases, other approaches may give a satisfactory analysis.⁸

Regression analysis, a process familiar to most scientists, concerns the fitting of data to linear models, ie, to models that are linear in the parameters. In particular, the fitting of straight lines is common to many different fields of scientific research. The process of least-squares fitting and the statistical properties of the slope and intercept are discussed, with applications to stability, dose-response relationships, calibration plots, kinetics, and so on. In particular, an analysis of stability data to predict shelf life is presented in some detail. This analysis includes hypothesis tests for the slope and intercept, as well as confidence limits for the line.

Regression is used when one of the variables (X) is measured with little or no error, and the other variable (Y) is measured with error. *Correlation* is related to linear regression. This analysis may be appropriate when both variables are subject to error, and an estimate of the degree of their association is desired. A correlation coefficient is calculated, which can have values between +1 and -1. A correlation coefficient of 0 suggests that the variables are not correlated. Care should be exercised in the interpretation of correlation coefficients. A value of the correlation coefficient close to 1 does not prove that the variables have a linear relationship.

The chapter concludes with a discussion of *transformations*, which are useful when data distributions do not conform to that assumed for the statistical analysis. In particular, a transformation may help to normalize data that are not normal (eg, skewed). The most common transformation is the logarithmic transformation, which will equalize variances for data that have a relatively constant relative standard deviation, S/\bar{X} .

This chapter covers a wide variety of material in a small space. Although the concepts here should provide a basic understanding, much effort is needed to understand and apply statistics in the real world. The chapter bibliography should help students in this endeavor.

VARIABILITY AND VARIABLES—The prime reason for the need of statistical approaches to the analysis of real-life data is the inherent variability present in experimental data, in particular, in biological material and laboratory processes. Variability has the same meaning in statistics as it does in

everyday usage. In its statistical sense, *variability* implies a lack of exact predictability of an experimental outcome. For example, although 50% of prescriptions are written for generic warfarin tablets, it cannot be predicted with certainty that a new prescription written by Dr Jones will be for the generic product. Conversely, the chance that the new prescription will be for the generic product is 1/2 or 0.50.

In statistical terms, variability commonly is called *error*. Measurement error does not mean that a mistake was made, but rather that the measurement yields inherently variable results.

A *variable*, simply put in statistical jargon, is a measurement that exhibits variability. Practically all measurements in scientific research and data collection are variable. Variables can be divided conveniently into two classes, discrete and continuous.

Discrete data have a countable number of possible outcomes. The number of animals that die when 12 animals are given 10 mg/kg of an experimental drug in an LD₅₀ experiment, or the number of bottles missing a label in a packaging run of tablets, or the proportion of patients with a successful outcome in a clinical study, are examples of discrete variables. In the former case, the number of animals that could die in the experiment could be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12. There are 13 possible outcomes. The number of dead animals is a discrete variable. Similarly, the number of bottles without labels is an integer that can vary between 0 and *N*, where *N* is the number of bottles in the run.

A *continuous* variable is one in which there are an unlimited (infinite) number of possible outcomes in some interval. The weight of a tablet may be any value between 180 and 220 mg, for example. The only limitation on the weight measurement is the accuracy and precision of the weighing device. Blood pressure measurement is a continuous variable. Although the actual measurement may appear to be limited to some countable number of outcomes—integers between 0 and 300, for example—this is due only to the approximate nature of the measuring instrument, the sphygmomanometer. With a more sophisticated device, one could expect a systolic blood pressure to be any value such as 160.629837465 torr. This exaggerated example is meant only to illustrate that the number of decimal places is limited only by the precision of the measuring device. To make this concept clearer, Table 12-1 gives further examples of discrete and continuous data encountered in pharmaceutical science.

SAMPLES AND POPULATIONS—Many experiments have as an objective the definition or comparison of two or more groups of data. For example, one may wish to compare the efficacy of two antihypertensive agents, or a new antipsychotic drug versus a placebo. Or it may be desired to estimate the average

drug content and variability of a batch of tablets. In virtually all such experiments, it is not realistic to observe all possible experimental units. In fact, sometimes the entire population of conceivable observations cannot be identified completely. The potential experimental material for a clinical study comparing an antipsychotic drug to a placebo would include not only patients but also persons with the disease who are not yet diagnosed. All of these people are the population or universe. Clearly, one would not perform an experiment that included the entire population for many reasons:

- All of these people could not be identified.
- The time or money to conduct such a huge experiment is not available.
- To include so many people in such an experiment could be dangerous or unethical.

It is not necessary to run such a large experiment to arrive at a fair conclusion regarding the efficacy of the drug. In fact, in most cases, the test consists of a relatively small *sample* taken from a relatively large *population*.

Another more concrete example is the process of sampling in quality control. It may be of interest to estimate the proportion of defective tablets or the average drug content and uniformity of tablets in a production batch. Certainly in the latter case every tablet in the batch would not be examined because the test is destructive, ie, the tablet is destroyed during the analysis for drug content. Rather, a sample of 20 tablets would be chosen to estimate the average drug content of the more than 1 million tablets in the batch.

Thus, in typical experiments in the pharmaceutical sciences, a small sample from the population is examined in order to make inferences about the large population.

THE AVERAGE OR MEAN—Suppose that a sample of *n* objects is taken from a population or universe in order to estimate some characteristic of the population, such as the average reduction of blood pressure after drug treatment, the average age of consumers purchasing an over-the-counter (OTC) acne product, or the average dissolution rate of drug from a tablet. The sample of *n* determinations can be designated by

$$x_1, x_2, x_3, \dots, x_n$$

The sample mean, \bar{x} , is calculated as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \sum x_i/n$$

where *i* goes from 1 to *n*.

The *sample mean*, \bar{x} , estimates the *actual* or *true population mean*, designated as the Greek letter mu (μ). That is, the sample mean would not be expected to exactly equal the population mean μ in any given experiment, but should equal the population mean on the average. Figure 12-1 illustrates this idea.

Example 1—The weights, in mg, of nine tablets are

201	204	200
203	202	207
209	206	207

The average, \bar{x} , is $\sum x_i/n = 1839/9 = 204.33$ mg.

The *average* is a measure of the center of a set of data. Another measure of central tendency is the median. The *median* divides the data set in half; that is half of the data is below the median and half is above. For a sample with an odd number of observations, the median is the middle number—the (*n* + 1)/2 data point—after the data has been listed in order of magnitude.

200, 201, 202, 203, 204, 206, 207, 207, 209

For the tablet weights in this example, the median is 204 mg, the 5th, (9+1)/2, ordered value.

For an even number of data points, the median is the average of the two middle values after the data have been ordered.

MEASURES OF VARIATION—The mean alone is not sufficient to describe a set of data. When describing data, in

Table 12-1. Examples of Discrete and Continuous Data

Measurement of LD₅₀—Although the number of animals dead at each of a series of doses is a discrete variable, the measurement of LD₅₀ is a continuous variable. For example, the LD₅₀ could take on any value between 1 and 100 mg, limited perhaps only by the precision of the analytical computations.

Preference Tests—If 100 consumers are asked for their preference for one of two products, the number who prefer one of the products is discrete.

Defects in Quality Control—The number of defects observed in a sample of 200 capsules sampled for quality control is a discrete variable.

Dissolution Test—The average time for 50% dissolution obtained from 12 tablets is a continuous variable. The 50% dissolution time is interpolated from the data. The dissolution time of the 12 tablets can have any number of possible outcomes of the average dissolution, limited only by the sensitivity of the measuring instruments, ie, the measurement of time and amount of drug dissolved.

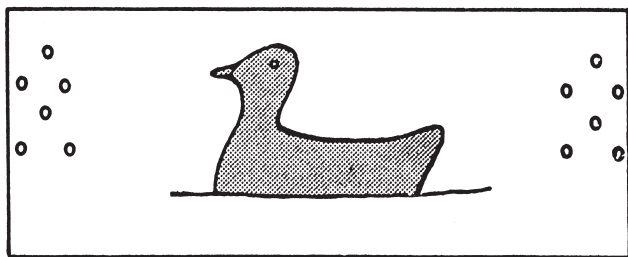


Figure 12-1. On an average the duck was dead. A hunter fired both barrels of a shotgun at a duck. The first hit 2 feet in front, the second hit 2 feet behind, and on an average the duck was dead. What the hunter really wanted was meat on the table. In duck hunting one wants to keep trying until a single shot hits the mark. But in estimating purity by a chemical test the best estimate is usually the average.

addition to the mean or average value, some measure of the variability or spread of the data should be calculated and reported. Two sets of data may have the same mean or average, but may have different distributions.

Example 2—The data in Example 1 have a mean of 204.33 mg. These data are reproduced below.

201	204	200
203	202	207
209	206	207

The following data set also has a mean of 204.33.

151	154	150
153	202	257
259	256	257

Clearly, the second set of data is spread out more, ie, it is more variable than the first set. The difference between the largest and smallest value in a data set is known as the *range*. For the first data set, the range is 209 – 200 = 9. In the second data set, the range is 259 – 150 = 109.

The standard deviation is a more common way of expressing the variability of data. The *standard deviation* of a sample of *n* values, designated as *S* or *SD*, is calculated as

$$S = SD = \sqrt{\sum(x_i - \bar{x})^2 / (n - 1)}$$

The standard deviation of the numbers 1, 3, 5, 9, and 12 is

$$\begin{aligned} &\sqrt{\sum(x_i - \bar{x})^2 / (n - 1)} \\ &= \sqrt{[(1 - 6)^2 + (3 - 6)^2 + (5 - 6)^2 + (9 - 6)^2 + (12 - 6)^2] / 4} \\ &= \sqrt{80 / 4} = \sqrt{20} = 4.47 \end{aligned}$$

Exercise 1—Calculate the SD of the two sets of data in Example 2 above.

Answer: 3.08 and 52.65, respectively.

A shortcut formula for computing the SD is

$$\sqrt{[\sum x_i^2 - (\sum x_i)^2 / n] / (n - 1)}$$

For the numbers 1, 3, 5, 7, 9, and 12, the computation is

$$\sqrt{[1^2 + 3^2 + 5^2 + 9^2 + 12^2 - 30^2 / 6] / 5} = \sqrt{20} = 4.47$$

The sample SD calculated as shown above is an estimate of the *population SD*, designated as the Greek letter sigma (σ). As with the mean, the population SD is usually unknown. One can obtain an estimate of σ from the sample SD.

The SD measures the spread of a data set, but it is more difficult to interpret than the range. When the normal distribution is introduced, the SD will have a more tangible interpretation. For the moment, it can be said that the larger the spread of numbers in a data set, the larger the SD and vice versa.

The *coefficient of variation* or *relative standard deviation* (RSD) is defined as SD/\bar{x} . This manner of expressing variability is useful when the SD is proportional to the magnitude of the measurement. This relationship often is seen in physical and biological measurements. For example, the analysis of large amounts of material often will have larger variability than the analysis of small quantities.

A very important concept in statistics is the *standard error of the mean*, designated as $s_{\bar{x}}$. Intuitively, one would expect that means of *n* observations would be less variable than the single, individual observations. The individual observations vary from extremes on the low side (below the average) to high values (above the average). When the means of 10 observations are taken, for example, the means will tend to be closer to the true average, μ , than the individual values. This can be better understood by visualizing the averaging effect of the mean, averaging extreme values with the other observations. In fact, the smaller variability of means can be proved mathematically; the SD of means of size *n* is equal to

$$s_{\bar{x}} = s / \sqrt{n}$$

For example, if the SD of individual values is 10, the standard error of means of size 25 is $10/\sqrt{25} = 2$. Thus, the means of size 25 are considerably less variable than the individual data points.

An examination of the equation for the standard error of the mean reveals that means constructed from very large sample sizes will be very stable, ie, nonvariable. If individual measurements are very variable, and a precise estimate of the mean is desired, this can be attained by making observations on a large number of samples. Of course, this is more easily said than done. Time and expense usually are limiting factors in data gathering and observation. However, it is true that the more observations, the more precise is the estimate of the mean (as well as estimates of other parameters such as the SD).

FREQUENCY DISTRIBUTIONS—A *frequency distribution* of a data set can be constructed by counting the number of data points falling into a series of intervals (usually of equal size). The frequency distribution and its corresponding graph, a *histogram* or *bar chart*, show the distribution of the data, its central value (eg, mean or median) and variability (eg, SD or range). Example 3 shows the weights of 50 weanling rats to be used in an experiment.

Example 3—The weights of 50 rats at weaning were as follows:

30g	47g	37g	29g	38g
32	42	32	30	34
34	32	33	37	36
39	33	45	40	35
43	41	35	32	41
36	27	28	35	30
38	28	41	37	34
41	36	32	30	37
31	31	35	28	25
26	49	34	34	33

Table 12-2 is a frequency distribution with 13 intervals derived from the data given in Example 3. A rule of thumb is to use 8 to 20 intervals, depending on the quantity and spread of the data. The histogram or bar chart of these data is shown in Figure 12-2.

BIAS, PRECISION, AND ACCURACY—*Precision* refers to the reproducibility of a series of measurements. If the values are very close to each other, the measurements are said to be precise. *Accuracy* refers to the closeness of measurements to the true value. For example, if a tablet contains exactly 200 mg of drug, and three analyses show a drug content of 205, 205, and 206 mg, it might be concluded that the analysis is precise, but not accurate. *Bias* refers to a systematic difference from the true value. Figure 12-3 illustrates these concepts.

The three assays observed above seem to be biased on the high side, ie, errors in the assay procedure result in too-high

Table 12-2. Frequency Distribution of Rat Weights

WEIGHT GROUP	FREQUENCY	WEIGHT GROUP	FREQUENCY
24–25 g	1	38–39	3
26–27	2	40–41	5
28–29	4	42–43	2
30–31	6	44–45	1
32–33	8	46–47	1
34–35	9	48–49	1
36–37	7		

values. Figure 12-3 shows that “precise” data need not be accurate. In fact, there need not be any relationship or correlation between the qualities of precision and accuracy. Note that biased data cannot be accurate but can be precise.

In addition to the concept of bias in the area of experimental measurements, it appears also in the field of experimental design. Bias can be introduced into an experiment, not because of an error in an experimental measurement, but because of poor judgment. For example, consider an experiment where the efficacy of oral and sublingual nitroglycerin are to be compared by administering both products to 20 patients on two different occasions and measuring the time to incidence of an angina attack in a treadmill test. Each of 20 patients will receive both the oral and buccal forms. If each patient receives the buccal drug on Monday and the oral drug on the following Sunday, a bias may be observed in the experimental results even if the measurements are not biased. This could be due to either the day of the week when the test was given (gloomy Monday versus a holiday weekend day) or an order effect where there is a different effect depending on which drug is given first. For example, there may be psychological factors causing the response to drug taken first to be systematically better (or worse) than that taken second, or the weather may be such as to cause more positive results on the first occasion. In the latter case, differences between the two dosage forms would be exaggerated (biased) in favor of the drug administered first, the buccal drug. To obviate this potential bias, we would give ten of the patients the oral drug first (Monday) and the buccal drug second (Sunday). The other ten patients would receive the products in opposite order. Perhaps, an improvement in this design would be to test the drugs on the same day of the week, eg, Monday.

DESIGN OF EXPERIMENTS AND COLLECTION OF DATA

The application of statistics in the analysis of data is optimal when the data are collected in a planned or designed manner.

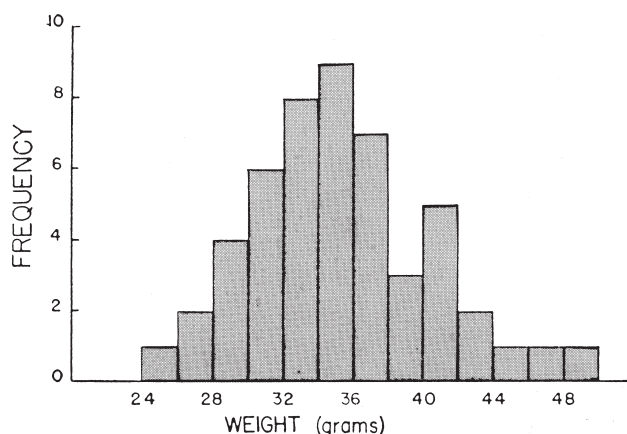


Figure 12-2. Bar chart showing frequency distribution of weights of 50 weaning rats (data in Example 3).

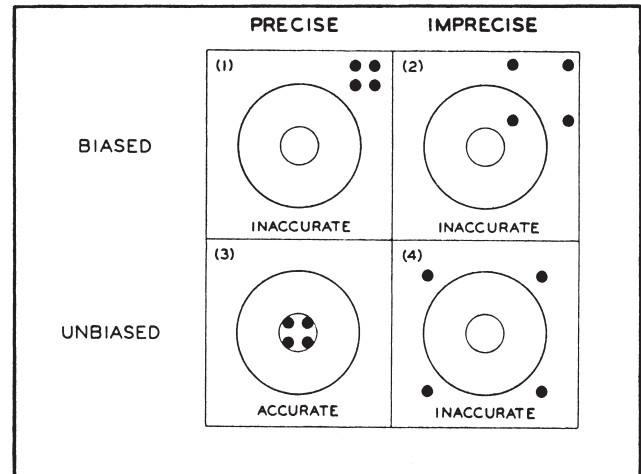


Figure 12-3. Diagram illustrating bias, precision, and accuracy. The shots on targets 1 and 2 are biased; in both cases the shots cluster away from the bull’s-eye. The clusters on targets 3 and 4 both are unbiased; the center of each cluster is on the bull’s-eye. The shots on targets 1 and 3 are precise; both sets are bunched together. The shots on targets 2 and 4 are scattered widely, hence imprecise. Only the shots on target 3 are accurate—precise and unbiased (courtesy, Lilly).

If data are analyzed after the fact (retrospective analysis), great care should be taken to examine the data for possible bias. For example, prescription-volume data gathered for the years 1970 to 1980 may be available only from cities with populations greater than 500,000 or from cities in the Western States. Clearly, conclusions from such data should not be applied indiscriminately to the entire country. Also, the information may have been gathered on a voluntary basis; without knowledge of the characteristics of those who did and did not supply the information, the conclusions could be tainted.

The manner in which data are collected is connected to the planning and design of experiments. In the collection of data, a small sample generally is taken from a large population or universe. Sometimes a sample is taken inadvertently when the original intention was to obtain data from the population. For example, when a questionnaire is sent to every pharmacist in the state, there always will be some people who do not respond to the questionnaire, and anything less than 100% response constitutes a sample. A variety of examples of sampling methods is illustrated below.

SAMPLING BY QUESTIONNAIRE—Suppose that questionnaires on the sales of certain drugs were sent to all pharmacists in a state and only 50% were returned. In this type of survey, the results tabulated from such a sample probably would be biased because those who did not return the questionnaire would not be represented in the sample.

It has been shown that persons who respond may have different characteristics from those who do not respond. In this hypothetical example, perhaps unanswered questionnaires were represented largely by pharmacists who had large drug sales and were too busy to answer. In another community, a pharmacist may have little or no sales of the drugs, resulting in a nonresponse. The reason for each unanswered questionnaire is unknown. These unreturned questionnaires cause a bias, the direction and magnitude of which is unknown.

Other potential errors in this type of response that may introduce bias include the way in which the question is asked, the order in which questions are asked, and the psychological interaction between the interviewer and respondent. Questionnaire and survey techniques that can be employed to reduce or eliminate bias in the sample of responses have been proposed by mathematical statisticians.²

For example, public opinion polls use certain statistical sampling techniques that not only reduce bias but also optimize the information gathered. The Census Bureau has information about the percentages of men, women, and children in the US in various income and nationality groups, in addition to many other detailed categorizations. A sample may be designed to contain the same proportion of particular group(s) as that in the population. Instead of mailing questionnaires, interviewers may be recruited and assigned quotas of the types of people to interview. The interviewers fill out the questionnaires for each respondent during the interview, ensuring a complete response.

It is not possible to elaborate fully on the various methods of sampling here. One should be aware of problems in sampling, and that a sampling design can be used that will give the limits of error of the resulting compilation for any given cost.²

SAMPLING IN THE CHEMICAL LABORATORY—The procedure for gathering data in the laboratory differs from that of the questionnaire. Different kinds of sampling processes include the sampling of material to be assayed chemically or physically, sampling of analytical reagents and instruments when multiple instruments are available, and sampling of analysts, the chemists who will perform the assay.

By way of illustration, several samples may be taken from a large lot of digitalis leaves for the chemical determination of acid-insoluble ash, or drug may be analyzed in samples taken from a blend. For the sample to be representative of the lot, the samples should be taken from different parts of the lot to ensure that every part of the lot is represented. Determinations from five samples taken from the same part of a lot (eg, the top of a container) probably will have values closer together than five samples taken from different parts of the lot (eg, the top, top-middle, middle, low-middle, and bottom of a container). Despite the good precision, the former five samples may give a biased estimate of the average value for the lot. The more heterogeneous the lot, the more effort should be expended in being sure that every part of the lot is represented by a sample. It might be that the granulation having the most drug is in the bottom of the lot; samples all taken from the top would give too low an estimate of average drug content of the lot in this example.

Another aspect of sampling in a chemical determination is the sampling of the chemists who perform the chemical analysis. If a single chemist makes several determinations on portions taken from the same sample of thoroughly mixed material, one expects the results to be more precise than if several chemists made these determinations. Probably the true reproducibility of a method can be indicated only in terms of how closely an analyst at one laboratory can check an analyst at another laboratory on exactly the same material. Thus, due to slight differences in technique, one chemist always might obtain higher results than another chemist. Thus, the technique of chemists will have an effect on the results and the reproducibility of the method.

SAMPLING IN BIOLOGICAL AND CLINICAL EXPERIMENTS—A typical animal experiment might involve determining the temperature response of rabbits to pyrogens. The results of such an experiment constitute a sample of all possible results that could be obtained from the population of all possible rabbits, laboratories, and technicians. Using different rabbits, laboratories, and technicians will give different results, all contributing to the variability or error in the experiment. The differences between results from two or more laboratories are usually greater than differences between results obtained by two or more technicians in the same laboratory.

Concurrent conditions—such as season of the year, temperature, and humidity—sometimes can contribute to the experimental variability. In biological experiments, differences between animals are relatively large, so experiments repeated in the same laboratory with different animals but under otherwise identical conditions will give different results. The use of statistical procedures gives an estimate of the amount of variation to be expected due to animal differences. The same can be said of clinical studies where more than one clinical site is needed to produce reliable, unbiased results.

Appropriate statistical designs and procedures will eliminate or account for potential bias in experiments. This point

Table 12-3. A Short Table of Random Numbers

39	61	09	51	68	81	26	30	52	20	61	41	25
89	35	48	61	72	10	84	34	10	44	72	94	77
37	98	37	56	40	30	70	31	75	03	68	32	15
20	55	68	05	53	73	60	28	96	48	91	81	18

may be illustrated by an extreme example, an illustration of what not to do. A technician wishes to compare two drugs as to their effects on the growth of rats. Thirty rats from a single cage are used; the first 15 rats caught are put on Drug 1 and the last 15 caught are put on Drug 2. The first 15 rats caught are less lively than the last 15, and because they are less lively they very likely differ in size and temperament from the last 15 rats. Thus, the results were biased from the very beginning, and one drug was favored merely because of the method of choosing the animals used for each drug.

Obviously, some method entirely free from subjective influences (unconscious or conscious) should be used. A table of random numbers³ or computer-generated random numbers commonly is used to assign animals or patients to treatments. Table 12-3 is a short table of random numbers.

Example 4 (The use of the random number table)—Suppose that 10 patients are to be assigned to two treatment groups, five in each group. Table 12-3 can be used to assign patients randomly to groups. Patients first are numbered from 1 to 10. One way of assigning treatments to patients is to read across Table 12-3, and the first five distinct numbers will be assigned to the first treatment. The remaining patients are assigned to the second treatment. A zero will correspond to patient number 10. The first five numbers are 3, 9, 6, 1, and 0. (Note that if a number repeats itself, we skip the number and proceed to the next one.) Therefore, patients numbered 3, 9, 6, 1, and 10 are assigned to the first treatment. If 100 patients are to be assigned to the two treatments, two-digit numbers would be used: reading across, the 50 patients assigned to the first group would be numbered 39, 61, 9, 51, and so on.

There are many ways of using random numbers to ensure randomness in statistical experiments. The number of ways is limited only by the ingenuity of the experimenter. For example, random assignment could be accomplished by assigning patients to Group 1 or 2 as they enter the study, according to the appearance of an odd or even number in the random-number table.

In a biological assay, it often is advantageous to design a dosage schedule to take advantage of the reduced within-animal variation compared to between-animal variation. Because more than one dose may be given to a single animal, the order of dosing also must be designed to account for possible trends in response to consecutive doses caused by changes in the animal with time or due to site of application. This can be illustrated by an epinephrine assay (see *Remington's Practice of Pharmacy*, 14th ed, page 633), where a single dog is given 16 consecutive doses, the order of which is determined by a Latin square design, illustrated by

A	D	B	C
D	C	A	B
B	A	C	D
C	B	D	A

Note that in a Latin square each letter occurs only once in each row and each column of the square. A Latin square design was applied to an assay involving two levels of doses of the standard (high and low doses, s_H and s_L , respectively), and two levels of doses of the unknown (u_H and u_L), where the four doses correspond to the letters, A, B, C and D. The dosage schedule is given in Table 12-4. In this type of design each dose occurs once in each order of administration (eg, each of

Table 12-4. Typical Dosage Schedule for an Epinephrine Assay Using a Latin Square Design

	FIRST DOSE	SECOND DOSE	THIRD DOSE	FOURTH DOSE
First group	u_L	s_H	u_H	s_L
Second group	s_H	s_L	u_L	u_H
Third group	u_H	u_L	s_L	s_H
Fourth group	s_L	u_H	s_H	u_L

the four preparations are represented once in each group). In such an assay equal doses of epinephrine elicit a smaller and smaller rise in blood pressure with each succeeding dose. Therefore, order is important.

In all biological experimentation the design should be planned so that differences in treatment do not coincide with factors that could influence the outcome such as differences in age, weight, sex, dates of administration, and so forth. This is known as *confounding* in statistical jargon. For example, if males are given a control treatment and females are given a comparative active treatment, the differences between treatments are said to be confounded by sex. That is, it cannot be determined if the outcomes observed are due to treatment, sex, or a combination of these factors.

Animals or patients should be assigned to doses or treatments at random, taking advantage of the availability of optimal experimental designs. Fisher⁴ has written an excellent book on planning or designing experiments, which explains fully the various types of designs mentioned here. Cochran and Cox⁵ detail useful experimental designs and provide complete directions for the analysis of data using these designs. Another book by Cox⁶ is less mathematically oriented and comprehended more easily.

DESIGN AND CONDUCT OF CLINICAL TRIALS—

Proof of the efficacy and safety of new drugs or treatments requires testing in human subjects. This is best achieved by carrying out *controlled clinical trials*. The use of a placebo treatment or an established treatment as a *control*, a basis of comparison, usually is necessary. Thus, the effects of treatment with those of a concurrently tested control or placebo are compared. The trial includes an adequate number of patients to allow a reliable projection of the results to future patients. Theoretically, the results cannot be projected beyond the types of severity of disease or the ages and sex of the patients included in the trial, although in practice this is not always the case.

The distribution of variables such as age, sex, differences in diagnosis, and initial severity of disease among treatments may be controlled by *stratification*. Usually patients are assigned to treatments at random, and allowances are made for the effects of the variables by using suitable statistical methods. A restricted randomization procedure is useful if it is desired to insure that about an equal number of patients enter the trial on each treatment. Table 12-5 illustrates a completely randomized design in which 15 patients are allocated at random, five to each of three treatments.

Note that the individual patients in each triad (the groups of three) are assigned randomly to one of the three treatments. Here the randomization is restricted in that each set of three patients when entered must be assigned to Treatments A, B and C. The patients are assigned to Treatments as they enter the trial. The first patient (#1) gets Treatment B. This scheme prevents runs in the randomization where a long consecutive number of patients are assigned to the same treatment. Another example is shown in Table 12-6 for a simple crossover design in which the individual patients take both Treatments consecutively, and are assigned randomly to one of two treatment order groups.

The latter design may be more efficient than a completely randomized design because each patient acts as his or her own control, thus eliminating patient-to-patient variability in the statistical analysis. However, this advantage may be offset if drug carryover effects are present, or if the severity of the dis-

Table 12-6. Crossover Design

GROUP	PATIENT	PERIOD 1	PERIOD 2
I	1		
	4		
	5	A	B
	7		
	10		
II	2		
	3		
	6	B	A
	8		
	9		

ease wanes in the second period to the point where treatment differences no longer can be demonstrated.

To be certain that the random allocation is followed strictly and to remove subjective bias on the part of both the patient and the clinical investigator in assessing the effects of the treatments, the clinical trial should be carried out blind. A *double-blind trial* is one in which neither the patient nor the investigator is made aware of the nature of treatment administered.

To ensure that the study remains blinded, all treatments must be packaged as identical-appearing dosage forms. This may require a great deal of ingenuity on the part of the packaging pharmacist, especially with respect to the taste of orally administered liquid products, the color and shape of tablets, and so on. In some cases, the characteristic side effects of the drugs make it difficult to keep a study blind. In these situations, one must rely more heavily on objective measures of response, and less on subjective measures. However, the *placebo effect* may also result in changes in so-called *objective* measure of response.

Federal regulations require that drugs shipped to clinical investigators must be labeled properly with the name of the drug. To keep the study blind, one suggested procedure is to use a two-part tear-off label.

One part is glued to the container and reveals only the patient's study number, the period number, and directions for taking the drug; the tear-off part shows the identity of the drug. The name of the drug is overlaid with a water-washable or erasable ink so as not to reveal the identity of the drug to the investigator. This portion of the label is torn off and stapled to the back of the clinical form. The investigator is instructed to break the code for an individual patient, if necessary, by washing off or erasing the overlaid ink.

Laboratory determinations in clinical studies usually include such measures as complete blood count, liver function tests, and analyses on urine and stool specimens. The occurrence of adverse effects may be recorded as ascertained by inquiry or as volunteered by the patient. It is informative also to determine the severity as well as the frequency of occurrence of adverse effects, and whether the investigator feels the effects were drug related.

Generally, it is more difficult to evaluate clinical data than laboratory animal data. Some of the contributing factors are

- The failure of patients to take the medication as directed and to report for examination at stated intervals.
- Patients' use of ancillary or concomitant medications.
- Incomplete data that may result from patients dropping out of the study for various reasons.

These factors are more prevalent among outpatients than among hospitalized patients. A trial secretary or Clinical Research Associate (CRA) can be of great help in assuring the completeness and accuracy of clinical forms.

Table 12-5. Allocation of Patients in Randomized Design

A	B	C	A	B	C
3	1	2	12	10	11
6	5	4	14	13	15
8	9	7			

THE BINOMIAL AND NORMAL PROBABILITY DISTRIBUTIONS

Statistical conclusions are based on *probability*. The process of *statistical inference* first considers an assumption about the

distribution of the population data. If the observed data from the sample collected do not conform reasonably to the assumed distribution, the results are viewed as *significant*, ie, the sample data show significant differences from the assumed distribution. For example, it may be assumed or hypothesized that an antibiotic will cure 80% of the patients treated. If three of six patients are cured with the drug, this is the question: What is the probability that three or fewer of six patients treated will be cured if the probability of a single patient being cured is 80%? If this calculated probability is small, the probability of a cure is probably not 80%, but rather some lesser value.

To compute these probabilities, the properties of the assumed probability distribution must be known. Two important and often-used distributions in statistical theory are the binomial and normal distributions, which are examples of a discrete and continuous probability distribution, respectively. The experiment discussed in the preceding paragraph that related to the cure of patients treated with an antibiotic is an example of one application of the binomial distribution.

THE BINOMIAL DISTRIBUTION—The binomial distribution is applicable to data where one of two mutually exclusive and independent outcomes are possible as a result of a single observation or experimental trial. A patient may be cured or not cured. Only one of these two mutually exclusive events can occur at the time of observation. *Independence*, in this context, means that the probability of a cure for any given patient is 80%, regardless of the experimental outcome of the other patients in the study.

The problem to be solved is to compute the probability that three (or less) of six patients will be cured if the probability of a cure for an individual patient is 0.8, or 80%. The general solution to this problem uses the binomial distribution. If two independent and mutually exclusive outcomes are possible as the result of an experimental trial, the probability of x outcomes of one kind (arbitrarily called *successes*) in n binomial trials (n patients in this example) is

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

where $P(x)$ is the probability of exactly x successes and n is the number of binomial trials.

$$\binom{n}{x} \text{ is } \frac{n!}{(x!(n-x)!)}$$

(! means factorial. For example $5! = 5 \times 4 \times 3 \times 2 \times 1$.
 $0! = 1$ by definition.)
 p = probability of success
 $q = 1 - p$ = probability of a failure
 (note that $p + q = 1$)

Now it is possible to calculate the probability of exactly three successes (cures) in six trials (patients) if $P = 0.8$; ie, the probability of a success or cure is 0.8.

$$\begin{aligned} P(3) &= \binom{6}{3} 0.8^3 0.2^3 \\ &= \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(3 \times 2 \times 1)} \times 0.512 \times 0.008 \\ &= 20 \times 0.512 \times 0.008 = 0.082 \end{aligned}$$

Thus, the probability of exactly three cures in six patients is 0.082. This is interpreted to mean that the chance of observing exactly three successes in six binomial trials with $P = 0.8$ is approximately 8 in 100.

There are seven possible outcomes for the treatment of six patients as shown in Table 12-7, an example of a binomial probability distribution defined by $n = 6$ and $P = 0.8$. It lists all the possible outcomes, with the probability of each outcome. The sum of all the probabilities is equal to 1. This distribution is shown graphically in Figure 12-4. A knowledge of this distribu-

Table 12-7. Binomial Distribution for $n = 6$ and $P = 0.8$

NUMBER OF SUCCESSSES	PROBABILITY OF OUTCOME	NUMBER OF SUCCESSSES	PROBABILITY OF OUTCOME
0	0.0000026	4	0.24576
1	0.001536	5	0.39322
2	0.01536	6	0.262144
3	0.08192		

tion allows a decision to be made as to whether three or fewer cures in six patients is a probable outcome for patients treated with a drug that has a cure rate of 80%. The probability of observing three or less successes (0, 1, 2, or 3 successes) is $0.08192 + 0.01536 + 0.001536 + 0.0000026 = 0.0988$, or about 1/10. Is this sufficient evidence to say that the true probability of a cure for the drug is less than 0.8? This question will be discussed in more detail in the section on *Statistical Inference*.

Table 12-8 lists individual probabilities for $P = 0.2, 0.5$, and 0.8 , for N equal to 6 to 10, inclusive. For probabilities not listed in this table, the student should consult tables of binomial probability distribution⁷ or use one of the statistical software packages listed at the end of this chapter.

Exercise 2—Calculate the probability of four successes in six trials for $P = 0.8$.

Answer: 0.246.

The mean of the binomial distribution can be expressed in two equivalent ways. In terms of probability (or proportions), the mean is equal to P , the probability of success. In terms of the number of successes in n trials, the mean is NP . Thus, for the binomial distribution with $P = 0.8$ and $N = 100$, the mean is $P = 0.8$ or $NP = 80$. That is, if 100 patients were treated with the antibiotic that has a cure rate of 80%, one could expect to see 80% or 80 patients cured of 100 treated on the average. The standard deviation of a binomial distribution is $\sqrt{pq/n}$ or \sqrt{npq} , depending on whether one is looking at P or NP , respectively. The standard deviation of the proportion of patients cured of 100 treated in the above example is

$$\sqrt{pq/n} = \sqrt{0.8 \times 0.2/100} = 0.04$$

The standard deviation of the number cured is

$$\sqrt{npq} = \sqrt{100 \times 0.2 \times 0.8} = 4$$

This can be interpreted as follows. If 100 patients are treated, it may be expected that 80 are cured on the average, but in any given experiment one probably would not see exactly 80 cured. The number cured will vary around 80, the mean, with a standard deviation equal to 4.

THE NORMAL DISTRIBUTION—The normal distribution can be considered as the underlying foundation of statistical theory and its applications. It is a continuous probability distribution with values ranging from $-\infty$ to $+\infty$. Each of the infinite number of different normal distributions is defined by its mean and standard deviation. The mean can be any positive or negative value, but the standard deviation must be a positive

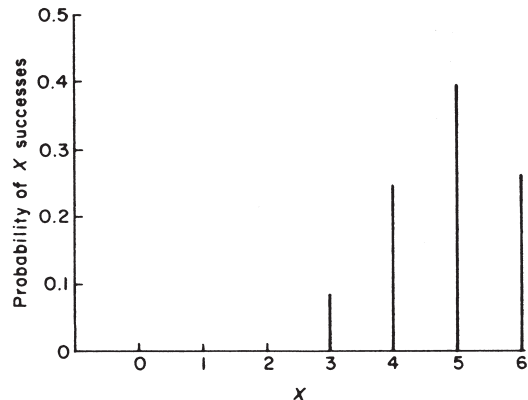


Figure 12-4. Binomial distribution for $n = 6$ and $P = 0.8$.

Table 12-8. Short Table of Binomial Probabilities

		$P = 0.2$ Probability of x successes in n trials										
x	n	0	1	2	3	4	5	6	7	8	9	10
6	6	0.262	0.393	0.246	0.082	0.015	0.002					
7	7	0.210	0.367	0.275	0.115	0.029	0.004					
8	8	0.168	0.336	0.294	0.147	0.046	0.009	0.001				
9	9	0.134	0.302	0.302	0.176	0.066	0.017	0.003				
10	10	0.107	0.268	0.302	0.201	0.088	0.026	0.006	0.001			

		$P = 0.5$ Probability of x successes in n trials										
x	n	0	1	2	3	4	5	6	7	8	9	10
6	6	0.016	0.094	0.234	0.313	0.234	0.094	0.016				
7	7	0.008	0.055	0.164	0.273	0.273	0.164	0.055	0.008			
8	8	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004		
9	9	0.002	0.018	0.070	0.164	0.246	0.246	0.164	0.070	0.018	0.002	
10	10	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

		$P = 0.8$ Probability of x successes in n trials										
x	n	0	1	2	3	4	5	6	7	8	9	10
6	6		0.002	0.015	0.082	0.246	0.393	0.262				
7	7			0.004	0.029	0.115	0.275	0.367	0.210			
8	8			0.001	0.009	0.046	0.147	0.294	0.336	0.168		
9	9				0.003	0.017	0.066	0.176	0.302	0.302	0.134	
10	10				0.001	0.006	0.026	0.088	0.201	0.302	0.268	0.107

value. Figure 12-5 shows two normal probability curves. The normal distribution is characterized by the symmetry about its mean; most of the data cluster around the mean. There are fewer values as the deviation is farther from the mean. The normal distribution is a theoretical probability distribution, not exactly observed in practical situations. However, much data approximate the normal distribution closely enough to make its application useful.

The Central Limit Theorem (CLT) is perhaps the most powerful theorem in statistics. It supports the pervasive use and importance of the normal distribution in statistical analyses. In simple terms, the CLT states that averages or means approach normality as n , the sample size, increases, no matter what the distribution of the individual variables. For data that are close to normal, means from even a small sample size will be approximately normal. For data that have distributions far from normal, larger sample sizes will be needed for the averages to be close to normal. The concept of the CLT is illustrated by the following example.⁸

The outcome of a disease after treatment can be (1) death = 1, (2) not cured but continue treatment = 2, and (3) cured = 3. The probabilities of these three outcomes are 0.1, 0.3, and 0.6, respectively. This distribution is shown in Figure 12-6. This is a discrete distribution (three pos-

sible outcomes in a single trial), and it clearly is not normal. Figure 12.7 shows the distribution of means of size 20 ($n = 20$). The means are obtained by treating 20 patients, assigning outcomes of 1, 2, or 3, according to the previous definition, and computing the mean. The distribution shown in Figure 12-7 was constructed from a computer simulation representing outcomes that can be expected in realistic situations. Note that the averages cluster almost symmetrically around 2.5 (the mean), and are beginning to look like a normal distribution. One also should note the small variability of the mean results, most values ranging between approximately 2.2 to 2.7. A single outcome varies from 1 to 3.

The CLT allows the use of statistical methods that assume an underlying normal distribution of the data when dealing with averages of data that do not come from a normal distribution.

COMPUTING PROBABILITIES FROM THE NORMAL DISTRIBUTION—The area under the normal curve is 1 and area represents probability. The probability of observing a single value from a continuous distribution such as the normal is 0. However, one can calculate the probability of observing values in any interval x_1, x_2 —designated as $P(x_1 \leq x \leq x_2)$ —by computing the area under the curve in that interval. Table 12-9 is a short compilation of cumulative probabilities from the standard normal curve (Fig 12-8) that has a mean of 0 and a standard deviation of 1. Table 12-9 shows the probability of

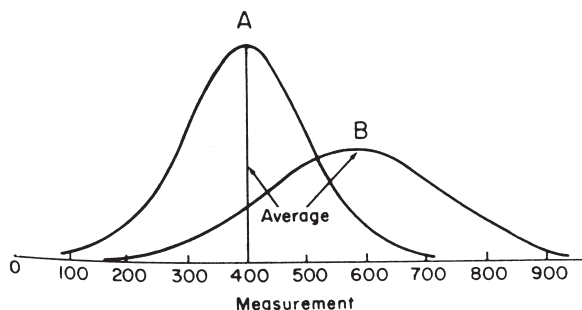


Figure 12-5. Normal probability curves.

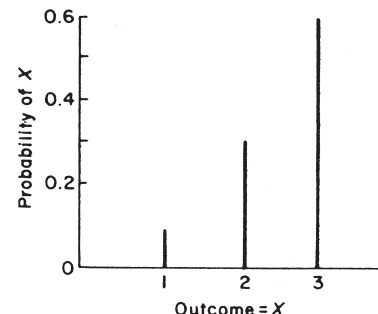


Figure 12-6. Probability distribution of outcomes after drug treatments.

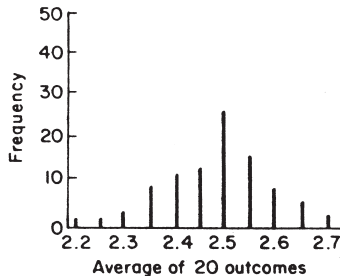


Figure 12-7. Simulation of distribution in Figure 12-6 with average of sample size of 20.

observing a value less than or equal to Z . For example, the probability of observing a value less than or equal to -1 from the standard normal distribution is 0.16. The symmetry of the normal curve indicates that the probability of a value being greater than or equal to $+1$ is also 0.16. Since the total area under the normal curve is 1 and area represents probability, the area less than or equal to $Z = +1$ is $1 - 0.16 = 0.84$. This relationship is illustrated in Figure 12-8. In general, to calculate the area in any interval, Z_1, Z_2 , look up the cumulative areas corresponding to Z_1 and Z_2 . The difference of the two areas is the area between Z_1 and Z_2 , or the probability of observing a value in that interval.

Exercise 3—Calculate the probability of a value falling between -1.96 and $+1.28$ for the standard normal curve.

Answer: The area corresponding to $Z = -1.96$ is 0.025. The area corresponding to $+1.28$ is 0.90. The difference is 0.875. Thus, the probability of observing a value between -1.96 and $+1.28$ is 0.875.

Since there are an infinite number of normal distributions (defined by their means and standard deviations) a reasonable question is, How would one calculate probabilities from a normal distribution that is different from the standard normal distribution? Fortunately, there is a simple transformation that converts data from any normal distribution into the standard normal distribution; Table 12-9 can then be used to compute the probabilities.

The Z transformation is

$$\frac{x - \mu}{\sigma} = Z$$

Example 5—What is the probability that a tablet will weigh between 185 and 210 mg if tablet weights have an approximately normal distribution with mean 200 mg and a standard deviation of 10? Using the Z transformation,

$$(185 - 200)/10 = -1.5$$

$$(210 - 200)/10 = +1.0$$

Table 12-9. Short Table of Areas for the Standard Normal Distribution (Area for Values Less Than Z)

Z	AREA	Z	AREA	Z	AREA
-3	0.0013	-1.28	0.1003	1.2	0.8849
-2.58	0.0049	-1.2	0.1151	1.28	0.8997
-2.3	0.0107	-1.0	0.1587	1.4	0.9192
-2.2	0.0139	-0.84	0.2005	1.5	0.9332
-2.1	0.0179	-0.8	0.2119	1.6	0.9452
-2.0	0.0228	-0.6	0.2743	1.645	0.950
-1.96	0.025	-0.4	0.3446	1.8	0.9641
-1.9	0.0287	-0.2	0.4207	1.9	0.9713
-1.8	0.0359	0.0	0.500	1.96	0.975
-1.7	0.0446	0.2	0.5793	2.00	0.9772
-1.645	0.050	0.4	0.6554	2.1	0.9821
-1.6	0.0548	0.6	0.7257	2.2	0.9861
-1.5	0.0668	0.8	0.7881	2.3	0.9893
-1.4	0.808	1.0	0.8413	2.58	0.9951
				3.00	0.9987

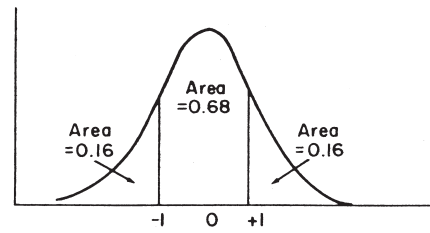


Figure 12-8. Standard normal distribution.

The cumulative areas corresponding to $Z = -1.5$ and $Z = 1.0$ are found from Table 12-9. These areas are 0.07 and 0.84, respectively. Therefore, the probability of finding a tablet weighing between 185 and 210 mg is $0.84 - 0.07 = 0.77$. Note carefully that the transformation is equivalent to finding a value that is between -1.5 and $+1.0$ standard deviations from the mean (ie, 185 is -15 mg from the mean, which is equal to -1.5 standard deviation units).

What is the probability that a tablet will weigh less than 180.4 or more than 219.6 mg? The Z transformation results in the values -1.96 and $+1.96$. The student can verify that 95% of the values are found in this interval.

Several Z values appear frequently when testing for statistical significance:

- 68% of the values are within ± 1 standard deviation of the mean value.
- 80% of the values are within ± 1.28 standard deviations of the mean value.
- 90% of the values are within ± 1.65 standard deviations of the mean value.
- 95% of the values are within ± 1.96 standard deviations of the mean value.
- 99% of the values are within ± 2.58 standard deviations of the mean value.

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION—The CLT can be applied to the binomial distribution if n , the number of binomial trials, is sufficiently large. As a general rule, if both np_0 and nq_0 are equal to or greater than 5 (p_0 is the true probability of success), the normal approximation can be used. For binomial distributions with p_0 close to 0.5, the approximation is good for values of np_0 and nq_0 smaller than 5. Under these conditions, $(p - p_0)/\sqrt{p_0q_0/n}$ is approximately normally distributed with mean 0 and standard deviation 1 (the standard normal distribution). This transformation allows easy calculation of binomial probabilities. The approximation is improved if $1/(2n)$ is subtracted from the absolute value of the numerator. This is known as the *Yates continuity correction*.

Example 6—When inspecting 100 tablets for quality, what is the probability of observing a proportion of defective tablets equal to or greater than 0.10, if the true proportion defective is 0.07? Using the continuity correction,

$$Z = \frac{|0.10 - 0.07| - 1/200}{\sqrt{0.07 \times 0.93/100}} = 0.98$$

From Table 12-9, the probability of a value less than $Z = 0.98$ is approximately 0.84. The probability of a value greater than $Z = 0.98$ is $1 - 0.84 = 0.16$. Therefore, the probability of observing a proportion greater than 0.10 when inspecting 100 tablets from this batch is approximately 0.16.

ESTIMATION AND STATISTICAL INFERENCE

ESTIMATION AND CONFIDENCE INTERVALS—After gathering data from, for example, a survey or an experiment, it is often of interest to estimate the mean value or aver-

age of the population. As has been noted, the sample average, \bar{x} , is not exactly equal to the population average, μ , but in a well-designed and implemented experiment, \bar{x} should be an unbiased estimate of the true mean. Thus, the best estimate of the true, but unknown, population average is the sample mean \bar{x} .

However, the mean of the sample gives no idea of the precision of this average. If the average assay of 10 tablets is 100 mg, it is not known how close this value is to the unknown true value. It would be important to have some estimate of the reliability of the result. *Confidence intervals*, or confidence limits, give an interval that may encompass the mean with a known probability. That is, a 95% confidence interval of 97 to 103 mg means that one would give 19 to 1 odds that the true mean is in this interval. It cannot be said for certain that the true mean is in the interval, but if the experiment were repeated many times and a 95% confidence interval constructed each time, then 19 of 20 such intervals would contain the true mean. For any given experiment, there is no way to tell if the true mean is in the interval, but it is known that the chances are 95% that the true mean is in the interval.

In statistical inference, no statements can be made with as-surity. Statistical proofs are not like mathematical proofs. Statistical conclusions are couched in terms of probability. A statement such as “The means are significantly different,” means that it is believed that the means are different but there is a chance, albeit a small one, that the conclusion is incorrect. However, the probability of making the wrong decision is known.

Symmetric confidence limits are computed as

$$\bar{x} \pm Z(\sigma_{\bar{x}})$$

where Z is an appropriate constant, depending on the probability statement (degree of confidence) associated with the confidence interval. For a normally distributed variable with standard deviation σ known, the value of Z is obtained from Table 12-9. For example, to obtain a 95% confidence interval, ± 1.96 standard deviations covers 95% of the area. For a 90% confidence interval, $Z = 1.65$; for a 99% confidence interval, $Z = 2.58$. Note that if the standard deviation is unknown but estimated from the sample data, the value of Z for normally distributed variables is replaced by t , obtained from the t distribution, which will be introduced in the next section.

Example 7—A drug shows an average blood pressure reduction of 9.8 torr when tested on 100 patients. The standard deviation is known to be 8 torr. A 95% confidence interval for the mean blood pressure reduction is

$$9.8 \pm 1.96 \times 8/\sqrt{100} = 9.8 \pm 1.57$$

A 99% confidence interval is

$$9.8 \pm 2.58 \times 8/\sqrt{100} = 9.8 \pm 2.06$$

Note that if the interval has a higher probability of containing the true mean, the confidence interval is wider.

Example 8—A survey of 1000 pharmacists showed that 30% have more than 15 yr of experience and 70% have less than 15 yr of experience. A 95% confidence interval on the proportion of pharmacists with more than 15 yr of experience is

$$\begin{aligned} p \pm 1.96 \sqrt{pq/n} &= 0.3 \pm 1.96 \sqrt{0.3 \times 0.7/1000} \\ &= 0.3 \pm 0.028 \end{aligned}$$

This means that the true proportion is between 0.272 and 0.328 with 95% probability.

On rare occasions, a *one-sided* or an *unsymmetrical* confidence interval may be appropriate. One use of a one-sided interval is described under linear regression as applied to stability prediction.

STATISTICAL INFERENCE AND THE T DISTRIBUTION—Statistics are used most often as a decision-making tool. The familiar phrase “the difference is statistically significant” results from the application of statistical inference to experimental data. The procedure allows a probability statement to be made about comparative data. Statements made using

this approach cannot be made with absolute certainty. Because experimental results generally come from sample data, one never can be sure of the exact properties of population data. However, decisions can be made with a known probability of error.

Example 9—Consider the estimation of the tablet potency of a batch of tablets based on an assay of 10 individual randomly selected tablets. The assay values (mg) are

98.6	99.3	97.9	100.3	99.6
98.0	100.1	97.5	98.4	99.1

The average is 98.88 mg and the standard deviation is 0.954. In this example, an estimate of the mean and standard deviation is obtained from a sample of size 10.

When the standard deviation is unknown but an estimate is available from a relatively small sample, the t distribution is used to describe the distribution of the means. The t distribution may be defined as the distribution of

$$\frac{\bar{x} - \mu}{SD/\sqrt{n}} = t$$

The t values show a symmetrical distribution centered at 0; ie, the mean is 0. The t distribution is spread out more than the standard normal distribution. Some commonly used points from the t distribution are shown in Table 12-10. The t distribution is defined by degrees of freedom (DF), which in Example 9 is $n - 1$. Note that when the DF are large (ie, n is very large), the values in the t table approach the corresponding values from the standard normal-curve table (see Table 12-9). For example, the value below which 97.5% of the area is found is 1.96 when the DF are infinite in the t table.

When the SD is unknown, values from the t table are used to construct confidence intervals in exactly the same manner as was done using Table 12-9. Table 12-10 shows t values that cut off areas of the t distribution in one tail or symmetrically in both tails of the distribution. For example, the two-tailed 5% points cut off 2.5% of the area in each tail. For DF = 9, t values below -2.262 and greater than $+2.262$ comprise 5% of the area. Conversely, it can be said that the probability of finding a t value between -2.262 and $+2.262$ is 95% for DF = 9.

Table 12-10 gives values for one-tailed probabilities for $P = 0.5\%$ and 2.5% . These values correspond to the two-tailed probabilities of 0.01 (1%) and 0.05 (5%). For example, for 9 DF, the probability of finding a t value greater than $+2.262$ (or -2.262) is 2.5%. Examples throughout the remainder of this chapter should make the use of the t table clear. In the current example of tablet assays, a 95% confidence interval can be constructed using the t distribution. The mean is 98.88 and the sample SD is 0.954. The t value for 95% of the area for 9 DF is 2.262. The 95% confidence interval is

$$\begin{aligned} 98.88 \pm 2.262 \times 0.954/\sqrt{10} &= 98.88 \pm 0.68 \\ &= 98.20 \text{ to } 99.56 \end{aligned}$$

This can be interpreted to mean that the probability is 95% that the true mean of the batch lies between 98.20 and 99.56 mg.

Are you surprised by the narrow limits of the interval based on only 10 tablets? The reason for the tight limits is the small standard deviation. Note that this does not guarantee that the true mean, μ , lies in this interval. As has been emphasized before, statistical statements and conclusions are probabilistic in nature.

T TEST—In addition to estimating the mean assay of a batch of 10 tablets, the 10 assay values were obtained to perform a statistical test comparing the average result to that expected based on the labeled potency of 100 mg. If every one of the 3,000,000 tablets in this batch were assayed, the average potency would be known. The random sample of 10 is representative of the entire batch, but it is extremely unlikely that the sample average exactly will equal the batch average. The question to be asked is, in view of the variability of the 10

Table 12-10. The *t* TableDistribution of *t* Giving Both the Two-Sided or Two-Tailed Probability and the One-Sided or One-Tailed Probability According to Degrees of Freedom

DF	ONE TAIL							
	<i>P</i> = 0.4	<i>P</i> = 0.3	<i>P</i> = 0.2	<i>P</i> = 0.1	<i>P</i> = 0.05	<i>P</i> = 0.025	<i>P</i> = 0.01	<i>P</i> = 0.005
	TWO TAILS							
	<i>P</i> = 0.8	<i>P</i> = 0.6	<i>P</i> = 0.4	<i>P</i> = 0.2	<i>P</i> = 0.1	<i>P</i> = 0.05	<i>P</i> = 0.02	<i>P</i> = 0.01
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

assays and the average result, can it be ascertained that these 10 tablets came from a population with an average of 100 mg? The solution to this question, an example of statistical inference, is obtained using a simple *t* test. This *t* test consists of the following steps, which can be considered typical in many designed experiments.

Construct a Null Hypothesis—A null hypothesis is an assumption about the parameter under investigation, which is the mean value in this example. The null hypothesis is a statement that assumes that the parameter is equal to some value, usually a null value. That is, the hypothetical value is considered to represent a situation of no change. How to construct the null hypothesis is not always obvious, but a few examples should make this concept clearer.

For the tablet assays, no change means that the population average, μ , is equal to the labeled potency, 100 mg. The null hypothesis is of the following form

$$H_0: \mu = 100 \text{ mg}$$

The statistical test allows a decision to be made: the sample of tablets are or are not representative of a population with mean 100 mg.

Construct an Alternative Hypothesis—An alternative hypothesis makes an assumption about alternative values of the parameter, usually encompassing complementary values. Thus, if H_0 is $\mu = 100$ mg, an alternative could include all values greater than or less than 100 mg. This is a two-sided alternative represented as $H_a: \mu \neq 100$ mg. In some cases, a one-sided alternative may be suitable. This may be expressed as $H_a: \mu > 100$ mg or $H_a: \mu < 100$ mg.

The process of statistical inference will result in one of two possible decisions: either accept or reject the null hypothesis. *Rejection* means the alternative is accepted. For a two-sided alternative, it is anticipated in advance that if the null hypothesis is not true, that the true average

could be either greater or smaller than the hypothetical or assumed value. A one-sided alternative is viable if the alternative only can take on either a lower or higher value than the hypothetical value, or only higher (or lower) values are of interest.

It is not clear always which alternative (one- or two-sided) is correct or appropriate for any given situation. Usually, two-sided alternatives must be considered because, in most situations, smaller and larger values of the parameter are possible and relevant. Some situations where one-sided alternatives may be best will be discussed.

Choose the Level of Significance—The level of significance also is known as the *alpha level* (α) or *error of the first kind*. This is the basis of the well-known statement (eg, the difference is significant at the 5% level). The α error is set in advance and has the following meaning. The level of significance or α error is the probability of erroneously stating that the difference between the observed value of the parameter (the mean in this example) and the hypothetical value is real or significant.

The α error commonly is chosen as 5%, although this is not obligatory. A more conservative approach would be to choose a level of 1%. This would mean that an error of the first kind, ie, erroneously declaring a difference, is only 1%. It will be seen that a larger difference is needed for significance if the α error is made smaller. That is, it is more difficult to find a significant difference.

Beta Error and Power—Usually, only the α error is chosen in advance of the experiment. However, it should be understood that there is a second kind of error that should be considered when making statistical decisions. This error, the *beta error* (β), is the probability of declaring no difference between the observed sample value and hypothesized value of the parameter when, in fact, a difference of size delta (δ) exists. The α level, β error, and sample size are related. Sample-size determination, an important topic, is discussed in most elementary statistics books.^{8,9} When it is declared that differences are (or are not) significant, only the α level is considered, and not the β error.

Choose a Sample—The choice of a proper sample and the size of the sample are very important considerations in statistical experimentation and experimental design. The number of objects to be included in the sample is a consequence of the α and β errors. The manner in which samples are chosen will dictate the statistical analysis.

In this simple example, the choice of experimental units (tablets to be analyzed) appears to be uncomplicated. However, further thought reveals many alternatives. Ten tablets are to be chosen from 3,000,000. Some possible sampling schemes include (1) take the first 10 tablets from the batch, (2) take the last 10 tablets, (3) take tablets at regular intervals during the run and select 10 of these tablets at random, or (4) take 10 tablets at random from the entire batch. A random sample is one in which each object has an equal probability of being chosen (see under Sampling for more detail). Random samples will assure a valid statistical analysis.

Random sampling can be visualized as a kind of lottery device, in which all of the tablets are mixed and one selected. In many cases, random sampling is not convenient, or the design can be improved by using variations of random sampling schemes. Although the statistical analysis in the present example assumes a random sample, it would not be convenient to implement this procedure for a batch of 3,000,000 tablets. Scheme 3, above, is a more realistic sampling scheme; although it is not truly random, one can proceed as though it were random for this example. In this case, a sample size of 10 tablets is chosen, not for statistical reasons, but because this number has been written into the quality-control procedure. A better procedure would be to base the number of samples on the α and β levels.⁸

Determine Whether the Test Should Be One- or Two-Sided—In this example, a two-sided test is chosen because the observed average potency could be either lower or higher than the hypothetical value of 100 mg. That is, there is no reason to believe, based on the manufacturing process, that the observed value should deviate on one side rather than the other of the labeled potency.

Make Observations and Construct a t Test—Having gathered the tablets and performed the assays, the value of t is computed. This allows one to make the decision, significant or not significant. For a two-sided test, the t value is computed as

$$t_{n-1} = \frac{|\bar{x} - \mu|}{SD/\sqrt{n}}$$

where μ is the hypothetical mean defined by the null hypothesis. In this example, t is

$$t_{n-1} = \frac{|\bar{x} - \mu|}{SD/\sqrt{n}} = \frac{|98.88 - 100|}{0.954/\sqrt{10}} = 3.71$$

The t value then is compared to the t values in Table 12-10 at the specified α level with $n - 1$ DF. For a two-sided test, the absolute value of t is noted, because either small or large values of the difference ($\bar{x} - \mu$) will lead to significance. If the observed value of t is equal to or greater than the value in the table, the difference between the observed and hypothetical values of the parameter, the mean in this example, is declared to be statistically significant. The value of t for a two-sided test at the 5% level for 9 DF is 2.262, the same value used for the 95% confidence interval. This is no coincidence, as will be shown below. Since the observed absolute value of t (3.71) is larger than the value in Table 12-10 (2.262), significance is declared. The true potency is apt not to be 100 mg, but rather some lower value, based on the observed value of 98.88.

An examination of the equation for t reveals that large differences between the observed and hypothetical mean coupled with a small SD of the mean lead to large values of t . This makes sense, from an intuitive point of view, as large differences with small variability suggest that the difference is real. Also, one should note that had the test showed a non-significant difference, it cannot be said with any assurance that the mean of the batch is 100 mg. In fact, it seems extremely unlikely that this should be true. The data simply do not provide sufficient evidence to show that the mean is different from 100 mg. In this case, the confidence interval would give a region in which the mean probably lies.

If the above test had been performed at the 1% level, it still might have been concluded that the mean of the batch was not 100 mg. The value of t at the 1% level with 9 DF from Table 12-10 is 3.25. As 3.71 is greater than 3.25, one would declare significance. A test significant at the 1% level gives greater assurance that the true mean differs from 100 mg, compared to a test significant at the 5% level.

There is a relationship between the two-sided t test and the confidence interval. For example, if the 95% confidence interval does not cover the hypothetical value defined by H_0 , the test will show significance and vice versa. This suggests that the true mean is different from

the hypothetical mean. In the example discussed above, the 95% confidence interval was calculated as 98.2 to 99.56, which does not cover the hypothetical value 100. Therefore, it may be concluded correctly that the t test will show a significant result at the 5% level. Had the confidence interval included 100, the t test would not be significant.

The example described above is known as a one-sample t test. In this test, the experimental design consists of determining the mean value of a random sample from a single population and comparing the mean to some hypothetical value. Thus, it may be of interest to compare the mean tail-flick value of an analgesic compound in rats to some value that represents activity based on previous experience, or to compare the average assay result of 10 tablets to the labeled amount or to a previously accumulated average, as may be available from quality-control records.

TWO INDEPENDENT SAMPLE T TEST—A common design in research involves the comparison of two treatments applied to two independent groups. For example, in a clinical study, a drug is compared to a placebo using 20 patients for the drug treatment and 20 different patients for the placebo treatment. Or the dissolution of tablets prepared by a marketed formulation is compared to the dissolution of tablets prepared from an experimental formulation. Note that this design differs from the one-sample test in that averages are obtained from two groups for purposes of comparison, whereas in the one-sample test, the average of a single group is compared to some hypothetical value.

Three key assumptions are necessary for the two independent sample t test to be valid: (1) each of the two groups are distributed normally, (2) each of the two groups are distributed with the same variance, and (3) the two samples are independent.

The independence assumption is very important. Independent samples mean that the results for any single individual do not influence the results of any other individual. In the case of a clinical trial, independence would mean that the treatment effect for one patient does not influence the result of a treatment for other patients. If one patient discussed the results of his or her medication with another patient in the study, their results would not be independent. If treatments are applied to more than one rat in a cage, their results would not be independent. In the latter case, competition for food and other animal interactions might favor the stronger animal and influence the treatment effect.

Equality of variance also is an important assumption. If the variances are reasonably close, the test should be conducted as usual. As a general rule, if the variances do not differ by more than a factor of four, no special procedure is needed. If the variances differ widely, a modified procedure should be used (the Behrens-Fisher test¹⁰). The normality assumption is less critical. The CLT results in approximate normality of means of non-normal variables.

This statistical design consists of randomly dividing n objects into two groups of size n_1 and n_2 . Treatment 1 is applied to the first group (n_1) and Treatment 2 is applied to the second group (n_2). Optimal treatment allocation in this design is to have an equal number of experimental units ($n/2$) in each group if the primary objective is to compare the means of the two groups. However, if n_1 is not equal to n_2 , the data are analyzed easily, and not much is lost if the two samples are close in size. In animal and human experiments, samples often are lost due to patient dropouts and animal deaths. An experiment that is carried out according to this plan sometimes is called a *parallel design*—two separate groups are treated in parallel.

The test is similar to the one-sample test. In a typical experiment to compare the mean results of the two samples, the null hypothesis is

$$H_0: \mu_1 = \mu_2$$

A two-sided alternative has an alternative hypothesis:

$$H_a: \mu_1 \neq \mu_2$$

Once the α level (usually 0.05) is specified and data are obtained, a t test is performed. This allows a decision to be made

about the equality of the underlying population averages. As in the one-sample case, a value of t is computed as

$$t = \frac{\text{Difference}}{\text{Standard error of difference}}$$

For calculation purposes,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where

- \bar{x}_1 = mean of first sample of n_1 observations.
- \bar{x}_2 = mean of second sample of n_2 observations.

and

$$s^2 = \frac{\sum x_{1i}^2 - (\sum x_{1i})^2/n_1 + \sum x_{2i}^2 - (\sum x_{2i})^2/n_2}{n_1 + n_2 - 2}$$

where

- $\sum x_{1i}^2$ is the sum of squares of observations in first sample.
- $\sum x_{1i}$ is the sum of observations in first sample.
- $\sum x_{2i}^2$ is the sum of squares of observations in second sample.
- $\sum x_{2i}$ is the sum of observations in second sample.
- s^2 is the pooled variance of the two samples.

Example 10—Suppose one sample of four and one sample of five are taken, respectively, from each of two lots of amobarbital capsules and the amount of amobarbital is determined in each capsule. It is desired to determine if there is a significant difference between the two samples.

$$H_0: \mu_1 = \mu_2$$

where μ_1 and μ_2 represent the true averages of the two lots of capsules. This is a two-sided test at the 5% level.

Sample 1	Sample 2
10.1	9.8
13.6	9.6
12.5	11.4
11.4	9.1
$\sum x_{1i} = 47.6$	$\sum x_{2i} = 50.0$
$\sum x_{1i}^2 = 573.18$	$\sum x_{2i}^2 = 502.98$
$\bar{x}_1 = 11.90$	$\bar{x}_2 = 10.00$
$n_1 = 4$	$n_2 = 5$
$s^2 = \frac{573.18 - (47.6)^2/4 + 502.98 - (50.0)^2/5}{4 + 5 - 2}$	
$= \frac{573.18 - 566.44 + 502.98 - 500.00}{7} = \frac{9.72}{7} = 1.3886$	
$s = 1.18$	
$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	
$= \frac{11.90 - 10.00}{1.18} \sqrt{\frac{4(5)}{4 + 5}}$	
$= 1.61(1.49) = 2.40$	

The degrees of freedom involved in the pooled standard deviation are 7, $DF = (n_1 - 1) + (n_2 - 1)$. In the t table (see Table 12-10), for $P = 0.05$ and $DF = 7$ (two tails) the value of t given is 2.365. The value of t calculated is greater than this. Therefore, since the probability of these two samples being drawn from the same population is less than 0.05, we conclude that they were drawn from different populations. (This conclusion may be wrong 5 times in 100.) It can be stated that there is a statistically significant difference between the two samples.

The examples illustrated so far have used a two-sided test. A one-sided test may be used when the difference can only occur in one direction or when only one direction is relevant.

Example 11—A drug is formulated to be dissolved more rapidly by substituting lactose for part of the lipoidal lubricant in the regular-release product. The formulator is convinced that this formulation change only could increase the rate of drug dissolution. A one-sided test at the 5% level is proposed when comparing the drug dissolution from the two

products. The time to 50% dissolution for six tablets of each product (minutes) is

- Original product: 25, 22, 29, 30, 26, 24
- Modified product: 18, 23, 24, 22, 19, 16

For this test, H_0 and H_a are defined as

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

where μ_1 is the 50% dissolution time for the original product.

If the test indicates rejection of the null hypothesis, it must be concluded that the new formulation has a faster dissolution time. If the test shows a nonsignificant difference, it is concluded that the data is insufficient to show that the new formulation reduces the dissolution time. Note that if the data show a longer dissolution time for the new formulation, a test would not be performed, but it would be concluded that the new formulation did not decrease the dissolution time. The average results and standard deviation for the two sets of data are

$$\begin{aligned} \bar{x}_1 &= 26 & \bar{x}_2 &= 20.33 \\ SD_1 &= 3.033 & SD_2 &= 3.141 \\ t &= \frac{|26 - 19.17|}{3.087/\sqrt{3}} = 3.18 \end{aligned}$$

Note that the pooled standard deviation is equal to

$$\sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

when the sample sizes are equal in the two groups.

For a one-sided test, refer to one tail of the t distribution. For 10 degrees of freedom ($6 + 6 - 2$), the t value, leaving 5% of the area in the upper tail, is 1.812 (see Table 12-10). Therefore, it is concluded that the new formulation causes faster dissolution of the drug. Note that it is easier to get significance with a one-sided test. Had the test been two-sided, the t would have had to exceed 2.228 at the 5% level for significance, according to Table 12-10.

PAIRED T TEST—In many situations, the scientist is interested in comparing the means of two experimental treatments using a paired-sample design. This differs from the independent two-sample design in that each of two different treatments may be applied to a single group of experimental units (eg, patients). In a bioavailability study, a generic drug is compared to a standard drug in each of 20 patients. A new analytical method is compared to a previously used method by comparing assay results on different concentrations of the same material divided into two parts.

The paired design has certain advantages over the two independent sample or parallel groups design. It has been noted that significance is determined by the ratio of the difference of the averages divided by the standard error. This ratio can be increased by reducing the standard error. One way is to increase the sample size. Another way of increasing the value of t is to reduce the variability.

In a two independent sample test, the variability is a result of the differences among different experimental units (differences among patients' responses to a drug, for example). In the paired test, the variability results from differences within experimental units. The within-individual variability should be less than the between-individual variability. (Theoretically, the measured between-individual variation includes the within-variation; therefore, the between-variation is larger than the within-variation.) Therefore, the paired-sample design has the advantage of reduced variability.

The paired-sample test also needs less experimental material. In a two independent sample design, comparing the response to two drugs, one might use 24 patients in each of two groups. In a paired design, each patient receives both drugs, on two different occasions if necessary. Thus, there is the need to recruit 24 patients rather than 48. For example, when testing a skin preparation, the products could be applied randomly to each arm of the 24 patients.

The paired design can be used only when there is a natural or easy way of pairing the experimental units. When comparing the dissolution of two different formulations, there seems to be no obvious way of pairing the tablets from the two different formulations, as is the case of applying two treatments to the same individual. In animal experiments, litter mates may be paired. Pairing implies that the paired units are more alike than are two different units. In clinical studies, test units may be paired or matched on the basis of certain characteristics such as sex, age, or severity of disease. Then each subject in the pair is assigned to one of the experimental treatments.

A disadvantage of the paired design is that if treatments cannot be applied concurrently, as may be the case where two drugs administered orally are to be compared, the time to complete the experiment can be extended. In the case of clinical studies, this may be an important detriment because time usually is of the essence. Also, as these studies are prolonged, the chances of patient dropouts increase, and time can influence the progress of the disease.

In the paired design, a missing value means that the single unpaired datum is of no value. In this design, each experimental unit (eg, each patient) essentially acts as its own control. That is, the comparison is made within each experimental unit. If one of the two paired values is missing, a comparison cannot be made.

Another potential disadvantage is that a *carryover effect* may be present. This means that effects from one treatment may affect the results of the other. For example, in a bioavailability study, if the first drug administered is not eliminated completely before the second drug is given, blood levels of the second drug will be contaminated. Or, in a clinical study, the first drug administered may modify the disease condition so that the effect of the second drug is not comparable directly with that of the first drug.

In any event, there are many situations where the advantages of the paired-sample design strongly suggest its use. For computational purposes, the formula is

$$t = \frac{\bar{d}}{s} \sqrt{n}$$

where

\bar{d} = mean of the differences, $x_1 - x_2$, of the n pairs of observations

$$s^2 = \frac{\sum d_i^2 - (\sum d_i)^2/n}{n - 1}$$

where

$\sum d_i^2$ = the sum-of-squares of the n differences

$\sum d_i$ = the sum of the n differences

n = the number of differences or pairs of observations

Example 12—The duration of loss of the righting reflex (minutes) was measured in 16 mice following treatment with a barbiturate. The drug was administered in the morning and the afternoon on two different occasions; the order of giving the morning or the afternoon dose was randomized in each mouse. It was desired to test the null hypothesis that the duration of loss of the righting reflex is the same in the morning and the afternoon (Table 12-11).

$$s^2 = \frac{354 - (40)^2/16}{16 - 1} = \frac{354 - 100}{15} = 16.9333$$

$$s = 4.11$$

$$t = \frac{\bar{d}}{s} \sqrt{n} = \frac{2.5}{4.11} \sqrt{16} = \frac{2.5(4)}{4.11} = 2.43$$

$$DF = n - 1 = 16 - 1 = 15$$

In the t table (see Table 12-10), for $P = 0.05$ and $DF = 15$ (two tails) the value of t is 2.131. The value of t calculated is greater than this. Therefore, as the probability of the morning and afternoon values being the same is less than 0.05, we conclude that they are different. Apparently, the duration of loss of the

Table 12-11. Loss of Righting Reflex on 16 Mice

MOUSE NO	AM x_1	PM x_2	DIFFERENCE $D = x_1 - x_2$
1	75	73	2
2	86	89	-3
3	93	89	4
4	87	79	8
5	91	95	-4
6	87	81	6
7	76	77	-1
8	83	89	-6
9	87	82	5
10	95	91	4
11	91	87	4
12	86	86	0
13	83	78	5
14	76	69	7
15	82	78	4
16	93	88	5
			$\sum d_i = 40$
			$\sum d_i^2 = 354$
			$\bar{d} = 2.5$
			$n = 16$

righting reflex in mice tested on the barbiturate in the morning was longer than when tested in the afternoon.

Note the similarity of the one-sample t test and the paired t test. The test is identical after differences between pairs have been calculated in the paired test. The null hypothesis in the paired test almost always is of the form $H_0: \delta = 0$, where δ is the hypothesized difference of the true means. It is hypothesized that the mean results of the two treatments are identical.

TESTS FOR PROPORTIONS—The t test is applicable for continuous data that is distributed normally. Much of the data that is seen in pharmaceutical experiments is dichotomous. That is, answers to a questionnaire regarding filling a prescription for a specified drug may be yes or no, or a bottle of tablets may be acceptable or not acceptable, or a patient may be cured or not cured. Tests similar to the t test may be constructed for binomial data. The principle is to compute proportions that are probable, based on the sample proportion. If the probable proportions do not include the hypothetical proportion, the null hypothesis is rejected.

For large sample sizes (n is large), such computations can be tedious and difficult. Therefore, the normal approximation to the binomial is used whenever possible. Fortunately, in most practical cases, the normal approximation is applicable. When comparing proportions from two independent samples when the normal approximation is clearly not applicable, the Fisher Exact test can be used.⁸ (Statistical software programs can compute exact probabilities.) In general, use the rule that np and nq should be equal to or greater than 5 in order to use the normal approximation. In practice, this rule may be relaxed somewhat. When in doubt, a professional statistician should be consulted.

Simple statistical tests for proportions are analogous to the t tests. For a one-sample test, where the hypothetical value defined by H_0 is P_0 , the ratio

$$Z = \frac{p - P_0}{\sqrt{P_0 Q_0/n}}$$

may be computed, where p is the observed proportion and n is the number of binomial trials, the sample size.

The calculated value of Z is compared to the standard normal distribution, rather than the t distribution. Refer to Table 12-9 or to the last line in the t table, Table 12-10. For a two-sided test at the 5% level, the test statistic, Z , must exceed 1.96 for the difference to be considered significant.

Example 13—A one-sample test for proportions. A questionnaire was sent to pharmacists asking which of two cold medications the phar-

macist would recommend to customers. A statistical test was proposed to decide which product was most recommended. The null hypothesis was that the two products, A and B, were recommended equally: $H_0: p_A = p_B = 0.50$. The test is two-sided at the 5% level. Two-hundred and fifty (250) pharmacists responded; Product A was recommended 145 times, and Product B was recommended 105 times.

The scientist conducting the experiment had sent out 400 questionnaires and was rightfully concerned about the nonresponders. However, she decided that there was no reason to suspect a bias because of the lack of 100% response and proceeded to analyze the data. The observed proportion of successes (A is a success) is $145/250 = 0.58$ (or the observed proportion could be 0.42 as well). The absolute value of the numerator of the Z statistic will be the same for $p = 0.42$ or $p = 0.58$.

$$Z = \frac{|0.58 - 0.50|}{\sqrt{0.5 \times 0.5/250}} = 2.53$$

Since 2.53 exceeds the tabled value for α of 5% (1.96), it is concluded that Product A is the more recommended product. The normal approximation is improved if $1/(2n)$ is subtracted from the absolute value of the numerator, although the effect of the continuity correction is more evident for small sample sizes. In the present example, the corrected value of Z is 2.47. A 95% confidence interval for the proportion recommending Product A also was reported.

$$0.58 \pm 1.96 \sqrt{0.58 \times 0.42/250} = 0.519 \text{ to } 0.641$$

Exercise 4—Compute Z, with and without the continuity correction, if 141 of 250 pharmacists recommended Product A. Determine whether the result is significant by using a two-sided test at the 5% level.

CHI-SQUARE TEST—To test for differences of two proportions from two independent samples, the chi-square test is used. Chi-square (χ^2) is a probability distribution derived from the sum of squares of normal variables. The chi-square distribution is not symmetrical and can have only positive values. Table 12-12³ is a short table of chi-square probabilities. This table is used in the same way as the normal and t tables: first compute a chi-square statistic and if the value exceeds the tabled value, a significant effect is declared.

Chi-square is calculated as

$$\chi^2 = \sum \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

Example 14—In tossing a coin, 50% tails and 50% heads are expected. Suppose a coin is tossed 40 times and 25 heads and 15 tails are obtained, whereas 20 heads and 20 tails are expected. Is the coin biased or weighted in some way?

$$\chi^2 = \frac{(25 - 20)^2}{20} + \frac{(15 - 20)^2}{20} = 2.5$$

The degrees of freedom (DF) associated with χ^2 are one less than the number of categories. Here $\chi^2 = 2.5$ with 1 DF. The greater the disagreement between expected and observed, the larger will be χ^2 . See Table 12-12 for probabilities of getting this value or larger. For 1 DF the probability of getting a value larger than 2.5 is somewhere between $P =$

Table 12-12. The Chi-Square Table^a Probability

DF	P = 0.20	P = 0.10	P = 0.05	P = 0.01
1	1.64	2.71	3.84	6.64
2	3.22	4.61	5.99	9.21
3	4.64	6.25	7.82	11.34
4	5.99	7.78	9.49	13.28
5	7.29	9.24	11.07	15.09
6	8.56	10.64	12.59	16.81
7	9.80	12.02	14.07	18.48
8	11.03	13.36	15.51	20.09
9	12.24	14.68	16.92	21.67
10	13.44	15.99	18.31	23.21
20	25.04	28.41	31.41	37.57
30	36.25	40.26	43.77	50.89

^a Adapted from Fisher RA, Yates F. *Statistical Tables for Biological, Agriculture and Medical Research*. New York: Hafner, 1963.

Table 12-13. Survival Rates in Swine Dysentery

TREATMENT	SURVIVED	DIED	TOTAL
Drug	a = 25	b = 14	a + b = 39
Controls	c = 21	d = 22	c + d = 43
Totals	a + c = 46	b + d = 36	N = 82

0.20 and $P = 0.10$. To say that there is a statistically significant departure from the expected values, χ^2 would have to be larger than 3.84, which is the value for $P = 0.05$ at 1 DF. A value of χ^2 larger than 6.64 for 1 DF would indicate a statistically highly significant ($P < 0.01$) departure of the observed from the expected values.

The chi-square test commonly is used for comparing two percentages in a 2×2 or fourfold contingency table (Table 12-13).

Example 15—Table 12-13 gives the survival rates for drug-treated and control pigs with swine dysentery. The survival rates for the drug-treated and control pigs are $p_D = 25/39 = 64\%$ and $p_C = 21/43 = 49\%$, respectively. To test the null hypothesis that there is no difference in the survival rates of drug-treated and control pigs, χ^2 is calculated:

$$\chi^2 = \frac{(\text{Observed frequency} - \text{Expectancy frequency})^2}{\text{Expected frequency}}$$

The expected values in each of the four cells can be obtained by multiplying the column total by the row total and dividing this result by the grand total. The expected value for Cell a is $46 \times 39/82 = 21.9$. The expected frequencies for Cells b, c, and d are 17.1, 24.1, and 18.9, respectively. Note that the sum of the expected frequencies across any row or column equal the totals for the row or column. For example, the expected frequencies for b and d are 17.1 and 18.9, which sums to 36, the total number who died. The calculation of chi-square is

$$\frac{(25 - 21.9)^2}{21.9} + \frac{(14 - 17.1)^2}{17.1} + \frac{(21 - 24.1)^2}{24.1} + \frac{(22 - 18.9)^2}{18.9} = 1.91$$

The DF associated with an $R \times C$ contingency table = $(R - 1)(C - 1)$, so that for a 2×2 contingency table we have 1 DF. Table 12.12 shows that for 1 DF the probability of getting a value of χ^2 larger than the calculated value 1.91 is greater than $P = 0.10$. Since P is not equal to or less than 0.05, we conclude that there is insufficient evidence to indicate that the survival rates for the drug-treated and control pigs are different.

The chi-square test for comparing two correlated percentages for paired data takes a somewhat different form.

Example 16—Two different types of penicillin were given to each of 22 patients in random order, on successive occasions, and the presence or absence of a detectable blood level was determined (Table 12-14). The percentage of patients with detectable blood levels for the two forms of penicillin are $p_I = 16/22 = 73\%$ and $p_{II} = 8/22 = 36\%$. To test the null hypothesis that there is no difference in the percentage of patients with detectable blood levels for the two forms of penicillin, we calculate

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} = \frac{(|10 - 2| - 1)^2}{10 + 2} = \frac{49}{12} = 4.08$$

In Table 12-13 for $P = 0.05$ and $DF = 1$, the value of χ^2 given is 3.84. The value of χ^2 calculated is greater than this. Therefore, as the probability of the percentages for Type I and Type II penicillin being the same is less than 0.05, we conclude that they are different.

Note that this test compares the number of patients who are positive on one test and negative on the other.

The chi-square distribution is an approximation of the discrete distribution represented by the fourfold table. The approximation can be improved by applying a *correction factor* for the Observed - Expected values. If the absolute difference is an exact integer (eg, 4.0), subtract 0.5 from the absolute difference; 4.0 would become 3.5. If the absolute difference has a decimal between 0.5 and 0 (eg, 3.8), change the decimal to 0.5; 3.8 would

Table 12-14. Data for Example 16

Type I	TYPE II		Totals
	+	-	
+	a = 6	b = 10	16
-	c = 2	d = 4	6
Totals	8	14	22

become 3.5. If the decimal is between 0 and 0.5, reduce the absolute difference to its integer value; 4.1 would become 4. In Example 15, the absolute difference of Observed – Expected would be reduced to 3.0. The corrected chi-square would be 1.79.

Exercise 5—Calculate the corrected chi-square for Example 15.
Answer: 1.79.

THE *F* DISTRIBUTION AND TESTS OF SIGNIFICANCE—The *t* distribution is suitable for a statistical test comparing two means. The *F* distribution is used to compare two variances, *F* being defined by the ratio of the variances, with $n_1 - 1$ DF in the numerator and $n_2 - 1$ DF in the denominator of the ratio

$$F_{n_1-1, n_2-1} = s_1^2/s_2^2$$

Similar to the chi-square distribution, the *F* distribution consists of only positive values and is a skewed distribution. The ratio of two variances is compared to values in the *F* table (Table 12-15¹¹) with the appropriate DF in the numerator and denominator to test for statistical significance. If the calculated ratio exceeds the value in the table at a given α level, the variances differ at the α level of significance. The following two examples describe the use of the *F* test for comparing variances for independent and dependent samples.

Example 17 shows a test to compare the variances of two independent samples using the *F* test. Example 18 shows the test for comparing variances in related or paired samples, which

uses the *t* statistic to determine significance. To compare the variances of samples from two independent populations, the calculation is

$$F = s_1^2/s_2^2 \text{ with } s_1^2 > s_2^2$$

where

$$s_1^2 = \frac{\sum x_{1i}^2 - (\sum x_{1i})^2/n_1}{n_1 - 1} = \text{larger variance}$$

$$s_2^2 = \frac{\sum x_{2i}^2 - (\sum x_{2i})^2/n_2}{n_2 - 1} = \text{smaller variance}$$

To test for significance, the *F* ratio is referred to the *F* table (see Table 12-15) with $f_1 = n_1 - 1$ and $f_2 = n_2 - 1$ DF. The null hypothesis that the two variances are the same is rejected at the $2P$ level of significance.

Example 17—Two treatments showed the results in Table 12-16. Entering Table 12-15 with $f_1 = 6$ and $f_2 = 5$ DF, we find that the tabulated values of *F* are 4.95 and 6.98 for $P = 2 (0.05) = 0.10$ and $P = 2 (0.025) = 0.05$, respectively. Thus, the probability of getting a value of *F* larger than the calculated value 5.75 is between $P = 0.05$ and $P = 0.10$. Since *P* is not equal to or less than 0.05, we conclude that there is insufficient evidence to indicate that the two variances are different.

If it is desired to compare the variances from paired data, the *F* test described above would be inappropriate. Instead, proceed as exemplified below.

Table 12-15. The *F* Table
 10%, 5%, 2.5%, and 1% Points for the Distribution of *F*

		<i>f</i> ₁ DEGREES OF FREEDOM (FOR GREATER MEAN SQUARE)															
<i>T</i> ₂	<i>P</i>	1	2	3	4	5	6	7	8	9	10	20	30	40	60	120	∞
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.21	3.17	3.16	3.14	3.12	3.10
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50	4.46	4.43	4.40	4.36
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.33	6.23	6.18	6.12	6.07	6.02
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.55	9.38	9.29	9.20	9.11	9.02
10	0.10	3.28	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.20	2.16	2.13	2.11	2.08	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.62	2.58	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.42	3.31	3.26	3.20	3.14	3.08
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.41	4.25	4.17	4.08	4.00	3.91
15	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.92	1.87	1.85	1.82	1.79	1.76
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.33	2.25	2.20	2.16	2.11	2.07
	0.025	6.20	4.76	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.76	2.64	2.58	2.52	2.46	2.40
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.36	3.20	3.12	3.05	2.96	2.87
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.79	1.74	1.71	1.68	1.64	1.61
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.95	1.90	1.84
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.46	2.35	2.29	2.22	2.16	2.09
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	2.94	2.77	2.69	2.61	2.52	2.42
25	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.72	1.66	1.63	1.59	1.56	1.52
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.01	1.92	1.87	1.82	1.77	1.71
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.30	2.18	2.12	2.05	1.98	1.91
	0.01	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	2.70	2.54	2.45	2.36	2.27	2.17
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.67	1.61	1.57	1.54	1.50	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.79	1.74	1.68	1.62
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.20	2.07	2.01	1.94	1.87	1.79
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.55	2.38	2.29	2.21	2.11	2.01
40	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.61	1.54	1.51	1.47	1.42	1.38
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74	1.69	1.64	1.58	1.51
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.07	1.94	1.88	1.80	1.72	1.64
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.37	2.20	2.11	2.02	1.92	1.81
60	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.54	1.48	1.44	1.40	1.35	1.29
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.75	1.65	1.59	1.53	1.47	1.39
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	1.94	1.82	1.74	1.67	1.58	1.48
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.20	2.03	1.93	1.84	1.73	1.60
120	0.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.48	1.41	1.37	1.32	1.26	1.19
	0.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.66	1.55	1.50	1.43	1.35	1.25
	0.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	1.82	1.69	1.61	1.53	1.43	1.31
	0.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.03	1.86	1.76	1.66	1.53	1.38
∞	0.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.42	1.34	1.30	1.24	1.17	1.00
	0.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.57	1.46	1.39	1.32	1.22	1.00
	0.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.71	1.57	1.48	1.39	1.27	1.00
	0.01	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	1.87	1.69	1.59	1.47	1.32	1.00

Adapted from Snedecor GW, Cochran WG. *Statistical Methods*, 7th ed. Ames: Iowa State University Press, 1980.

Table 12-16. Data for Example 17

	A	B
	6	15
	4	4
	3	10
	7	10
	6	5
	4	11
		9
$\sum x_i$	30	64
$\sum x_i^2$	162	668
n_i	6	7
s_i^2	2.40	13.81
f_i	5	6
$F = s_1^2/s_2^2 = 13.81/2.40 = 5.75$		
$f_1 = n_1 - 1 = 7 - 1 = 6$		
$f_2 = n_2 - 1 = 6 - 1 = 5$		

Example 18—A characteristic was measured before and after aging for each of 10 items (Table 12-17). Has the variability changed with aging?

$$\sum x_B^2 = 2,393.81 \quad \sum x_A^2 = 2,252.72$$

$$\sum x_B x_A = 2,298.92$$

$$[x_B]^2 = 2,393.81 - (148.5)^2/10$$

$$= 188.59$$

$$[x_A]^2 = 2,252.72 - (147.2)^2/10$$

$$= 85.94$$

$$[x_B x_A] = 2,298.92 - (148.5)(147.2)/10$$

$$= 113.00$$

$$t = \frac{([x_B]^2 - [x_A]^2)\sqrt{n-2}}{2\sqrt{[x_B][x_A] - [x_B x_A]^2}} = \frac{(188.59 - 85.94)\sqrt{8}}{2\sqrt{(188.59)(85.94) - (113.00)^2}} = 2.476$$

$$DF = n - 2 = 10 - 2 = 8$$

In the *t* table (see Table 12-10), for *P* = 0.05 and DF = 8 (two tails), the value of *t* given is 2.306. The value of *t* calculated is greater than this. Therefore, because the probability of the variance before and after aging being the same is less than 0.05 it is concluded that they are different. Apparently, the variability decreased after aging.

The *F* distribution is used most often for the comparison of more than two means through the analysis of variance, and is equivalent to the *t* test if used to compare two means.

ANALYSIS OF VARIANCE (ANOVA) AND EXPERIMENTAL DESIGN—There is almost always more than one way to conduct an experiment to achieve a given objective. It was seen that when comparing the means of two treatment groups, two independent groups or a paired design could be used. For example, in a clinical study, two treatments could be

applied to two separate and independent groups of patients, or each patient could take both treatments.

In the paired experiment, a further refinement could be added with regard to treatment order. For example, if drugs cannot be administered concurrently, the order of drug administration can be balanced. Half of the patients receive Drug A at the first administration and the other half receive A on the second administration. The patients who receive A on the first occasion will receive B on the second occasion and vice versa. This is known as a *crossover design*.

An alternative design is to assign each patient an order of administration randomly. In the latter case, it is likely that a balanced allocation would not be achieved, as in the crossover design. In fact, there is always a possibility, albeit a small one, that all patients will receive one and the same treatment first and the other treatment second, a situation that intentionally is avoided in the crossover design.

When more than two treatments are to be used in an experiment, a variety of designs are possible. In these cases, one design usually will be optimal, depending on the nature of the experiment, the treatments, and the experimental units or subjects. A common feature of most good designs is *symmetry*. That is not to say that all good designs are symmetrical. In some special cases, an asymmetrical design may be optimal, but this is not the usual circumstance.

The most simple analysis of variance design is known as a *one-way analysis of variance* (one-way ANOVA) or completely randomized design. This is the ANOVA analogy of the two independent sample *t* test. In the ANOVA design, there is interest in comparing the means of two or more treatment groups. As has been noted before, in the jargon of clinical trials this design often is one of a class known as parallel-groups designs.

In the following description, an example from clinical trials will be used. However, one should understand that tablets, bottles, or consumers could be substituted for patients and the process is the same: *n* patients are available for the experiment with *t* treatments. For example, 150 patients are to be assigned to three treatment groups, one placebo and two actives. The *n* patients are assigned randomly to the three groups (see the discussion on random assignment). The optimal assignment in the examples discussed in this chapter will result in equal numbers in each group, *n/t* units per group. Note that *n* is chosen to be divisible by *t*. If there are three treatments and *n* = 150, we randomly would assign 50 units per treatment. A loss of observations will not invalidate the analysis, as is also true in the two independent groups *t* test. Observations are made, and the null hypothesis that all *t*-means are equal is tested by an ANOVA procedure.

The ANOVA separates the total sum-of-squares, $\sum(x_i - \bar{x})^2$, into parts determined by the experimental structure. For the one-way ANOVA, the sum-of-squares consists of the between (among) and within sum-of-squares. The between sum-of-squares (BSS) represents differences among treatments, large values indicating large treatment differences (eg, if the treatment means are identical, the BSS will be 0 on the average). The within sum-of-squares (WSS) represents differences within treatments, or error; ie, the differences among objects (subjects in clinical studies) within a treatment is a measure of the variability of the observations.

An ANOVA table is prepared consisting of source of variation, degrees of freedom, sums-of-squares, and mean square. In the one-way ANOVA, the sources consist of the between, within, and total terms. The sum-of-squares divided by the DF is known as the mean square, between mean square (BMS) and within mean square (WMS) in the one-way ANOVA (Table 12-18).

For a one-way ANOVA, the DF for treatments is *t* - 1. The DF for error (within treatments) is *n* - *t*, where *n* is the total number of observations. The total sum-of-squares (SS) is exactly the sum of the between and within sums of squares. The error mean square (WMS) corresponds to the variance for the test, and in the case of two treatments, corresponds to the pooled variance in the *t* test.

Table 12-17. Measurement Before and After Aging

ITEM NO	BEFORE AGING	AFTER AGING
1	8.3	9.3
2	8.4	10.9
3	14.9	13.2
4	12.2	12.8
5	12.5	16.0
6	15.0	15.2
7	17.1	16.8
8	19.2	16.2
9	22.0	17.9
10	18.9	18.9
	$\sum x_B = 148.5$	$\sum x_A = 147.2$

Table 12-18. ANOVA for Example 19

ANALYSIS OF VARIANCE				
SOURCE OF VARIATION	DEGREES OF FREEDOM	SUMS-OF-SQUARES	MEAN SQUARES	F RATIO
Between regimens	$t - 1 = 9$	160.54	17.81	8.22
Within regimens	$\sum (n_i - 1)^a = 20$	43.33	$s^2 = 2.17$	
Total	$N - 1 = 29$	203.87		

^a $\sum (n_i - 1) = N - t$.

The ratio BMS/WMS has an F distribution under the null hypothesis, with $(t - 1)$ DF in the numerator and $(N - t)$ DF in the denominator. If the ratio exceeds the appropriate F value found in the table, then at least two of the treatments tested are significantly different. The computations consist of simple arithmetic, summing individual values and their squares. The following numerical example illustrates the computations and should clarify these concepts. Although it always is useful to practice some calculations, computer programs are available that should be used for most practical situations.

Example 19—Groups of three subjects each were given one of 10 food regimens and showed the weight gains (lb) in Table 12-19. These are unpaired data, and this type of study is referred to as a completely randomized experiment. There are only two sources of variation; the variation between regimens and the variation within regimens, as indicated in Table 12-18. The sums-of-squares are obtained as

$$\text{Total SS} = \sum x^2 - (\sum x)^2/N = 934 - (148)^2/30 = 203.87$$

$$\begin{aligned} \text{Between regimens SS} &= \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \dots + \frac{(\sum x_{10})^2}{n_{10}} - \frac{(\sum x)^2}{N} \\ &= \frac{(7)^2}{3} + \frac{(3)^2}{3} + \dots + \frac{(16)^2}{3} - \frac{(148)^2}{30} \\ &= 160.54 \end{aligned}$$

$$\text{Within regimens SS} = 203.87 - 160.54 = 43.33$$

The mean squares are obtained by dividing the sums-of-squares by their corresponding DF. The mean square within regimens, s^2 , is the pooled variance for the 10 samples. Since this is the only variance that can be identified as random sampling error (the mean square between regimens has in addition a component due to the variability among regimens), it becomes the denominator in the F ratio, so that

$$F = \frac{\text{mean square between regimens}}{\text{mean square within regimens}} = \frac{17.81}{2.17} = 8.22$$

To test for significance, the F ratio is referred to the F table (see Table 12-15) with $f_1 = t - 1 = 9$ and $f_2 = \sum (n_i - 1) = 20$ DF. We find that the calculated value 8.22 is larger than the tabulated value 3.45 for $P = 0.01$. Therefore, as the probability of these 10 samples being drawn from the same population is less than 0.05 (actually, it is less than 0.01), it is concluded that they are not all the same (ie, not all the means are equal).

MULTIPLE COMPARISONS IN ANOVA—If the F test is significant and more than two treatments are included in the experiment ($t > 2$), it may not be obvious immediately which

treatments are different. Some or all of the treatments may be different. Various multiple-comparison procedures have been proposed to solve this problem. It is not always apparent when a particular procedure is best, given the variety of procedures available. Several of these tests are described here, with discussion of their application. The general procedure is to list the ranked means from lowest to highest and underline the means that are not statistically significantly different from each other. Sometimes brackets or parentheses are used instead of an underline. The procedure is carried out by calculating a 5% allowance, which is defined as the critical difference between means which allows one to reject the null hypothesis ($\mu_i = \mu_j$) and accept the alternative hypothesis ($\mu_i \neq \mu_j$) for any two sample means \bar{x}_i and \bar{x}_j at $P = 0.05$. To calculate the 5% allowance the following data is required.

- s^2 = pooled variance from the analysis of variance.
- DF = degrees of freedom for the pooled variance from the analysis of variance.
- n_i, n_j = the number of observations from which the means \bar{x}_i and \bar{x}_j were determined, respectively.
- t = a critical value at $P = 0.05$ which depends upon the DF and the degree of conservatism desired as exemplified by the multiple comparison procedures described below.

Least Significant Difference Procedure—For this procedure

$$5\% \text{ allowance} = t \sqrt{s^2(1/n_i + 1/n_j)}$$

where t is the value of t from Table 12-10 (two tails). This is the least conservative procedure, and assures that the probability that any one comparison is judged to be significant by chance alone is 5%. However, the probability of one or more comparisons being judged significant would be greater than 5%. Applied to the results of Example 19,

$$\begin{aligned} s^2 &= 2.17 \\ n_i, n_j &= 3, 3 \\ DF &= 20 \end{aligned}$$

and $t = 2.086$ from Table 12-10 for 20 DF and $P = 0.05$ (two tails).

$$5\% \text{ allowance} = t \sqrt{s^2(1/n_i + 1/n_j)} = 2.086 \sqrt{2.17(1/3 + 1/3)} = 2.51$$

Thus, any two means differing by 2.51 or more are judged to be different.

Ranked Means

<u>B</u>	<u>A, C</u>	<u>I</u>	<u>J</u>	<u>F, G</u>	<u>D, H</u>	<u>E</u>
1.0	2.3	5.0	5.3	5.7	6.3	9.3

or, (BAC) (IJFGDH) (E).

Any two means underscored by the same line (or included in the same parentheses) do not differ statistically at $P = 0.05$.

Any two means not underscored by the same line (or not included in the same parentheses) are statistically significantly different at $P \leq 0.05$.

Table 12-19. Weight Gains in Ten Food Regimens

	FOOD REGIMEN										(t = 10 REGIMENS)
	A	B	C	D	E	F	G	H	I	J	
	2	1	2	4	9	3	6	7	4	4	
	3	2	4	8	8	8	5	6	4	6	
	2	0	1	7	11	6	6	6	7	6	
$\sum x_i$	7	3	7	19	28	17	17	19	15	16	Sums
$\sum x_i^2$	17	5	21	129	266	109	97	121	81	88	$\sum x = 148$
n_i	3	3	3	3	3	3	3	3	3	3	$\sum x^2 = 934$
$n_i - 1$	2	2	2	2	2	2	2	2	2	2	$N = 30$
\bar{x}_i	2.3	1.0	2.3	6.3	9.3	5.7	5.7	6.3	5.0	5.3	$\sum (n_i - 1) = 20$

Table 12-20. The Q Table
Upper 5% Points, Q, in the Studentized Range

DF	k (NUMBER OF TREATMENTS)																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.12	6.20	6.27	6.34	6.41	6.47	
11	3.11	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.14	6.20	6.27	6.33	
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.06	6.11	
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.56	5.64	5.72	5.79	5.86	5.92	5.98	6.03	
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.79	5.85	5.91	5.96	
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90	
17	2.98	3.62	4.02	4.31	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84	
18	2.97	3.61	4.00	4.28	4.49	4.67	4.83	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75	
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.50	5.56	5.61	5.66	5.71	
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.55	5.59	
30	2.89	3.48	3.84	4.11	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48	
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.90	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.36	
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24	
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13	
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.84	4.89	4.93	4.97	5.01	

Adapted from Snedecor GW, Cochran WG. *Statistical Methods*, 7th ed. Ames: Iowa State University Press, 1980.

Studentized Range Procedure—For this method

$$5\% \text{ allowance} = \frac{Q}{\sqrt{2}} \sqrt{s^2(1/n_i + 1/n_j)}$$

where *Q* is the Studentized Range value for *k* treatments from Table 12-20.¹² This is one of the more conservative procedures, and it ensures that the probability of one or more comparisons being judged significant by chance alone is 5%. Applied to the results of Example 19,

Q = 5.01 from Table 12-20

for *k* = 10 treatments, 20 *DF* and *P* = 0.05.

$$5\% \text{ allowance} = \frac{Q}{\sqrt{2}} \sqrt{s^2(1/n_i + 1/n_j)} = \frac{5.01}{\sqrt{2}} \sqrt{2.17(1/3 + 1/3)} = 4.26$$

Thus, any two means differing by 4.26 or more are judged to be different.

Ranked Means

<i>B</i>	<i>A, C</i>	<i>I</i>	<i>J</i>	<i>F, G</i>	<i>D, H</i>	<i>E</i>
1.0	2.3	5.0	5.3	5.7	6.3	9.3

or, (*BAC*) (*ACIJFGDH*) (*JFGDHE*).

Duncan's New Multiple Range Procedure—For this method:

$$5\% \text{ allowance} = \frac{t_k}{\sqrt{2}} \sqrt{s^2(1/n_i + 1/n_j)}$$

where *t_k* are values for 2, 3, . . . , *k* treatments obtained from Table 12-21.¹³ The critical values will be *A*₂, *A*₃, . . . , *A_k*, depending upon how many means are included in the range of ranked means being compared. This is next to the least conservative procedure. Applied to the results of Example 19,

$$5\% \text{ allowance} = \frac{t_k}{\sqrt{2}} \sqrt{s^2(1/n_i + 1/n_j)} = \frac{t_k}{\sqrt{2}} \sqrt{2.17(1/3 + 1/3)}$$

Values of *t_k* from Table 12-21 for *k* = 2 to 10 treatments, 20 *DF*, and *P* = 0.05 give the allowances in Table 12-22. Thus, the critical difference between *E* and *B* is 2.89 because the range includes 10 means, the critical difference between *E* and *H* is 2.64 because the range includes three means, and so on.

Ranked Means

<i>B</i>	<i>A, C</i>	<i>I</i>	<i>J</i>	<i>F, G</i>	<i>D, H</i>	<i>E</i>
1.0	2.3	5.0	5.3	5.7	6.3	9.3

or, (*BAC*) (*IJFGDH*) (*E*).

Dunnnett's Procedure—The three procedures previously described are appropriate when it is desired to compare all possible pairs of means. Dunnnett¹⁴ considered the problem when the objective of the study is to compare several treatments with a standard or control. In his method,

$$5\% \text{ allowance} = t_d \sqrt{s^2(1/n_i + 1/n_j)}$$

Table 12-21. The Multiple Range Table
Values of *t_k* for Duncan's New Multiple Range Test at the 5% Level of Significance

DF	κ (NUMBER OF TREATMENTS)									
	2	3	4	5	6	8	10	14	20	
10	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.48	
12	3.08	3.23	3.33	3.36	3.40	3.44	3.46	3.46	3.48	
14	3.03	3.18	3.27	3.33	3.37	3.41	3.44	3.46	3.47	
16	3.00	3.15	3.23	3.30	3.34	3.39	3.43	3.45	3.47	
18	2.97	3.12	3.21	3.27	3.32	3.37	3.41	3.45	3.47	
20	2.95	3.10	3.18	3.25	3.30	3.36	3.40	3.44	3.47	
24	2.92	3.07	3.15	3.22	3.28	3.34	3.38	3.44	3.47	
30	2.89	3.04	3.12	3.20	3.25	3.32	3.37	3.43	3.47	
60	2.83	2.98	3.08	3.14	3.20	3.28	3.33	3.40	3.47	
100	2.80	2.95	3.05	3.12	3.18	3.26	3.32	3.40	3.47	
∞	2.77	2.92	3.02	3.09	3.15	3.23	3.29	3.38	3.47	

Adapted from Duncan DB. *Biometrics* 11 1948; 1.

Table 12-22. Critical Values using Duncan's Test for Example 19

<i>k</i>	<i>T_k</i>	<i>A_k</i>	<i>k</i>	<i>T_k</i>	<i>A_k</i>
2	2.95	2.51	7	3.34	2.84
3	3.10	2.64	8	3.36	2.86
4	3.18	2.70	9	3.38	2.87
5	3.25	2.76	10	3.40	2.89
6	3.30	2.81			

where *t_D* is Dunnett's *t_D* value for *k* treatments (excluding the standard or control) obtained from Table 12-23.

Like the Studentized Range procedure, this is one of the most conservative procedures, and it ensures that the probability of one or more comparisons between treatments and a standard or control being judged significant by chance alone is 5%. The one-tail values (listed in tables for *P* = 0.10) are used when the objective of the study is to select only those treatments that have higher (or lower) means than the standard or control. The two-tail values (listed in the table for *P* = 0.05) are used when the objective of the study is to select those treatments that are either higher or lower than the standard or control. Of course, the decision to carry out a one-tailed or a two-tailed test must be made before the study begins.

In Example 19, suppose *J* is a standard regimen, and it is desired to determine which regimens show different weight gains

from *J*. Here, *t_D* = 3.07 from Table 12-23 for *k* = 9 treatments, 20 DF, and *P* = 0.05 (two-tails).

$$5\% \text{ allowance} = t_D \sqrt{s^2(1/n_i + 1/n_j)} = 3.07 \sqrt{2.17(1/3 + 1/3)} = 3.68$$

Thus, any regimen mean that differs from the mean for Regimen *J* by 3.68 or more is judged to be different from *J*.

Ranked Means

<i>B</i>	<i>A, C</i>	<i>I</i>	<i>J</i>	<i>F, G</i>	<i>D, H</i>	<i>E</i>
1.0	2.3	5.0	5.3	5.7	6.3	9.3

It would be concluded that *B* showed a statistically significant smaller weight gain than *J*, *E* showed a statistically significantly larger weight gain than *J*, and there was insufficient evidence to indicate that the other regimens were different from *J*.

In the same example, if Regimen *A* is a control group and we knew beforehand that all of the other regimens had to be at least as good as the control or better, it may be desired to select those regimens that are statistically significantly better. We would proceed as follows:

$$t_D = 2.60 \text{ from Table 12.23 for } k = 9 \text{ treatments, } 20 \text{ DF, and } P = 0.10 \text{ (this corresponds to a one-tail } P = 0.05)$$

$$5\% \text{ allowance} = t_D \sqrt{s^2(1/n_i + 1/n_j)} = 2.60 \sqrt{2.17(1/3 + 1/3)} = 3.12$$

Thus, any regimen mean that is larger than the mean for Regimen *A* by 3.12 or more is judged to be better than *A*.

Table 12-23. The *t_D* Table

Values of *t_D* for Dunnett's Procedure for Comparing Several Treatments With a Control at the 5% Level of Significance (Use *P* = 0.10 Values for a One-Tailed Test and *P* = 0.05 Values for a Two-Tailed Test.)

DF	<i>k</i> (NUMBER OF TREATMENTS, EXCLUDING THE CONTROL)									
	<i>P</i>	2	3	4	5	6	7	8	9	
10	0.10	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	
	0.05	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	
11	0.10	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	
	0.05	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	
12	0.10	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	
	0.05	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	
13	0.10	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	
	0.05	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	
14	0.10	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	
	0.05	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	
15	0.10	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	
	0.05	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	
16	0.10	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	
	0.05	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	
17	0.10	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	
	0.05	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	
18	0.10	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	
	0.05	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	
19	0.10	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	
	0.05	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	
20	0.10	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	
	0.05	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	
24	0.10	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	
	0.05	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	
30	0.10	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	
	0.05	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	
40	0.10	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	
	0.05	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	
60	0.10	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	
	0.05	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	
120	0.10	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	
	0.05	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	
∞	0.10	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	
	0.05	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	

Adapted from Dunnett CW. *Am Stat Assoc J* 1955; 50: 1096.

Ranked Means

<i>B</i>	<i>A, C</i>	<i>I</i>	<i>J</i>	<i>F, G</i>	<i>D, H</i>	<i>E</i>
1.0	2.3	5.0	5.3	5.7	6.3	9.3

It can be concluded that *F, G, D, H,* and *E* showed a statistically significantly better weight gain than *A*, and that there is insufficient evidence to indicate that *B, C, I,* and *J* were any better than *A*.

OTHER ANOVA DESIGNS COMMON TO PHARMACEUTICAL PROBLEMS—A somewhat more complex design is the *two-way ANOVA*. This design is analogous to the paired *t* test, but consists of more than two treatments; ie, more than one treatment is applied to the same experimental unit (eg, patient) or related units (eg, litter mates, males between 50 and 60 yr, etc). This design has the same advantages and disadvantages as the paired *t* test described earlier in this chapter. The ANOVA table is similar to the one-way table, but includes some new terms. The between-treatments term has the same interpretation as that in the one-way analysis, representing differences between treatments. A new term, between rows, represents the variability of the units to which the treatments have been applied (eg, patients). Finally, the table contains an error term, sometimes referred to as row \times treatment interaction (patient \times drug in a clinical trial).

The treatment mean square is divided by the error mean square (EMS) to form an *F* ratio, for purposes of performing a statistical test. Some complications can exist in the interpretation of this table and the *F* ratios. The examples here consider treatments as including all treatments of interest, and rows as a random selection of experimental units taken from a large population of such units.

For example, to compare a placebo, a generic drug, and a standard drug (three treatments) use a random selection of patients as the experimental units, with each patient to take each of the three treatments. Another example is the comparison of five analytical methods (five treatments) where 10 analysts, selected at random, each perform assays with each method.

Example 20—Three variations of an acne preparation and a control are to be tested for skin irritation. The four products, *A, B, C,* and the control, each are applied to sites on the backs of eight patients. The assignment of the four products to the four sites on the patient is random; ie, a random assignment of treatments to the four sites on each patient is done for each patient, using a random-number table. The products are applied, and after 24 hr, the degree of irritation is determined by assessing irritation subjectively on a scale of 1 to 10. A value of 1 means no irritation and a value of 10 means extreme irritation. The results are shown in Table 12-24.

The computations are similar to those for the one-way ANOVA. The sum-of-squares for treatments is obtained as before. The sum-of-squares for patients is determined exactly as for treatments except the operation is across rows. This is the same as rotating the table 90° and treating the rows as columns in the table matrix. The EMS (expected sum of squares) is obtained by subtracting the row and column sum-of-squares from the total sum-of-squares. The student may wish to follow the computations for this example, in general, however, the use of a statistical computer program is encouraged, as it is much quicker and eliminates potential arithmetical errors.

Table 12-24. Skin Irritation Test

RABBIT	TREATMENT				Σx	Σx^2
	A	B	C	CONTROL		
1	7	5	5	4	21	115
2	4	3	5	2	14	54
3	8	9	7	6	30	230
4	8	6	4	5	23	141
5	7	7	4	2	20	118
6	6	7	5	4	22	126
7	5	6	4	5	20	102
8	4	7	5	4	20	106
Totals	49	50	39	32	170	992

$$\begin{aligned} \text{Total SS} &= \Sigma x_i^2 - (\Sigma x_i)^2/n, \text{ where } (\Sigma x_i)^2/n = CT \\ &= 992 - 170^2/32 = 88.875 \end{aligned}$$

$$\begin{aligned} \text{Between treatments SS} &= [49^2 = 50^2 + 39^2 + 32^2]/8 - 170^2/32 \\ &= 27.625 \end{aligned}$$

$$\begin{aligned} \text{Between rabbits SS} &= [21^2 + 14^2 + \dots + 20^2]/4 - CT \ 3705/4 - CT \\ &= 23.125 \end{aligned}$$

$$\begin{aligned} \text{Error} &= \text{Total SS} - \text{Between treatments SS} \\ &\quad - \text{Between rabbits SS} = 88.875 - 27.625 \\ &\quad - 23.125 = 38.125 \end{aligned}$$

Table 12-25 shows the ANOVA. Since the *F* ratio for treatments (5.07) exceeds the tabled *F* value with 3 and 21 DF at the 5% level, it can be concluded that at least two of the treatments differ. Although one may apply one of the a posteriori tests discussed under one-way ANOVA, inspection of the results suggests that results for Treatments *A* and *B* are similar and both are greater in magnitude than Treatments *C* and the control.

CROSSOVER DESIGN—A design that is popular in experimental research is the crossover design. This is in the class of paired-sample or two-way designs in that all treatments are applied to each experimental unit. For example, in practically all human bioequivalence studies, each subject takes all of the treatments. That is, if a control marketed drug is to be compared to two new formulations, each subject takes all three products.

The difference between the crossover and the two-way design (also known as a randomized block design) is that in the two-way design, the order or placement of treatments are assigned randomly to each patient. In the crossover design, an additional constraint, *order* or *balance*, is imposed on the experiment. For example, in a bioequivalence study of three products, these are taken sequentially during three periods. In the crossover design, each product appears an equal number of times in each period.

Table 12-26 shows how three products, *A, B,* and *C,* may be assigned to nine subjects in a bioavailability study. Note that Treatments *A, B,* and *C* appear exactly three times in each period and that each subject takes all three products. The balancing of order of administration compensates for period effects. If any extraneous variables affect the outcome differently in one period compared to another, all treatments may be affected equally. This would result in a fair comparison of the different treatments. In a purely random assignment of treatments, it would be unlikely that treatments would be assigned in such a balanced order. In an unbalanced design, differences due to periods would not affect treatments equally, resulting in a potential bias and a larger experimental error—the experimental error would include the usual causes of variability plus variability due to period effects. Thus, the crossover design can be considered an improvement over the two-way design in that the error has been reduced and the experiment made more efficient.

Many such designs are available, but care should be exercised to apply the correct design to each experimental situation. The crossover design is related to the Latin square design. Several very good references are available on principles of experimental design. In particular, the book by Cox⁶ is recommended

Table 12-25. ANOVA for Data of Table 12-24

SOURCE OF VARIATION	DF	ANALYSIS OF VARIANCE		
		SUMS-OF-SQUARES	MEAN SQUARE	F RATIO
Between treatments	3	27.625	9.208	5.07
Between rabbits	7	23.125	3.304	
Error	21	38.125	1.815	
Total	31	88.875		

Table 12-26. Example of Crossover Design

SUBJECT	PERIOD 1	PERIOD 2	PERIOD 3
1	B	C	A
2	A	C	B
3	B	A	C
4	C	B	A
5	A	B	C
6	C	A	B
7	B	A	C
8	C	B	A
9	A	C	B

because it is not overly technical and can be understood without resorting to too much mathematics.

Example 21—Three drug formulations were administered to nine subjects in a bioavailability study according to the crossover design illustrated in Table 12-26. The area under the blood level curves were computed for each dosing, and the results are shown in Table 12-27.

The ANOVA (Table 12-28) separates the total variance into four parts: subjects, period (order of administration), treatments, and error.

$$\sum x_i = 2992 \quad \sum x_i^2 = 364,720$$

$$\text{Total SS} = \sum x_i^2 - (\sum x_i)^2/n = 33,162.1$$

$$\text{Subject SS} = \sum (\text{row}^2)/3 - (\sum x_i)^2/n = 29,834.1$$

$$\text{Treatment SS} = \sum (\text{treat. sum}^2)/9 - (\sum x_i)^2/n = 1116.5$$

$$\text{Order SS} = (\sum I^2 + \sum II^2 + \sum III^2)/9 - (\sum x_i)^2/n = 264.3$$

$$\text{Error SS} = \text{Total SS} - \text{Subject SS} - \text{Treatment SS} - \text{Order SS} = 1947.2$$

Neither treatments nor order are significant (see Table 12-15). For 2 and 14 DF, an *F* value of 3.70 is needed for significance. Treatment C has a higher average result, but fails to reach significance in this study. In the early days of bioequivalence testing, bioequivalence studies were designed to have a power of 0.8 to detect a difference of 20% between treatments. This means that a sufficient number of subjects should be included in the study so that if a true difference of 20% or more exists between two treatments, there will be at least an 80% chance of finding a significant difference. This method of evaluating equivalence has been replaced by a more meaningful confidence interval approach.⁸

If the crossover design becomes unbalanced, due to dropouts, or other conditions, a computer analysis can be used (eg, SAS).

Another experimental design common in clinical trials is the repeated-measures design, often called a *split-plot design*. For example, two treatments are compared by making observations in two independent groups of patients over time. Although an equal number of patients in each group is desirable, it is not necessary for the data analysis. The observations are made at

Table 12-27. Results of Bioavailability Study

SUBJECT	PERIOD 1	PERIOD 2	PERIOD 3	SUM
1	B = 107	C = 102	A = 99	308
2	A = 100	C = 106	B = 89	295
3	B = 98	A = 90	C = 128	316
4	C = 71	B = 54	A = 63	188
5	A = 92	B = 111	C = 107	310
6	C = 113	A = 115	B = 91	319
7	B = 169	A = 187	C = 195	551
8	C = 88	B = 95	A = 77	260
9	A = 122	C = 168	B = 155	445
Period sum	I: 960	II: 1028	III: 1004	2992
Treatment sum	A: 945	B: 969	C: 1078	
Treatment average	105	107.7	119.8	

Table 12-28. ANOVA for Bioavailability Study

SOURCE OF VARIATION	ANALYSIS OF VARIANCE			F RATIO
	DF	SUMS-OF-SQUARES	MEAN SQUARE	
Between subjects	8	29,834.1	3729.3	
Between treatments	2	1,116.5	558.3	3.15
Order	2	264.3	132.1	0.75
Error	14	1,947.2	177.0	
Total	26	33,162.1		

the same time periods in both groups. The example shows the basic design and ANOVA table. The details of the calculations are not shown. Usually, a software program is used to analyze and summarize the data. The details of the analysis are given in Bolton⁸ and Winer.¹⁵

Example 22—A pilot study to compare the effects of an antihypertensive drug versus placebo was designed with four patients on drug and four on placebo. Blood pressure changes from baseline were measured for 6 weeks at biweekly intervals. The results are shown in Table 12-29.

The ANOVA is shown in Table 12-30. The terms of interest are Treatments and Treatment × Times. The former term measures differences of the overall average results of the two treatments. The error term for Treatments is the mean square for Patients. The Treatment × Times term compares the time trends for the two treatments. The error term for the Treatment × Times effect is Patient × Times (treatments). If the trends are parallel, this term will not be significant. Significance for this term indicates a lack of parallelism, suggesting that differences between treatments depend on the time of observation.

As with most experimental data, a graphic display is recommended. Figure 12-9 is a plot of the average results versus time. The significant difference between treatments (*P* < 0.05) is apparent from the plot and the ANOVA. The time trends of both treatments are similar, and can be explained by the experimental variability (Treatment × Times is not significant).

NONPARAMETRIC TESTS OF SIGNIFICANCE—The validity of the *t* test for comparing two means depends to some extent (especially for small samples) on the assumptions that the two populations sampled are distributed approximately normally and have essentially equal variances. A procedure for testing the equality of variances has been discussed previously. Statistical procedures that do not depend on the assumption of normality are called nonparametric tests. Three commonly used procedures are the Rank Sum test for unpaired data, and the Signed-Rank Sum and Sign tests for paired data.

Rank Sum Test of Significance—The rank sum test of significance is the nonparametric analog of the two-independent sample *t* test. The *n*₁ and *n*₂ observations are taken from two independent groups. After the *n*₁ and *n*₂ observations are arranged in order of size, the combined values are ranked from 1, for the lowest, to (*n*₁ + *n*₂) for the highest, and the sum of the ranks *T* of the *n*₁ observations in the smaller sample is computed. Values that are tied are given average ranks. Also calculate *T'* = *n*₁(*n*₁ + *n*₂ + 1) - *T*, and enter Table 12-31¹⁶ with *n*₁, *n*₂, and *T* or *T'*, whichever is smaller. If the calculated *T* (or *T'*) is equal to or less than the tabled value, the null hypothesis is rejected at the significance level *P*.

Example 23—Data were available on the duration of loss of the righting reflex (min) for 10 mice given a standard barbiturate and for 11 mice given a test barbiturate (Table 12-32). Entering Table 12-31 with

Table 12-29. Reduction in Diastolic Blood Pressure from Baseline

PATIENT	DRUG WEEK			PATIENT	PLACEBO WEEK		
	2	4	6		2	4	6
1	10	8	12	2	10	8	12
3	8	6	14	5	6	2	10
4	12	14	8	6	4	0	2
7	10	10	14	8	0	4	10
Average	10.0	9.5	12.0		5.0	3.5	8.5

Table 12-30. ANOVA for Example 22

SOURCE	DF	SS	MS	F
Patients	6	109	18.2	
Treatments	1	140.2	140.2	7.7
Times	2	60.3	30.2	3.5
Treatment × times	2	6.3	3.2	0.4
Patient × times (treatments)	12	104	8.7	
Total	23	419.8		

$n_1 = 10$, $n_2 = 11$, and $T' = 69.5$, we find that the calculated T' value 69.5 is less than the tabulated value 73 for $P = 0.01$. Therefore, because the probability of the standard drug and test drug values being the same is less than 0.05 (actually, it is less than 0.01), it is concluded that they are different. This test compares the medians of the two-populations sampled. The median of an ordered set of observations is defined as the middlemost value for an odd number of observations, and as the average of the two middlemost values for an even number of observations. Thus, the median for the standard drug is $(130 + 148)/2 = 139$ and the median for the test drug is 103.

Signed-Rank Sum Test of Significance—The signed-rank sum test of significance is the nonparametric analog of the paired t test. The differences between the n paired values are ranked in order of absolute size from 1, for the lowest, to n , for the highest, ignoring zero differences. Tied values are assigned an average rank. After the differences are ranked, the signs of the differences are attached to the ranks, and the sum of the positive ranks and of the negative ranks are obtained. Enter Table 12-33 with $n =$ the number of non-zero differences and the sum T of positive or negative ranks, whichever is smaller. When the calculated T is equal to or less than the tabled T , the null hypothesis is rejected at the significance level P .

Example 24—The procedure is illustrated for data given in Example 12 (Table 12-34). Entering Table 12-33 with $n = 15$ and $T = 22.5$, we find that the calculated T value 22.5 is less than the tabulated value 25 for $P = 0.05$. Therefore, because the probability of the morning and afternoon values being the same is less than 0.05, it is concluded that they are different.

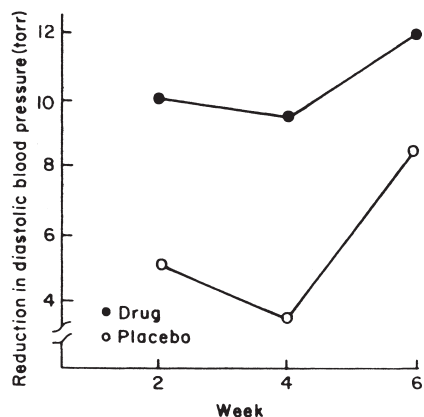
Sign Test—The sign test also is used for paired data, but it is not as powerful as the signed-rank test; it is more difficult to find significant differences when they exist with the sign test. Count the number of positive differences (b) and the number of negative differences (c), ignoring zero differences, and calculate

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

where $|b - c|$ is the absolute (ie, positive) difference $b - c$.

This is referred to the chi-square table (see Table 12-12) with $DF = 1$, the test being essentially the same as the chi-square test illustrated in Example 16.

Example 25—The procedure is illustrated for the data given in Examples 12 and 24.

**Figure 12-9.** Plot of average results for Example 22.

$$b = \text{number of positive differences} = 11$$

$$c = \text{number of negative differences} = 4$$

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} = \frac{(|11 - 4| - 1)^2}{11 + 4} = \frac{36}{15} = 2.40$$

Table 12-12 shows that for 1 DF the probability of getting a value of χ^2 larger than the calculated value 2.40 is between $P = 0.10$ and $P = 0.20$. Since P is not equal to or less than 0.05, it is concluded that there is insufficient evidence to indicate that the morning and afternoon values are different. This conclusion is not in agreement with that of the t test and the signed-rank test. The reason for this is that the statistical sign test considers only the sign of the difference and not the magnitude, and thus is a less-sensitive test in borderline situations such as this one.

REJECTION OF ABERRANT OBSERVATIONS—It is common practice among chemists and others working in the physical sciences to make observations in duplicate or triplicate. This is usually done for the purpose both of obtaining a more accurate result and also detecting mistakes in dilution, weighing, and so on. It is quite a common practice to reject the most extreme of the three results if it appears to disagree with the others.

Youden,^{17,18} a chemist as well as a statistician, made a study of the problem of rejection of observations in an attempt to answer three questions:

1. If the extreme observation of triplicates is always rejected when only normal variation is present, how accurate is the result?
2. Is the average of the two closest observations as good an estimate as the average of all three?
3. By how much should the outlying observation of triplicates differ from the other two in order to be reasonably assured that this difference is due to a blunder rather than normal variation?

He found that rejection of the outlying observation resulted not only in the variation being greatly underestimated but the mean was biased.

If one wished to follow a simple rule of rejection^{17,18} of observations in samples of three so as to reject not more than 5% of the extreme observations arising from normal variation, a rejection ratio of D/d greater than 20 would be required.

$$D/d = 20$$

where

D = difference between the most extreme observation and its closest neighbor

d = difference between two closest observations

In the USP there is an excellent chapter on the design and analysis of biological assays in which are included some tests for rejection of outlying observations. These and other tests can be applied to chemical, as well as, biological assays.¹⁹ Two criteria are presented here, one for rejecting single suspect observations in one group and the other for rejecting a whole group of observations.

To use the first criterion, arrange the observations in the group in order of their magnitude and number them from 1 to n beginning with the supposedly erratic or outlying observation, thus

$$y_1, y_2, y_3, \dots, y_n$$

where y_1 is the suspect observation. If there are 3 to 7 observations in the group, calculate

$$G_1 = \frac{y_2 - y_1}{y_n - y_1}$$

If there are 8 to 10 observations in the group, and the smallest value seems suspect, again arrange them in order from lowest to highest and calculate

$$G_2 = \frac{y_2 - y_1}{y_{n-1} - y_1}$$

Table 12-31. The Rank Sum Table

Values of T or T', Whichever Is Smaller, Significant at the 10%, 5%, and 1% Levels

N ₂	P	N ₁ (SMALLER SAMPLE)																	
		4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
8	0.10	15	23	31	41	51													
	0.05	14	21	29	38	49													
	0.01	11	17	25	34	43													
9	0.10	16	24	33	43	54	66												
	0.05	14	22	31	40	51	62												
	0.01	11	18	26	35	45	56												
10	0.10	17	26	35	45	56	69	82											
	0.05	15	23	32	42	53	65	78											
	0.01	12	19	27	37	47	58	71											
11	0.10	18	27	37	47	59	72	86	100										
	0.05	16	24	34	44	55	68	81	96										
	0.01	12	20	28	38	49	61	73	87										
12	0.10	19	28	38	49	62	75	89	104	120									
	0.05	17	26	35	46	58	71	84	99	115									
	0.01	13	21	30	40	51	63	76	90	105									
13	0.10	20	30	40	52	64	78	92	108	125	142								
	0.05	18	27	37	48	60	73	88	103	119	136								
	0.01	14	22	31	41	53	65	79	93	109	125								
14	0.10	21	31	42	54	67	81	96	112	129	147	166							
	0.05	19	28	38	50	62	76	91	106	123	141	160							
	0.01	14	22	32	43	54	67	81	96	112	129	147							
15	0.10	22	33	44	56	69	84	99	116	133	152	171	192						
	0.05	20	29	40	52	65	79	94	110	127	145	164	184						
	0.01	15	23	33	44	56	69	84	99	115	133	151	171						
16	0.10	24	34	46	58	72	87	103	120	138	156	176	197	219					
	0.05	21	30	42	54	67	82	97	113	131	150	169	190	211					
	0.01	15	24	34	46	58	72	86	102	119	136	155	175	196					
17	0.10	25	35	47	61	75	90	106	123	142	161	182	203	225	249				
	0.05	21	32	43	56	70	84	100	117	135	154	174	195	217	240				
	0.01	16	25	36	47	60	74	89	105	122	140	159	180	201	223				
18	0.10	26	37	49	63	77	93	110	127	146	166	187	208	231	255	280			
	0.05	22	33	45	58	72	87	103	121	139	158	179	200	222	246	270			
	0.01	16	26	37	49	62	76	92	108	125	144	163	184	206	228	252			
19	0.10	27	38	51	65	80	96	113	131	150	171	192	214	237	262	287	313		
	0.05	23	34	46	60	74	90	107	124	143	163	182	205	228	252	277	303		
	0.01	17	27	38	50	64	78	94	111	129	147	168	189	210	234	258	283		
20	0.10	28	40	53	67	83	99	117	135	155	175	197	220	243	268	294	320	348	
	0.05	24	35	48	62	77	93	110	128	147	167	188	210	234	258	283	309	337	
	0.01	18	28	39	52	66	81	97	114	132	151	172	193	215	239	263	289	315	

Adapted from Tate MW, Clelland RC. *Nonparametric and Shortcut Statistics*. Danville IL: Interstate Print, 1957.

Table 12-32. Data for Example 23

STANDARD DRUG	RANK	TEST DRUG	RANK
96	4.5	0	1
109	8	91	2
126	13	92	3
130	15	96	4.5
130	15	99	6
148	17	103	7
153	18	117	9
158	19	118	10
169	20	119	11
Died	21	120	12
		130	15
	$T = 150.5$		$n_2 = 11$
	$n_1 = 10$		
	$T' = n_1(n_1 + n_2 + 1) - T = 10(10 + 11 + 1) - 150.5 = 69.5$		

Table 12-33. The Signed-Rank Sum Table

Values of T for Signed-Rank Test, Significant at the 10%, 5%, and 1% Levels

n	P			n	P		
	0.10	0.05	0.01		0.10	0.05	0.01
5	0			18	47	40	27
6	2	0		19	53	46	32
7	3	2		20	60	52	37
8	5	3	0	21	67	58	43
9	8	5	1	22	75	65	49
10	10	8	3	23	83	73	55
11	14	10	5	24	91	81	61
12	17	13	7	25	100	89	68
13	21	17	9	26	110	97	75
14	25	21	12	27	120	106	83
15	30	25	16	28	130	116	91
16	35	29	19	29	141	126	100
17	41	34	23	30	152	136	109

Table 12-34. Signed Ranks from Example 24

DIFFERENCES	SIGNED-RANKS
2	2
-3	-3
4	6
8	15
-4	-6
6	12.5
-1	-1
-6	-12.5
5	10
4	6
4	6
0	ignore
5	10
7	14
4	6
5	10
Sum of positive ranks = 97.5	
Sum of negative ranks = 22.5 = T	
n = 15	

If there are 11 to 13 observations, follow the same procedure, but use the statistic

$$G_1 = \frac{y_3 - y_1}{y_{n-1} - y_1}$$

If there are 14-25 observations, follow the same procedure, but use the statistic

$$G_4 = \frac{y_3 - y_1}{y_{n-2} - y_1}$$

If the largest value is open to suspicion as possibly being aberrant, arrange the observations in order from highest to lowest and number them, always labeling the suspect observation y_1 .

If the calculated value of G_1 , G_2 , G_3 , or G_4 is larger than the tabled value (which gives the probability of a value being so extreme as that observed), it can be assumed that the observation

Table 12-35. Criteria for Testing Extreme Value

STATISTIC	n, NUMBER OF OBSERVATIONS	CRITICAL VALUES
$G_1 = \frac{y_2 - y_1}{y_n - y_1}$	3	.988
	4	.889
	5	.780
	6	.698
	7	.637
$G_2 = \frac{y_2 - y_1}{y_{n-1} - y_1}$	8	.683
	9	.635
	10	.597
$G_3 = \frac{y_3 - y_1}{y_{n-2} - y_1}$	11	.679
	12	.642
	13	.615
$G_4 = \frac{y_3 - y_1}{y_{n-2} - y_1}$	14	.641
	15	.616
	16	.595
	17	.577
	18	.561
	19	.547
	20	.535
	21	.524
	22	.514
	23	.505
	24	.497
	25	.489

truly does not belong to the group and the observation is rejected. The values of G for a probability $P = 0.01$, that an outlier could occur at either end are shown in Table 12-35. This same criterion could be used for testing whether the largest or smallest average in a group of averages differs significantly from the remainder of the averages (Table 12-35).

Example 26—Suppose among the gains in weight of six rats after a feeding experiment, one weight was found to be much less than the other five. Can that observation be discarded? The six gains in weight are 36, 40, 38, 42, 20, and 39.

Rearrange these in order from smallest to largest and label y_1, \dots, y_6 , where $n = 6$.

y_1	20
y_2	36
y_3	38
y_4	39
y_5	40
y_6	42

$$G_1 = \frac{y_2 - y_1}{y_6 - y_1} = \frac{36 - 20}{42 - 20} = \frac{16}{22} = 0.727$$

Referring to the value of G_1 for $n = 6$ in the table, $G_1 = 0.698$ for $P = 0.01$. Since the calculated value of G_1 is larger than this value, reject the value of 20 and work with the remaining five values.

The second criterion for an aberrant observation as given in the USP compares the variation or range between various groups. It is a test for the homogeneity of the ranges (the range is again the highest value in a group minus the lowest value) and is for the purpose of locating outliers within one group of values. This method and its accompanying table are presented in considerable detail in the USP. The rejection of outliers using only statistical criteria is controversial. A knowledge of the characteristics or properties of the chemical or biological systems being studied should be used when making decisions to reject outlying data.

QUALITY CONTROL METHODS—A very short explanation is given here regarding the quality control methods that were developed primarily by Dr Walter Shewhart of the Bell Telephone Laboratories. A more complete explanation can be found in two short publications of the American Standards Association^{20,21} and many texts, including Dixon and Massey.²²

The quality control method for variables involves plotting the data as dots on a graph with the variable measured on the vertical axis and time (hours, days, etc) on the horizontal axis. The control is maintained by inserting on the chart the grand average and control limits that have been calculated from accumulated experience and drawn on the chart as parallel horizontal lines as shown in Figure 12-10. When all the dots fall within the limits, the results are said to be in a state of statistical control. When a dot falls outside the limits, a potential problem is indicated.

In a control chart, usually each dot is an average for a sample consisting of, say, four observations. The standard error of the average then is calculated for each group of four observations, and an average value for the standard error of the average is obtained. This is designated by s_x . The grand average of all the averages plotted also is calculated and is labeled \bar{x} . The 3-sigma

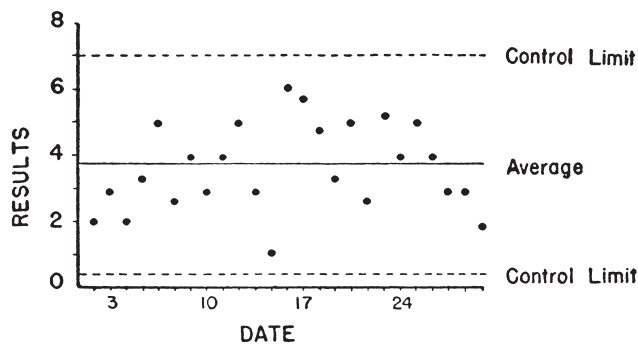


Figure 12-10. A typical quality control chart.

Table 12-36. Calculation of Standard Deviation from Range

SIZE OF SAMPLE (n)	AVERAGE NUMBER OF STANDARD DEVIATIONS IN THE AVERAGE RANGE (d)	SIZE OF SAMPLE (n)	AVERAGE NUMBER OF STANDARD DEVIATIONS IN THE AVERAGE RANGE (d)
2	1.128	7	2.704
3	1.693	8	2.847
4	2.059	9	2.970
5	2.326	10	3.078
6	2.534		

control limits used on the control chart can be obtained from

$$\begin{aligned} \text{Upper limit} &= \bar{\bar{x}} + 3s_{\bar{x}} \\ \text{Lower limit} &= \bar{\bar{x}} - 3s_{\bar{x}} \end{aligned}$$

Thus, it can be seen that the control-chart technique is a graphic means of investigating whether the variation exhibited over a very short period of time is the same as the variation that occurs over a long period of time. If the two variations are identical and all of the plotted dots fall within the control limits, the experiments or processes that produced the data are said to be in a state of *statistical control*.

For many pharmaceutical processes, particularly heterogeneous processes, typical Shewhart control charts do not describe the process adequately. The process seems to not be in control. In these cases, alternative methods should be considered.²³ Control charts are a valuable tool for process validation.

It is possible to calculate the control limits by using the range in each group of four, instead of calculating the standard deviation. This is because, on the average, for samples of less than 10, the range and the standard deviation are related very closely. Given the number of observations in the sample, the standard deviation can be calculated by dividing the range by the appropriate figure given in Table 12-36 for the size of the sample, *n*. The factors for calculating 3-sigma limits from the range are given as Column *A*₂ in Table 12-37.

Control charts using 3-sigma limits can be obtained by the use of figures given in Table 12-37. The formulas are

$$\begin{aligned} \text{Upper limit for averages} &= \bar{\bar{x}} + A_2 \bar{R} \\ \text{Lower limit for averages} &= \bar{\bar{x}} - A_2 \bar{R} \\ \text{Upper limit for ranges} &= D_4 \bar{R} \\ \text{Lower limit for ranges} &= D_3 \bar{R} \end{aligned}$$

Where \bar{R} = average range

These calculated limits are drawn on the charts as described above.

Example 27—A drug manufacturer keeps a record of the uniformity of the machine that is filling a given weight of a drug into ampuls. Samples of the finished product are taken at definite time intervals. The data are accumulated and arranged into groups of four ampuls according to the order in which they were taken from a filling machine. The av-

Table 12-37. Factors for 3-Sigma Limits^a

SIZE OF SAMPLE (n)	FACTORS FOR \bar{R} CHART		FACTOR FOR \bar{X} CHART <i>A</i> ₂
	<i>D</i> ₃	<i>D</i> ₄	
2	0	3.27	1.880
3	0	2.57	1.023
4	0	2.28	0.729
5	0	2.11	0.577
6	0	2.00	0.483
7	0.08	1.92	0.419
8	0.14	1.86	0.373
9	0.18	1.82	0.337
10	0.22	1.78	0.308

^a This table contains data from the tables in Appendix 1 of Z13—1958.²⁰

Table 12-38. Calculations for a Quality-Control Chart On Averages and Ranges for Samples of 4 from a Filling Machine

TIME	AVERAGE (g)	RANGE (g)	TIME	AVERAGE (g)	RANGE (g)
Jan 6					
8 AM	38.1	1.5	Jan 7		
9 AM	37.6	2.1	8 AM	37.6	2.1
10 AM	38.3	1.1	9 AM	39.1	1.4
11 AM	36.5	2.4	10 AM	38.5	1.1
12 M	38.9	3.1	11 AM	37.7	1.9
1 PM	37.8	2.8	12 M	38.1	2.3
2 PM	38.5	1.7	1 PM	38.5	2.4
3 PM	39.4	1.6	2 PM	37.6	1.6
4 PM	36.4	2.5	3 PM	37.9	1.8
			4 PM	38.6	1.0

Grand average = $\bar{\bar{x}} = 38.1$
 Average range = $\bar{R} = 1.9$
 Control limits^a for average = $\bar{\bar{x}} \pm A_2 \bar{R} = 38.1 \pm 0.729(1.9)$
 Upper limit = 39.49
 Lower limit = 36.71
 Control limits^b for range, are $D_3 \bar{R}$ and $D_4 \bar{R}$ or $0(1.9)$ and $2.28(1.9)$ which equal 0 and 4.33, respectively.

^a *A*₂ is the factor for using the range to calculate 3-sigma limits for the average (ie, 3 times the standard error of the average). See Table 12-37 for *N* = 4.

^b *D*₃ and *D*₄ are the factors for using the range to calculate 3-sigma limits for the range (ie, 3 times the standard error of the range). These values are taken from Table 12-37. In two instances the point plottings fell below the lower control limit, indicating a lower average fill than one might expect, ie, there is a lack of statistical control.

erage and the range are computed for each group of four as given in Table 12-38 according to the time the samples are taken. The resulting quality-control charts are shown in Figure 12-11.

CONTROL CHART FOR FRACTION DEFECTIVE—

The control chart for fraction defective may be applied to results of an inspection that accepts or rejects individual items of a product. It is designed with the same objectives in mind as the \bar{x} and \bar{R} charts. Its most effective use is in the improvement of quality, although it also discloses the presence of assignable causes of variation. It provides management with an effective quality history. Fraction defective, *p*, may be defined as the ratio of the number of defective articles found in any inspection or series of inspections to the total number of articles actually inspected. This is expressed nearly always as a decimal fraction (Fig 12-12). The formula for the control limits on a fraction defective chart is

$$\bar{p} \pm 3 \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

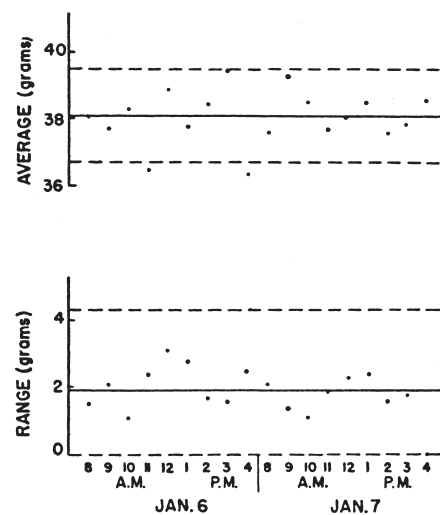


Figure 12-11. Quality control charts for data from Table 12-38.

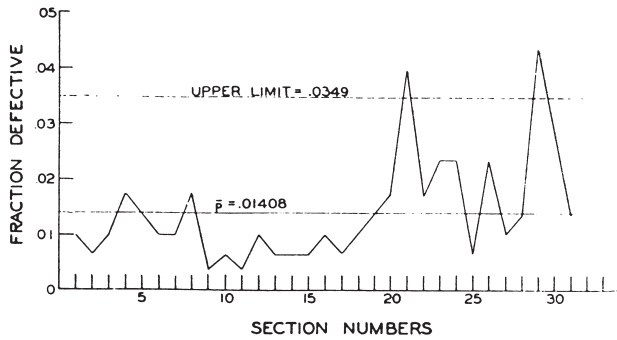


Figure 12-12. Control chart for fraction defective. (Courtesy Lilly)

Example 28—A department head in the capsule department of a pharmaceutical company keeps a record of the number of defective capsules found in sections of large lots of capsules (Table 12-39). Each section consists of approximately 19,000 capsules. In Table 12-39 and Figure 12-12, where points fall above the upper control limit, a greater number of defects are present than may be expected—there is a lack of statistical control. These sections are reinspected carefully and action is taken at the machine to correct the causes of bad quality.

The sample size, *n*, from each section is 300 capsules and typical data are shown in Table 12-39, plotted in Figure 12-12. Note that Sections 21 and 29 appear to be out of control. These sections were subjected to 100% reinspection. Approximately 4.5% of the capsules were defective and were removed.

ACCEPTANCE SAMPLING—Acceptance sampling has become one of the major fields of statistical quality control. It is used in many phases of manufacturing such as inspection of incoming materials, process inspection at various points in the manufacturing operations, and final inspection of the finished product. Sampling inspection usually is used in lieu of 100% inspection for several reasons:

1. The cost of 100% inspection is prohibitive.
2. 100% inspection is fatiguing and may result in the inspectors making errors.

Table 12-39. Data Collected from the Process in Example 28

SECTION NUMBER	NUMBER DEFECTIVES	FRACTION DEFECTIVE	SECTION NUMBER	NUMBER DEFECTIVES	FRACTION DEFECTIVE
1	3	0.01	17	2	0.0067
2	2	0.0067	18	3	0.01
3	3	0.01	19	4	0.0133
4	5	0.0167	20	5	0.0167
5	4	0.0133	21	12	0.04
6	3	0.01	22	5	0.0167
7	3	0.01	23	7	0.0233
8	5	0.0167	24	7	0.0233
9	1	0.0033	25	2	0.0067
10	2	0.0067	26	7	0.0233
11	1	0.0033	27	3	0.01
12	3	0.01	28	4	0.0133
13	2	0.0067	29	13	0.0433
14	2	0.0067	30	9	0.03
15	2	0.0067	31	4	0.0133
16	3	0.01			

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number inspected}} = \frac{131}{31 \times 300} = \frac{131}{9300} = 0.01408$$

$$\begin{aligned} \text{Control limits for } \bar{p} &= \bar{p} \pm 3 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ &= 0.01408 \pm 3 \sqrt{\frac{0.01408(1-0.01408)}{300}} \end{aligned}$$

Upper limit = 0.0349
Lower limit = 0

3. The inspection operation may involve destructive testing.
4. A statistical sampling plan well applied may give better quality assurance than 100% inspection.

In sampling one must consider the laws of probability. The risk of rejecting good-quality material and the risk of accepting bad merchandise should be appraised. Sampling plans can be designed and applied in such a manner as to reduce these risks to a minimum and, over a period of time, give assurance of quality products.

The graph illustrating the performance of a sampling plan (ie, ability to discriminate between acceptable and unacceptable lots) is called an *operating characteristic curve* (OC curve). For any given quality of submitted material it is possible to determine the probability of acceptance.

Figure 12-13 is an example of an OC curve for the sampling plan described in Example 29. The government publication MIL-STD-105E²⁴ gives many different sampling plans with their corresponding OC curves. A plan that is appropriate for a product is chosen depending on lot size and seriousness of the defect.

Example 29—Example of a *statistical sampling plan*. A pharmaceutical manufacturer receives empty bottles of a particular size from a supplier in lots of 20,000 bottles each. The drug firm would like the producer to submit material that is not more than 1.0% defective most of the time, or specifically 95% of the time. See point A, Figure 12-13. However, the pharmaceutical firm has agreed to take one chance in 10 of accepting a lot that is 2.6% defective. See point B, Figure 12-13.

The acceptance sampling plan that complies with these specifications is as follows. Take a random sample of 540 bottles. Inspect the bottles for defectives. If zero to nine bottles are found defective, accept the lot; if 10 or more defectives are found, reject the lot. The operating characteristic curve for this plan is illustrated in Figure 12-13.²⁴

One also can see that, using this sampling plan, submitted lots having 0.5% defective will be accepted about 99 times in 100 (probability of acceptance = 0.99) and thus rejected about one time in 100. Submitted lots having 1.75% defective will be accepted 50 times in 100 (probability of acceptance = 0.50) and rejected half the time.

STATISTICS OF THE STRAIGHT LINE—The use of straight lines to illustrate and define relationships or to help interpret data is common in research investigations. In phar-

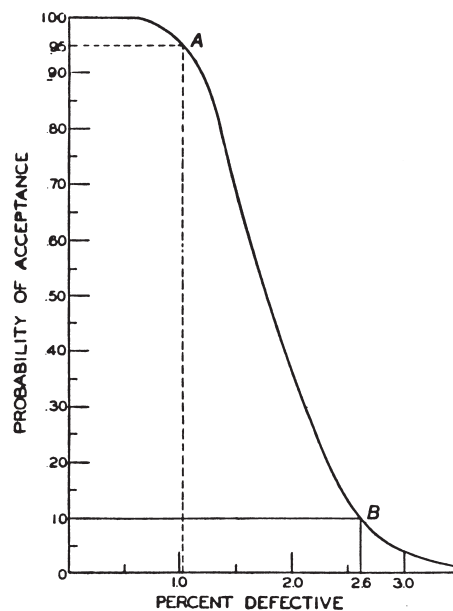


Figure 12-13. Operating characteristic curve. (From *Military Sampling Procedures and Tables for Inspection by Attributes*. MIL-STD-105E. Washington DC: USGPO, 1989.)

maceutical research, straight lines may be used for predictive purposes in stability studies, or to estimate future events such as sales figures in market research studies. Straight lines are found in many theoretical relationships in physical and biological chemistry. First-order and zero-order kinetics can be expressed in a linear fashion. Michaelis-Menten kinetics and the Arrhenius relationship can be transformed to a linear form. Dose-response curves often are linearized if the response is plotted versus log dose. In fact, it is almost always desirable to express a relationship in the form of a straight line, if at all possible.

Reasons for the desirability of straight-line relationships include the ease of extrapolation and interpolation as well as the simplification of the determination of the parameters of the line, the slope and the intercept. The straight line is defined by these two parameters and these often have biological and/or physical significance. Consider the example of a first-order kinetic relationship

$$C = C_0 e^{-kt}$$

where

- C = concentration at time t
- C_0 = concentration at time 0
- k = first-order rate constant

This equation is not linear—a plot of C versus t will not result in a straight line. If the experimental data are gathered for C as a function of time, one usually is interested in defining the first-order relationship, in particular to evaluate the constants (sometimes called parameters) k and C_0 . This is done most easily by linearizing the equation using a logarithmic (log) relationship. Using log to the base 10, the following linear relationship is obtained.

$$\log C = \log C_0 - kt/2.3$$

This has the form of a straight line. The general equation of a straight line can be expressed as

$$y = a + bx$$

where

- y is the dependent variable.
- a is the Y intercept (the value of y when $x = 0$).
- b is the slope of the line.
- x is the independent variable.

Figure 12-14 shows this linear relationship and calculation of the parameters.

The linearized first-order kinetic equation will show a straight line when $\log C$ is plotted versus time, with intercept $\log C_0$ and slope $-k/2.3$. The linearized form makes it easy to obtain the values of C_0 and k . C_0 is the antilog of the intercept, $\log C_0$, and $k = -2.3 \times \text{slope}$.

One of the problems in estimating these values from real data is the variability; a plot does not clearly define a straight line. If variability is large, it may be very difficult to decide how to draw the line. Figure 12-15 shows real data from a pharma-

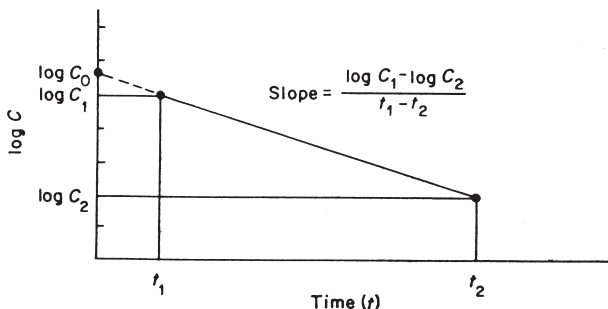


Figure 12-14. Plot of $\log C$ versus time.

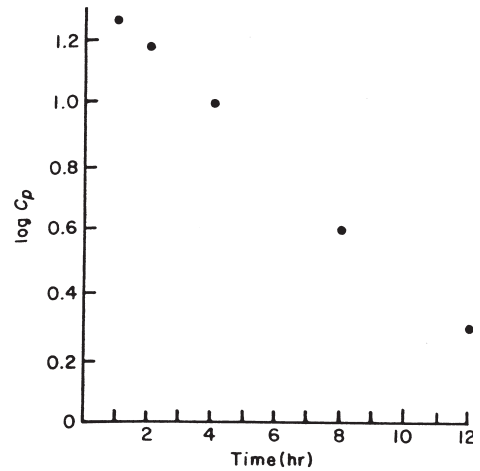


Figure 12-15. Drug plasma levels following an intravenous bolus dose of a drug.

cokinetic study where plasma drug concentrations are measured following an intravenous bolus injection of drug (a one-compartment model).

When confronted with a relationship that should be linear from a theoretical viewpoint but where the x,y values do not lie exactly on a single line, the lack of an exact fit will be considered to be due to variability (error) in y (the dependent variable). In most cases that are encountered, the x variable (the independent variable) tends to have little error relative to the y variable. For example, in a dose-response relationship, the drug is carefully prepared so that an almost unerring dose is administered. However, the response is unpredictable due to the biological variability of the natural material (eg, animals or bacteria). In a kinetic study, the x variable, time, can be measured with great accuracy. The dependent variable, concentration, is variable due to analytical error, for example. The best line for such variable data is called the least-squares (LS) line. This line is such that the sum of the squared deviations of each point from the line is minimized. That is, if the vertical distance from each point to the LS line is calculated, and the squares of these distances are summed, the LS line would minimize the sum-of-squares. Using methods of calculus, one easily can show that the slope and intercept of the LS line are as follows.²⁵

$$b = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b \bar{x}$$

Example 30—Consider the data from Figure 12-15. See Table 12-40. The equation of the LS line is

$$\text{Log (Concentration)} = 1.345 - 0.0886 (\text{Time})$$

$$k = -2.3(-0.0886) = 0.204$$

$$\therefore C = 22.1 e^{-0.204t}$$

Table 12-40. Concentration vs Time for Example 30

Time (x)	1 hr	2 hr	4 hr	8 hr	12 hr
Concentration ($\mu\text{g/mL}$)	18	15	10	4	2
Log concentration	1.255	1.176	1.000	0.602	0.301

$$b = \frac{16.036 - 27(4.334)/5}{229 - 27^2/5} = -0.0886$$

$$a = 0.8669 + 0.0886 (5.4) = 1.34534$$

This procedure can be used to fit a line for any two variables. If statistical inference procedures are to be applied to the line, certain assumptions about the data are necessary.

1. The x variable is measured without error. In practical situations, the error in x should be small compared to the error in the y variable.
2. The y variable is distributed normally with a true mean equal to $A + Bx$ (A and B are the true values of the intercept and slope) and with the same variance, σ^2 , at all values of x .

With these assumptions, the confidence intervals for the line can be computed and statistical tests performed on the parameter estimates, a and b .

Example 31—In analytical procedures for drugs, a calibration curve often is constructed using known concentrations of the material to be analyzed. The relationship of drug concentration and the analytical measurement usually is linear. In spectrometric methods, absorption usually is proportional to concentration. The data in Table 12-41 were obtained for the construction of such a calibration curve. These data and the LS line are plotted in Figure 12-16. The LS slope, b , is

$$(72.67 - 2.421 \times 100/4)/500 = 0.02429$$

The LS intercept, a , is

$$0.60525 - 0.02429(25) = -0.002$$

The estimate of the variance of y , $s_{y,x}^2$ is

$$\begin{aligned} s_{y,x}^2 &= \frac{\sum y_i^2 - (\sum y_i)^2/n - b^2[\sum x_i^2 - (\sum x_i)^2/n]}{n - 2} \\ &= \frac{1.7607 - (2.421)^2/4 - 0.02429^2[3000 - (100)^2/4]}{2} \\ &= [0.2954 - (0.02429)^2[500]]/2 = 0.000208 \end{aligned}$$

The value of the numerator is the sum-of-squares of the difference between the actual values of y and the value of y on the LS line, for each y . The divisor, $n - 2$, is the number of data pairs minus 2, the DF. Thus, the estimate of the variance of y in this example has 2 DF. The reason that 2 is subtracted from the number of data points to obtain the DF is that two parameters are being estimated in the case of a straight line. In previous examples, such as the t test, only the mean is estimated for a treatment group, and $DF = n - 1$.

With an estimate of the variance, statistical procedures can be applied to these data if the assumptions, stated above, hold. Concentration is measured with little error, whereas the spectrometric readings, y , have error due to instrumental variability, sample processing, and handling (diluting, pipetting, etc), among other sources of variability. If it is assumed that the variance is the same at each concentration value and that the concentration values are distributed normally, the following statistical procedures can be used.

Confidence Limits and Test of the Slope—As in the statistical-hypothesis testing procedures described previously in this chapter, a test of the slope versus a hypothetical value can be performed. Also, confidence limits can be placed on the slope.

Example 32—Suppose that a value of 0.025 for the slope of the line is reported in an authoritative publication on this assay procedure. It is desired to determine if the slope in the experiment of Example 31 is different from 0.025 (ie, $H_0: B = 0.025$). The estimate of the variance of a slope is

$$s_b^2 = s_{y,x}^2 / \sum (x_i - \bar{x})^2$$

The test is a two-sided t test with $n - 2$ DF of the following form:

$$\begin{aligned} t &= \frac{|b - B|}{\sqrt{s_b^2}} \\ t &= \frac{|0.02429 - 0.025|}{\sqrt{0.000208/500}} = 1.10 \end{aligned}$$

Table 12-41. Absorbance vs Concentration

Concentration	10 mg/L	20 mg/L	30 mg/L	40 mg/L
Absorbance	0.241	0.492	0.710	0.978

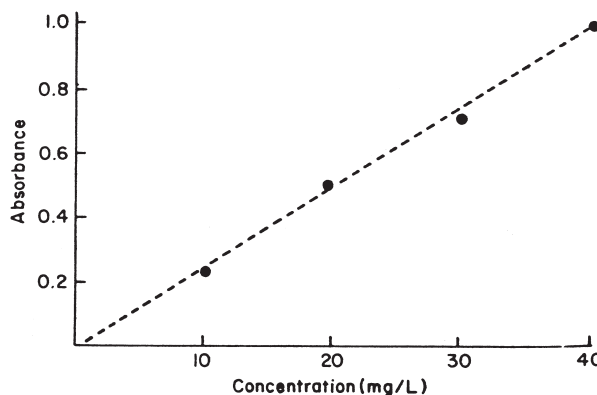


Figure 12-16. Beer's law plot.

Since t is less than the t value in the tables with 2 DF at the 5% level (see Table 12-10), it is concluded that the observed slope is not significantly different from 0.025. Note that a relatively large difference from 0.025 would be necessary to obtain significance because of the few DF in this test (the test is not very powerful). To increase the DF, more observations would be needed.

A confidence interval for the slope can be constructed in a manner similar to that described for means. A 95% confidence interval is

$$\begin{aligned} b \pm t \sqrt{s_b^2} &= 0.02429 \pm 4.30 \sqrt{0.000208/500} \\ &= 0.00243 \pm 0.0028 = 0.0215 \text{ to } 0.0271 \end{aligned}$$

Confidence Limits and Test of the Intercept—Tests for the intercept and confidence limits are analogous to those presented immediately above for the slope. The variance estimate of the intercept is

$$S_a^2 = s_{y,x}^2 [(1/n + \bar{x}^2 / \sum (x_i - \bar{x})^2)]$$

In Example 32, the calibration curve, a reasonable test would be to compare the intercept to zero. That is, one discovers whether zero concentration could correspond to a reading of zero. This would be a reasonable assumption if no interfering substances are present and if the optical density versus concentration relationship is a straight line from 0 to the highest concentration tested.

$$t = \frac{|-0.002 - 0|}{\sqrt{0.00028(1/4 + 625/500)}} = 0.1132$$

Since 0.1132 is less than the tabled value at the 5% level (see Table 12-10) with 2 DF, it is concluded that the intercept is not significantly different from zero.

A 95% confidence interval for the intercept is

$$-0.002 \pm 4.3 \sqrt{0.00028(1/4 + 625/500)} = -0.002 \pm 0.082$$

These ideas as applied to analytical data are discussed in some detail by Youden.²⁶

FITTING A LINE WITH AN INTERCEPT OF ZERO—In some situations, it is desirable to force the LS line to have a y intercept equal to zero.

Example 33—In the Beer's Law line in Example 32, if it is known that there are no interfering substances and that the relationship is linear throughout the region of concentration being tested, the assumption that the line must pass through the origin is valid. The slope of this line is calculated as

$$\begin{aligned} b &= \sum x_i y_i / \sum x_i^2 \\ &= 72.67 / 3000 = 0.02422 \end{aligned}$$

The slope of the line with 0 intercept is very close to that obtained above where the intercept was computed with no constraints on the value of the intercept.

CONFIDENCE INTERVAL FOR Y AND X—Many situations arise where a confidence interval for y at some specified x is of interest.

Example 34—The data from Figure 12-17 show the results of a kinetic stability study, where drug content in tablets is measured as a function of time. The labeled content is 100 mg. The LS line was

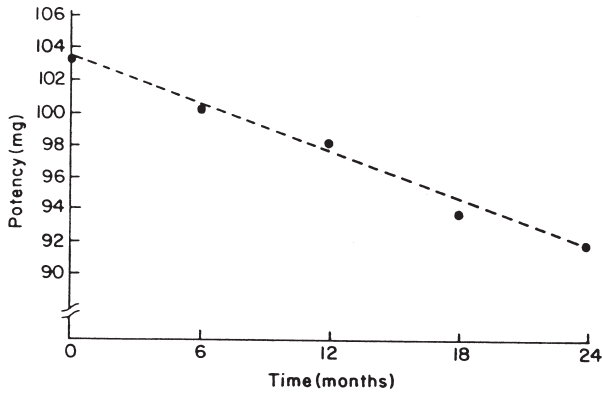


Figure 12-17. Tablet stability study.

calculated as $p = 103.3 - 0.483t$, where p is tablet potency and t is time. Note that the intercept is greater than 100 mg because a slight overage is built into the manufacturing process. The variance estimate, $S^2_{y,x}$, with 3 DF is equal to 0.367. In such stability studies, it often is of interest to predict the time for drug potency to reach 90% of the labeled amount in order to estimate shelf life or an expiration period. Substituting 90 for p (potency) and solving for time,

$$t = \frac{103.3 - 90}{0.483} = 27.54 \text{ months}$$

Therefore, the best estimate of the time to 90% potency is 27.54 months.

When establishing an expiration date, a conservative approach would take into account the error in the estimated values. A two-sided confidence interval can be constructed for the true value of y at a given x using

$$y \pm t \sqrt{S^2_{y,x} [1/n + (x - \bar{x})^2 / \sum(x_i - \bar{x})^2]}$$

where y is a point on the LS line. The value of t (3 DF) for a two-sided 95% interval is 3.182. The width of the confidence interval depends on the value of x , being minimal when $x = \bar{x}$. The value of y when $x = \bar{x}$ is

$$y = 103.3 - 0.483(12) = 97.5$$

The 95% confidence interval when $x = \bar{x} = 12$ is

$$97.5 \pm 3.18 \sqrt{0.367[1/5 + 0/360]} = 96.64 - 98.36$$

Exercise 6—Calculate the 95% confidence interval for potency when $t = 24$ months.

Answer: 90.21 to 93.19.

Figure 12-18 shows 95% confidence intervals (confidence band) for the line calculated from the data of Figure 12-17. Note the hyperbolic shape, the interval being smallest at \bar{x} and wider as x deviates more from its mean value. Using the lower line of the

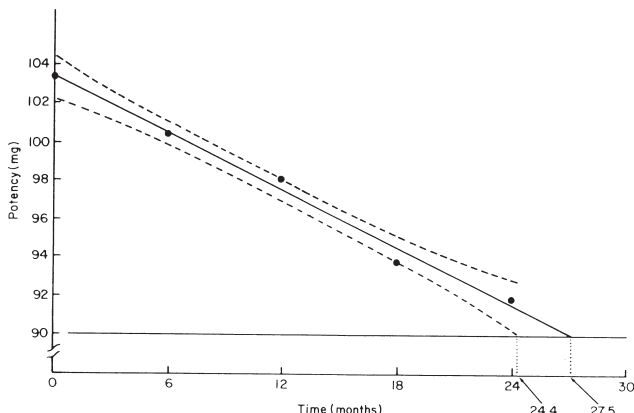


Figure 12-18. Two-sided confidence interval for stability data.

confidence interval to compute the time to 90% potency yields a conservative estimate. In the example in Figure 12-18, a reasonable estimate of the expiration date would be 24.4 months. A one-sided interval (below the line) has been proposed as being more appropriate for stability data, as one usually is concerned with the loss of potency. For a 95% one-sided confidence interval, for 3 DF, the value of t from Table 12-10 is 2.353. Using this value of t to calculate the one-sided confidence interval, the one-sided confidence band shown in Figure 12-19 is obtained. For example, when $x = 12$ (months), the lower limit has the value 96.86 months.

With this approach, the expiration date would be set at 25.1 months (see Fig 12-19).

A confidence interval can be computed for x at a given value of y . This is sometimes known as inverse estimation. In the stability example, interest would be in computing a confidence interval for the time at which 90% of the potency remains. This time was estimated as 27.5 months. The formula for the confidence interval for x is more complex than that for y , but the computations are relatively simple.

$$\frac{(x - c^2\bar{x}) \pm t[s_{y,x}/b]\sqrt{(1 - c^2)/n + (x - \bar{x})^2/\sum(x_i - \bar{x})^2}}{1 - c^2}$$

where

$$c^2 = [t \cdot s]^2 / [b^2 \sum(x_i - \bar{x})^2]$$

Exercise 7—Use the above formula to show that a one-sided lower interval for x (time); 90% potency is 25.1.

Answer: This answer corresponds to the value of time taken from Figure 12-19.

COMPARISON OF THE SLOPES OF TWO LINES—A statistical test may be performed to compare the slopes of two lines, using a t test. The null hypothesis is

$$H_0: B_1 = B_2 \text{ or } B_1 - B_2 = 0$$

The t test compares the difference of the two slopes to the standard error of the difference. The variances of y for the two lines are assumed to be equal, and the estimates are pooled as in the two-sample t test

$$s^2 \text{ pooled} = \frac{s^2_{y,x}(n_1 - 2) + s^2_{y,x}(n_2 - 2)}{(n_1 + n_2 - 4)}$$

A two-sided t test with $(n_1 + n_2 - 4)$ DF is

$$t = \frac{|b_2 - b_1|}{\sqrt{s^2 \text{ pooled}(1/x_1^2 + 1/x_2^2)}}$$

where, x_1^2 and x_2^2 are $\sum(x_i - \bar{x})^2_1$ and $\sum(x_i - \bar{x})^2_2$, respectively.

Example 35—The line for the stability data depicted in Figure 12-17 has a slope of -0.483 , with a variance estimate of 0.367 with 3 DF.

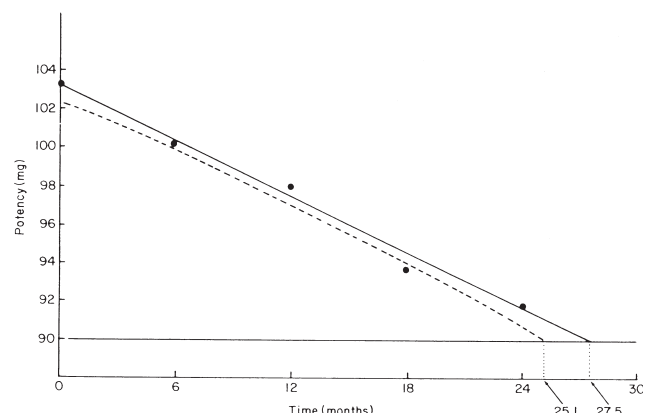


Figure 12-19. One-sided confidence interval for stability study.

The value of $\sum(x_i - \bar{x})^2$ is 360. Another formulation was prepared and tested for stability. Ten sampling times were used for the stability study, and the slope was determined to be -0.533 . The variance estimate (8 DF) was 0.289, and $\sum(x_i - \bar{x})^2$ was equal to 2565. The test for equality of the slopes (rate of decomposition) is

$$t = \frac{|0.533 - 0.483|}{\sqrt{0.310(1/2565 + 1/500)}} = 1.84$$

Since 1.84 is less than the tabled t value for the 5% significance level with 11 DF, 2.20 (3 from one line and 8 from the other), it can be concluded that the slopes of the two lines are not significantly different.

In a biological assay, a common procedure is to determine the relative potency of two or more substances using the *parallel line assay*. In this procedure, the lines from a plot of response versus log dose are forced to be parallel and the distance between the lines is a measure of the relative potency. Before performing this procedure, a test is made to ensure that the lines are parallel. Nonparallel lines will cross, suggesting that at low doses one product gives a greater response, whereas at higher doses the other product gives the greater response. Figure 12-20²³ illustrates the principle of this assay. The computations are tedious, and the book by Finney²⁷ should be consulted for those who wish more detail on the statistical treatment of this and other biological assay methods.

CORRELATION—Correlation is related to, but should not be confused with, linear regression. It is a measure of the linear relationship between two variables but does not prove linearity. In fact, the usual formulas for determining the significance of the correlation assume that the variables already are related linearly. The question that is usually posed indirectly when testing the correlation is, can the value of one of the variables be used to predict the value of the second variable? This amounts to testing the slope of the line relating the variables versus 0. If the slope is significantly different from 0, then the variables have a *significant correlation*. Correlation is used when both variables are subject to error. If one variable is not subject to error (fixed), the linear regression approach to establish the relationship of the variables is more appropriate.

The measure of association is the correlation coefficient, r .

$$r = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The correlation coefficient can vary between +1 and -1 . A correlation coefficient of +1 would result if all points fall exactly on a single line with positive slope; this is a perfect positive correlation. Similarly, if all points fall on a line with negative slope, $r = -1$, a perfect negative correlation is observed. If $r = 0$, the variables are not correlated. These three cases are shown in Figure 12-21.

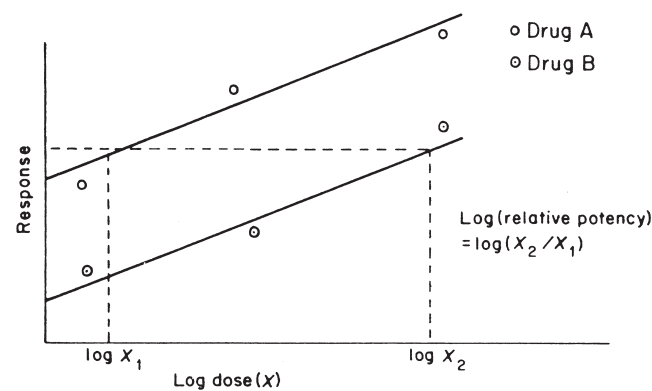


Figure 12-20. Relative potency estimate using a parallel line assay. Doses shown for $\log x_1$ and $\log x_2$ give the same response for Products A and B, respectively. (From Bolton S. *Pharmaceutical Statistics*. New York: Marcel Dekker, 1984, pp 416, 463.)

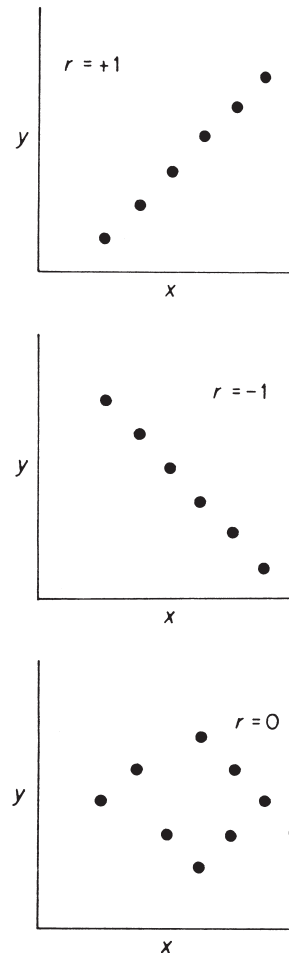


Figure 12-21. Correlation diagrams (scatter plots).

In real situations, these extreme results are seen rarely, but rather some intermediate value of r is observed. The statistical question of interest usually is concerned with the significance of the correlation—a test of r versus 0. One should appreciate, however, that the meaning of the correlation should be considered carefully. For example, if n , the number of data pairs, is large, correlation coefficients that are very small (practically insignificant) will be deemed statistically significant. Also, data that is not linear, but clearly related, may show small correlation coefficients.

Exercise 8—Compute the correlation between x and y for $x = -2, -1, 0, +1$ and $+2$, for the relationship $y = x^2$. *Answer:* $r = 0$.

The test of the correlation coefficient versus 0 is

$$t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}} \quad (\text{DF} = n - 2)$$

Example 36—An experiment was performed to examine the relationship of tablet hardness to tablet dissolution. Dissolution was measured as the time (minutes) for 50% of the drug to dissolve in the USP Dissolution Test. Hardness was measured in kilograms. The following results were obtained for 12 tablets:

Hardness:	6.8	5.3	5.8	7.2	6.9	6.0	6.8	8.1	7.5	6.3
Dissolution:	18	17	21	26	28	20	25	29	31	18

These data are plotted in Figure 12-22, known as a *scatter plot*. This plot suggests a trend toward slower dissolution as hardness increases. In this example, r is equal to

$$\frac{1585.5 - 233(66.7/10)}{\sqrt{236.1 \cdot 6.321}} = 0.81$$

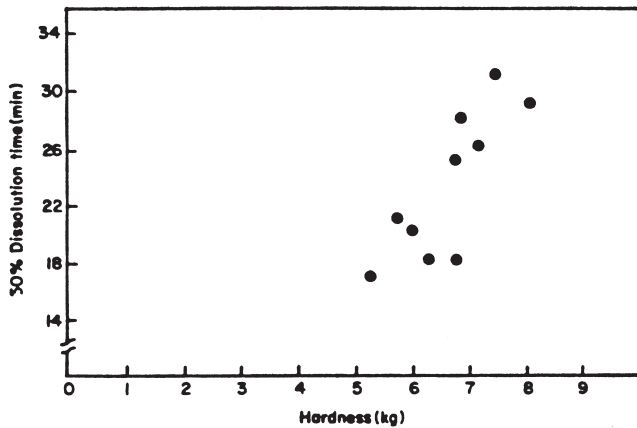


Figure 12-22. Scatter plot of hardness versus dissolution data (Example 36).

The test of the significance of the correlation coefficient shows $t = 3.94$ with 8 DF.

It is concluded that r is significantly different from 0, and hardness and dissolution are correlated ($P < 0.05$; see Table 12-10; $t = 2.228$ for significance at $P = 0.05$).

DATA TRANSFORMATIONS

Probabilities calculated from statistical analyses are based on assumptions underlying the nature of the data. The typical analyses presented in this chapter often assume normality of data and variance homogeneity. When dealing with means of a sufficiently large sample size, the assumption of normality is not critical. However, small sample sizes and a large deviation from normality can result in a significant violation of the normality assumption. When comparing samples from two or more groups, lack of homogeneity of variance (heteroscedasticity) is an important problem that can result in an unreliable analysis. One way of overcoming these problems is the use of *transformations*. Each data point is transformed, resulting in data that more closely fits the normality and variance homogeneity assumptions.⁹

The logarithmic, square root, and arcsine transformations will be presented here as examples of the more popular data transformations.

LOGARITHMIC (LOG) TRANSFORMATION—This transformation (log to the base 10 or log to the base e , \ln , may be used) is most applicable for skewed data of the form illustrated in Figure 12-23. These data typically show a relatively constant coefficient of variation (CV). That is, the larger the

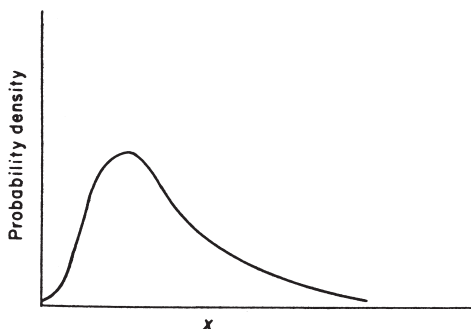


Figure 12-23. Example of a skewed distribution.

value, the larger is the SD; the SD is proportional to the mean (SD/\bar{x} is constant). This transformation is applicable to data that meet the above conditions and also are greater than 0; the log of 0 or a negative number is undefined.

This probably is the most common transformation for data in the pharmaceutical sciences. Many physical and biological measurements show larger variability as the size of the measurement increases. This is logical for many types of data. For example, the measurement of a large value such as the assay of a concentrated solution may be expected to show considerable variability about its mean (eg, 1000 mg/mL \pm 50, a 5% variability). The assay of a dilute solution, 10 mg/mL, cannot show very large variation, particularly on the low side where zero (0) is the lower limit. If the CV were 5% (10 mg/mL \pm 0.5), a log transformation would be suitable. If the data are skewed (see Fig 12-23) and the CV is constant, a log transformation will tend to normalize the data distribution and equalize the variances.

When data are presented as ratios, a log transformation often is appropriate. Unless the data are extremely variable, the conclusions using the original or transformed data should be similar. The conclusions using the transformed data, however, will be more reliable if the transformation is appropriate.

Care should be taken that the log transformation does not help to improve one assumption while making another less valid. For data that are skewed but have constant variance, the normality assumption may improve while causing problems with the variance homogeneity assumption. Fortunately, the log transformation, when indicated, does not seem to cause such difficult and perplexing problems.

Example 37—The means of two treatment groups are to be compared where it is known that large values are associated with proportionally larger standard deviations. The measurements are 50% dissolution time in minutes (Table 12-42).

A two independent sample t test (two-sided) comparing the means shows

$$t = \frac{|43.17 - 61.17|}{19.11 \sqrt{1/3}} = 1.75$$

A log transformation results in the data in Table 12-43.

Neither test is significant at the 5% level but the log-transformed values in this example result in a test with a lower probability level.

Exercise 9—A bioequivalence study comparing two dosage forms, A and B, with six subjects in a paired design resulted in the following ratios of AUC_a/AUC_b :

1.27, 1.06, 0.90, 1.30, 1.15, 0.96

Calculate the mean, standard deviations and a 95% confidence interval for the data using the untransformed data and a log transformation. For the log transformation, calculate the antilogs for the lower and upper limit of the confidence interval. Repeat the calculations for the mean and standard deviation of the ratio of AUC_b/AUC_a . (Note that this is the reciprocal of the data presented above.) What can be said about the confidence intervals for the two kinds of ratios, A/B and B/A ?

Answer:

Mean = 1.107; SD = 0.1627; CI = 1.107 \pm 0.171. log transformation: Mean = 0.0400; SD = 0.0646; CI = 0.04 \pm 0.0678

Table 12-42. 50% Dissolution Time for Two Formulations

	FORMULATION A	FORMULATION B
	27	65
	55	60
	33	98
	69	47
	36	57
	39	43
Mean	43.17	61.67
SD	15.75	19.59

Table 12-43. Log Transformation of Data of Table 12-42

	FORMULATION A	FORMULATION B
	1.431	1.813
	1.740	1.778
	1.519	1.991
	1.839	1.672
	1.556	1.756
	1.591	1.633
Mean	1.613	1.774
SD	0.150	0.126
$t = \frac{1.613 - 1.774}{0.139 \sqrt{1/3}} = 2.01$		

CI = 0.938 – 1.282

Reciprocal: Mean = 0.92; SD = 0.1376

log transformation: Mean = –0.0400; SD = 0.0646

For bioequivalence data, a log transformation of AUC and C_{max} is currently recommended.

SQUARE-ROOT TRANSFORMATION—A square-root transformation is useful for data where the sample means are proportional or equal to the variances (s^2). The transformation will cause the data to have approximately homogeneous variance. This transformation may be used to replace the log transformation when the data consist of small numbers. If the numbers are less than 10 and zeros are present, $\sqrt{x+1}$ may be an appropriate transformation.²⁸ This transformation, like the log transformation, will tend to normalize distributions skewed to the right (distributions with a relative small number of very large values).

Exercise 10—Compute the mean and standard deviation of the following data before and after applying the square-root transformation (\sqrt{x}). Draw a histogram of the original and transformed values. Note the greater symmetry of the transformed data.

0, 11, 7, 3, 0, 15, 4, 2, 6, 9, 3, 0, 12, 5, 3, 6

Answer:

Original data: $\bar{x} = 5.625$ SD = 4.272
 Transformed data: $\bar{x} = 2.00$ SD = 1.206

ARCSINE (INVERSE SINE) TRANSFORMATION—The arcsine (inverse sine) transformation is used for binomial data or data expressed as percentages or proportions. The transformation is arcsine \sqrt{p} , where p , the proportion or probability, is expressed as a decimal. The variance of a binomial proportion is pq/n , where p is the proportion of successes and q the proportion of failures in n binomial observations. If p varies in different treatment groups, the variance will vary. The arcsine transformation applied to the proportions tends to equalize the variances and normalize the data.⁹ The variance of the transformed proportion is $821/n$ when the transformed data are in degrees. This transformation assumes that all proportions transformed have the same value of n . If n is approximately equal for the different groups, the transformation may still be used.

Example 38—Use a normal test to compare the proportion of rats who developed tumors in placebo-control and active-drug groups. In the placebo group, 15 of 100 animals developed tumors, whereas in the drug group, 22 of 100 developed tumors. The arcsines of $\sqrt{0.15}$ and $\sqrt{0.22}$ are 22.786 and 27.972, respectively. The normal test is

$$Z = \frac{|27.972 - 22.786|}{\sqrt{(821/100) + (821/100)}} = 1.28$$

The proportions are not significantly different.

Exercise 11—Calculate the value of chi-square for the test of these two proportions.

Answer: 1.625. Note that chi-square = Z^2 in this example.

REFERENCES

1. Siegal S. *Nonparametric Statistics*. New York: McGraw-Hill, 1956.
2. Kish L. *Survey Sampling*. New York: Wiley, 1995.
3. Fisher RA, Yates F. *Statistical Tables for Biological, Agriculture and Medical Research*. New York: Hafner, 1963, p 134 (Table 38).
4. Fisher RA. *The Design of Experiments*, 5th ed. Edinburgh: Oliver & Boyd, 1986.
5. Cochran WG, Cox GM. *Experimental Design*, 2nd ed. New York: Wiley, 1957.
6. Cox DR. *Planning of Experiments*. New York: Wiley, 1958.
7. United States Department of the Army. *Tables of the Binomial Probability Distribution*. Applied Mathematics Series No. 6 Washington, DC: USGPO, 1952.
8. Bolton S. *Pharmaceutical Statistics*, 3rd ed. New York: Marcel Dekker, 1997.
9. Dixon WJ, Massey FJ Jr. *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw-Hill, 1969, p 324.
10. Snedecor GW, Cochran WG. *Statistical Methods*, 8th ed. Ames: Iowa State University Press, 1989, p 97.
11. Snedecor GW, Cochran WG. *Statistical Methods*, 7th ed. Ames: Iowa State University Press, 1980, p 476 (Table A14).
12. Snedecor GW, Cochran WG. *Statistical Methods*, 7th ed. Ames: Iowa State University Press, 1980, p 480 (Table A15).
13. Duncan DB. *Biometrics II* 1948; 1.
14. Dunnett CW. *Am Stat Assoc J* 1955; 50: 1096.
15. Winer BJ. *Statistical Principles in Experimental Design*, 2nd ed. New York: McGraw-Hill, 1971.
16. Tate MW, Clelland RC. *Nonparametric and Shortcut Statistics*. Danville IL: Interstate Print, 1957: p 137 (Table L).
17. Youden WJ. *Sci Monthly* 1953; 77: 143.
18. Youden WJ. *Natl Bur Std (US) Tech News Bull* 1949: 33 (July).
19. Dixon WJ, Massey FJ Jr. *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw-Hill, 1969, p 328.
20. *Control Chart Method of Controlling Quality During Production (Std Z1.3)*. New York: American Standards Association, 1958.
21. *Guide for Quality Control (Std Z1. 3)*. New York: American Standards Association, 1958.
22. Dixon WJ, Massey FJ Jr. *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw-Hill, 1969, p 142.
23. Bolton S. *Pharmaceutical Statistics*. New York: Marcel Dekker, 1984, pp 416, 463.
24. *Military Sampling Procedures and Tables for Inspection by Attributes*. MIL-STD-105E. Washington DC: USGPO, 1989.
25. Snedecor GW, Cochran WG. *Statistical Methods*, 8th ed. Ames: Iowa State University Press, 1989, p 151.
26. Youden WJ. *Statistical Methods for Chemists*. New York: Wiley, 1951.
27. Finney DJ. *Statistical Method in Biological Assay*, 4th ed. New York: Hafner, 1980.
28. Steel RGD, Torrie JH. *Principles and Procedures of Statistics*. New York: McGraw-Hill, 1960.

BIBLIOGRAPHY

Experimental Design

- Cox DR. *Planning of Experiments*. New York: Wiley, 1992.
- Fisher RA. *Statistical Methods for Research Workers*, 13th ed. Edinburgh: Oliver & Boyd, 1970.
- Montgomery DC. *Design and Analysis of Experiments*, 4th ed. New York: Wiley, 1996.
- Chow S-C, Liu J-P. *Statistical Design and Analysis in Pharmaceutical Science*. New York: Dekker, 1995.

Statistical Quality Control

- Grant EL. *Statistical Quality Control*, 5th ed. New York: McGraw-Hill, 1980.
- Mandel J. *The Statistical Analysis of Experimental Data*. New York: Dover, 1984.
- Montgomery DC. *Introduction to Statistical Quality Control*, 3rd ed. New York: Wiley, 1996.
- Weber RT. *An Easy Approach to Acceptance Sampling: How to Use MIL-STD-105E*. Milwaukee, WI: American Society for Quality Control, 1991.

Sampling

- Cochran WG. *Sampling Techniques*, 3rd ed. New York: Wiley, 1997.
- Deming WE. *Some Theory of Sampling*. New York: Wiley, 1984.

Yates F. *Sampling Methods for Censuses and Surveys*. New York: Hafner, 1981.

Biological Assay

Bliss CI. *The Statistics of Bioassay with Special Reference to the Vitamins*. New York: Academic Press, 1952.

Bliss CI. *Am Sci* 1957; 45: 449.

Finney DJ. *Statistical Method in Biological Assay*, 3rd ed. New York: Hafner, 1978.

Finney DJ. *Probit Analysis*, 4th ed. London, Cambridge University Press, 1980.

General

Bennett CA, Franklin NL. *Statistical Analysis in Chemistry and the Chemical Industry*. New York: Wiley, 1954.

Bolton S. *Pharmaceutical Statistics*. Third Edition, New York: Marcel Dekker, 1997

Brownlee KA. *Statistical Theory and Methods in Science and Engineering*. New York: Wiley, 1960.

Buncher CR, Tsay J. *Statistics in the Pharmaceutical Industry*. New York: Marcel Dekker, 1981.

Chow, Shein-Chung, Editor, *Encyclopedia of Biopharmaceutical Statistics*: New York: Marcel Dekker, 2000.

Davies OL. *The Design and Analysis of Industrial Experiments*. New York: Hafner, 1954.

Peace KE. *Biopharmaceutical Statistics for Drug Development*. New York: Marcel Dekker, 1988.

Snedecor GW, Cochran WG. *Statistical Methods*, 8th ed. Ames, Iowa State University Press, 1989.

Statistical Software Packages (Examples)

BMDP, Biomedical Computer Programs, University of California, Los Angeles, CA.

NCSS, NCSS Statistical Software, 329 North 1000 East, Kaysville, Utah 84037 (website: <http://www.ncss.com>, phone: 800-898-6109).

SAS, SAS Institute Inc, Cary, NC (website: <http://www.sas.com/>, e-mail: software@sas.com)

Molecular Structure, Properties, and States of Matter

Eric J Lien, PhD



The many significant advances in the pharmaceutical sciences in recent years are in large part attributable to the accumulation of knowledge of the molecular structure and physicochemical properties of drugs, and to the correlation of this knowledge with that of the nature of biological reactions of drugs. This chapter discusses fundamental principles of atomic and molecular structure and certain physicochemical properties that are important in the pharmaceutical sciences, to aid in the understanding of drug action at the molecular level.

ATOMIC STRUCTURE

ATOMS AND ELEMENTARY PARTICLES—The atoms (from the Greek *atomos*, indivisible) were believed to be the minute, indivisible particles of which all material things were made. The search for the ultimate particle has been a continuous effort since the time of Democritus (about 460–370 BCE). Before the discovery of mesons and hyperons, the structure of matter was believed to be much simpler. The nucleus was thought to consist of protons and neutrons; and to form an atom, only electrons needed to be added in external shells. Therefore, protons, neutrons, and electrons were considered as the elementary particles. In theory, all the elements in the periodic table can be made by splitting neutrons into electrons and protons, and by combining these particles in proper ratios.

During the past three decades, nuclear physics progressively has probed atoms from their periphery to their center. The search for ultimate units of nuclear structure, by means of experiments consisting in large part of bombarding nuclei with high-energy particles, has revealed a spectrum of over 100 species, most of them unstable. Some of these particles are listed in Table 13-1. The proton is no longer considered an ultimate particle, but is believed to be made up of particles called *quarks* (from “three quarks for Muster Mark,” in James Joyce’s *Finnegans Wake*). One theory of quark structure of protons calls for nine kinds of quarks (along with antiquarks) and eight kinds of *gluons* (analogous to photons) to hold the quarks together. Whether these and other elementary particles are all composed of yet simpler elements remains to be investigated.¹

In 1924, de Broglie raised the question that if light waves show corpuscular character, should not particles also show wave character? Now it generally is accepted that in the case of a photon there are two fundamental equations to be obeyed: $E = h\nu$ and $E = mc^2$, where E is the energy, h is Planck’s constant, ν is the frequency and c is the speed of light. Combining both equations gives $h\nu = mc^2$ or $\lambda = c/\nu = h/mc = h/p$, where p is the momentum of the photon.

De Broglie proposed that a similar equation should govern the wavelength of the electron wave. It is interesting to note that x-ray diffraction is a good example of the use of the wave property of electromagnetic radiation.

Scattering of slow neutrons has been employed to provide information about the structure and dynamic properties of biological structures, for example, myoglobin and membranes.²

DALTON’S ATOMIC THEORY—In 1808, Dalton proposed his atomic theory on the basis of three generalizations: the Law of Conservation of Mass, the Law of Definite Proportions, and the Law of Multiple Proportions. The essential parts of the theory can be summarized as

1. All elements are composed of very small, discrete, indivisible particles called atoms.
2. All atoms of any one element are identical. Modern structural theory tells us that electronic differences between the atoms of an element may occur, but these differences arise as a consequence of electronic excitation. The lowest energy state of an atom is more appropriate for purposes of classification.
3. The atoms of no two elements are alike.
4. Atoms undergo no fundamental change during chemical reaction. There are subtle changes in the electronic character of atoms, although this does not change the identity of an atom.
5. Compounds are formed when atoms of two or more different elements combine to form a molecule.
6. In general, atoms combine in simple, integral ratios.

PERIODIC TABLE—The periodic classification of the elements is one of the most striking advances in generalizing many isolated facts; moreover, it contributes tremendously to the strength of the atomic theory and extends it to new sets of facts. The periodic table serves as an easily learned summary of almost limitless information about the chemical nature of the elements; it is of prime importance to students of pharmaceutical sciences as well as to students of chemistry.

After the publication of the independent researches of Mendeleev and Meyer in 1869, the *periodic law* was well-established. The *periodic table* is an arrangement of the elements in accordance with the periodic law (see Periodic Chart of the Elements). The present arrangement is essentially the same as that of Mendeleev, although there are now minor variations due to the incorporation of new elements and modern data. A few terms should be carried in mind for a thorough understanding of the table.

- *Atomic number (Z)* is the positive charge of the nucleus expressed as multiples of the electronic charge e .
- *Atomic weight* is the average weight expressed in atomic weight units of the natural atoms of an element existing as a mixture of isotopes in the same ratio as found in nature. An atomic weight unit, used in chemistry, is exactly 1/16 the average mass of the oxygen isotopes taken in the same ratio as they occur in nature. One atomic weight unit is equivalent to 1.000272 atomic mass units.
- An *isotope* is one of a group of nuclides of the same element (same Z), having the same number of protons in the nucleus but differing in the number of neutrons, resulting in different mass numbers.
- A *nuclide* is any one of the more than 1000 species of atoms and is characterized by the number of protons and neutrons in the nucleus.

Table 13-1. Subatomic Particles

GROUP	PARTICLES	RELATIVE MASS (ELECTRON = 1)	ELECTRIC CHARGE	MEAN LIFE-TIME (sec)
Heavy particles	α -Particle (He^{2+} , α)	7348	+2	Stable
	Triton (T , ${}^3\text{H}$)	5451	+1	3.8×10^8
	Deuteron (D , d , ${}^2\text{H}$)	3674	+1	Stable
	Neutron (n)	1837	0	7.2×10^2
Hyperons	Proton (p , ${}^1\text{H}$)	1837	+1	Stable
	Λ^0 Particle	~ 2181	0	2.5×10^{-10} $\Sigma^+ 0.8 \times 10^{-10}$
	Σ^\pm Particle	~ 2326	± 1	$\Sigma^- 1.6 \times 10^{-10}$
Mesons	Ξ^\pm Particle	~ 2580	± 1	1.3×10^{-10}
	K meson (K^\pm)	966	± 1	1.2×10^{-8}
	K meson (K^0)	974	0	$10^{-9} - 10^{-10}$
	Pi meson (π^\pm)	273		2.6×10^{-8}
	Pi meson (π^0)	264	0	1.9×10^{-16}
	Mu (μ^\pm)	209 ± 2	± 1	2.2×10^{-6}
Leptons	Electrons (e^- , β^-)	1	-1	Stable
	Positron (e^+ , β^+)	1	+1	Stable
	Neutrino (ν)	0.01	0	Stable
	Photons (γ)	0	0	Stable

BOHR'S THEORY OF ATOMIC STRUCTURE—In 1913 Bohr proposed a theory of atomic structure for the interpretation of atomic spectra. His picture of the atom had the extranuclear electrons revolving around the nucleus in definite orbits. These orbits were assigned principal quantum numbers 1, 2, 3, . . . , n , counting outward from the nucleus.

When an electron absorbs a definite increment (quantum) of energy, it is promoted to an orbit of higher energy (excited state), and when it falls back to the original orbit, it emits radiation energy. The energy of the various levels in the atom can be related to the frequency of radiation that is emitted from or absorbed by the atom. This relationship is expressed by

$$\Delta E = E_2 - E_1 = h\nu \quad (1)$$

where ΔE is the difference of the energy in ergs between two levels, h is Planck's constant (6.624×10^{-27} erg sec) and ν is the frequency. Because the frequency is equivalent to the speed of light, c , divided by the wavelength, Equation 1 can be written as

$$\Delta E = hc/\lambda \quad (2)$$

When the electrons possess the lowest energy possible, the atom is said to be in its *ground state*.

The energy of an electron in an orbit is given by

$$E = \frac{-2\pi^2 Z^2 m e^4}{n^2 h^2} \quad (3)$$

where Z is the atomic number, m is the mass of the electron (9.1×10^{-28} g), e is the charge of the electron in electrostatic units (4.8×10^{-10} esu), n is the principal quantum number, and h is Planck's constant. One can calculate the radiation energy emitted when an electron falls from n_2 orbit to n_1 orbit by

$$E_2 - E_1 = \frac{2\pi^2 Z^2 m e^4}{h^2} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \quad (4)$$

When n_2 is ∞ , Equation 4 gives the energy required for ionization; for example, the ionization potential of the hydrogen atom

can be calculated as

$$E_\infty - E_1 = \frac{2 \times (3.14)^2 \times (1)^2 \times 9.1 \times 10^{-28} \times (4.8 \times 10^{-10})^4}{(6.624 \times 10^{-27})^2} \\ \times \left(\frac{1}{(1)^2} - \frac{1}{(\infty)^2} \right) \\ = 2.18 \times 10^{-11} \text{ erg} \\ = \frac{2.18 \times 10^{-11} \text{ erg}}{1.60 \times 10^{-12} \text{ erg/electron volt (ev)}} \\ = 13.6 \text{ ev}$$

It is interesting to note that the quantum theory is founded on the principle that the energy of an atom or molecule does not change continuously but only by some definite whole number unit of energy referred to as a quantum.

MODERN MODEL OF ATOMIC STRUCTURE—After Bohr published his theory, there was a period of intense activity by theoreticians and experimental physicists. Based on mathematical principles and considerable experimental data, a more definite picture of atomic structure emerged. The modern interpretation of the atom is more elaborate than the original idea of Bohr. Four quantum numbers are used to describe the energy levels or orbitals of each electron.

The *principal quantum number*, n , is an approximate measure of the size of the electron cloud—that is, the order of magnitude of the potential energy. It has the values 1, 2, 3, . . . , 7, corresponding to the K, L, M, . . . , Q shells of electrons.

The *azimuthal quantum number*, l , is related to the shape of the electron cloud, indicating whether it is spherical, dumbbell-shaped, or of more complex geometry. It may have values of 0, 1, 2, . . . , $(n - 1)$, corresponding, respectively, to the terms s , p , d , or f used by spectroscopists; for example, a $4d$ electron would have an n number of 4 and an l value of 2.

The *magnetic quantum number*, m_l , is related to the orientation of the electron cloud in space. It has values of 0, ± 1 , ± 2 , . . . , $\pm l$. For a spherical cloud there is only one orientation. However, the dumbbell-shaped orbital, for example, could be oriented in three different directions corresponding to the x , y , and z axes of a set of Cartesian coordinates.

The *spin quantum number*, s (or m_s), gives the orientation of the magnetic component of an electron. There are only two discrete ways an electron can interact with an external magnetic field. Like a tiny magnet, it either can line up in the direction of the field or orient itself in the opposite direction. The electron's magnetic moment was at first pictured as being due to the rotation of the electron on its axis, and for this reason an electron was said to exhibit spin. The two spin quantum numbers, $s = +1/2$ and $s = -1/2$, were used to describe the two observable spin states.

Considerable progress has been made in the recent years in the application of quantum mechanical and molecular orbital theories in studying drug-receptor interactions and in correlating chemical structure with pharmacological activities of drugs (see Chapter 27).

ELECTRONIC CONFIGURATION OF THE ELEMENTS—Two rules are of extreme importance in explaining the building up of electronic shells of elements (Fig 13-1 and Table 13-2).

The *Pauli exclusion principle* states that an atom cannot exist in a state where two electrons in the same energy level or orbital have the same set of four quantum numbers. This is analogous to the principle in classical physics that no two bodies can be in the same place at the same time. Thus, two electrons in the K shell may have the same principal, azimuthal, and magnetic quantum numbers ($n = 1$, $l = 0$, $m_l = 0$), but different spin quantum numbers ($s = +1/2$ and $-1/2$).

Hund's rule of maximum multiplicity states that when orbitals are of the same energy, electrons distribute themselves one to each orbital so as to maintain parallel spins; for example, oxygen, with an atomic number of eight, possesses eight elec-

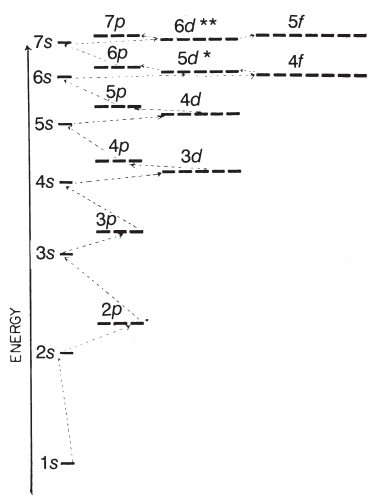


Figure 13-1. Atomic energy levels and the order of filling of orbitals: (*) a single 5d electron is added before the 4f orbitals can be filled; (**) one or more 6d electrons must be added before the 5f orbitals can be filled.

trons. Two electrons are in the K shell ($1s^2$), and six are in the L shell. In the L shell, two electrons fill the 2s orbitals ($2s^2$) and the remaining four fill the 2p orbitals ($2p^4$).

According to Hund's rule, three electrons occupy $2p_x$, $2p_y$, and $2p_z$ orbitals and spin in the same direction (see the direction of the arrow in Fig 13-2); the fourth electron can pair up with any one of these three electrons (say $2p_x$). The electronic configuration for oxygen atom can be expressed as $1s^2 2s^2 2p_x^2 2p_y 2p_z$.

MOLECULAR STRUCTURE

A molecule is the smallest possible quantity of a substance. It is composed of two or more atoms, for example, N_2 , O_2 , $CHCl_3$, or H_2SO_4 . There is a chemical bond between atoms when the forces acting between them are strong enough to give an aggregate with sufficient stability to make it convenient for the chemist to consider the aggregate as an independent molecular species. Different types of chemical bonds will be discussed in the following sections.

Table 13-2. Electronic Configurations of Some Elements in Their Ground States

ATOMIC NO	ELEMENT	1		2		3		4		
		K	L	M	N	O	P	Q	R	S
1	H	1								
2	He	2								
3	Li	2	1							
4	Be	2	2							
5	B	2	2	1						
6	C	2	2	2						
7	N	2	2	3						
8	O	2	2	4						
9	F	2	2	5						
10	Ne	2	2	6						
11	Na				1					
12	Mg				2					
13	Al				2	1				
14	Si		Neon core		2	2				
15	P				2	3				
16	S				2	4				
17	Cl				2	5				
18	Ar				2	6				

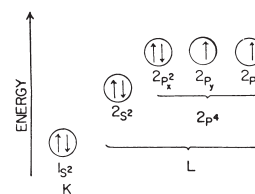


Figure 13-2. The electronic configuration of an oxygen atom.

COVALENT BONDS—When two electrons of two atoms are paired and localized in the space between the two atoms, a *covalent bond* results. The paired electrons (with opposed spins) then will occupy the new molecular orbital encompassing the two atoms. It should be noted that the electron pair held jointly by two atoms is considered to do double duty by completing a stable electronic configuration for each atom.

For instance, in the case of methane, the carbon atom, with its two inner electrons and its outer shell of eight shared electrons, has assumed the stable 10-electron configuration of neon; and the hydrogen atoms have achieved the configuration of helium. Covalent and ionic bonds are found in both organic and inorganic chemistry.

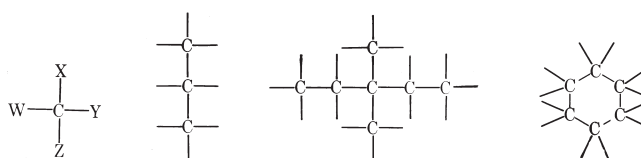
THE UNIQUENESS OF CARBON

Since organic chemistry is concerned mainly with carbon and its compounds, closer attention is warranted to the kinds of bonds exhibited by the carbon atom.

Carbon (and, to a much lesser extent, boron and beryllium) is in a special class. Although only the twelfth most abundant element on earth, its compounds far outnumber those of the remainder of the periodic table combined. The exact number of existing carbon compounds is probably unknown, and the theoretical number is infinite. This uniqueness stems from the simple fact that carbon is capable of bonding with itself in many unusual modes.

Carbon-Carbon Bonds

Ordinarily, carbon is said to exhibit a valence of four. Thus, it can combine with four other monovalent atoms or groups or with four other carbon atoms in a linear or cyclic fashion, with or without branching, or any combination thereof.



Also, carbon atoms can unite to each other or to other atoms such as nitrogen, oxygen, or sulfur by means of multiple bonds.

To compound the situation further, the structural diagrams just presented are not flat objects, but are three-dimensional. For example, a six-membered carbon ring may have several configurations, such as



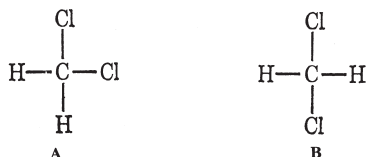
This feature alone could essentially double the number of possible compounds of this type.

HYBRIDIZATION—What is so unusual about the constitution of the carbon atom that allows so many diverse compounds? Simply stated, the reason is *hybridization*, and a review of the electronic configuration of the atom is required to explain what hybridization is and how it is attained. The extra-nuclear configuration of an isolated carbon atom is $1s^2 2s^2 2p_x^1 2p_y^1 2p_z^0$, which means that there are two electrons in the $1s$ level, two in the $2s$ level, and two in the $2p$ level, but since the two $2p$ electrons reside in different subshells (p_x and p_y), they are unpaired. As only unpaired valence electrons are capable of bonding, it would be expected that carbon should exhibit a valence of two. However, in every instance (except for possibly carbon monoxide), carbon combines with four univalent atoms or groups.

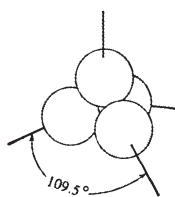
Bond formation is a stabilizing (exothermic) process, and there is a tendency to form as many bonds as possible, even if the resulting molecular orbitals bear little resemblance to the atomic orbitals that exist in the isolated or *ground* state of an atom. A carbon atom must be elevated or *excited* (energetically) to assume a valence state of four; to do this, four unpaired electrons must be created. This feat can be accomplished by promoting one electron from the $2s$ level to the vacant $2p_z$ level; thus, the resulting extranuclear electronic configuration becomes $1s^2 2s^1 2p_x^1 2p_y^1 2p_z^1$. More than enough energy is available during the process of bond formation to excite the $2s$ electron. Four unpaired electrons are now available for bonding purposes.

It might now be expected that carbon could form two different types of bonds, such as three bonds of a type using p orbitals ($2p_x, 2p_y, 2p_z$) and a fourth bond using the $2s$ orbital. But this is contrary to known fact—all four bonds are equivalent so far as bond energy and bond length are concerned.

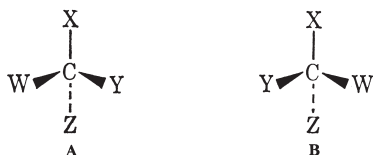
The simplest two-dimensional picture of such a carbon atom, as noted in the diagram of the molecule dichloromethane, CH_2Cl_2 , would be as in **A**.



However, it readily can be observed that if the molecule were flat, it should exist in the two isomeric forms, **A** and **B**. As only one dichloromethane is known (and for other, more convincing, reasons) the structure as depicted is spatially incorrect. In 1874, LeBel and van't Hoff demonstrated, using the concept of *stereoisomerism*, that a carbon atom assumes a *tetrahedral* configuration. That is, each covalent bond is directed to a corner of a regular tetrahedron.



To more clearly illustrate the three-dimensional aspect of this arrangement, the usual two-dimensional diagram is better shown by



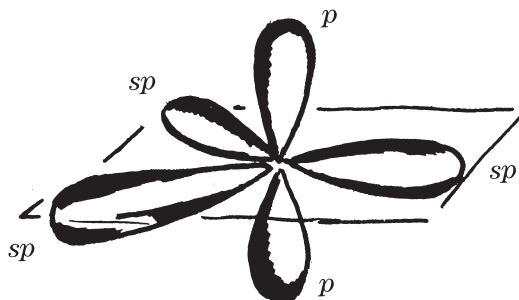
in which a solid line is understood to be in the plane of the paper, a broken line extends behind the plane, and solid arrowheads extend in front of the plane.

Study of the many kinds of three-dimensional organic models is very beneficial in the understanding of this concept. A cursory look at such models (or diagrams) indicates that **A** and **B** are not identical (not superimposable), but rather are in reality *isomers*. This situation, *stereoisomerism*, is a phenomenon that essentially doubles the number of possible compounds of this particular type.

Since the resultant bonds are comprised of one s and three p electrons, and neither are of the spherical s or linear p configuration but rather some combination thereof, they are said to be *hybridized*. This tetrahedral or sp^3 hybridization can be explained by the tendency for unshared electrons to get as far from each other as possible (the Pauli *exclusion* principle); for four bonds, the tetrahedral configuration satisfies this requirement. Covalent bonds, besides having characteristic bond length and energy, also are associated with direction in space.

Another peculiarity is associated with carbon-to-carbon bonding. In addition to the aforementioned tetrahedral, or sp^3 , hybridization, two other possibilities are known to occur in the bonding of two carbon atoms: trigonal or sp^2 , and linear (diagonal) or sp hybridization.

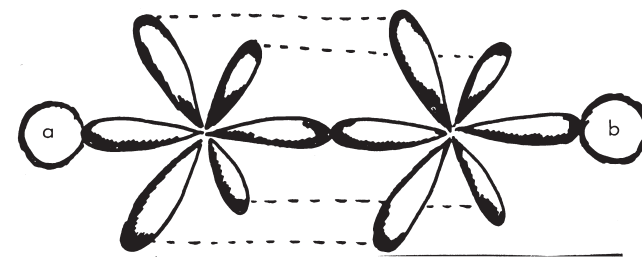
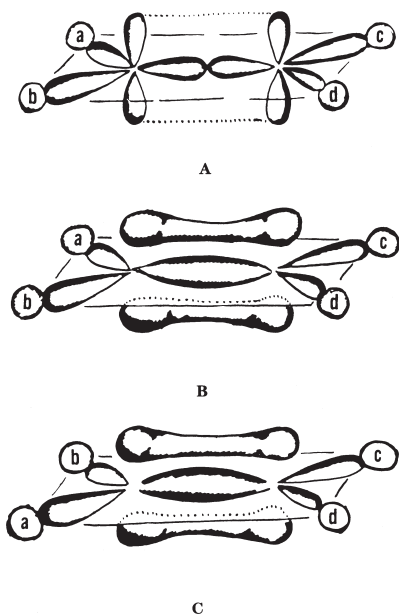
SIGMA (σ) AND PI (π) BONDS—Alkenes are examples of the sp^2 type of carbon-to-carbon bonding: the hybrid orbitals are directed toward the corners of an equilateral triangle. This permits the hybrid orbitals to be as far removed from each other as possible. An unhybridized p orbital also exists perpendicular to the plane of the sp^2 orbitals.



The union of two carbon atoms of this type produces a multiple bond involving two electron pairs (a double bond), as shown in the next set of figures. Overlap of the sp^2 orbitals forms a sigma (σ) bond and the p orbital overlap produces a pi (π) bond. A carbon-carbon double bond is not composed of two similar bonds, as might be interpreted from the usual notation of $\text{C}=\text{C}$ that is used. Rather, each bond is a distinct and separate entity and many physical and chemical properties confirm this feature.

All of the sigma bonds lie in the same plane, but the pi bonds project above and below the plane, as is evident from the previous diagram. As might be expected, because of the added "cementing" properties of the extra electrons, the carbon atoms of a multiple bond are held more closely. Thus the carbon-carbon bond distance for a double bond is 1.34 Å in ethylene, compared to 1.54 Å for the single carbon-carbon bond of ethane.

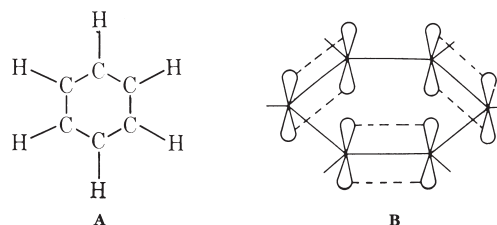
Another situation occurs due to the configuration of the sigma-pi double bond. Reference to the following illustration of the completed molecule shows that groups *a*, *b*, *c*, and *d* are in the same plane, and, by reversing the two substituent groups at either end of the molecule (as in **B** and **C**), an isomer is generated—a *geometric isomer*.



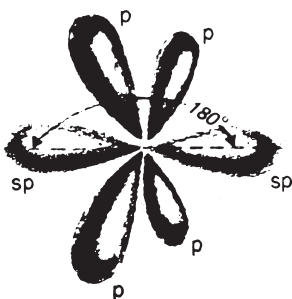
DELOCALIZATION AND RESONANCE—Benzene represents a large series of compounds exhibiting a kind of bonding that is perhaps as unique and different from the usual carbon-carbon bond types as is carbon from the rest of the periodic table. Although the six annular carbon atoms are bonded to each other via sp^2 orbitals (as with ethylene), the resultant molecule does not behave as an unsaturated compound. The compound is depicted as having a conjugate system of three double bonds (**B**, **C**, and **D**).

Again, this phenomenon leads to a doubling of the number of possible compounds of this particular type.

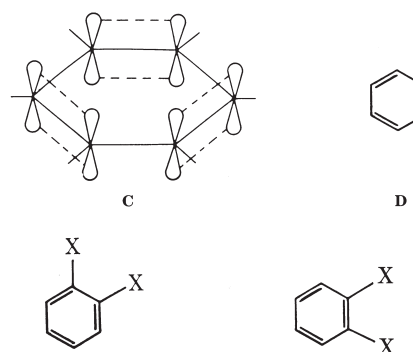
A third variety of hybridization that exists involves the coalition of one s and one p electron (sp). The resulting two sp orbitals produced are directed axially, 180° apart and 90° removed from the plane of the unhybridized p orbitals.



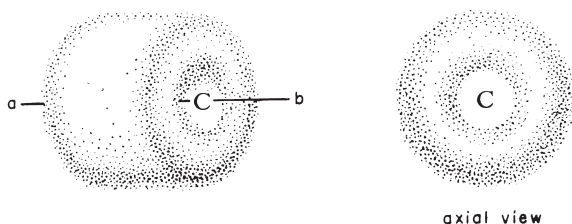
However, the benzene molecule does not behave chemically like a simple conjugated triene. Reactions normally occur by substitution of a hydrogen atom, rather than by the expected addition to the double bond. Also, two simple disubstitution products would be expected.



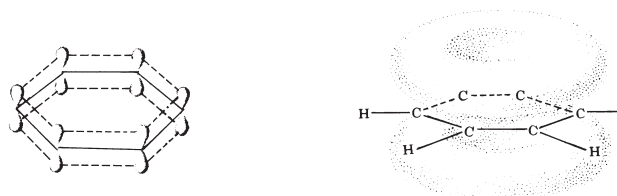
A combination of two carbon atoms exhibiting sp hybridization along the sp axis will yield carbon-carbon triple bonds.



However, only one disubstitution product is known. Benzene, therefore, must exhibit an entirely different kind of bonding than those previously discussed. It is believed that the p orbitals, above and below the plane of the benzene ring, overlap in both directions and each electron can participate in several bonds. The ability of the π electrons to be active in joining several atoms results in stronger bonds and a more stable molecule. This phenomenon of *delocalization* of electrons results in a delocalization, or *resonance*, energy of stability.



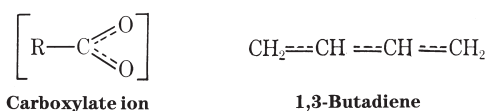
The p orbitals form a cylindrical sheath about the sigma bond. For a carbon-carbon triple bond the interatomic distance is smaller than a single or double bond, being 1.20 \AA . Isomerism (geometric or stereoisomerism) is not possible with a triple bond as the substituents, **a** and **b**, are located axially.



Due to the delocalization of the electrons only one type of bond exists and the classic alternate arrangement of single and double bonds between carbon atoms of the benzene molecule is misleading and incorrect. The carbon-carbon bond distance for benzene is 1.39 Å, lying between the single- and double-bond interatomic distance. The term *delocalization* better describes the resultant molecular orbital picture of benzene, as opposed to the concept of *resonance*, which may imply a rapid alternation between two or among several structural forms, which is totally incorrect.

Delocalization (resonance) stabilization is evidenced by many organic compounds that contain multiple bonds. Just as a lowering of energy results from the formation of molecular orbitals, whereby electrons are associated with two positive nuclei, a further lowering results if a molecular orbital is formed by using several nuclei. This extra energy-lowering increases the stability of a compound; the net energy difference derived from summing bond energies and that of the heat of combustion of the molecule is termed *resonance or delocalization energy*.

Several types of organic compounds, other than benzene, exhibit delocalization.



Delocalization accounts for the stability of *aromatic* compounds such as naphthalene, anthracene, pyridine, pyrimidine, thiophene, furan, etc. *Aromaticity* has become synonymous with the unusual stability and chemical behavior of benzene-like compounds. A quantum mechanical treatment of cyclic, conjugated systems indicates that aromaticity exists in those rings associated with $(4n + 2)\pi$ electrons, where n is an integer. Thus, rings having 6, 10, or 14π electrons may be aromatic (if they are planar), whereas those of 4, 8, or 12π electrons cannot be. The supporting mathematical theory is beyond the scope of this chapter, but chemical evidence easily suggests that compounds such as pyridine, thiophene, or furan do behave as benzene, while cyclooctatetraene—although a cyclic, conjugate system—behaves merely as a typical conjugated alkene and does not show the exceptional stability of an aromatic compound.

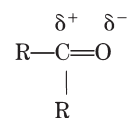
Carbon-Heteroatom Bonds

Practically all of the foregoing material pertains to the structure of a carbon-to-carbon or carbon-to-hydrogen bond. A majority of the compounds normally included in the area of organic chemistry also contain *heteroatoms* (atoms other than carbon and hydrogen), and the mode of bonding between carbon and the heteroatoms is of great importance. A rigorous treatment of this subject is beyond the limits of this chapter, but several general observations are in order.

Carbon forms a typical sigma bond with the univalent nonmetals (halogens) and with other electronegative polyvalent elements such as oxygen, nitrogen, sulfur, and phosphorus. Because of the differing electronegativities of the atoms on either side of the sigma bond, the bond is not entirely symmetrical and the slightly uneven distribution of bonding electrons causes an asymmetry leading to increased values of dipole moments with increased difference in electronegativities.

Multiple bonds also can exist between the polyvalent elements and carbon. Typical of this group is the carbonyl function ($\text{C}=\text{O}$), an example of sp^2 hybridization. The carbon atom is joined to two other atoms and the oxygen atom by sigma bonds; the remaining p orbital of the carbon overlaps a p orbital of oxygen to form a typical pi bond. Thus, carbon and oxygen are joined by a double bond. Each of the three sigma bonds radiating from the carbon atom is at an angle of 120° , and the carbonyl portion and the two atoms to which it is attached lie in the same plane.

The electrons of the carbonyl double bond join two elements of quite different electronegativity, and hence are not shared equally, the electron cloud being pulled more strongly toward the electronegative oxygen atom. As the π electrons are of a lower energy than σ electrons, they are influenced more easily by the electronegative oxygen atom. This effect is much more pronounced with multiple bonds than for a single (sigma) bond and results in the occurrence of a permanent polarity. Therefore, aldehydes and ketones (which contain the carbonyl function) exhibit fairly large dipole moments (2.30–2.75 D) because of the polarity of the carbonyl group, as shown below. A lower-case delta (δ) indicates that a fractional charge of appropriate sign resides on the designated atoms.



The structure of the carbonyl group largely determines the physical and chemical properties of aldehydes and ketones. Similar analogies can be drawn for carbon-to-sulfur and carbon-to-nitrogen multiple bonds.

Although carbon usually bonds to other elements by covalent-type linkages, several examples of ionic-type bonds are known (carbanion, R_3C^- ; and carbonium ion or carbocation, R_3C^+), but these are very short lived and are primarily useful in explaining the *mechanisms* of various organic reactions via intermediates of transient existence.

Noncarbon Bonds

The magnitude of the number of organic compounds is not due solely to the intricacies shown in carbon-to-carbon and carbon-to-heteroatom bonds. The electronegative elements, especially nitrogen and oxygen, impart their individualities such that a carbon-to-oxygen or carbon-to-nitrogen bond can participate in new types of bonds not discussed previously. As an example, the *hydrogen bond* or *bridge* can cause intermolecular association which can lead to an apparent increase in molecular weight. The hydrogen bond also may be the reason for a drug binding to certain sites of activity. Formation of *chelates*, *clathrates*, coordination complexes, and so on also extends the number of compounds that would be possible if only classic types of bonding existed between elements. Chapter 14 deals in depth with the concepts mentioned in this paragraph.

Interatomic distances decrease appreciably to achieve the overlap needed to form pi bonds between atoms. The bond distance is characteristic of the atoms involved and the type of bond between them. Table 13-3 gives the bond energy and the bond distance for some covalent bonds.

Table 13-3. Covalent Bond Energy

BOND	BOND ENERGY, ΔH KCAL/MOL	BOND DISTANCE Å
H—H	103.2 ^a	0.74 ^c
H—Cl	102.1 ^a	1.27 ^c
O—H	109.4 ^a	0.96 ^b
N—H	92.2 ^a	1.01 ^b
C—H	98.2 ^a	1.09 ^b
C—Cl	78.0 ^a	1.77 ^b
Cl—Cl	57.8 ^a	1.99 ^c
C—C	80.0 ^a	1.54 ^b
C=C	130.0 ^a	1.33 ^b
C≡C	193.0 ^a	1.20 ^b
C=O	152.0 ^b	1.21 ^b

^aData from Pitzer KS. *J Am Chem Soc* 1948; 70: 2140.

^bData from Fieser LF, Fieser M. *Introduction to Organic Chemistry*. Boston: DC Heath, 1957.

^cData from Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960.

Table 13-4. Electronegativity Values for Some Elements

F	4.0	I	2.4	Be	1.5
O	3.5	P	2.1	Mg	1.2
N	3.0	H	2.1	Li	1.0
Cl	3.0	B	2.0	Ca	1.0
Br	2.8	Si	1.8	Na	0.9
S	2.5	Sn	1.7	K	0.8
C	2.5	Al	1.5	Cs	0.7

Adapted from Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, arranged in decreasing order.

POLAR BONDS: PARTIAL IONIC BOND AND IONIC BOND—There are many different types of partial ionic bonds between the two extremes of a covalent bond and an ionic bond. The tendency of a pair of atoms to form an ionic or a partial ionic bond is measured by the difference in their abilities to attract an electron, or in their *electronegativities*.

If a molecule acts as if it has a positive and negative pole (ie, has a partial separation of charge), it is called a *dipole*. A molecule with a dipolar bond is said to be *polar*, while an electrically symmetric molecule is designated as *nonpolar*.

The electronegativity values for some common elements are listed in Table 13-4. The relationship between electronegativity differences and the partial ionic character is shown in Table 13-5. It is interesting to point out that fluorine, the most electronegative of all elements, has not only unique chemical qualities but also important physiological properties. In very low doses fluorides can reduce the number of dental caries by well over 50%, while in excessive doses mottled enamel may result during the period of tooth formation. Lithium, a metal of very low electronegativity has been used in the treatment of manic depressive disorders; both the carbonate and citrate are the salt forms used clinically (see Chapters 24 and 82).

DIPOLE MOMENT—The process by which dipoles arise is known as *polarization*. The total polarization P can be written as

$$P = P_i + P_0 + P_a \quad (5)$$

The induced or electronic polarization, P_i , represents the shift of the electron cloud due to the influence of an electric field or an electromagnetic wave such as light. The induced molar polarization P_i can be determined from molar refraction measurements using the D-line of a sodium lamp, as the permanent dipole cannot follow an electromagnetic wave of such high frequency.

$$P_i = \frac{n_D^2 - 1}{n_D^2 + 2} \times \frac{M}{d} = MR \quad (6)$$

Equation 6 is known as the Lorentz-Lorenz equation, where n_D is the refractive index of the liquid measured with the D-line of a sodium lamp, M is the molecular weight, d is the density, and MR is the molar refraction (refractivity).

Table 13-5. The Difference in Electronegativities and Ionic Character of Some Chemical Bonds^a

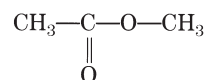
BOND	ELECTRONEGATIVITY DIFFERENCE, $X_a - Z_b$	PARTIAL IONIC CHARACTER, %
C—H	0.4	4
I—Br	0.4	4
I—Cl	0.6	9
O—H	1.4	30
C—F	1.5	44
Si—F	2.2	70
Be—F	2.5	79
K—F	3.2	92

Data from Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, arranged in increasing order.

Table 13-6. Atomic and Group Refractions for Sodium-D Light

ELEMENT	NA _D CC	ELEMENT	NA _D CC
C	2.42	N in	
H	1.10	Aliphatic oximes	3.93
O in OH	1.52	R—CONH ₂	2.65
O in ester OR	1.64	R—CONHR'	2.27
O=	2.21	R—CONR'R''	2.71
F	1.22	NO ₂ group in	
Cl	5.96	Alkyl nitrates	7.59
Br	8.86	Alkyl nitrites	7.44
I	13.90	Nitroparaffins	6.72
S in SH	7.69	Aromatic nitro compounds	7.30
S in RS	7.97		7.30
S in RCNS	7.91	Nitramines	7.51
S in RS	8.11	NO group in	
N in		Nitrites	5.92
Hydroxylamines	2.48	Nitrosamines	5.37
Hydrazines	2.47	Structural units	
RNH ₂	2.32	Double bond	1.73
RNHR'	2.49	Triple bond	2.40
RNR'R''	2.84	3-membered ring	0.71
ArNH ₂	3.21	4-membered ring	0.48
ArNHR	3.59	Oxirane	
ArNRR'	4.36	Terminal	2.02
R—C≡N	3.05	Nonterminal	1.85
Ar—C≡N	3.79	Conjugation—(see Ref 6)	

One also can calculate the induced molar polarization from the electron-group refractions given by Smyth, or from the Atomic Refraactivities compiled by Fajans (see Fajans⁸ and Table 13-6). For example, the molar refraction of methyl acetate,



can be calculated as

$$\begin{aligned} \text{Na}_D & \\ 3 \times \text{C} &= 3 \times 2.42 = 7.26 \\ 6 \times \text{H} &= 6 \times 1.10 = 6.60 \\ 1 \times \text{=O} &= 1 \times 2.21 = 2.21 \\ 1 \times \text{—O—} &= 1 \times 1.64 = 1.64 \\ \text{Total} &= 17.71 \end{aligned}$$

or

$$\begin{aligned} MR &= \frac{n_D^2 - 1}{n_D^2 + 2} \times \frac{M}{d} \\ &= \left[\frac{(1.3593)^2 - 1}{(1.3593)^2 + 2} \right] \times \left[\frac{74.08}{0.928} \right] \text{ (at } 20^\circ) \\ &= 17.57 \end{aligned}$$

An apparent correlation between the activity of chloramphenicol analogs, as determined by microbial kinetics, and the group refraction of their aromatic substituents has been reported.⁹

In Equation 5, P_0 is the orientation polarization due to the permanent dipole and P_a is the atomic polarization, which may be neglected for practical purposes because it is only 5 to 10% of P_i . The orientation polarization, P_0 , arises from the separation of charges due to the difference in electronegativities of the atoms.

Using an electromagnetic wave of much lower frequency than the frequency of light, such as a radio wave, one can measure the total polarization, as the permanent dipole as well as the electron cloud can follow the alternation of direction of the radio wave. In other words, one can calculate P from dielectric constant and molar volume (M/d) measurements.

$$P = \frac{\epsilon - 1}{\epsilon + 2} \times \frac{M}{d} \quad (7)$$

Combining Equations 5 through 7, Debye's equation (Eq 8) for a pure compound, and the Clausius-Mossotti equation (Eq 9), and neglecting P_a , gives Equation 10:

$$P = \frac{4}{3} \pi N_A \left(\alpha + \frac{\mu^2}{3kT} \right) \quad (8)$$

$$P_i = \frac{4}{3} \pi N_A \alpha \quad (9)$$

$$P_o = P - P_i = \frac{4}{3} \pi N_A \frac{\mu^2}{3kT} \quad (10)$$

where N_A is Avogadro's number, α is the induced polarizability (a measure of the ease of polarization by an electric field), μ is the dipole moment (esu · cm), k is the Boltzmann constant, and T is the absolute temperature. It should be noted that molar refraction is a molar property and induced polarizability is a molecular property.

Equation 8 can be written as

$$P = a + b/T \quad (11)$$

where $a = 4\pi N_A \alpha / 3 = P_i$ and $b = 4\pi N_A \mu^2 / 9k$. Because Equation 11 is a linear equation, by plotting values of P at several temperatures (calculated from dielectric constant measurements) versus $1/T$, one can compute α and P_i from the intercept and the permanent dipole moment (μ) of the compound from the slope, b . This procedure usually is applied to gases.

For a pure liquid, one can obtain the total polarization, P , according to Equation 7 and the induced polarization, P_i , from refractive index and molar-volume measurements at a constant temperature (Eq 6). Regardless of the manner of obtaining P_i , the final equation for the calculation of the dipole moment, usually expressed in Debye units, is the same (Eq 12). One Debye unit (D) is equivalent to $10^{-18} \times$ esu · cm.

$$\begin{aligned} \mu &= \sqrt{\frac{9kb}{4\pi N_A}} = 0.0128 \times 10^{-18} \sqrt{b} \\ &= 0.0128 \times 10^{-18} \sqrt{(P - P_i)T} \text{ (esu · cm)} \\ &= 0.0128 \sqrt{(P - P_i)T} \text{ (Debye units)} \end{aligned} \quad (12)$$

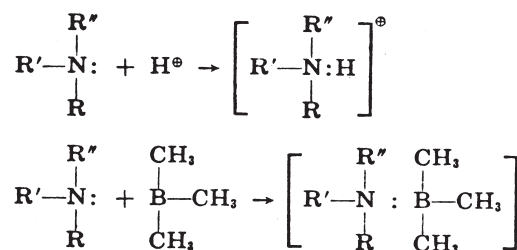
There are other equations for calculating dipole moments from measured values of the dielectric constant, refractive index, and density of liquids. However, for pure liquids the results are not very satisfactory. The dipole moment of medicinal substances usually is measured in a nonpolar solvent (eg, benzene, cyclohexane, or heptane) or in a solvent with some polarity but without resultant moment (eg, dioxane).

It has been suggested that, to eliminate the inaccuracies that arise from treating the solvent in a different way than solutions, only the results of measurements on dilute solutions be used.

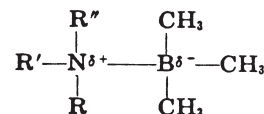
Correlations of biological activity with dipole moment have been reported for the insecticidal activity of chlorophenothane (DDT) isomers, the cholinesterase inhibitory activity of *N*-alkyl substituted amides, and the respiratory stimulation activity of cyclic ureas and cyclic thioureas. Investigations have shown that high dipole moment enhances central nervous system (CNS) stimulatory activity or toxicity, whereas low dipole moment favors anticonvulsant or CNS-depression activity. The use of dipole moment as a parameter in drug-receptor interaction and quantitative structure-activity relationship studies has been reviewed by Lien et al.¹⁰⁻¹²

When the electronegativities of the bonded atoms are quite different, a formal electron-pair bond can no longer exist. The bonding electron pair is now associated exclusively with the more electronegative atom, and an *anion* is formed. The atom that has lost its electron becomes positively charged, and a *cation* is formed.

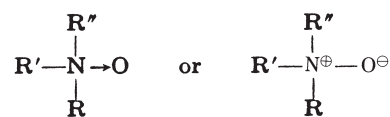
COORDINATE COVALENT BONDS—A *coordinate covalent bond* is formed when only one atom donates both electrons; for example, the unshared electron pair on the nitrogen atom of an amine (a Lewis base) can serve to form such a bond with a proton or trimethyl boron (a Lewis acid).



Because the nitrogen suffers a loss of negative charge and the boron atom gains an equivalent negative charge, it is more realistic to depict the complex molecule as the *adduct*.



Amine oxides are other examples of coordinate covalent compounds.



Because oxygen is much more electronegative than boron (see Table 13-4), the ionic character of the N-oxide is more pronounced than that of the N-B bond. This is evidenced by the relatively high melting point, high solubility in water, and low solubility in nonpolar solvents of the amine oxides. One also can infer the polar character by a comparison of dipole moments: 6.2 D for KCl (ion pairs), 5.02 D for trimethylamine oxide, and 3.92 D for the trimethylamine-trimethylboron complex.

CHELATES—The term *chelate* (from the Greek *chela*, claw) describes this class of compounds appropriately. *Chelates* consist of a partial ring of atoms that close up by holding a given atom, usually a metal, in a molecular claw. The compounds capable of forming a ring structure with a metal are designated as *ligands* (see Chapter 14 for a thorough discussion of complex formation).

Cisplatin, a platinum coordination compound with the amine groups at the *cis* position, has been used in the treatment of testicular and ovarian tumors in combination with other anticancer drugs (see Chapter 86).

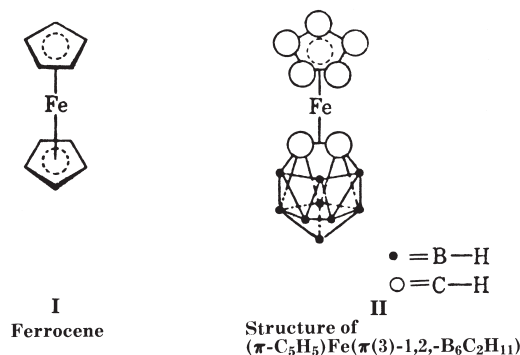
Some biologically important compounds (eg, chlorophyll, hemoglobin, peroxidases, cytochromes, oxidases, ascorbic acid oxidase, tyrosinase, polyphenoloxidase, lactase, phosphatase,

carboxylases, insulin, and cyanocobalamin) are naturally occurring chelates. Tetracyclines also are capable of forming chelates with metals. Chelating agents may be used for a number of purposes, such as sequestration of metals, stabilization of drug preparations vulnerable to oxidation in the presence of trace-metals, and the treatment of heavy metal poisoning.

MOLECULAR BONDS—Several classes of compounds contain *intermolecular coordinate covalent bonds* (eg, sandwich compounds, charge-transfer complexes, and the molecular-addition compounds). These types of bonds are referred to as *molecular bonds* for brevity.

METALLOCENES—In 1951 Kealy and Pauson accidentally discovered ferrocene by oxidizing cyclopentadienemagnesium bromide with anhydrous ferric chloride in ether solution. Ferrocene has aromatic character and is an unusually stable iron-containing orange product, formula $C_{10}H_{10}Fe$, that melts at 174° and boils at 249° ; it is soluble in common organic solvents but insoluble in water. The generally accepted structure of ferrocene was first proposed by Woodward et al in 1952. X-ray and electron diffraction studies have shown that the iron is packed between two parallel cyclopentadienyl rings like a *sandwich* (I, below).

The solubility, volatility, and other properties of metallocenes are due to the covalent character of the molecular bonds. This indicates that each cyclopentadienyl ion donates an electron pair to the metal ion. Ferrocene is diamagnetic, hence the six $3d$ electrons of iron are paired up to make available two open $3d$ orbitals. A large number of metallocenes have been prepared and studied since the discovery of ferrocene.



Several different aromatic rings (eg, indene, azulene, and benzene) also will form metallocenes. In many metallocenes, CO or NO molecules are found in place of one of the aromatic rings, and the metal may be Cr or Mn, as well as Fe. Metallocenes undergo most of the typical aromatic reactions.

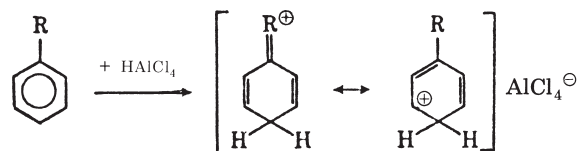
It has been shown that daily oral administration of ferrocene produced hemosiderosis with an unusually high, dose-related accumulation of iron in dogs. A decrease in hemoglobin, packed-cell volume and erythrocyte count occurred within 4 weeks in dogs receiving 300 mg/kg of ferrocene. This and higher dosage resulted in cirrhosis, which was considered to be an effect of the hydrocarbon moiety.

There is a field of research that combines polyhedral carbene and transition metal chemistry. Several families of polyhedral species now are known in which a transition metal resides in the polyhedral surface (II, III).

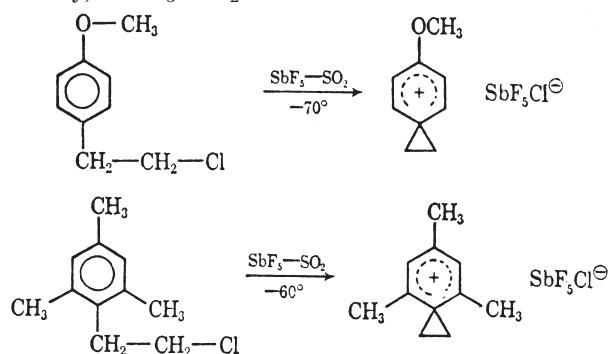
CHARGE-TRANSFER COMPLEXES—Certain substances combine in a 1:1 molar ratio to form crystalline addition products. The molecular addition compound is held together by weak forces, such as van der Waals (dipole-dipole, dipole-induced dipole, induced dipole-induced dipole), ion-dipole, and even hydrogen bonds. Polynitroaromatic compounds, such as trinitrobenzene and picric acid, are well known for their ability to form charge-transfer complexes (pi complexes) (see Chapter 14).

Caffeine complexes with various drugs such as sodium benzoate, sodium salicylate, sulfonamides, barbiturates, and 5-chlorosalicylic acid.

AROMATIC SIGMA (σ) BOND COMPLEXES—Aromatic compounds react with $HCl \cdot AlCl_3$ or $HF \cdot BF_3$ to produce salts that ionize in highly polar nonaqueous solvents, for example, liquid hydrogen fluoride or sulfuric acid.



Using NMR spectrometry, Olah et al¹³ detected the *p*-anisonium and the 2,4,6-trimethylphenonium ions produced by ionizing β -*p*-anisylethyl chloride and β -mesitylethyl chloride, respectively, in SbF_5-SO_2 at -70° to -60° .



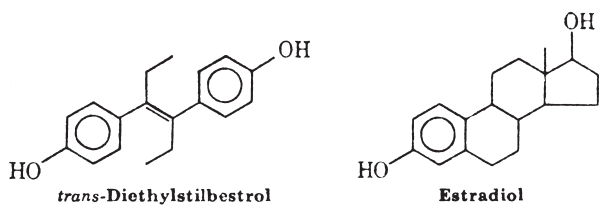
Sigma complexes are molecular complexes resulting from the rupture of a sigma bond (eg, $H-AlCl_4$, $ArCH_2CH_2-Cl$); they also occur in Friedel-Crafts reactions. Because they are reactive toward water, no practical pharmaceutical uses have been made of these complexes. Refer to Chapter 14 for a more extensive treatment of complexation.

STEREOISOMERISM—Early in 1874 van't Hoff envisaged a double bond by joining two tetrahedrons at two corners and correctly predicted that unsymmetrically substituted derivatives of ethylene should exist in two stereochemical forms, or as a pair of *cis* and *trans* isomers.



As in the previous discussion of sigma and pi bonds, it was shown that in an alkene, rotation about the sigma bond is restricted by the overlap of *p* orbitals comprising the pi bond.

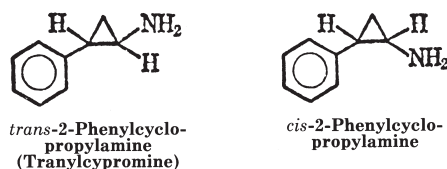
Stereoisomerism as a result of the rigid configuration about a double bond, or other rigid structure such as a ring, is known as *geometric isomerism*. It is interesting to note that in the case of the synthetic estrogens, the *cis* isomer of diethylstilbestrol is unstable and has less than one-tenth the activity of the *trans* isomer. One should note the structural similarity between *trans*-diethylstilbestrol and estradiol.



It has been reported that tamoxifen, a drug that structurally resembles *trans*-diethylstilbestrol, arrested or reduced breast tumor growth rate in 77% of patients given 20 mg of the drug orally twice a day. The compound is believed to block estrogen receptor sites.

Due to the presence of symmetry, the type of geometric isomerism occurring in substituted ethylenes usually is not associated with optical activity; some other site in the molecule ordinarily gives rise to optical isomerism.

Another type of geometric isomerism is found in ring compounds, the ring taking the place of the rigid double bond. For example, *trans*-2-phenylcyclopropylamine is more stable than the *cis* isomer and is a potent monoamine oxidase inhibitor.



A substance that rotates the plane of polarized light is said to be *optically active*. *Optical rotation* may be considered as a consequence of the phenomenon of circular double refraction in which a beam of polarized rays is resolved into two circularly polarized rays, one turning clockwise and the other counterclockwise as the beam advances. In an optically active medium these rays have different velocities and on recombination they vibrate in a plane different from that of the incident ray. Refer to Chapter 35 for a more complete discussion.

The necessary and sufficient condition for a molecule to show optical activity is that the molecule should be asymmetric (ie, the molecule should not be superimposable with its mirror image); in other words, it should be *chiral* (from the Greek *cheir*, hand; thus, right- or left-handedness). Although many optically active compounds have asymmetric carbon atoms (carbon atoms bearing four different groups), not all compounds possessing asymmetric carbon atoms are optically active; for example, *meso*-tartaric acid has two asymmetric carbon atoms, but it is optically inactive due to the presence of a plane of symmetry within the molecule (*internal compensation*).

Optical isomerism due to restricted rotation (as with *ortho*-substituted biphenyls and dissymmetric polyphenyls) is well documented in Eliel's book (see the bibliography). Atoms other than carbon can serve as a center of asymmetry. For instance, optically active *N*-oxides, quaternary ammonium compounds, sulfonium and selenium salts, and sulfoxides and sulfonic esters have been resolved. As living organisms are made of numerous chiral macromolecules, stereoselectivity commonly is observed for stereoisomers. Increasing emphasis is being placed on the use of the more active enantiomer (eutomer), instead of the racemic mixture (equal amounts of both eutomer and distomer) as the therapeutic agent (see Ariëns et al in the bibliography).

ENANTIOMERS—Molecules whose mirror images are nonsuperimposable are called *enantiomorphs*, *enantiomers*, or *optical antipodes*. Enantiomers have identical physicochemical properties in an optically inactive environment—they rotate the plane of polarized light to the same degree, but in opposite directions. The measurement of optical rotation is useful for the purpose of identifying and/or assaying an optically active

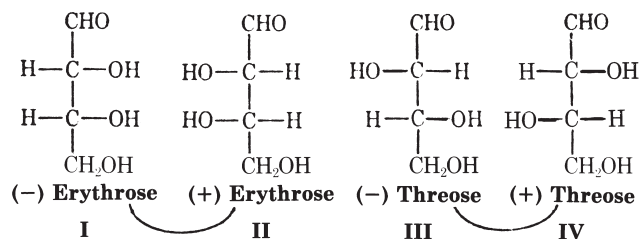
substance. The *specific rotation* is defined as

$$[\alpha]_D^t = \frac{\alpha}{l(g/v)}$$

where *D* is the D-line of sodium vapor lamp, *t* is the temperature, α is the observed rotation in degrees, *l* is the length of the cell in decimeters (1 dm = 10 cm) and *g/v* is the concentration in g/100 mL of solvent.

When equal amounts of *dextro* (+) and *levo* (−) isomers are mixed, a *racemic modification* arises. They are denoted as *d*, *l* (which no longer are used to designate direction of rotation of light) or \pm . Racemic modifications are the products of most organic syntheses that involve a chiral center; they also may be obtained by racemization of a pure enantiomer. In a racemic modification the substance in bulk is not optically active, even though an individual molecule is optically active. The resultant rotation is zero as the concentration of the molecules that rotate light to the left is equal to those that rotate it to the right.

DIASTEREOISOMERS—Stereoisomers that are not mirror images of each other are called *diastereoisomers* or *diastereomers*. Diastereoisomerism exists when a given structural formula has at least two asymmetric atoms. Diastereoisomers should have different physicochemical properties like melting points, solubilities, and optical rotation.



For one above example, compounds I and II or III and IV are enantiomers; compounds I and III, I and IV, II and III or II and IV are diastereomers.

ABSOLUTE CONFIGURATION—The designations (+), (−), *d*, and *l* refer to the rotation of plane polarized light by a molecule, but the actual three-dimensional arrangement in space of atoms in a chiral molecule may bear no relation to these descriptors. Even with the carbohydrates, the small capital D or L refers to the configuration of but a portion of a molecule relative to a reference compound, glyceraldehyde. Fortunately, the selection of the reference configurations for absolute and assumed configurations happened to coincide.

With the improvement of x-ray crystallographic techniques in the 1950s, it became possible to reveal the actual three-dimensional arrangement of atoms and the absolute configuration of (+)-tartaric acid was determined. This became a reference point to which other chiral molecules could be related, by chemical conversions, that previously had been demonstrated to retain or invert a configuration.

From these studies the *R* and *S*, or Cahn-Ingold-Prelog system (named after the chemists who devised the method), was developed. A series of *sequence rules* was promulgated, which is beyond the scope of this chapter. These rules accommodate geometric *cis* (*zusammen* or *Z*) and *trans* (*entgegen* or *E*) structures, as well as *R* (*rectus* or right) and *S* (*sinister* or left). The symbols *R* and *S* refer only to the right- or left-handedness of the chiral centers and not to the direction of rotation of polarized light.

OPTICAL ROTATORY DISPERSION (ORD)—*Optical rotatory dispersion* (ORD) involves the measurement of the angle of optical rotation of linearly polarized light at various wavelengths. Usually greater rotational angles are obtained at shorter wavelengths. The source of energy consists of a xenon arc and a monochromator to isolate the desired wavelength in the ultraviolet region. A photomultiplier and photometer are used to measure the intensity after the light has passed through the polarimeter.

As the wavelength of the polarized light is varied, the absolute value of rotation may increase continuously so that the plot of $[\alpha]$ versus λ is a plain curve (Line A, Fig 13-3). On the other hand, the rotation may change direction either from left to right or right to left, and show one or more maxima and minima.

The appearance of a maximum and a minimum in a plot of specific rotation versus wavelength is referred to as a *single Cotton effect* (Line B, Fig 13-3), whereas the appearance of several maximums and several minimums is referred to as a *multiple Cotton effect*. If, in approaching the region of the Cotton effect from long wavelengths, one passes first through a maximum and then through a minimum, the Cotton effect is called *positive*. If the minimum is reached first and then the maximum at shorter wavelength, it is called a *negative* Cotton effect.

The Cotton effect is due to the presence of an asymmetric center near a chromophoric group, such as $\text{C}=\text{O}$, in the optically active molecule which has unequal absorption of right and left circularly polarized light. The concept of ORD is useful for the study of the stereochemistry of natural products, ketosteroids, and the analysis of randomly coiled and helical configurations of polypeptide chains.

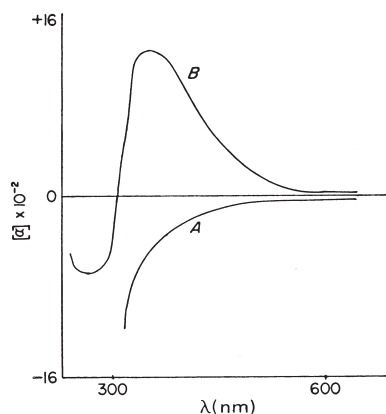


Figure 13-3. Rotatory dispersion curves: (A) levorotatory plain curve, (B) positive simple Cotton effect.

CIRCULAR DICHROISM (CD)—A *circular dichroic* curve is a plot of the molecular ellipticity $[\theta]$ versus the wavelength λ . The CD effect results from the fact that the right circularly polarized ray is *absorbed* differently from the left circularly polarized beam of light. The molecular ellipticity is defined as

$$[\theta] = 3300 \cdot \Delta\epsilon, \Delta\epsilon = \epsilon_L - \epsilon_R$$

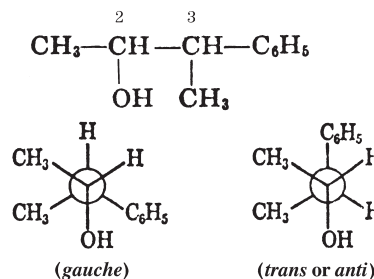
where $\Delta\epsilon$ is the differential dichroic absorption and ϵ_L and ϵ_R are the molar extinction coefficients for the left and right rays.

If in a dichrograph the oscillating crystal is oriented correctly, the plane-polarized beam of light passed through the instrument can be resolved into right and left components. These are passed through the optically active medium. When these unequally absorbed circular components are recombined in the region of electronic absorption, they give elliptically polarized light. Measurements involving CD have been used for studying drug-protein binding with 52 analgesic, sedative, and antidepressive drugs.¹⁴ From this study it was suggested that a planar system with high electron density (eg, benzodiazepine and dibenzazepine derivatives) appeared to be an essential factor for strong binding to human serum albumin.

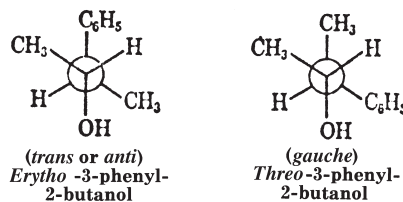
CONFIGURATION AND CONFORMATION—The spatial arrangement of the groups about a central atom is referred to as the *configuration* of the atom. Three-dimensional models, their projections, or perspective drawings must be used to illustrate the difference between stereoisomers. The particular shape that a molecule assumes by free rotation about single bonds is referred to as its *conformation*.

An ethane molecule may have an infinite number of conformations because of rotation about the C—C bond; however, only a few conformations are possible that will make the molecular energy a minimum. The conformational preferences of some diastereoisomers have been determined from nuclear magnetic resonance (NMR) studies.

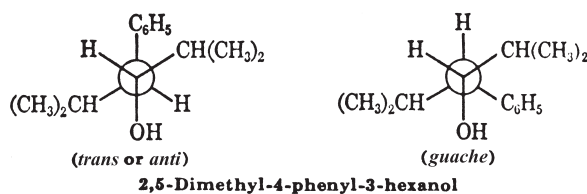
For a series of diastereoisomers involving a substituted phenylethyl skeleton, when the alkyl groups attached to each asymmetric center are small (eg, methyl), both *gauche*- and *trans*-conformers (rotamers)



have substantial populations because of the relatively low rotational barriers. Newman projection formulas are used for the illustrations, in which the molecules are viewed from front to back in the direction of the bond linking the asymmetric carbon atoms. In the following formulas, the center of the circle represents C-2 and the circle itself represents C-3 of 3-phenyl-2-butanol.



When the alkyl groups are bulky (eg, isopropyl), steric interactions cause these groups to prefer a *trans* orientation; the vicinal hydrogens are then *trans* in the *erythro* but *gauche* in the *threo* isomers.



For a more detailed discussion of potential-energy barriers in various systems, consult Eliel's book (see the bibliography).

The preferred conformation of serotonin has been calculated using molecular orbital theory. Complementary features of the serotonin receptor have been postulated, and the relationship of serotonin in its preferred conformation to the serotonin antagonist, lysergic acid diethylamide (LSD), has been presented as an explanation of LSD's antagonism.

INTERMOLECULAR BINDING FORCES

An understanding of intermolecular and intramolecular binding forces is very important in many different aspects of pharmaceutical sciences, such as in the manufacture of various preparations, in stability studies, and in the design of new drugs. A knowledge of these forces is not only essential for predicting some physicochemical properties of various dosage forms but also indispensable for the interpretation of drug ac-

tion at the molecular level and for structure–activity correlations. Martin’s classification¹⁵ for various types of forces will be used in the following discussion.

REPULSIVE AND ATTRACTIVE FORCES—Intermolecular repulsive forces exist when two dipolar molecules are brought close together *head-to-head* or *tail-to-tail*, or when any two molecules are brought so close that their nonbonding electronic clouds interpenetrate. Otherwise, two molecules having opposite charges closer together than the like charges will attract each other. When the repulsive and the attractive forces are equal, the potential energy of the two molecules is a minimum and an equilibrium will be established. Similar forces may exist in the same molecule (intramolecular) as well as between different molecules. Only intermolecular forces will be discussed here.

VAN DER WAALS’ FORCES—Due to electrostatic attraction, dipolar molecules tend to align themselves with neighboring molecules so that the negative pole of one molecule points towards the positive pole of the next, for example,



This type of attraction is known as a *dipole–dipole* interaction and has a force of 1 to 7 kcal/mol. Dipole–dipole forces vary inversely as the fourth power of the distance between molecules, $F \propto (1/d^4)$.

The importance of the permanent dipole attractions in the stabilization of an α -helix has been pointed out. The electric dipoles in an α -helix add to one another along the direction of the axis. Two helices that wind in the same direction will, therefore, repel each other and two that wind in opposite directions will attract each other, as in DNA.

Permanent dipoles can induce a transient electric dipole in nonpolar molecules and produce *dipole-induced dipole*, or *Debye, forces*. These interactions involve an energy of about 1 to 3 kcal/mol.

When any two atoms belonging to different molecules are brought sufficiently close together, *induced dipole–induced dipole*, or *London, attractions* arise. In this case, the energy is about 0.5 to 1 kcal/mol. These forces originate from molecular internal vibrations. The temporary dipoles that this vibration creates in the constituent atoms induce dipoles in neighboring atoms of other molecules, and this process results in a net attraction. This type of force is responsible for the liquefaction of nonpolar gases. London forces vary inversely as the seventh power of the distance between molecules, $F \propto (1/d^7)$.

HYDROGEN BONDS—When a hydrogen atom holds two other atoms, a *hydrogen bond* (hydrogen bridge or H-bond) is formed. The two bonds attached to the same hydrogen cannot both be covalent bonds. The H-bond must be in part ionic. Indeed, the hydrogen bond usually is formed only between hydrogen and electronegative atoms. In addition, the atoms capable of forming H-bonds have at least one unshared electron pair.

Without hydrogen bonds this world would be much different, as water would boil at a temperature far below 0°. The surprisingly high boiling point of H₂O (100°), compared to H₂S (–60.7°) and H₂Se (–41.5°), can be attributed to the higher H-bonding ability of oxygen, which in turn is due to its smaller volume and higher electron density as compared to S and Se.

The most common atoms capable of forming H-bonds are F, O, N and, to a lesser degree, Cl and S. There also is some evidence that hydrogen attached to a triply-bound carbon (eg, HCN, HC≡CH, or CHCl₃) forms H-bonds. The strength of most H-bonds ranges from 1 to 7 kcal/mol.

H-BOND	BOND STRENGTH (kcal/mol)
F–H . . . F	7
O–H . . . O	4.5–7.6
O–H . . . N	4–7
C–H . . . π electrons	2–4
C–H . . . O	2–3
N–H . . . O	2–3
N–H . . . N	1.3

The strength of the H-bond depends on the solvent as well as the state. For instance, the H-bond strength of O–H . . . O for (CH₃COOH)₂ as a vapor is 7.64 kcal/mol, while that of (CH₃COOH)₂ in benzene is 4.85 kcal/mol. In water the H-bond has been estimated to have an energy of 4.5 kcal/mol; in ice, the bond strength is 6 kcal/mol. Hydrogen-bonding is responsible for the higher boiling point of a carboxylic acid compared with that of its ester. This is because in the free acid dimerization can occur by H-bonding, while this is impossible for an ester.

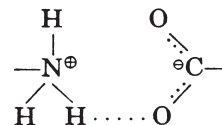
Hydrogen-bonding is also responsible for the high solubility of polyhydroxy compounds, such as sugars, in water. During the replication of DNA molecules, hydrogen bonds between base pairs are broken and rematched.

Various physical methods may be used to study H-bonding, such as molecular-weight determination, and infrared (IR) and NMR spectrometry.

ION-DIPOLE AND ION-INDUCED DIPOLE FORCES—

Ion pairs in the solid state have bond strengths comparable to or even stronger than covalent bonds (100–200 versus 50–150 kcal/mol). However, in a biological system, due to hydration and the large amount of inorganic salts present for ion exchange, the bond strength would be weakened substantially to the neighborhood of 5 kcal/mol.

When an ionic bond is reinforced by the simultaneous presence of other forces, such as hydrogen-bonding, the bond becomes stronger (10 kcal/mol).



An ion pair can attract a dipole or induce a dipole in a neighboring nonpolar molecule. The strength of an *ion-dipole* bond (eg, R₄N⁺ . . . $\overleftarrow{\text{NR}_3}$) is about 1 to 7 kcal/mol, and that of an *ion-induced dipole* (eg, $\overset{+}{\text{K}}-\overset{-}{\text{I}} \dots \text{I}-\text{I}$) would be somewhat weaker.

HYDROPHOBIC INTERACTIONS—The association of nonpolar groups with each other in aqueous solution, arising because of the tendency of water molecules to exclude nonpolar molecules, is known as a *hydrophobic interaction*, or *hydrophobic bonding*. The word *hydrophobic* really is a misnomer, because it implies that the nonpolar molecule dislikes water—in fact, it is water that dislikes the nonpolar molecule.

The formation of hydrophobic bonds is favored because of an entropy effect. Before the formation of a hydrophobic bond, water molecules are arranged in an ordered fashion around exposed nonpolar groups. When hydrophobic interactions occur, the order is disrupted and results in a favorable entropy change, which is great enough to overcome the enthalpy for the interaction of the nonpolar groups; hence, the free energy is negative and the process is spontaneous. The strength of hydrophobic interactions has been reported to be 0.37 kcal/mol per CH₂ group.

A chain of 14 carbon atoms that binds with another nonpolar counterpart would have a bond strength of 5.2 kcal/mol. This bond, being stronger than an ionic bond or other weak forces in the biological system, then may dominate the mode of binding of a complicated drug molecule. The importance of hydrophobic interactions in stabilizing protein structure, drug-protein binding, transport, and storage of drugs and drug-receptor interaction has been noted in recent years. A summary of the different types of intermolecular forces and molecular recognition is given in Table 13-7.^{10–12}

ADDITIVE PHYSICAL PROPERTIES

The division of physical properties into additive, constitutive, and colligative can be found in many textbooks. Additive physical properties depend on the number and kind of atoms in a

Table 13-7. Intermolecular Forces and Molecular Recognition

		Principles	
Like dissolves like (polar vs nonpolar)		II. Opposite charges attract each other (acids and bases; cations and I. anions)	
PROCESSES		FORCES INVOLVED (ALL INTERMOLECULAR FORCES ARE ELECTROSTATIC IN ORIGIN.)	
Non-Specific	<ul style="list-style-type: none"> Diffusion Dispersion Solution Mixing Phase transfer Passive absorption Excretion Non-chiral chromatographic separation 	Non-stereospecific in general	<ul style="list-style-type: none"> Ionic, ion-dipole, ion-induced dipole Dipolar (dipole-dipole, Keesom forces; dipole-induced dipole, Debye forces; induced dipole-induced dipole, London forces) (<i>van der Waals' forces</i>) H-bonding Hydrophobic interactions
Self-Association	<ul style="list-style-type: none"> Crystallization Formation of 4° protein structure, DNA, RNA 	All of the above, size, shape, complementarity, and group asymmetry are important	
Specific	<ul style="list-style-type: none"> Enzyme-substrate Drug-receptor Antigen-antibody DNA replication Transcription (DNA/RNA) Translation Active transport Facilitated transport Active secretion Chiral chromatographic separation 	Stereospecific in general	<ul style="list-style-type: none"> All of the above Complementarity is involved Shape, group symmetry as well as size are important

molecule. Such additivity enables one to calculate many molecular values from a few fundamental constants. The best example is the calculation of molecular weights from atomic weights. The additive nature of molar refractions has been used for the calculation of induced polarization (see the discussion of dipole moment in Chapter 15).

MOLAR VOLUME—This term is self-explanatory. It is defined as the molecular weight divided by the density of a liquid (molar volume = MW/d). By using statistical analysis, it has been shown that the additivity of molar volume is better fulfilled at ordinary temperatures (20°) than at the boiling point of each individual substance. This is an interesting result, as from the *principle of corresponding states* it might be expected that additivity would hold better at the boiling point.

In the homologous series of nonbranched primary derivatives, the accuracy of a calculation of molar volume is relatively good. The deviations increase gradually with poly-substituted derivatives, 1,1-bis-derivatives, *ortho* derivatives, and branched isomers; nevertheless, the additivity scheme can serve as a first approximation.

PARTITION COEFFICIENTS AND THE π CONSTANT—In the early theory of narcosis, lipid solubility was regarded as the most important factor for the inhibition of cell activity. At the beginning of the 20th century Meyer and Overton proposed that narcotic efficiency parallels the coefficient for the partition of a drug between oil and water. Although this theory cannot explain the mechanism of narcotic action, it does explain the role of transport to nerve tissues.

It is more logical to use partition coefficients than solubility in a single solvent for structure–activity correlations since, in a biological system, one is dealing with a heterogeneous system rather than a simple solution. Partition coefficients have been used in the study of drug absorption, distribution, metabolism, toxicity, and structure–activity correlation.

It has been shown that the partition coefficients for a given compound in two different solvent systems (eg, ether/water, octanol/water) are related as follows:

$$\log P_1 = a \log P_2 + b$$

where a and b are constants. This suggests that one can use the results from one set of solvents to predict results in a second set.

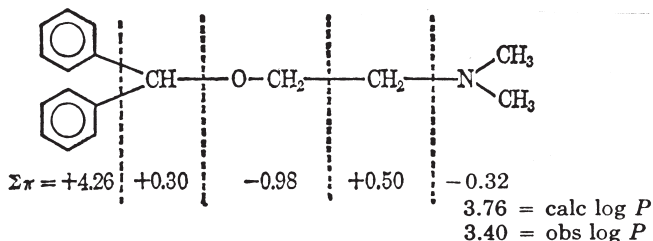
Hansch's group^{16–20} systematically has extended the use of partition coefficients, measured from octanol/water, to serve as a measure of the ease of passage of organic molecules through various lipoprotein barriers and/or as a measure of the hydrophobic binding with protein (such as bovine serum albumin). From the partition coefficients of a variety of derivatives of the type $X-C_6H_4OCH_2COOH$, $X-C_6H_5$, and $C_6H_5(CH_2)_n-X$, the substituent constants (π) for the aromatic and the aliphatic function (X) have been determined.

The π constant is defined as

$$\pi = \log P_X - \log P_H$$

where P_X is the partition coefficient of a derivative, and P_H is that of the parent compound. Although π varies continuously for a given function depending on its electronic environment, the variation generally is small; therefore, it is called *additive-constitutive*.

The application of $\log P$ and the additive-constitutive nature of constants for the correlation of biological activity with chemical structure has been illustrated in many cases. (See Chapter 28 for a discussion of the Hansch equation.) Table 13-8 lists the constants for some important functional groups.^{16–18,21,22} One can calculate many $\log P$ values from a few constants. The method of calculation can be illustrated with diphenhydramine.



Heightened interest in the structures and properties of proteins as drugs or drug targets has been greatly stimulated by the recent developments of pharmacogenomics and proteomics. Of particular importance are the hydrophobic contribution con-

Table 13-8. π Constants for Some Functional Groups

FUNCTION X	AROMATIC SYSTEM ^a	ALIPHATIC SYSTEM
H—	0	0
F—	0.13	-0.17
Cl—	0.76	0.39
Br—	0.94	0.60
I—	1.15	1.00
CH ₃ —	0.50	0.50
CH≡C—		0.48
CH ₂ =CH—		0.70
C ₂ H ₅ —	1.00	1.00
CH ₂ =CCH ₃		1.00
CH ₂ =CHCH ₂ —		1.20
<i>n</i> -C ₃ H ₇ —	1.50	1.50
<i>i</i> -C ₃ H ₇ —	1.30	1.30
<i>n</i> -C ₄ H ₉ —	2.00	2.00
<i>sec</i> -C ₄ H ₉ —	1.80	1.80
<i>t</i> -C ₄ H ₉ —	1.68	1.68
cyclo-C ₃ H ₅ —		1.21
cyclo-C ₅ H ₉ —	2.14	2.14
cyclo-C ₆ H ₁₁ —	2.51	2.51
Adamantyl	3.30	
C ₆ H ₅ —	2.13	2.13
—(CH ₂) ₃ —	1.04	
—(CH ₂) ₄ —	1.39	
—CF ₃	1.07	
—CH ₂ OH	-1.03	-0.66
—CH ₂ COOH	-0.72	-0.76
—COOH	-0.32	-1.26
—COO—	-4.36	
—CONH ₂	-1.49	-1.71
—COOCH ₃	-0.01	-0.27
—COCH ₃	-0.55	-0.71
—CN	-0.57	-0.84
—OH	-0.67	-1.16
—OCH ₃	-0.02	-0.47
—OCH ₂ COOH	-0.86	
—OCOCH ₃	-0.64	-0.91
CH=NNHCONH ₂	-0.85	
CH=NNHCSNH ₂	-0.27	
— <i>O</i> - β -glucose	-2.84	
—NH ₂	-1.23	-1.19
—N(CH ₃) ₂	-0.18	-0.32
—NO	-0.12	
—NO ₂	-0.28	-0.82
—NHCOCH ₃	-0.97	
—NHCOC ₆ H ₅	0.72	
—N=NC ₆ H ₅	1.69	
—NHCONH ₂	-1.01	
—N(CH ₃) ₃ ⁺	-5.96	
—SCH ₃	0.62	
—SCF ₃	1.58	
—SO ₂ CH ₃	-1.26	
—SO ₂ CF ₃	0.93	
—SF ₅	1.50	
—SO ₂ NH ₂	-1.82	

^aData from Hansch C, Anderson SM. *J Org Chem* 1967; 32:2583¹⁶; Hansch C, Anderson SM. *J Med Chem* 1967; 10:745¹⁷; Hansch C, et al. *J Med Chem* 1973; 16:1207¹⁸; Fujita T, et al. *J Am Chem Soc* 1964; 86:5175²¹; Iwasa J, et al. *J Med Chem* 1965; 8:150.²²

^bFrom X—C₆H₅ or X—C₆H₄OCH₂COOH system. For different positions in the latter system slightly different values were reported in the original paper.²¹ In cases where a strong interaction between two functions can occur (eg, in phenol or aniline series), different values should be used.

stants (faa) of different amino acids when they are incorporated into peptides or proteins. Table 13-9 summarizes the faa values of 21 common amino acids. The faa values range from -2.43 to + 1.47. This means when the most hydrophilic lysine is substituted with the most hydrophobic tryptophan, the logP in octanol/water is increased by 3.90 log units, and the partition coefficient P is increased by 7.94×10^3 fold.

X-RAY ANALYSES

In recent years the number of compounds of medicinal value that have been isolated from plant and animal sources and prepared by purely synthetic means has increased astronomically. In addition to the many compounds isolated, the more sophisticated isolation techniques now available have extended the capabilities of exploring biological molecules heretofore thought too complex to understand or investigate. The pharmaceutical chemist thus is faced with the task of identifying the chemical structure of a large number of complex materials in order to understand their biological functions.

For many of the compounds the chemist may rely on standard spectrometric methods (ie, IR, UV, NMR, and ORD), together with other chemical measurements, to elucidate molecular structure. Newer methods, especially mass spectrometry, have emerged as useful means of elucidating the structures of complex organic materials. In many instances these approaches have shortcomings, as they provide only fragmentary evidence about various portions of the molecule, which must be pieced together to get the picture of the whole compound (see also Chapter 34).

One of the most powerful of all techniques, when it is applicable, is that of x-ray crystallographic analysis. Using this method, the three-dimensional structure of a molecule can be determined without relying on any chemical information.

The maximum resolution that can be obtained through an ordinary light microscope under the most favorable conditions is about 2000 Å. This limitation is imposed primarily by the wavelength of the illumination. However, other forms of radiation capable of giving atomic resolution (1 Å or less) exist, namely electron beams, neutrons, and x-rays. Lenses have been constructed only for the first of these kinds of radiation, and at best they have a resolving power of about 6 Å. This resolution is insufficient to measure the distances between atoms. It is possible, however, to study the details of molecules without lenses, by means of diffraction experiments. Of the three types of radiation, x-rays have proved to be the most useful and fruitful for studying molecular structure.

Crystalline State

Atoms and molecules tend to organize themselves into their most favorable thermodynamic state, which under certain conditions results in their appearance as crystals. This form is characterized by a highly ordered arrangement of the molecules, associated with which is a three-dimensional periodicity. The repeating three-dimensional patterns, ideally depicted as *lattices*, are essential for x-ray structural analysis.

X-ray Diffraction

In 1912 von Laue and two of his students, Friedrich and Knipping, carried out an experiment with x-rays that opened the door to crystallographic structural analysis. They allowed a beam of nonhomogeneous x-rays to pass through a crystal of copper sulfate pentahydrate; they recorded, by means of photographic plates, the diffracted x-ray beam. A diagram of the experiment is shown in Figure 13-4.

The results showed that x-rays, which had been discovered by Roentgen less than two decades earlier, had wave characteristics (wavelength: approximately 1 Å). As a crystal is composed of a regular array of atoms with interatomic separations of the angstrom (Å) range, they were able to show that the diffraction pattern obtained on the plates was due to the crystal acting as a three-dimensional diffraction grating toward the x-rays.

This discovery led Bragg to make use of x-rays for the study of the internal structures of crystals. He considered that x-rays are reflected from planes of atoms within the crystal lattice. The reflections from a particular family of planes will occur

Table 13-9. The Hydrophobic Contribution Constants of Amino Acid Residues in Peptides and Proteins

AMINO ACID	HYDROPHOBIC CONTRIBUTION CONSTANT (faa) (measured in octanol/water, pH7)	COMMENTS
Lys (K)	- 2.43	Basic
Glu (E)	- 2.41	Acidic
Orn (O)	- 2.33	Basic
Asp (D)	- 2.32	Acidic
Arg (R)	- 1.86	Basic, guanadine group
Asn (N)	- 1.10	Monoamide group
Gln (Q)	- 1.09	Monoamide group
Ser (S)	- 0.78	Neutral, alcohol group
His (H)	- 0.54	Heterocyclic, imidazole
Gly (G)	- 0.51	Neutral
Thr (T)	- 0.50	Neutral alcohol group
Ala (A)	- 0.34	Neutral
Cys (C)	- 0.29	SH group
Pro (P)	- 0.18	Heterocyclic
Met (M)	0.43	Sulfur-containing
Val (V)	0.44	Neutral
Tyr (Y)	0.55	Aromatic, phenolic OH group
Ile (I)	0.87	Neutral
Leu (L)	0.97	Neutral
Phe (F)	1.23	Aromatic
Trp (W)	1.47	Heterocyclic, indole group

Adapted from Gao H, et al. *Pharma Res* 1995;12:1279 and Lien EJ, et al. *Prog Drug Res* 1997;48:9.

only at a particular angle of incidence and reflection. The essential condition for reflection is diagramed in Figure 13-5. In this figure the *crests* of the two incident waves will stay in phase if the thickened portion of the path (as shown in the diagram) of one wave is an integral multiple (n) of the wavelength (λ). The condition for reflection is given by the well-known Bragg equation:

$$\frac{\lambda}{2} = d_{nk,nk,nl} \sin \theta$$

The equation is satisfied only when $n = 1, 2, 3, \dots$ If n is not a whole number, there will be destructive interference between the diffracted waves.

In any crystal there are an infinite number of families of planes that can be constructed. These planes usually are denoted by their Miller indices (hkl), as shown in Figure 13-6. These indices dictate the spacing between the planes (d_{hkl}) for a particular crystal. Because the highest value of θ that is theoretically possible to measure is 90° (reflected beam comes back along the incident beam's path), the number of planes (highest order) that one is capable of orienting in a diffracting position is limited by the wavelength of the radiation.

The planes that are accessible for a particular wavelength (x-ray) can be brought into a diffracting position by the proper orientation of the crystal relative to the collimated beam. In turn, many sets of planes can be recorded on a photographic plate by the movement of the crystal, when each of the planes will come into its diffracting position. In diffraction photographs, in which the crystal has been oscillated about an axis relative to the incident radiation, the various spots on the film arise from reflections from different planes; each spot can be indexed, according to the Miller indices of the respective plane, by its location on the film. The spacing between the various spots enables one to derive the distances and angles between the primitive translations—that is, the unit-cell dimensions.

In most cases little information can be gleaned from a knowledge of the unit-cell dimensions alone. To learn about the crystal and molecular structure, it is necessary to consider the intensities of the Bragg reflections.

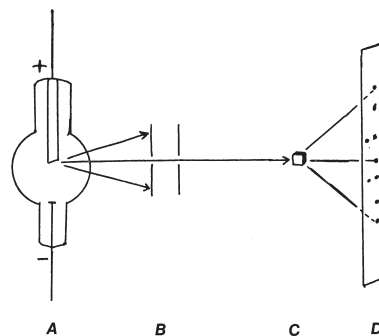


Figure 13-4. Diagram of Laue experiment:(A) x-ray tube, (B) lead slits, (C) crystal, (D) photographic plate.

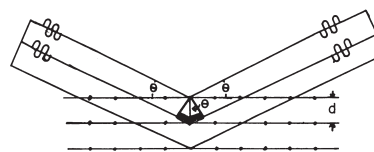


Figure 13-5. Bragg condition for reflection.

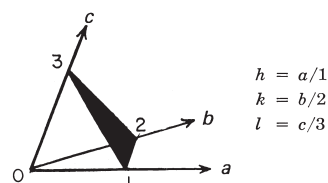


Figure 13-6. Crystal axes intercepted by a crystal plane.

Application of X-ray Diffraction

MOLECULAR WEIGHT—The measurement of the unit-cell parameters provides a means of accurately determining molecular weights of compounds. The density of a crystal can be obtained by means of flotation in mixtures of suitable liquids, the density of which may be altered by dilution until it matches that of the crystal.

The density (g/cm^3) is proportional to the molecular weight of the material in the unit cell.

The relationship is

$$\text{Mol wt} = \frac{\text{Density} \times V_{\text{cell}} \times N_a}{Z}$$

where N_a is Avogadro's number (6.023×10^{23}) and Z is the number of molecules in the unit cell. The unit-cell volume (V_{cell}) can be measured to a very high degree of accuracy. The number of molecules in the unit cell (Z) must be a whole number, with values of 1, 2, 4, and 8 being the most common among organic materials. When there is a high degree of solvation, it is necessary to approximate the amount of liquid bound by another means.

IDENTIFICATION OF MATERIALS—Every compound that is crystalline will give a characteristic x-ray diffraction pattern. These patterns can be very useful for identification purposes, and also for quantitative analysis of solid mixtures (see Chapter 34). They also have been used to a great extent by the pharmaceutical industry for the identification and classification of polymorphic and solvated forms of drugs. The *powder method*, in which the specimen is ground to a fine powder containing minute crystals oriented in every possible direction and a large number with their Bragg planes in correct orientation for reflection, is a valuable technique when quick comparisons of different forms are to be made and also when quantitative work is done. An example of such a comparison between the hydrated and anhydrous form of theophylline is shown in Figure 13-7.

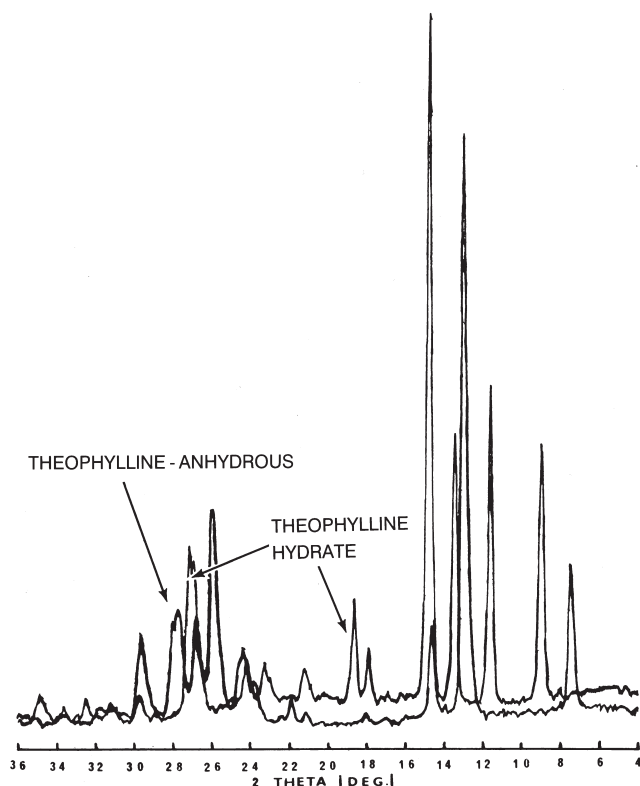


Figure 13-7. A tracing of the powder-diffraction patterns of theophylline monohydrate and an anhydrous form.

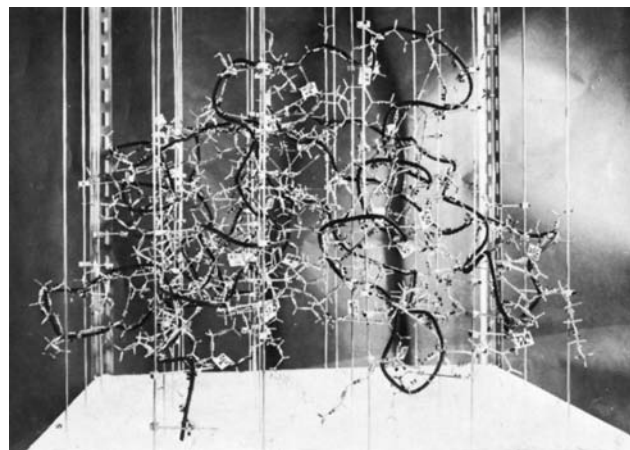


Figure 13-8. Model of bovine ribonuclease derived from x-ray data. The snakelike tube marks the backbone of the protein (courtesy, Dr G Kartha).

Extraction of quantitative information from diffraction patterns permits measurements of the physical and chemical stability of solid dosage forms. The kinetics of phase transformations are obtained easily by following the disappearance and/or appearance of various diffraction maxima corresponding to certain solid states as a function of time. One easily can visualize how this can be accomplished for theophylline hydrate by looking at the patterns in Figure 13-7.

STRUCTURE DETERMINATION—The body of substances of medicinal value whose structures were elucidated primarily by x-ray diffraction techniques is quite large. They range in molecular size from penicillin to vitamin B₁₂, and on up to the globular proteins. The structural determinations, in most instances, have played a major role in uncovering the secrets associated with the biological functions of the various molecules. A photograph of the ribonuclease molecule as determined by the x-ray studies of Kartha, Bello, and Harker is shown in Figure 13-8. This enzyme catalyzes the hydrolysis of phosphodiester bonds in RNA chains.

There also are large numbers of macromolecules of biological importance that do not form three-dimensional crystals in the usual sense, but will form fibers. The bundles of molecules in the fiber are aligned with respect to one another in a somewhat crystalline manner. These materials give x-ray diffraction patterns that have proved very useful in deriving molecular information. By fitting models to the x-ray pattern, many valuable biological polymers have had their secrets exposed. The two best examples are the α -helices of keratin and the double helix of deoxyribonucleic acid.

In recent years x-ray studies have been coupled with computer graphic and quantitative structure-activity relationship (QSAR) approaches in computer-assisted drug design (CADD) (see Chapter 28 for a more detailed discussion).

INTRAMOLECULAR BONDING AND CONFIGURATIONS—The precise determination of a crystal structure enables the bond lengths and angles between the various atoms to be determined accurately. This information is extremely valuable in the further understanding of how various chemical substituents influence the valence states and configurations of a molecule. With such knowledge, structure-activity relationships, which are of fundamental interest to the medicinal chemist, have much more depth. The observed bond orders also serve as experimental criteria by which theoretical models can be judged. It also is possible to compare quantum mechanical calculations relating drug interaction with actual observation.

Intramolecular steric effects, which tend to distort molecules, are unraveled easily by the scrutiny of their structures. It is possible to distinguish between repulsive and attractive effects of substituents. The torsional angles about var-

ious bonds can be calculated from the atomic positions and are extremely helpful in correlating NMR data to structure.

In recent years the combination of x-ray and neutron-diffraction studies has enabled information on the bonding and non-bonding electrons within a molecule to be delineated clearly. Neutron-diffraction experiments enable atomic nuclei in a crystal to be positioned accurately; on the other hand, x-rays locate the electron clouds. Both types of data can be combined to calculate three-dimensional electron density maps with the inner-core electrons around each atom subtracted; this makes the unshared pairs and bonding electrons clearly visible. The atomic positions derived from neutron data are used for phases in calculating electron density maps with the x-ray data.

Refer to Chapter 34 for additional information on the physical methods discussed in this chapter.

STATES OF MATTER

The aim of this section is to discuss both generalities and specifics, most of which are not related explicitly to dosage forms, because the latter will be discussed in other chapters. Some of the principles should be useful to have in mind when dosage forms and their manufacture and processing are studied by the product-development pharmacist. It should be noted that due to the range of subjects covered by the section title it was necessary to take an eclectic approach in developing mostly qualitative discussions. The goal has been not to produce a difficult, in-depth section, but rather one that presents a mostly macroscopic overview of the significant states of matter.

Normally, matter exists in one of three states: solid, liquid, or gas. Although it is not pharmaceutically important, two other states of matter exist: the plasma state, in which matter exists as a hot gaseous cloud of atoms and electrons; and a more speculative state, possibly having only a momentary existence, is one which has characteristics of a superdense supermetal. The latter transient state is produced when material is subjected to very high pressures such as those used to make diamonds when compressing graphite.

To avoid the pitfalls of semantics, there is no need to call attention to other systems of classification, because for all practical purposes it is convenient to think only of the three most obvious states. These states are actually a continuum, with two common factors determining the position on the *scale of states*.

The first factor is the *intensity of intermolecular forces* of all kinds: solids have the strongest forces, and gases have the weakest. The other common factor is *temperature*. Obviously, as the temperature of a substance is raised, it tends to pass from a solid to a liquid to a gas. When the phrase “as temperature is increased” is used, it should be remembered that this is a relative phrase. Even at what is called room temperature, some of the effects of a temperature increase are present because room temperature is far above absolute zero.

SOLVATES AND HYDRATES—During the process of crystallization, some compounds have a tendency to trap a fixed molar ratio of solvent molecules in the crystalline (solid) state. These are called *solvates*. When water is used as the solvent, *hydrates* may be formed. Some recent pharmaceutical examples include gallium nitrate ($\text{Ga}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$) and nafarelin acetate, where each decapeptide contains 1–2 molecules of acetic acid and 2–8 molecules of water.

As a point of historical interest, note that Lavoisier, the great “father of modern chemistry,” thought of heat as a type of matter; the view even as late as the 18th century was that the three states of aggregation differ only with respect to how much heat they contain. Thus, although not all are satisfied with this phraseology, the term *enthalpy* (or *heat content*) is still used in thermodynamics.

Thinking further back to the ancient Greek philosophers and their original four elements (earth, air, fire, and water), note again the great significance attached to heat. Although the ancient philosophers’ concepts of the nature of matter were not

correct, they did recognize heat as an integral part of the scheme of things, and nothing could be truer. Heat, a vital form of energy, the mirror of molecular motion, is *the* form of energy of greatest importance to mankind.

As alluded to above, there is no clear line of demarcation between the states of matter, but the following arbitrary division may make the approach this section takes more coherent.

Changes of State

As a solid becomes a liquid and then a gas, heat is absorbed and the *enthalpy* (*heat content*) increases as the material passes through these phase changes. Thus, the enthalpy of a liquid is greater than that of its solid form, and the enthalpy of a gas is greater than that of its liquid form, because heat is absorbed when melting and vaporization occur. The *entropy* (a measure of the degree of total molecular randomness) also increases as materials go from solid to liquid to gas.

It is the balance of enthalpy, entropy, and temperature that determines if changes proceed spontaneously. Obviously, if systems tend to settle to states of lowest energy, it means that enthalpy and entropy considerations may counteract each other. Much of thermodynamics is concerned with explaining and quantitating the changes that systems undergo.

Latent heat is heat absorbed when a change of state takes place without a temperature change, as when ice turns to water at 0°. This example is one in which the heat required to produce the change of state is designated the *heat of fusion*. The counterpart, the *heat of vaporization*, is used when a change of state from liquid to gas is involved.

As molecules of a liquid in a closed, evacuated container continually leave the surface and go into the free space above it, some molecules return to the surface, depending on their concentration in the vapor. Ultimately, a condition of *equilibrium* is established, and the rate of escape equals the rate of return. The vapor then is saturated and the pressure is known as the *vapor pressure*.

Vapor pressure depends on the temperature, but not on the amounts of liquid and vapor, so long as equilibrium is established and both liquid and vapor are present. Heat is absorbed in the vaporization process, and therefore the vapor pressure increases with temperature. As the temperature is raised further, the density of the vapor increases, and that of the liquid decreases. Ultimately, the densities equal each other and liquid and vapor cannot be distinguished. The temperature at which this happens is called the *critical temperature*, and above it there can be no liquid phase.

A very important process that involves a change of state from liquid to vapor and back to liquid is that of *distillation*.

Solids also have vapor pressures that depend on temperature. When a solid is converted directly into gas, it is said to *sublime*. Sublimation pressures of solids are much lower than those of liquids at any given temperature. When a solid is transformed directly into a liquid, two types of melting may be distinguished. In the first type, *crystalline melting*, a rigid solid becomes a liquid, during which procedure two phases are present—the bulk of the solid or its inner parts are not really changing. The second type is *amorphous melting*. This involves an intermediate plastic-like condition that envelops the whole mass; the viscosity decreases and a state of liquidity follows. Crystalline melting involves more definite melting points and latent heats than does amorphous melting.

Sublimation

All solids have some tendency to pass directly into the vapor state. At a given temperature each solid has a definite, though generally small, vapor pressure; the latter increases with a rise in temperature. *Sublimation* is the term applied to the process of transforming a solid to vapor without intermediate passage through the liquid state. In pharmaceutical manufacturing the

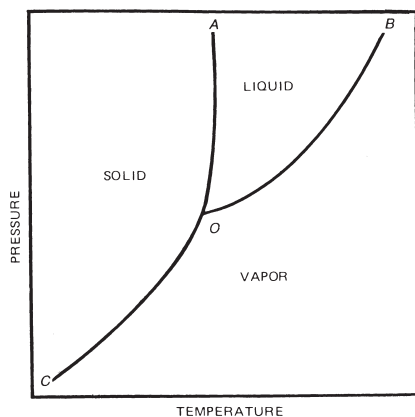


Figure 13-9. Phase diagram to illustrate the principle of sublimation.

process commonly includes also the condensation of the vapor back to the solid state.

A solid sublimes only when the pressure of its vapor is below that of the triple point for that substance. The *triple point* is the point, having a definite pressure and temperature, at which the solid, liquid, and vapor phases of a chemical entity are able to coexist indefinitely. If the pressure of vapor over the solid is above that of the triple point, the liquid phase will be produced before transformation to vapor can proceed.

Figure 13-9 depicts a phase diagram illustrating the principle involved. The line *OA* indicates the melting point of a substance at various pressures; only along this line can both solid and liquid forms exist together in equilibrium. To the left only the solid form is stable; to the right only the liquid form remains permanently. The line *OB* shows the vapor pressure of the liquid form of the substance at various temperatures. It is called the *vapor-pressure curve* of the liquid and represents the conditions of temperature and vapor pressure for coexistence of liquid and vapor phases. Above this line only the liquid phase exists permanently; below it only vapor occurs. The line *OC* represents the vapor pressure of the solid at various temperatures. It is designated as the *sublimation curve* of the solid and represents the conditions of temperature and vapor pressure for the coexistence of solid and vapor phases. To the left of this line only solid can exist; to the right only the vapor form is stable. The intersection of the three lines, point *O*, is the triple point. It is apparent from the diagram that at pressures of vapor below that of the triple point it is possible to pass directly from the vapor to the solid state, and vice versa, simply by changing the temperature.

At pressures above the triple point the liquid phase must intervene in transformations between solid and vapor phases, in a closed system. Because the melting point of a solid commonly is taken at 1 atm (atmosphere) of pressure, it is evident that if the triple-point pressure is less than 1 atm, fusion of the solid form will occur on heating in a closed vessel. If, on the other hand, the triple-point pressure is greater than 1 atm, the solid form cannot be melted by heating at atmospheric pressure.

In a current of air, however, the conditions are somewhat different; some solids that melt when heated in a closed system now sublime appreciably even at ordinary temperatures, because the vapor pressure of the solid does not attain the triple-point pressure. Thus, camphor, naphthalene, *p*-dichlorobenzene, and iodine, all of which have a triple-point pressure below 1 atm, will vaporize in a current of air but melt when heated in a closed system.

Critical Point

The critical point is expressed as a certain value of temperature or pressure (or molar volume) above which or below which cer-

tain physical changes will not take place or certain states of being will not exist. At these points, some properties are constant and are referred to as the critical temperature, pressure, or volume. At the usual critical point, the properties of liquid and gas are identical and the phase diagram curve of *P* versus *T* ends. (Phase diagrams will be discussed later.) When a liquid changes to a vapor, increased disorder or randomness—and therefore increased entropy—results. At the critical temperature, the entropy of vaporization is zero, as is the enthalpy of vaporization, as the gas and liquid are indistinguishable.

Although the gas–liquid critical point is the one most discussed, others do occur. Each critical point marks the disappearance of a state. Note that most liquids behave similarly not only at their critical temperatures, but also at equal fractions of their critical temperatures. For example, the normal boiling points of many liquids are approximately equal fractions (about 60%) of their critical temperatures (in absolute temperature degrees).

Supercritical Fluid

Over the last decade, supercritical fluid chromatography (SFC) and related unified chromatography techniques continue to grow, especially in food and natural products extractions and analysis.^{25,26}

When the temperature and pressure of a liquid go beyond the critical points, a *supercritical fluid* may form. Under these stressed conditions, polar and nonpolar compounds are completely miscible. For example, dense fluid solvents, like supercritical CO₂ ($T_c = 31.1^\circ$, $P_c = 73.8$ bar) and ethane ($T_c = 32.3^\circ$, $P_c = 48.8$ bar) have been shown to offer advantages for the solubilization of amino acids. Other applications of supercritical fluids include chromatography of polar drugs and elimination of toxic wastes.^{27,28}

Visualization of Changes of State

This section is to serve as an introduction to the following one on eutectics. When a pure substance cools and is transformed from a liquid to a solid, a graph (Fig 13-10) of decreasing temperature versus time is continuous. At the temperature at which solid crystallizes (ie, the *melting point*), the cooling curve becomes horizontal. The same is true at the *boiling point*—the temperature of a liquid at which the continuing application of heat no longer raises the temperature, but rather converts the liquid into vapor. It is the point where the vapor pressure of the liquid (or the sum of its components) equals that of the atmosphere above the liquid.

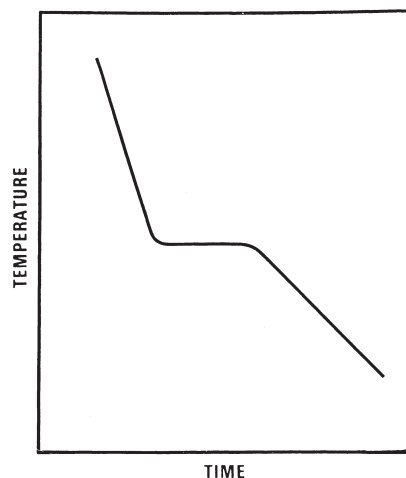


Figure 13-10. A single change of state as shown by a slowing of the cooling rate.

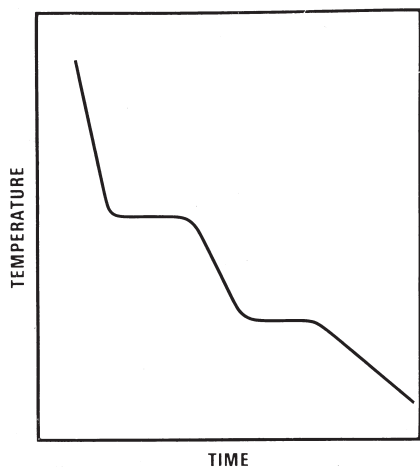


Figure 13-11. Two changes of state with resulting temporary decreases in cooling rate.

Increasing the pressure above the liquid or adding solutes raises the boiling point and *vice versa*. These plateaus observed at certain specific temperatures are due to the release of the heats of fusion or vaporization. Similarly, when solutions are cooled, the slope of the *cooling curve* (Fig 3-11) changes when one of the components starts to crystallize. Although a truly horizontal plateau may not be formed, as in the case of pure materials, the change in slope indicates precipitation of one of the components. If the same plateaus are formed when binary solutions of varied composition are cooled, it indicates that both components of the binary solution are coming out together. The temperature at which this occurs is the *eutectic temperature*, and the composition is generally called a *eutectic*.

Normally, cooling curves per se are converted to phase diagrams to facilitate visualization of the interrelationships as phase changes take place. If, instead of a minimum point or eutectic, a maximum point is observed, it may indicate that the components are reacting to form a solid compound that can exist in equilibrium with the melt over a range of compositions.

It is undoubtedly true that many unknown phase equilibria exist. Thus, when conditions are changed (eg, when a process is scaled up in a manufacturing process), different phase changes may take place and produce different final products. The pharmaceutical use of heterogeneous materials such as waxes and fats certainly provides ample opportunity for these changes to occur.

Eutectics

Although many very complex and complicated diagrams, including some three-dimensional models, are needed to characterize certain systems, most interesting to pharmacy are the diagrams (Fig 13-12) indicating eutectic formation. This section will only briefly describe this area of technology.

Phase diagrams are constructed by determining the melting points and cooling rates of a series of binary liquid solutions of compositions varying from pure A to pure B. This will be illustrated shortly—first, consider the Figure 13-12 phase diagram. The points where the V-shaped boundary of the melt intersect the right and left vertical axes are the melting points of the pure materials. To the left of the base of the V (ie, when solutions rich in A are cooled) solid A separates as the temperature falls; to the right, solid B separates as shown. Thus, the left arm of the V is the curve that represents the temperature conditions under which various liquid mixtures are in equilibrium with solid A, and the right arm of the V is that curve that shows which mixtures are in equilibrium with solid B.

At the point of the V, both solid A and solid B are in equilibrium with the liquid; this point, the lowest temperature at which any of the infinite possible combinations of liquid solutions of A and B will freeze (or the lowest melting point of any possible mixture of solids A and B) is called the *eutectic point*. Only at this point is the composition of the solid the same as that of the solution from which it is separating; this does *not* mean necessarily that the composition of the eutectic is a chemical compound of A and B. Thus, at the eutectic point, both A and B come out together in a constant proportion.

The eutectic composition is a simple two-phase mixture, but when made in situ it has a very fine-grained structure that could impart to it different properties (eg, solubility or gastrointestinal absorption rate), compared to a gross mixture of the same composition. The structure is very fine-grained because the crystallization was very intimate, because crystals of both phases were formed simultaneously. This is quite a different situation than one in which only one component is separating. It is important to remember that one can be only at one place on a phase diagram at any one time; that is, the diagram describes what a *particular* system is like at a certain temperature, which components are in the liquid and/or solid state, and the proportions of each.

The diagrams are constructed from information obtained on the cooling rates of binary solutions. Consider again a cooling curve analysis in which temperature versus time are plotted. The curves change slope to form plateaus when any solid phase separates; the plateaus tend to become more horizontal as absolute temperatures are lower because the intensity of radiation and conduction is lessened. A final plateau results when the whole liquid mass (or the last of it) solidifies. Thus, if a molten liquid having a composition lying *between*, for example, pure A and the eutectic were cooled, the following would be observed in a plot of temperature versus time (see Fig 13-11).

First, *T* drops with time; then solid A will come out of solution, release its heat of fusion, and thus slow the cooling rate to produce the first (upper) plateau. The temperature then starts to drop more sharply again as enough A comes out of solution, and the system changes composition until it contains only the eutectic composition.

When the eutectic composition is reached, the second solid (B) also coprecipitates, and the temperature remains constant (lower plateau) until all of A and B have solidified, after which, of course, the temperature will be able to drop further.

If the system being cooled started as the eutectic composition, only the lower break and plateau would be observed; that is, a pure material and a eutectic would have similarly shaped cooling diagrams.

Note then that a phase diagram can be constructed by studying a number of cooling curves made on a series of mixtures of known composition. To do this, the temperatures at which cool-

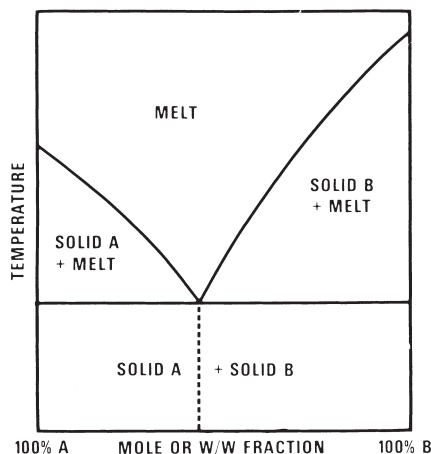


Figure 13-12. Simple phase diagram of system showing eutectic point.

ing-rate changes are plotted against each particular composition studied.

Note that Figure 13-12 is idealized in that no solid–solid solution of A and B is formed. If the two components are somewhat soluble in each other, the diagram would differ by having two thin solution areas along the left and right axes; such are partly in evidence in Figure 13-13.

Two pharmaceutical examples of eutectic formation are

1. A mixture of two common antipyretic-analgesic compounds: aspirin and acetaminophen. There always has been some “magic” associated with eutectic formation; indeed, as such a binary composition does melt at a lower temperature than other combinations, the eutectic probably does have weaker bonding forces, if any. And, being very fine grained, it dissolves more rapidly. It is known that many drug compounds form eutectics, and the aspirin-acetaminophen (APAP) eutectic (37% APAP by weight) does dissolve more quickly than a simple mixture of the two of the same composition. Because a formed eutectic is created under equilibrium conditions of intimate mixing as noted, the contact of the two compounds is much closer than that achievable by simply mixing the dry powders. The increase in dissolution rate obtained by using the eutectic may result in a greater speed of physiological absorption.
2. This example is illustrated in Figure 13-13. It was found that urea and acetaminophen formed a eutectic containing approximately 46% urea and 54% acetaminophen (by weight) which melted in the 110° to 115° range.

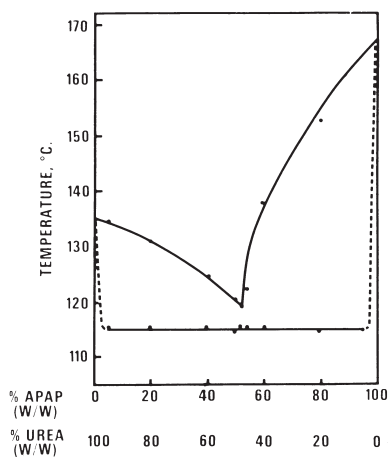


Figure 13-13. Phase diagram of the urea:acetaminophen (APAP) (46%:54%) eutectic melting in the 110 to 115° range. (From Goldberg AH, et al. *J Pharm Sci* 1966; 55:482.)

Gases

AEROSOLS—Gases are used directly in dosage forms in the field of aerosols. Although this subject, including the use of the so-called liquefied propellants, is covered elsewhere, note that pressure packs often use nitrogen, nitrous oxide, or carbon dioxide to expel the contents from their containers. The latter two gases are much more soluble in water, so some aeration (which may be desired) of the material discharged will take place.

Carbon dioxide is about six times as soluble in water as nitrogen, and nitrous oxide about four times as soluble as nitrogen. Thus, if it is desired to have some of the gas dissolved in the product, either nitrous oxide or carbon dioxide can be used. In organic solvents and in fatty materials, such as found in emulsions, nitrous oxide is somewhat more soluble than carbon dioxide. There is not a great deal of difference in solubility properties; however, the possibility exists that the pH-lowering effect of carbon dioxide as it forms carbonic acid may be just as undesirable, as it may cause precipitation of a carbonate in an alkaline product.

INHALERS—Inhalers are classified as being one of two types, surface or solution.

Surface-Type Inhaler—The volatile material resides on the surface of the pledget (cotton or other cellulosic material, usually). This represents a conventional adsorption situation; it is easy to appreciate the fact that the more surface area the pledget has, the greater the surface area of the material exposed to the airflow and the greater the opportunity for volatilization. Hence, a larger or more loosely packed pledget will cause a larger dose to emanate from an inhaler than will a smaller or tightly packed pledget.

It is convenient to make this type of inhaler if the volatile material itself is a liquid. The doses produced stay relatively high because the pledget charge is being depleted according to a zero-order scheme. This is reasonable because the volatile material has formed a multimolecular (as distinguished from a monomolecular) layer on the pledget surfaces. Thus, even though molecules are stripped off, the surface area—and hence the dose—remain essentially unchanged. However, as some areas of the pledget are denuded, the total exposed surface area of the volatile material decreases and so does the dose during successive uses.

Solution-Type Inhaler—The volatile material is dissolved in a suitable nonvolatile solvent, and this solution is placed on the pledget. The situation may be taken as an example of the operation of Raoult’s and Henry’s laws; that is, the vapor pressures of the components are proportional in some way to their concentrations. To keep the vapor-pressure contribution of the solvent low in order to enhance the vapor pressure of the solute, a solvent of very low vapor pressure is used as the vehicle.

In this inhaler type the exposed surface area of the material does not change as the inhaler is used; what does change is the concentration of the volatile material in the solvent. Thus, the dose gradually decreases according to a first-order scheme as the drug concentration decreases. Of course, the nature of the pledget and the inhaler body exert some effect here also, because if the airflow through the inhaler and the pledget does not permit volatilization of the material, insignificant, low doses will result.

If the drug is a volatile solid, the solution-type inhaler should be made because solids do not lend themselves to easy pledget-charging procedures even if a volatile solvent such as ether is used to deliver the material to the pledget during manufacturing.

Further amplification and clarification of the surface- and solution-type classification of inhalers might be achieved by considering the existing analogy to chromatographic systems. The surface-type inhaler corresponds to adsorption chromatography, with the material being adsorbed initially on a carrier and then desorbed by a passing stream of liquid or gas. The solution-type inhaler corresponds to partition chromatography, in which material in a solvent is supported by some medium, is partitioned between its original solvent and a passing stream of gas or liquid, and thus is removed.

Another point of significance concerns the relationship of the volatile active ingredient to the solvent. An increase in dose should result when the active ingredient is dissolved in solvents that cause it to deviate more positively from Raoult’s law. Thus, the less the solute–solvent interaction and the greater the solute–solute interaction, the more pronounced will be the tendency toward volatilization of the solute. Using relative solubility as a gauge of such interaction, one would expect delivery of larger doses of a volatile solid from dibutyl phthalate (if the solute was less soluble in it) than from benzyl salicylate (if it was more soluble in it) at the same concentrations.

Although it might seem that the vapor pressure of the drug and additives would assume a position of primary importance, this does not appear to be the case. Vapor-pressure values represent an equilibrium situation, whereas what is involved in the inhaler situation is a process controlled by factors affecting rates of volatilization.

Although it is true that volatile materials usually have appreciable vapor pressures, it generally is not true that a compound with a vapor pressure value of twice that of another compound will volatilize twice as fast. Besides this fact, inhaler recovery times may be essentially zero and no equilibration time may be needed. Also, no decrease in dosage would be noted with the surface-type inhaler and no regular (ie, linear with concentration) decrease in dose would be noted with the solution-type inhaler if the vapor pressure was the controlling factor.

Unfortunately (from the standpoint of not having a more straightforward system to analyze), equilibrium and rate concepts are inextricably mixed in the present situation. This easily can lead to the basically incorrect tendency to try to predict kinetic data from thermodynamic values. However, because vaporization relatively is unencumbered with entropy and orientation factors, rates of volatilization often are qualitatively proportional to the equilibrium properties of the materials involved.

Equimolar quantities of the following compounds, allowed to evaporate at room temperature under the same conditions, will complete the evaporation process in this order: ether, acetone, chloroform, carbon tetrachloride, ethyl acetate, and water. This order corresponds both to the vapor pressures of the materials and their boiling points.

To further cloud the cause-and-effect relationship, the very magnitude of the numbers (the concentrations in mole fractions) is such that the partial vapor pressure of a volatile solid may increase proportionately with the mole fraction. Hence, although vapor-pressure concepts should not be neglected in in-haler development, it is the rates of volatilization that must be controlled or modified. For more information and experimental data on inhalers see Kennon and Gulesich.³⁰ Various drug-delivery systems for use with metered-dose inhalers (MDIs) are commercially available. They are intended for delivering oral aerosolized medication from MDIs to the lungs.

RELATIVE HUMIDITY—In the production of effervescent products, one of the most vital factors to be considered is the use of controlled-humidity conditions. It is well-known that the effective control of humidity is related closely to the success or failure of attempts to produce effervescent products.

It is useful to bring to light some of the facets of this area of technology. Two factors predominate when one views the situation: the effective concentration of water in the air and the temperature. In chemical reactions, particularly the kind involved here, the effect of temperature on an equilibrium condition is not very significant when compared to the influence manifested by concentration. Certainly, water of hydration, crystallization, or simple adsorption (which is tenaciously held at room temperature) does not disappear at temperatures under 100°F. What is effective and influential, however, in keeping and increasing such additional moisture on solids, is the *concentration* of water in the air.

The concept proposed here is that considerations based purely on relative humidity probably will be unfruitful. For purposes of illustration, Table 13-10 shows the amounts of water that are found under conditions encountered during the development of effervescent products. The following points may be drawn from this information. A 10% relative humidity (RH) at 36°F is equivalent to 25% RH at room temperature. Either of these conditions represents a fairly dry day, but certainly not a very dry day. Therefore, although heating the air surely lowers the RH, it probably does not lower the ability of the water in the air to cause trouble. Regardless of the temperature of the processing rooms, experience has shown that for water concentrations present at 72°F, the range of 10 to 15% RH should not be exceeded if minimum difficulties are desired.

Liquids

The liquid state may be considered an intermediate in the phase transitions from solid to gas. Liquids have neither the

strong cohesive forces of solids nor the weak ones of gases. They are also intermediate in that they have neither the orderliness of a crystal nor the randomness of a gas. One then might consider a liquid a highly compressed gas or slightly released solid.

Due to the concept of molecular motion, there must be some free space in liquids. Also, if the motion is completely random, some spaces may be larger than others at a particular point in time. Thus, liquids may have holes, and this concept has explained phenomena such as the expansion of volume that materials undergo upon fusion (holes are created), diffusion in liquids, viscosity (movement of holes in the opposite direction of the viscous flow), and density decreases as temperature rises (the solubility of holes increases). It might be said that liquids are solutions of holes in material, whereas gases are solutions of matter in free space.

With respect to fluid mechanics, a fluid can be considered a material that cannot sustain shear forces when in static equilibrium. This is the factor distinguishing solids from fluids, the latter of which may be gases or liquids. This movement under the slightest stress sometimes is referred to as “no sideways friction.” It can be seen in operation in the case where a sailor standing watch near the gangplank of a docked ship can step on a mooring rope and cause the ship to move toward the dock.

Liquids, just like gases, take the shape of their container, but only the lower part of it, as the liquid occupies a definite volume; gases, on the other hand, expand to fill their entire container. Intermolecular spaces are greater in a gas than in a liquid, thus they can be compressed. Relative to gases, both liquids and solids are quite incompressible. They can be considered already compressed due to the stronger intermolecular forces.

After a fluid is set in motion, it comes to rest because of the internal friction caused by the molecules sliding over each other; this resistance to flow is called *viscosity*, and it can be quantified. To effect good quantification with viscometers, a normal, smooth (laminar or layer) flow is needed. With excessive stirring, at a so-called *critical velocity*, the fluid becomes turbulent, and instrumental measurements are difficult to effect. As the temperature of fluids increases, viscosity decreases. In general, also, as pressure increases, viscosity increases.

Because fluids have some structure, they may change upon standing so that, when one is considering viscous behavior, the recent past history of the sample may have great effects. *Thixotropy* is the term used for liquids that flow freely if recently stirred, but gel when left undisturbed. Solids also flow, but more slowly, even under minor stresses including those produced by their own weight. The wavy, bumpy surface of tarred roads, particularly seen on hills, is a result of a flow phenomenon.

Of interest also is the *cluster* theory of liquids, the main concept being that localized order exists but does not extend to a great distance. One property explained by this visualization is that, as the temperature rises, the clusters disintegrate and viscosity decreases. Another is that transmitting momentum through a liquid is due not only to molecular movement, but also to the transmissions of elastic waves through the groups of semistationary clusters. It is possible that the cluster theory affords another way of looking at pharmaceutical complexes in solution.

Complexes

In addition to structure in solvents, it also is possible for solutes to create a structure of a sort within the solvent. Thus, it has been shown that benzocaine in water solution with caffeine exhibits a much-reduced rate of hydrolysis. In a somewhat similar vein, it also has been noted that different salts of the same compound (eg, hydrochloride versus nitrate) may exhibit different stability characteristics. Similarly, it has been shown that saccharin in certain chlorpromazine hydrochloride solutions enhances the light-stability of the drug. It appears that such changes are because the ionic environment may form a protective molecular overcoat or loose ionic atmosphere complex around the drug.

Table 13-10. Moisture Content (g/m³) Existing at the Conditions Noted

TEMPERATURE	RELATIVE HUMIDITY (%)			
	10	15	25	40
RT (22°C or 72°F)	1.9	2.9	4.8	7.7
Hot (36°C or 97°F)	4.1	6.2	10.3	16.5

Liquid Crystals

Lipids, when heated, usually do not pass directly from a crystalline to an isotropic structure, but rather assume intermediate liquid crystal phases. Of most interest pharmaceutically and physiologically is the concept that these structures are undoubtedly involved intimately in the structure, and hence in the function, of membranes and cells.

All biological systems are basically aqueous, and it is particularly in such systems that lyotropic mesomorphism (the formation of liquid-crystal phases in the presence of water) takes place; that is, the lipid phases undergo transformations involving crystal, liquid-crystal, and liquid forms. It is these changes that are mediators of the various physiological absorption, transport, storage, and excretory functions of cells. Many *in vitro* studies of biologically significant lipids have been performed in an attempt to elucidate the mechanisms of their interaction and behavioral properties in aqueous systems.

Liquid-crystals differ from solids and gases in that they have some freedom to move and to take on many different shapes while maintaining a high degree of order through quite long distances, relatively speaking. In the laboratory, liquid-crystals can be prepared from one component by heat treatment (thermotropic systems) or from one or more components by adding controlled amounts of water or other polar solvents (lyotropic mesomorphism). Note that the only molecules of significance here are asymmetric and have a definite long direction, so their tridimensional orientation is essential. This should be remembered throughout the discussion.

For present purposes, three types of liquid-crystal phases will be described briefly so that at least some appreciation for this particular state of matter may be gained. The phases generally are characterized as being nematic, smectic, or cholesteric.

Nematic Phase—Nematic molecules (Fig 13-14) are set in parallel arrangements and have restricted rotation about at least one axis. The molecules are parallel or nearly so. One might picture this as a long box filled with pencils with the latter being able to roll. Overall, the system might be considered to be thread- or cable-like. Another picture would be that of a group of logs going through a pipe. There is overlap of the pencils or logs somewhat, as there is with cars in an auto race.

Smectic Phase—The smectic or “two-dimensional” crystal (Fig 13-15) has its molecules arranged in layers with their long axes essentially normal (ie, at right angles) to the plane of the layers. Their centers of gravity are then mobile in two directions in their plane, and the molecules can rotate about one axis. Overall, one could consider the arrangement layer-like, with the degree of order just described in each layer.

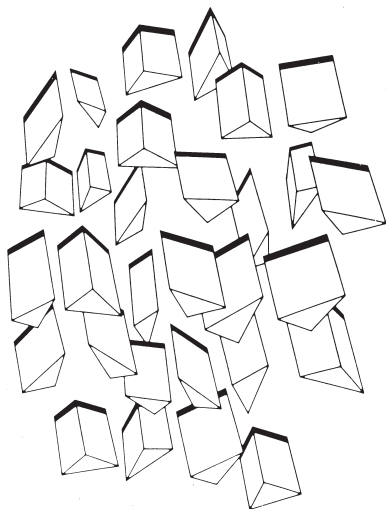


Figure 13-14. The nematic phase of a liquid crystal. (Adapted from Ferguson JL, Brown GH. *J Am Oil Chem Soc* 1968; 45:120.)

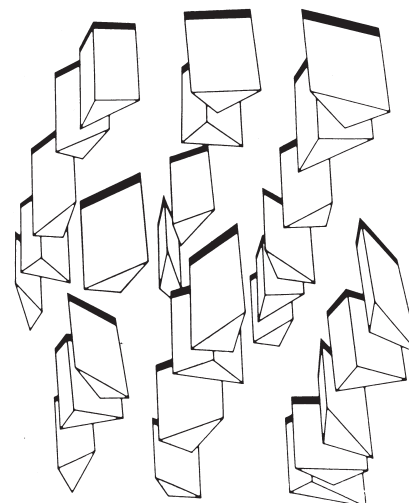


Figure 13-15. The smectic liquid-crystal phase. (Adapted from Ferguson JL, Brown GH. *J Am Oil Chem Soc* 1968; 45:120.)

The smectic arrangement is similar to the nematic in that there is still essentially only one axis of rotation, except in this case there is no overlap. The logs go through the pipe as a member of a group—it would be like a series of drag races in which no one wins and all are tied. Each successive group, however, does not follow the same paths as the others; within any one group there may, or may not, be equal spacings sideways between the long axes. Note also that the thickness of the layers is about the same as the length of the molecules.

Cholesteric Phase—The cholesteric arrangement (Fig 13-16) is to some extent a combination of the nematic and smectic; the layers are nematic, but in addition certain layering formations that resemble the smectic phase are incorporated. In essence, the result is a helical, twisting repetition of the nematic phase that, corkscrew-like, slowly changes head direction (eg, the lead end of the pencil) as one proceeds to examine underlying layers of molecules. The cholesteric arrangement is, *in toto*, much thicker than a smectic layer.

All three structures are involved in building cells, and each type can (when viewed totally) form curved surfaces, membranes, or any other required micelle-like shapes. Some researchers have constructed cell models using these structures and have shown how the mechanics of many cellular functions can be visualized using the known properties of liquid crystals.

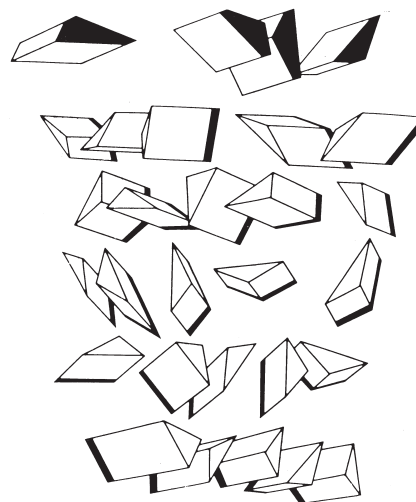


Figure 13-16. A 180° turn of the molecules in the cholesteric liquid-crystal phase. (Adapted from Ferguson JL, Brown GH. *J Am Oil Chem Soc* 1968; 45:120.)

The Glassy State

Although glass usually is thought of as a specific, nonconducting, transparent solid, it actually is a type of solid matter. It can be considered neither a typical solid nor liquid. The atoms of most solid states generally are strictly ordered structurally, whereas glassy materials are highly disordered. Glasses may, however, have some short-range order, just as do polymers. Another characteristic of glasses is that they do not have specific melting points, but rather slowly and gradually become liquids when they are heated. Sometimes glasses are considered supercooled liquids, but this is not strictly accurate.

A graph of volume versus temperature for most substances shows that the volume of a liquid decreases as the crystallization temperature is approached. If solidification is accomplished by crystallization, the volume decreases sharply at the freezing point, after which it continues to decrease gradually depending on its coefficient of thermal expansion. This type of behavior is not exhibited when solidification is followed by glass formation.

The uniqueness of the glassy state is evident in its cooling curve. As indicated in Figure 13-17, as a glass-former is cooled, it does not suddenly undergo a large drop in volume (or density, or index of refraction) at any particular temperature or as it passes through the melting point, nor does its volume decrease as rapidly as that of a supercooled liquid, although it follows the curve of the latter initially during cooling. With supercooled liquids, the cooling curve is a simple continuation of the liquid curve itself, with no melting or transition points.

Atomically, the structure of the glassy state is marked by a random selection of polyhedral molecules considered to be linked together at their corners. Certain materials are easy to cast into a glassy state, others can be made glassy with great difficulty, and some seemingly not at all. At present there seems to be no specific theory to help predict this behavior. Materials that do form glasses appear, however, to have a very high viscosity at their melting point; this inhibits the formation of an ordered structure. In addition, non-glass-formers tend to exhibit large energy differences between the ordered form of the solid and the disordered liquid. Thus, the low-energy, ordered form of the solid tends to be developed. Obviously, the energetic tendencies here are balanced by entropy factors, which tend to favor states of minimum order.

Although the most well-known glass-formers are the metal oxides, many other materials can exist in the glassy state; even steel can be so cast if it is cooled very, very quickly. This technique produces glasses as the materials become solid before they have a chance to develop a crystalline structure. With regard to crystal formation, note that, in a crystallization process,

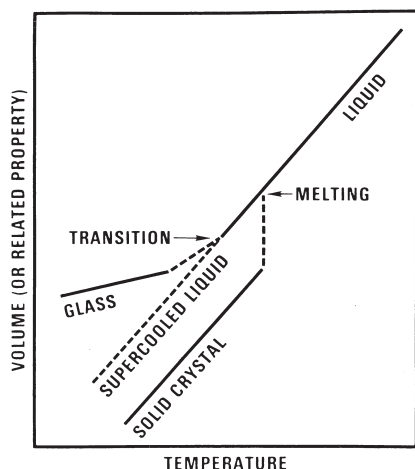


Figure 13-17. Composite cooling curve of liquids forming glass, supercooled liquid and solid-crystal states.

when concentrated solutions of the material to be crystallized are cooled slowly, larger and more perfect crystals form.

Incomplete or imperfect crystallization, whether due to technique or to the nature of the material itself (eg, natural and synthetic high polymers), often causes the formation of crystallites, glasses, or liquid crystals. Crystallites have no recognizable regular crystal pattern; rather, they are, in a sense, incipient crystals. Many shapes and arrangements are possible such as globular, rows or clouds of globules, threads, cylinders, or rods.

Solids

The most significant physical property of the solid state is the high degree of order in which substances such as metals and minerals exist. The structure may be crystalline and lattice-like or noncrystalline, such as in plastic, glass, or gels, which are not lattice-like or only partly so. These latter materials do have much more order than liquids and gases. These materials also have, in varying degrees, some plastic and elastic properties, wherein some resistance to applied stresses exists, but when the stress reaches a certain intensity either flow or fracture ensues.

Although different classifications exist, four major different types of bonds hold solids together; the strong bonds impart higher melting points to substances. In order of decreasing strength, the bond types are *metallic*, *ionic* (salts), *valence* (diamond), and *molecular* (many organic compounds). Thus, in some solids, the atoms or molecules or ions may be arranged in a regularly repeating pattern (crystalline state), whereas other solids are considered noncrystalline or amorphous if they do not have this characteristic of regularity. There is some blurring of the division, but in general, metals, minerals, rocks, and alloys are examples of the former class; glass, wood, ceramics, and plastics are examples of the latter.

Alloys are an example of a mixed solid having characteristics of regularity but being intermediate between the strictly crystalline and amorphous states. They are metal substances consisting of two or more elements, not counting the trace amounts of materials which make any element less than 100% pure. Alloys are solid solutions of one of two types. In the *interstitial* type, the smaller solute atoms occupy the interstices between the solvent atoms; the overall structure is quite like the parent or solvent metal. In the other type, *substitutional*, all atoms occupy (ie, contribute to building) a common lattice.

In general, alloys are stronger and harder than pure metals. This is probably because both dislocations in the crystalline lattice and the perfectly regular crystal structure of pure metals permit the planes of the crystals to slip over each other. These processes are inhibited in alloys because the resident or solute atoms interact with the dislocations and with the regular sections, so any lattice distortions produced make slipping more difficult.

A process that also depends on the internal structure, and possibilities for partial shifting of it, is *annealing*. This is based on the concept that a ductile metal becomes harder and less workable as cold work is done on it. Finally, a point is reached where cracking is imminent. To restore the original ductility, the metal is heated and slowly cooled. The temperatures used just permit the relaxation of the overstrained areas. A visualization might consider this a type of partial recrystallization or atomic rearrangement.

Polymorphism

Polymorphism, the existence of one or more crystalline and/or amorphous forms, is a characteristic of most solid substances. As applied to crystals, it refers to the different crystal structures the same chemical compound may have. The various forms also usually have different x-ray diffraction patterns, melting points, infrared spectra, and, most importantly from a pharmaceutical standpoint, different solubilities.

Particularly, in many cases in which dissolution in the gastrointestinal tract is the rate-limiting factor in absorption, differing solubilities may have a great effect, either good or bad. Different polymorphic forms are produced, depending on such factors as storage temperature, recrystallization solvent, and the rate of cooling (and, hence, the rate of crystallization) of the solvent. It appears that all organic materials exist in several polymorphic forms with the number of forms found depending on the effort spent searching.

In drugs, polymorphs of such diverse molecules from cortisone and prednisolone to aspirin have been found. As an example of the latter case, two different aspirin polymorphs form, depending on whether the material is crystallized from 95% alcohol or *n*-hexane. The two forms have different melting points, but, most importantly, the form produced from the hexane dissolves in water much more quickly. Toscani et al.³² have reported the stability hierarchy of three polymorphic forms of sulfanilamide.

REFERENCES

1. Feynman RP. *Science* 1974; 183:601.
2. Schoenborn BP. *Chem Eng News* 31, Jan 24, 1977.
3. Pitzer KS. *J Am Chem Soc* 1948; 70:2140.
4. Fieser LF, Fieser M. *Introduction to Organic Chemistry*. Boston: DC Heath, 1957, inside back cover.
5. Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, pp 225–226.
6. Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, p 93.
7. Pauling LC. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, Chap 3.
8. Fajans K. *Physical Methods of Organic Chemistry*, 2nd ed. Vol 1, Part II. New York: Wiley Interscience, 1949, p 1162.
9. Cammarata A. *J Med Chem* 1967; 10:525.
10. Lien EJ, et al. *J Pharm Sci* 1982; 71:641.

11. Lien EJ, et al. *J Pharm Sci* 1984; 73:553.
12. Lien EJ, et al. *Prog Drug Res* 1997; 48:9.
13. Olah GA, et al. *J Am Chem Soc* 1967; 89:711.
14. Sjöholm I, Szodin T. *Biochem Pharmacol* 1972; 21:3041.
15. Martin AN, et al. *Physical Pharmacy*, 3rd ed. Philadelphia: Lea & Febiger, 1983, 58–61.
16. Hansch C, Anderson SM. *J Org Chem* 1967; 32:2583.
17. Hansch C, Anderson SM. *J Med Chem* 1967; 10:745.
18. Hansch C, et al. *J Med Chem* 1973; 16:1207.
19. Hansch C, et al. *J Med Chem* 1977; 20:304.
20. Hansch C. *Farmaco Sci* 1968; 23:293.
21. Fujita T, et al. *J Am Chem Soc* 1964; 86:5175.
22. Iwasa J, et al. *J Med Chem* 1965; 8:150.
23. Gao H, et al. *Pharma Res* 1995; 12:1279.
24. Lien EJ, et al. *Prog Drug Res* 1997; 48:9.
25. Chester TL, et al. *Anal Chem* 2002; 74:2801.
26. Yang C, et al. *J Agr Food Chem* 2002; 50:846.
27. Lemert RM, et al. *J Phys Chem* 1990; 94:6021.
28. Crowther JB, Henion JD. *Anal Chem* 1985; 57:2711.
29. Goldberg AH, et al. *J Pharm Sci* 1966; 55:482.
30. Kennon L, Gulesich JJ. *J Pharm Sci* 1962; 51:278.
31. Ferguson JL, Brown GH. *J Am Oil Chem Soc* 1968; 45:120.
32. Toscani S. *Pharm Res* 1996; 13:151.

BIBLIOGRAPHY

- Ariëns EJ, et al. *Stereochemistry and Biological Activity of Drugs*. Boston: Blackwell, 1983.
- Eliel EL. *Stereochemistry of Carbon Compounds*. New York: McGraw-Hill, 1962.
- Hansch C, Leo A. Exploring QSAR. *Fundamentals and Applications in Chemistry and Biology*. Washington DC, ACS Professional Reference Book, 1995.
- Hansch C, et al. Exploring QSAR. Hydrophobic, Electronic and Steric Constants, *ibid*, 1995.
- Leo JA. *Chem Rev* 1993; 93: 1281.
- Lien EJ. *SAR Side Effects and Drug Design*. New York: Dekker, 1987.

Complex Formation



The word *complex* has many meanings in chemistry, so it is necessary at the outset to describe the types of systems that are included in this chapter. A complex is a species formed by the association of two or more interacting molecules or ions. To sharpen this concept the following definitions are provided:

- A *substrate*, S , is the interactant whose physical or chemical properties are observed experimentally.
- A *ligand*, L , is the second interactant whose concentration may be varied independently in an experimental study.
- A *complex* is a species of definite substrate-to-ligand stoichiometry that can be formed in an equilibrium process in solution, and also may exist in the solid state.

It is obvious that the complex must possess some properties that are different from those of its constituents; otherwise, there would be no evidence for its existence. Among the properties that may be altered upon complex formation are solubility, energy absorption, conductance, partitioning behavior, and chemical reactivity. It is by studying such properties of the substrate, as a function of ligand concentration, that complex formation may be recognized and described quantitatively. The terms *complex formation*, *complexation*, *binding*, and *association* are synonymous in the context of this chapter. Because complex formation is an equilibrium process, the methods of thermodynamics can be applied to describe it in the state of equilibrium. Moreover, the methods of chemical kinetics can be used to study the rate of approach to equilibrium. Finally, there may be interest in establishing the structure and properties of the complex.

These definitions are expressed succinctly in the following chemical equation for the formation of a complex S_mL_n .



This shows that the distinction between substrate and ligand is arbitrary and is made solely for experimental convenience. The definition omits any consideration of the forces acting between substrate and ligand in the complex; thus, it is very general. Therefore, the phenomena of interest may be restricted further by specifying that complexes are not formed with classic covalent bonds.

TYPES OF COMPLEXES—The definition of a complex leads to a classification into two groups based on type of chemical bonding.

Coordination Complexes—These complexes are formed by coordinate bonds in which a pair of electrons is, in some degree, transferred from one interactant to the other. The most important examples are the metal-ion coordination complexes between metal ions and bases. Such complexes can be viewed as products of Lewis acid–base reactions. Proton acids then constitute a special case of this type.

Molecular Complexes—These species are formed by noncovalent interactions between the substrate and ligand. The noncovalent forces

arise from electrostatic, induction and dispersion interactions, and they include, or give rise to, hydrogen-bonding, charge-transfer, and hydrophobic effects. Among the kinds of complex species that are included in this class are small molecule–small molecule complexes, small molecule–macromolecule species (eg, drug-protein and enzyme-substrate complexes), ion-pairs, dimers, and other self-associated species, inclusion complexes, intramolecular interactions (such as base–base interactions in the DNA helix), and clathrate complexes, in which the crystal structure of one interactant encloses molecules of the second interactant.

The following sections amplify these brief descriptions of coordination complexes and molecular complexes.

METAL-ION COORDINATION COMPLEXES

DESCRIPTIVE COORDINATION CHEMISTRY—Coordination complexes consist of a central metal ion (the substrate) bonded to an electron-pair donor (a base, the ligand). The ligand may be a conventional Brønsted base such as ammonia, an ion such as chloride ion, or even an aromatic compound. The complex may be neutral or charged. Coordination complexes also are called coordination compounds.

The number of bonds from the metal ion to the ligand (or ligands) is called the *coordination number* of the complex, and the maximum coordination number is evidently the largest possible number of such bonds. The maximum coordination number is determined by the electronic structure of the metal ion; numbers of 4 and 6 are most common, but other coordination numbers are possible. In solutions of Cu(II) in the presence of ammonia, these coordination complexes can form: $\text{Cu}(\text{NH}_3)_2^{2+}$, $\text{Cu}(\text{NH}_3)_3^{2+}$, $\text{Cu}(\text{NH}_3)_4^{2+}$. The maximum coordination number of Cu(II) is 4.

A ligand, like ammonia, that has a single basic group capable of bonding to the metal ion is a *unidentate* ligand. A ligand having more than one accessible basic binding site is *multidentate*; for example, ethylenediamine, $\text{H}_2\text{NCH}_2\text{CH}_2\text{NH}_2$, is a bidentate ligand. If a metal ion binds to two or more sites on a multidentate ligand, a cyclic complex is formed necessarily; this cyclic complex is a *chelate*. Thus, ethylenediamine forms a chelate with Cu(II):

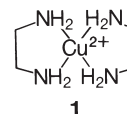
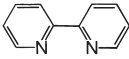
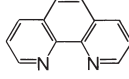

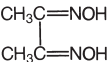
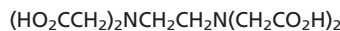


Table 14-1 shows several common multidentate ligands, and Table 14-2 lists abbreviations for some ligands. Thus, the complex shown in Structure 1 may be written $\text{Cu}(\text{en})_2^{2+}$. Of course, this complex ion must be associated with an appropriate number of anions.

Table 14-1. Some Important Multidentate Ligands^a

$\text{H}_2\text{NCH}_2\text{CH}_2\text{NH}_2$	Ethylenediamine
	2,2'-Bipyridine
	1,10-Phenanthroline
	8-Hydroxyquinoline (oxine)
	Dimethylglyoxime
	Ethylenediaminetetraacetic acid



^a Proton acid groups in these ligands are converted to basic groups upon the dissociation of the proton.

The nomenclature of coordination complexes is fairly complicated, and only the simplest features are reviewed here.¹

1. If the complex is an ion, the cation is listed first, then the anion.
2. Ligands (names): neutral ligands are named as the molecule, except for H_2O (aquo) and NH_3 (ammine). Positive ligands end in -ium (eg, hydrazinium, H_2NNH_3^+) and negative ligands in -o (eg, acetato). Some exceptions are chloro, fluoro, cyano, oxo, and hydroxo (OH^-).
3. Ligands (order): the order is anionic, neutral, and cationic. There are subrules within these categories; for example, simple ions generally precede polyatomic ions, and organic ions appear last.
4. Complex names (endings): anionic complexes end in -ate or -ic (if named as the acid). Cationic or neutral complexes do not have characteristic endings.
5. Central atom or ion (oxidation state): given by a Roman numeral in parentheses; no sign is used for positive oxidation states, but a negative sign indicates a negative oxidation state.

Examples:

$[\text{Pt}(\text{en})(\text{NH}_3)_2\text{NO}_2\text{Cl}]\text{SO}_4$	Chloronitrodiammine ethylenediamine-platinum(IV) sulfate
$\text{NH}_4[\text{Cr}(\text{SCN})_4(\text{NH}_3)_2]$	Ammonium tetrathiocyanatodiammine-chromate (III)
$[\text{Co}(\text{en})_3]_2(\text{SO}_4)_3$	Tris(ethylenediamine)cobalt(III) sulfate
$\text{K}_4[\text{Fe}(\text{CN})_6]$	Potassium hexacyanoferrate(II)
$\text{K}[\text{CrOF}_4]$	Potassium oxotetrafluorochromate(V)

Not all coordination complexes can be formed simply by mixing the reactants in solution. It has been found convenient to classify coordination complexes as either *labile* or *inert* complexes:

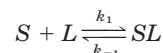
A labile complex is one whose rates of formation and dissociation are faster than, or comparable to, the typical time of mixing of the reactant solutions. An inert complex is one whose formation and dissociation rates are slower than the typical time of mixing of the reactant solutions.

Table 14-2. Common Abbreviations of Some Ligands

LIGAND	ABBREVIATION
Pyridine	<i>py</i>
Thiourea	<i>tu</i>
Ethylenediamine	<i>en</i>
Glycine	<i>gly</i>
Oxalate	<i>ox</i>
2,4-Pentanedione (acetylacetonate)	<i>acac</i>
1,10-Phenanthroline	<i>phen</i>
2,2'-Bipyridine	<i>bipy</i>
Ethylenediaminetetraacetate	<i>EDTA, Y</i>

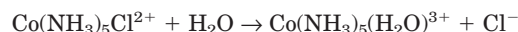
Clearly, the classification of labile versus inert is arbitrary, but it has experimental utility because inert complexes can be investigated by conventional chemical techniques, as they may persist long enough to be studied as isolated species; however, labile complexes tend to dissociate upon perturbation of the chemical system. At a more fundamental level, the lability or inertness of a complex can be related to its electronic configuration.²

It is important to note that the labile or inert classification is a kinetic one and generally is distinct from a consideration of complex stability, which is a thermodynamic concept (to be treated subsequently). To express this distinction more concretely, consider the example of complex formation



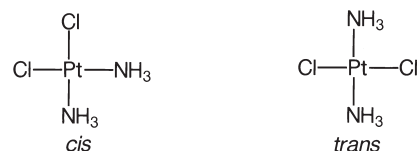
where k_1 is the rate constant for association and k_{-1} is the dissociation rate constant. Then, approximately, if $(k_1[L] + k_{-1})$ is greater than the rate of mixing, the complex is labile. The stability of the complex, however, is described by the equilibrium constant for its formation, which is equal to the ratio k_1/k_{-1} .

Although labile complexes form and dissociate rapidly, even inert complexes can undergo reactions in which one or more ligands are replaced, thus forming a new complex. Such reactions are called substitution reactions, and because ligands are bases, these are nucleophilic substitutions. A nucleophile, or *nucleus-lover*, is an electron-rich species that reacts with an electrophilic site; nucleophilicity refers to reactivity, ie, kinetics. Basicity refers to equilibrium behavior. The following equation is a typical nucleophilic substitution reaction (a hydrolysis reaction) in which water is the nucleophile.



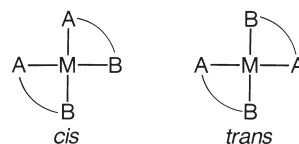
ISOMERISM AND STEREOCHEMISTRY—From organic chemistry it is known that the geometry of bonding about the saturated carbon atom is that of a regular tetrahedron (the coordination number of carbon being 4). As a consequence, there is only one substance with the formula CA_2B_2 , where C is carbon and A or B represent atoms or groups bonded to the carbon. For example, there is only one compound (methylene chloride) with the formula CH_2Cl_2 .

It is otherwise with metal-ion coordination complexes having coordination number 4, for which it has been found that there may be two compounds of structure MA_2B_2 , where M represents the metal ion. These two compounds are geometrical isomers, and their existence means that they have a square planar structure. For example, the two dichlorodiammineplatinum(II) isomers have these structures:



In the *cis* isomer, two like ligands are adjacent; in the *trans* isomer, they are opposite each other. The metal and the four ligand groups all lie in the same plane. Figure 14-1 shows alternative representations of the square planar complex structure. The demonstration of geometrical isomerism by chemical methods was based on the isolation of both isomers, which is possible if the complexes are inert.

There exists also the possibility of *cis* and *trans* isomerism in square planar complexes of the structure $\text{M}(\text{AB})_2$, where AB is an unsymmetrical bidentate ligand, such as glycinate.



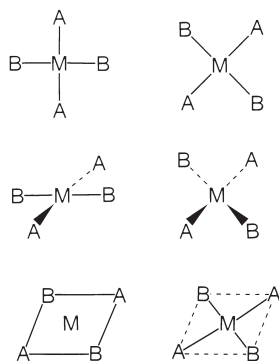


Figure 14-1. Equivalent representations of the square planar complex $trans\text{-MA}_2\text{B}_2$.

Most complexes of coordination number 4 have the square planar structure, but some are tetrahedral. Nearly all complexes with coordination number 6 are octahedral; ie, the coordinate bonds lie along the x, y, and z axes of a Cartesian coordinate system with the metal ion at the origin. This structure is consistent with the experimental observations that only two isomers can be isolated of each of the structures MA_4B_2 and MA_3B_3 . The *cis* and *trans* isomers of the octahedral dichlorotetraamminecobalt(III) chloride have these structures:

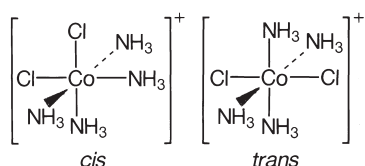


Figure 14-2 shows equivalent ways to draw an octahedral complex.

It should be noted that chloride in the above cobalt compounds plays two different roles; two chlorides are ligands, being coordinately bound to the cobalt, whereas the other chloride serves as a counterion to the complex cation.

Octahedral complexes can exhibit optical isomerism when two structures are related as nonsuperimposable mirror images. Such isomers are called *enantiomers*. The optical isomers of $\text{M}(\text{AA})_3$, where AA is a symmetrical bidentate ligand, are shown in Figure 14-3, which also shows the specific example $[\text{Pt}(\text{en})_3]^{4+}$.

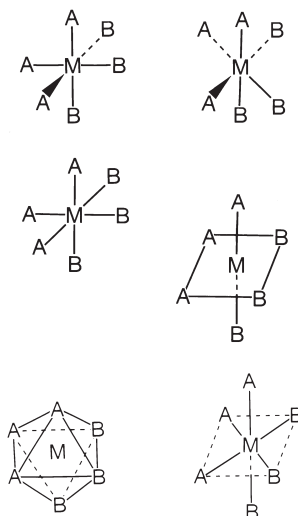


Figure 14-2. Equivalent representations of the octahedral complex $cis\text{-MA}_3\text{B}_3$.

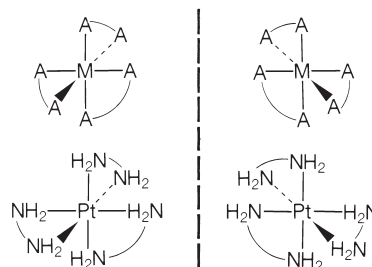
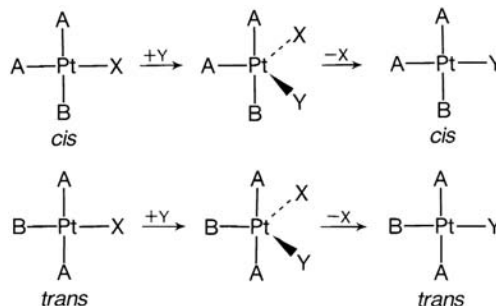


Figure 14-3. Optical isomers of $\text{M}(\text{AA})_3$ (top pair) and $[\text{Pt}(\text{en})_3]^{4+}$ (bottom pair). Each enantiomer is a nonsuperimposable mirror image of the other as reflected in the central vertical plane.

The existence of geometrical and optical isomers of coordination complexes has provided valuable insight into possible complex structures, as noted above; but, in addition, these isomers, when subjected to substitution reactions, have led to important inferences concerning the mechanisms of these reactions. For example, nucleophilic substitution reactions of square planar complexes are known to be bimolecular displacement processes, on the basis (in part) of complete retention of configuration; *cis* reactants yield *cis* products, and *trans* reactants yield *trans* products.³ This rules out a dissociation ($\text{S}_{\text{N}}1$) mechanism. The reaction is believed to take place via a trigonal bipyramidal structure in which the metal-ion coordination number is increased as shown below.



THEORIES OF COORDINATE BONDING—A great range in complexing behavior is observed in the interactions of different metal ions with different ligands. A successful theory of coordinate bonding should be able to describe and predict the chemistry of coordination complexes given the identities of the metal ion and the ligand. Developments in this field have been concerned particularly with the transition elements, which may be defined as those elements having partly filled *d* or *f* shells in any of their common oxidation states⁴; with this definition, slightly more than half of the elements are transition elements. In addition, of course, some main group elements may form complexes.

A theory of coordinate complexing should be able to account for the coordination numbers of ions and the stereochemistry of their complexes. It should explain commonly observed regularities in complex stability, such as the *chelate effect*: the greater the number of sites of bonding of each ligand to the metal ion, the greater the complex stability. Another pattern is that of the complexes of certain divalent metal ions, whose stabilities vary in the order $\text{Mn} < \text{Fe} < \text{Co} < \text{Ni} < \text{Cu} > \text{Zn}$. The electronic absorption spectra (ie, the allowed electronic transitions) of complexes are a readily observed property that a theory should describe. Many metal coordination complexes absorb strongly in the visible region. Metal ions and their complexes also may have magnetic properties that can be accounted for theoretically. Substances having no unpaired electrons are diamagnetic, whereas those with unpaired electrons are paramagnetic,

and these properties easily are distinguished experimentally. Thus, a theory should be able to predict the number of unpaired electrons in the coordination complex.

Many theories have been developed, and they are essentially all different in concept. It is not possible here to treat any of them in detail, but their basic approaches will be outlined.

The *electrostatic* theory is completely classical (ie, nonquantum mechanical).⁵ Ions are treated as spherical charges and molecules are treated as dipoles; the energy of a complex is calculated as a sum of charge–charge, charge–dipole, and charge-induced dipole terms and repulsive forces. Experimental values of dipole moments and intermolecular distances are employed in the calculations, which yield results for bond energies in remarkably good agreement with experimental values for many complexes. However, the theory necessarily is approximate, because it does not include quantum mechanical effects and it oversimplifies the structural differences among metal ions and ligands.

The *valence bond* theory of Pauling⁶ is a quantum mechanical theory. A coordinate bond is formed when a pair of electrons on a ligand is donated to a vacant orbital on the metal ion. The coordination number is determined by the number of available orbitals, and the geometry of the complex is determined by the directional properties of the hybrid orbitals formed by combination of the atomic orbitals (the tetrahedral arrangement of hybrid sp^3 orbitals of carbon).

This theory has been quite successful in accounting for complex stereochemistry. It also can incorporate observations on magnetic type, as illustrated by the electronic configurations in Table 14-3.⁷ From the vacant atomic orbitals of Fe^{2+} or Fe^{3+} there can be formed six equivalent hybrid orbitals of composition $3d^24s4p^3$; thus, octahedral complexes are anticipated. Each ligand contributes two electrons to a hybrid orbital, resulting, in the case of $Fe(CN)_6^{4-}$, in a complex having no unpaired electrons and, therefore, diamagnetic; $Fe(CN)_6^{3-}$, on the other hand, possesses one unpaired electron, in agreement with experimental conclusions.

The valence bond theory is useful mainly in this qualitative pictorial way. In principle, bond energies can be calculated; in practice, this is extremely difficult.

As the coordinate bond has been treated thus far, it consists entirely of a pair of electrons donated by the ligand to a vacant metal orbital. Another type of donation is sometimes possible (as in the case of the two hexacyanato complexes shown in Table 14-3). If the ligand possesses vacant orbitals, the metal may contribute electrons from its d orbitals to vacant p or d orbitals on the ligand, thus producing a bond with double-bond character. This phenomenon is called *back-bonding*.

The *valence shell electron-pair repulsion* theory is a very simple approach to the prediction of complex geometry. This is based on the principle that the valence shell electrons of the metal are directed in space so as to minimize their total repulsive energy. Thus, if there were two electron pairs, they will distribute themselves on opposite sides of the central ion, and a linear complex will be formed. This theory is not able to calculate bond energies.

The *crystal field* theory has been very fruitful in the study of coordination complexes. (The word “crystal” in this context is a historical accident; the theory is applicable to complexes in so-

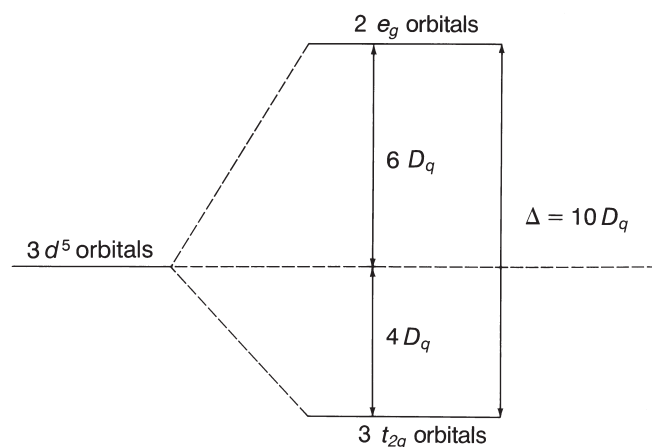


Figure 14-4. Energy-level diagram showing crystal-field splitting of the 5-fold degeneracy of metal-ion $3d$ orbitals in an octahedral complex.

lution as well as in the solid state.) The basis of the theory is seen readily with the example of an octahedral complex of a metal ion, such as iron. The five $3d$ orbitals are of equal energy (they are said to be 5-fold degenerate). According to crystal field theory, arranging the ligands colinear with d orbitals requires more energy (because of electron–electron repulsion) than does the approach of ligands between d orbitals. Two d orbitals (d_z and $d_{x^2-y^2}$) have lobes along the three Cartesian coordinates that define the geometry of the octahedral complex; thus, the electrical field of the ligands destabilizes (raises the energy of) these two orbitals. The other three orbitals (d_{xy} , d_{yz} , d_{xz}) are directed between the axes, so they are stabilized by the field of the ligands. Thus, the 5-fold degeneracy is split to produce two doubly degenerate orbitals (labeled e_g) and three triply degenerate orbitals (labeled t_{2g}), with no net energy change. This crystal-field splitting is shown in Figure 14-4. The total-energy difference Δ is conventionally labeled $10D_q$. It, therefore, follows that the e_g orbitals are destabilized by $6D_q$, and the t_{2g} orbitals are stabilized by $4D_q$.⁸

Now, the first orbitals to be filled upon formation of the complex will tend to be the lower energy t_{2g} orbitals, unless the stabilization is slight, in which case normal *Hund's rule* behavior will be observed, the electrons tending to remain unpaired. Thus, large splitting will lead to the formation of paired electrons (low-spin complexes), whereas small splitting will lead to more unpaired electrons (high-spin complexes).

A further subtlety can occur in which a distortion of the regular octahedral geometry takes place to lower the total energy of the system. This is known as the *Jahn-Teller effect*, with the result that for many octahedral complexes four of the ligands are coplanar with, and equidistant from, the metal ions; the other two ligands lie at a greater distance from the metal ion.

The crystal field theory has been developed in great detail, and many explanations and predictions have been achieved successfully. It especially is useful for explaining complex absorption spectra, and spectral measurements can be used to obtain values of the crystal field splitting, Δ .

The *molecular orbital* theory (which also is called the *ligand field* theory) is a quantum mechanical description in which molecular orbitals are constructed mathematically by the linear combination of atomic orbitals (MO-LCAO). The number of molecular orbitals (MOs) formed is equal to the number of atomic orbitals (AOs) taken, but the MOs are formed in pairs; one member of each pair is a symmetric, lower energy, bonding MO, and the other is an antisymmetric, higher energy, antibonding MO. The complex electronic configuration and energy are established by assigning electrons to the bonding MOs.

This concept is illustrated in Figure 14-5, shows a schematic MO diagram for an octahedral complex in which the ligand forms only single coordinate bonds (no back-bonding).⁹ The

Table 14-3. Electronic Configurations of Some Iron Species According to Valence Bond Theory^a

SPECIES	3D					4S	4P		
Fe^0	$\uparrow\downarrow$	\uparrow	\uparrow	\uparrow	\uparrow	$\uparrow\downarrow$	—	—	—
Fe^{2+}	$\uparrow\downarrow$	\uparrow	\uparrow	\uparrow	\uparrow	—	—	—	—
Fe^{3+}	\uparrow	\uparrow	\uparrow	\uparrow	\uparrow	—	—	—	—
$Fe(CN)_6^{4-}$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$
$Fe(CN)_6^{3-}$	$\uparrow\downarrow$	$\uparrow\downarrow$	\uparrow	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$

^a Electrons in closed shells are not shown; thus, the electron configuration of Fe^0 is $1s^22s^22p^63s^23d^64s^2$.

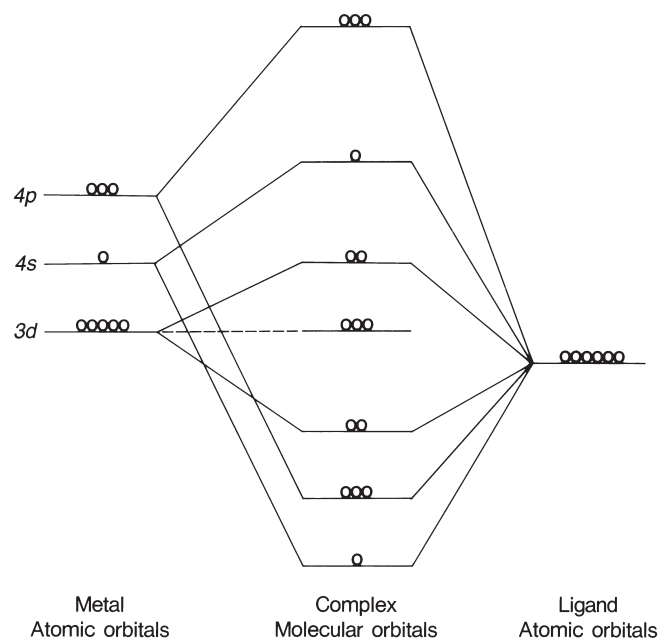


Figure 14-5. Schematic molecular orbital diagram for an octahedral complex. The vertical distance represents energy. Each circle denotes an orbital.

combination of AOs must take place according to certain quantum mechanical rules. For example, the metal s orbital combines with a ligand σ orbital to generate a bonding σ orbital and an antibonding σ^* orbital. The nine metal AOs combine with six ligand AOs to produce 15 MOs. The octahedral complex is formed by using the six bonding MOs (lowest energy MOs).

The MO theory is the most powerful of the theories of coordinate bonding, although quantitative calculations may be extremely difficult to make. Basolo and Pearson¹⁰ have presented a comparison of the several theories.

Another view that has been found useful for its explanatory and predictive power is the *hard and soft acid–base* (HSAB) concept. A *hard acid* is defined as one in which the electron-pair acceptor atom is small in size, with high positive-charge density and low polarizability. A *soft acid* is large and polarizable. A hard base has high electronegativity and low polarizability, whereas a soft base is easily polarizable. Examples of these classes are listed in Table 14-4. Polarizability is a measure of the ease with which the electron cloud can be deformed under the influence of a field. Hardness and softness are related inversely.

The HSAB principle states that hard acids prefer to coordinate to hard bases and soft acids to soft bases. This empirical generalization can account qualitatively for much coordinate-complex chemistry. The HSAB concept has been extended by the introduction of a quantitative definition of hardness¹¹ as

$$\eta = \frac{(I - A)}{2}$$

where η is the hardness, I is the ionization potential (a measure of the ease with which an electron can be lost), and A is the electron affinity, which measures the ease with which an electron

Table 14-4. Examples of the Hard-Soft Classification of Lewis Acids and Bases

	ACIDS	BASES
Hard	H^+ , Li^+ , Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Mn^{2+} , Al^{3+}	H_2O , OH^- , F^- , Cl^- , PO_4^{3-} , SO_4^{2-} , ClO_4^- , NO_3^- , NH_3
Soft	Cu^+ , Ag^+ , Au^+ , Hg_2^{2+} , Pd^{2+} , Pt^{2+}	I^- , SCN^- , CN^-

combines with the species. Pearson has related hardness to the MO theory and developed the quantitative aspects of HSAB theory.¹¹

At the start of this chapter it was specified that complexes are not formed with covalent bonds, but in the case of coordination complexes it is seen that a coordinate bond may have extensive covalent character, even though both electrons are donated by one of the reactants. One of the goals of theory is to be able to calculate the fractions of ionic and covalent character of the coordinate bond. Very roughly, it may be expected that when the bond is between atoms that differ greatly in their electronegativities (propensities for attracting negative charge), the bond will be largely ionic, whereas if the atoms have similar electronegativities, the bond will be largely covalent.

MOLECULAR COMPLEXES

NONCOVALENT INTERMOLECULAR FORCES—

Molecules in condensed systems (liquids and solids) experience mutual forces of attraction, which is why the systems are condensed. These forces are much weaker than those of “chemical” (ie, covalent) bonds, as shown by the ease with which they can be broken, such as by vaporization or dissolution. These are the noncovalent intermolecular forces.¹²

Two different kinds of solute molecules or ions are to be noted, labeled S (substrate) and L (ligand), in a solvent that is thought of conveniently (but somewhat artificially) as a homogeneous continuum; ie, for the present, neglect the molecular nature of the solvent. The intermolecular forces between S and L are of interest. The force of interaction F is related to the potential energy of interaction V by

$$F = -\frac{dV}{dr}$$

where r is the distance between the interacting species. It is conventional to express the intermolecular forces in terms of the corresponding energies. The most important noncovalent potential energy functions, as established by theoretical arguments, are listed in Table 14-5.

Table 14-5. Potential-Energy Functions for Noncovalent Interactions^a

TYPE OF INTERACTION	POTENTIAL-ENERGY FUNCTION
<i>Electrostatic</i>	
Charge–charge	$+ \frac{C_S C_L}{r}$
Charge–dipole	$- \frac{1}{3kT} \cdot \frac{C_S^2 \cdot \mu_L^2}{r^4}$
Dipole–dipole	$- \frac{2}{3kT} \cdot \frac{\mu_S^2 \cdot \mu_L^2}{r^6}$
<i>Induction</i>	
Charge-induced dipole	$- \frac{C_S^2 \cdot \alpha_L}{2r^4}$
Dipole-induced dipole	$- \frac{\mu_S^2 \cdot \alpha_L}{r^6}$
<i>Dispersion</i>	
Induced dipole–induced dipole	$- \frac{3}{4} \left[\frac{\epsilon_S \cdot \epsilon_L}{\epsilon_S + \epsilon_L} \right] \frac{\alpha_S \cdot \alpha_L}{r^6}$

^a C is the charge on an ion, μ is permanent dipole moment, α is polarizability, r is intermolecular distance, ϵ is a specific energy term, T is absolute temperature, and k is Boltzmann's constant, where $k = R/N_A$.

The noncovalent forces are of three broad types:

- The *electrostatic* forces among ions and molecules possessing permanent dipole moments.
- The *induction* (or polarization) forces between an ion and a nonpolar molecule or a polar molecule and a nonpolar molecule.
- The *dispersion* (London) force, which operates between all molecules.

The *electrostatic* forces are the consequence of classical attraction and repulsion effects between charges. In the potential-energy terms in Table 14-5, the magnitudes of the charges are to be accompanied by their signs; a negative value for the energy is attractive, whereas a positive value is repulsive. Note that charges and dipole moments always appear as squared quantities.

The *induction* forces arise as a result of an ion or a polar molecule inducing a dipole in a neighboring molecule. Thus, their strength depends upon the ionic charge or the dipole moment of the inducing species and the polarizability (a measure of electron-cloud deformability) of the induced species.

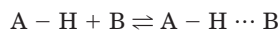
The *dispersion* force is nonclassical in origin; ie, it is a quantum mechanical effect. At any moment the electronic distribution in one molecule, such as *S*, may result in the production of a dipole moment in *S*, even if it is a nonpolar molecule. This instantaneous dipole then can induce a dipole in *L*. The dispersion force, therefore, is general and acts among all molecules, both polar and nonpolar. (The term *van der Waals' force* sometimes is used to describe the dispersion force, but some authors use this term to include all noncovalent forces.)

It is important to notice that, for neutral molecules, the electrostatic, induction, and dispersion-energy terms all possess an intermolecular-distance dependence of r^{-6} . As two molecules approach each other they will experience a force of attraction that varies with distance as r^{-7} . They cannot continue to approach closely indefinitely because ultimately they experience repulsive forces as their electron clouds tend to repel each other, and at an even closer distance there is an internuclear repulsive force. The net force between molecules is a balance between the attractive and repulsive forces. This often is described by the potential-energy function below, which is called the Lennard-Jones 6-12 potential,

$$V = 4V_{\min} \left[\left(\frac{r_0}{r} \right)^{12} - \left(\frac{r_0}{r} \right)^6 \right] \quad (1)$$

where V_{\min} is the value of V at the minimum in the "potential well," ie, where $r = r_{eq}$, the equilibrium intermolecular distance. This is the distance at which the attractive and repulsive forces are balanced. The term in r^{-12} is the repulsive term, that in r^{-6} is the attractive term, and r_0 is the value of r when $V = 0$. Figure 14-6 shows a plot of the Lennard-Jones 6-12 potential for a hypothetical system to illustrate the qualitative features of noncovalent interaction. Values of V_{\min} are typically 5 kcal/mol, or less, which are much smaller than typical covalent bond energies.

Although Table 14-5 includes the most important noncovalent interactions, additional types of bonding often are invoked when describing complex formation. One of these is *hydrogen bonding*. The formation of a hydrogen bond (H-bond) between a proton-donor HA and a proton-acceptor B can be represented formally as



The strength of the hydrogen bond is controlled, in part, by the acid strength of HA and the base strength of B, but the solvent also is very important. The A—H bond is mainly covalent in character, and the hydrogen bond $H \cdots B$ is predominantly electrostatic.¹³ Structure **2** shows intermolecular hydrogen bonding in a dimer of acetic acid, and Structure **3** shows an intramolecular hydrogen bond of the salicylic acid anion.

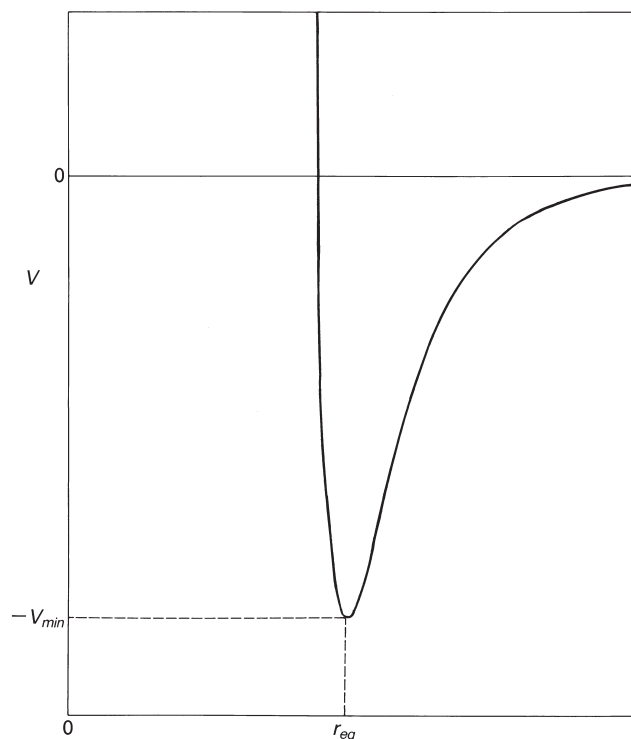
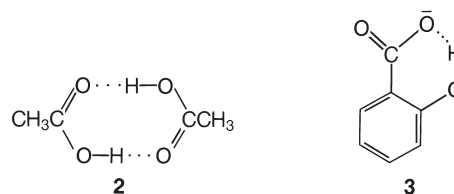


Figure 14-6. Potential-energy diagram according to Equation 1, the Lennard-Jones potential; r_{eq} is the equilibrium intermolecular distance, which minimizes the potential energy of the system.



Another type of bonding is *charge-transfer bonding*, which is a consequence of the transfer of an electron from a molecular orbital on an electron-donor molecule to an orbital on an electron-acceptor molecule. Because charge-transfer results in a change in electron configuration, it produces a change in the electronic energy levels and, therefore, in the ultraviolet-visible absorption spectra of the interacting molecules. The appearance of new absorption bands often is cited as evidence for charge-transfer bonding.

Since an electron transfer is involved in this type of bonding, the resulting bond may have some covalent character. In addition, the usual noncovalent forces of Table 14-4 also are present, and the contribution of the charge-transfer phenomenon to the overall stability of the complex depends upon the particular molecules involved. A classification of the types of electron donors and acceptors, with some examples, is given in Table 14-6.¹⁶

ROLE OF THE SOLVENT—The preceding discussion of intermolecular forces focused on the interactions between two molecules or ions that may be viewed as the substrate and the ligand in a complex-formation reaction. The solvent was ignored in this treatment, which strictly is applicable only in the vapor state. However, in this discussion, interest in complexes arises largely from their occurrence in solutions and solids, so the solvent must be introduced as a component of the system. Table 14-7 presents some useful groupings of solvents according to structure and chemical behavior.

Table 14-6. Classification of Charge-Transfer Donors and Acceptors

TYPE	ORBITAL INVOLVED	EXAMPLE
<i>Electron donors</i>		
n	Nonbonding pair	:NR ₃
$b\pi$	Bonding π orbital	Benzene
<i>Electron acceptors</i>		
v	Vacant orbital	BCl ₃
$a\pi$	Antibonding π orbital	TCNE ^a
$a\sigma$	Antibonding σ orbital	I ₂

^a Tetracyanoethylene, (N=C)₂C=C(C=N)₂.

As the solvent is a molecular species, it is subject to the same types of intermolecular forces as the solute species. Thus, a competition is set up among the several species, and the net effect, as manifested by the stability of the complex, is a consequence of this competition. The role of the solvent on complex stability may be expressed as

$$\Delta G_{\text{net}} = \Delta G_{MM} + \Delta G_{MS} + \Delta G_{SS} \quad (2)$$

where ΔG_{net} is the overall free-energy change for complex formation, ΔG_{MM} is a contribution from medium–medium (ie, solvent–solvent) interactions, ΔG_{MS} describes medium–solute interactions, and ΔG_{SS} includes all solute–solute interactions. The value of ΔG_{SS} is determined by the substrate–ligand intermolecular interactions, but Equation 2 shows that the net stability of the complex also can be influenced by the solvent. The term ΔG_{MS} represents a solvation contribution, and its effect can be either stabilizing (if the complex is solvated more extensively than the reactants) or destabilizing (if the converse is applicable). The ΔG_{MM} term represents another way in which the solvent can influence complex stability.

Consider a nonpolar solvent, in which the MM interactions are weak (arising only from the dispersion force). In such a solvent the ΔG_{MS} term probably also will be small, and then the ΔG_{SS} term will make the major contribution to ΔG_{net} . If, on the other hand, the solvent is polar (such as water, in particular), the solvent–solute interaction may be the predominant contributor to ΔG_{net} . The ΔG_{MM} term in water may be identified as the *hydrophobic effect*, which will be considered in more detail. There are two points of view from which the hydrophobic effect can be discussed.

One of these theories takes as its key feature the structure of water, ie, the intermolecular network of water molecules generated by their mutual hydrogen-bonding.¹⁵ When a nonpolar solute dissolves in water, no H-bonds from water to the solute can form, so the water structure in the vicinity of the solute must be modified to compensate for the water–water H-bonds that were broken upon insertion of the solute into the water. The number of possible orientations of water molecules is decreased in the presence of the solute, so its dissolution is unfavorable entropically; this is why nonpolar compounds have low aqueous solubilities, according to this view.

Table 14-7. Classification of Solvents

SOLVENT CLASS	EXAMPLES
Hydroxylic H-bond donors	Water, alcohols, glycols
H-bond acceptors Dipolar aprotic ^a	Water, alcohols, glycols, carboxylic acids, amides, imides, chloroform Amines, ethers, aldehydes, ketones Acetonitrile, dimethylsulfoxide, acetone, <i>N,N</i> -dimethylformamide
Nonpolar	Hydrocarbons, halogenated hydrocarbons

^a These are solvents with large dipole moments and no readily donated proton.

When two such dissolved solute molecules come into contact, some of the *structured* water surrounding them must be released into the bulk medium, resulting in an increase in entropy, which (through its contribution to the ΔG_{MM} term) is the main driving force in the hydrophobic interaction of nonpolar molecules in water. Although this description is acceptable for nonpolar solutes, it must be modified for polar solutes, for which the main driving force may be either a favorable entropy change or a favorable enthalpy change.¹⁶

The second theory of the hydrophobic effect is the cavity model, which treats the solvent as a continuum. The surface tension γ of a solvent is a measure of its surface energy, and in water, whose surface tension is unusually high (72 dynes/cm), there is a strong driving force for the minimization of surface area. In order to dissolve a solute molecule in a solvent, a cavity must be created in the solvent, and then the solute is inserted in the cavity. This can be thought of as “digging a hole in the solvent,” and it takes an energy equal to the product of the surface area of the cavity (which is determined by the size of the solute molecule) and the surface tension of the solvent. Some of this energy cost may be offset by the subsequent interaction energy, through solvation, of the molecule with the solvent.

When two dissolved molecules unite to form a complex, the two cavities containing the separated species coalesce into a single cavity holding the complex. There is a net decrease in surface area (ΔA) in this process, and the product $\Delta A\gamma$ is the driving force for the complex formation. Figure 14-7 is a representation of this cavity model of the hydrophobic effect.

It now can be anticipated that if the hydrophobic effect makes a major contribution to the stability of a complex in water, the incorporation of an organic cosolvent into the medium (resulting in a lower surface tension) will decrease the stability of the complex. On the other hand, if the interest is in a complex (in a nonhydroxylic solvent) whose stability is derived largely from strong intermolecular substrate–ligand H-bonding, then incorporation of water or an alcohol will reduce the stability of the complex due to competition by the hydroxylic solvent.

Thus, it is seen that solvent effects on complex-formation can be varied and complicated, but that their study may offer insight into the nature of the intermolecular interactions responsible for the formation of the complex. A quantitative theory of solvent effects on complex formation has been developed.¹⁷

EXAMPLES OF MOLECULAR COMPLEXES—There is no systematic classification of molecular complexes, nor has a system of nomenclature been developed to describe complexes. Particular types may be classified in terms of the kinds of interactions involved in their formation, the kinds of interactants involved, or the kinds of complexes formed. Table 14-8 gives an outline of molecular complexes according to this classification.

Since molecular complexes are formed by noncovalent interactions, their bonding is localized less than that observed with covalent and coordinate bonds, which are highly directional. As a consequence, for many of these complexes it is not possible to indicate a specific complex structure. Hydrogen-bonded complexes are exceptions because of the requirements for the existence of the H-bond, and many H-bonded structures are known.

Studies by x-ray diffraction on crystalline complexes can reveal the mutual orientations of interactants in the solid state, but this knowledge does not indicate the nature or location of the noncovalent bonding directly. Theoretical calculations have

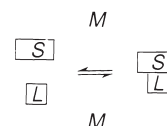


Figure 14-7. Representation of complex formation between S and L (planar molecules viewed in cross section through the molecular planes) to form complex SL . The total surface area exposed to solvent M is less for the complex than for the separated species.

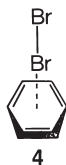
Table 14-8. Classification of Molecular Complexes

I. Type of bonding or interaction	
Charge-transfer	
Hydrogen-bonding	
Hydrophobic interaction	
Stacking interaction	
II. Type or structure of interactants	
Small molecule–small molecule complex	
Small molecule–macromolecule binding	
Drug–protein binding	
Enzyme–substrate complex	
Drug–receptor complex	
Antigen–antibody complex	
III. Type or structure of complex	
Self-associated aggregate	
Micelle	
Inclusion complex	
Clathrate	

been helpful in suggesting how the substrate and ligand are positioned in the complex, and this approach is being used to design new drugs that can bind specifically to biological receptors.

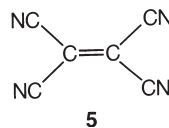
Some examples of complexes, or of substrates and ligands that form complexes, will follow, using the outline in Table 14-8 as a guide.

Charge-transfer (CT) complexes, also called electron donor–acceptor (EDA) complexes, may be formed when one interactant can perform as the electron donor and the other as the electron acceptor. The appearance of a new electronic absorption band, not attributable to either the donor or the acceptor, often is taken as evidence for charge-transfer complexing. A classic example is provided by solutions of iodine in organic solvents. When I_2 is dissolved in aliphatic hydrocarbons or carbon tetrachloride, the solution has a violet color characteristic of iodine, but solutions in aromatic hydrocarbons, alcohols, or ethers are brown. It is inferred that, in these latter solvents, a complex is formed and, because of the color (spectral) change, charge-transfer is implicated. In solvents in which iodine forms a complex (“brown” solvents), the solvent is the electron donor and the iodine is the acceptor. Thus, from Table 14-6, the benzene–iodine complex may be described as a $b\pi$ – $a\sigma$ CT complex, whereas the ethanol–iodine complex is an n – $a\sigma$ complex. Investigation of the benzene–bromine complex by X-ray crystallography shows that in the solid state the axis of the halogen molecule is perpendicular to the plane of the aromatic ring, as in Structure 4. The structure of the complex in solution may, however, be different from this.

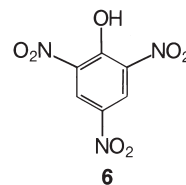


4

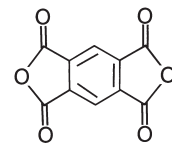
Referring to Table 14-8, it is noted that electron donors of the n type will be found among the amines, ethers, alcohols, and sulfides, whereas $b\pi$ donors include alkenes, alkynes, and aromatic hydrocarbons. Substitution on the donors by alkyl groups (which are electron-releasing) enhances their donor properties, unless the bulkiness of the substituent leads to steric hindrance. Hexamethylbenzene is a good electron donor. Lewis acids are electron acceptors, but among organic compounds the most important acceptors are unsaturated and aromatic compounds substituted with electron-withdrawing groups, exemplified by tetracyanoethylene, Structure 5, picric acid, Structure 6, and pyromellitic dianhydride, Structure 7.



5

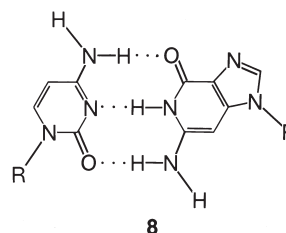


6



7

Structures 2 and 3 show hydrogen-bonding interactions. H-bonded complexes are observed readily in solvents that do not compete as H-bond donors or acceptors. The complex between phenol and pyridine in inert solvents is an H-bond complex. Perhaps the most famous hydrogen-bonded complexes are those of adenine-to-thymine and guanine-to-cytosine, which, as constituents of desoxynucleic acid, are responsible for the double-helix structure of the DNA molecule. Structure 8 shows the hydrogen bonds linking the cytosine of one polynucleotide strand to the guanine of a second strand.¹⁸



8

Any complex in aqueous solution may receive some portion of its stability from the hydrophobic effect, and for nonpolar interactants the hydrophobic contribution probably is the major one. According to the cavity model (see Fig 14-7), solvents other than water also can lead to complex formation via this surface-energy effect, although less effectively than in water because the surface tensions are lower; in solvents other than water this is called the *solvophobic effect*.

When two planar molecules undergo a primarily hydrophobic association, the total surface area of the complex exposed to the solvent can be minimized if the molecules are in plane-to-plane contact, as suggested in Fig 14-7. This plane-to-plane orientation is called a *stacking interaction*. The purine–pyrimidine H-bonded base pairs in DNA (Structure 8) are planar assemblies that undergo stacking interactions with adjacent pairs.

Aside from the interactions specifically mentioned in Table 14-8, complex formation also can be the result of the several types of noncovalent interactions depicted in Table 14-5, and most complexes probably involve a combination of interactions.

The third class listed in Table 14-8 is not depicted so easily as is the second class. *Self-association* is a type of complexation in which a molecule forms complexes with others of its own species. If S represents a molecule capable of self-association, then S_2 is called its dimer, S_3 its trimer, S_4 its tetramer, and so on. Structure 2 shows a hydrogen-bonded dimer of acetic acid, which can exist in the vapor phase and in inert solvents. Benzene forms dimers in aqueous solution, as does caffeine; these planar molecules probably undergo hydrophobic stacking interactions in water.

A *micelle* is a special form of self-aggregated complex in which the interactant is a surfactant, a molecule possessing both a nonpolar and a polar portion. See Chapter 20 for an in-depth treatment of micelles.

Inclusion complexes are formed when a macrocyclic compound, possessing an intramolecular cavity of molecular dimensions, interacts with a small molecule that can enter the cavity. The macrocyclic molecule is called the *host*, the small, included molecule is the *guest*, and the inclusion process gives rise to *host–guest chemistry*. Both synthetic and naturally occurring

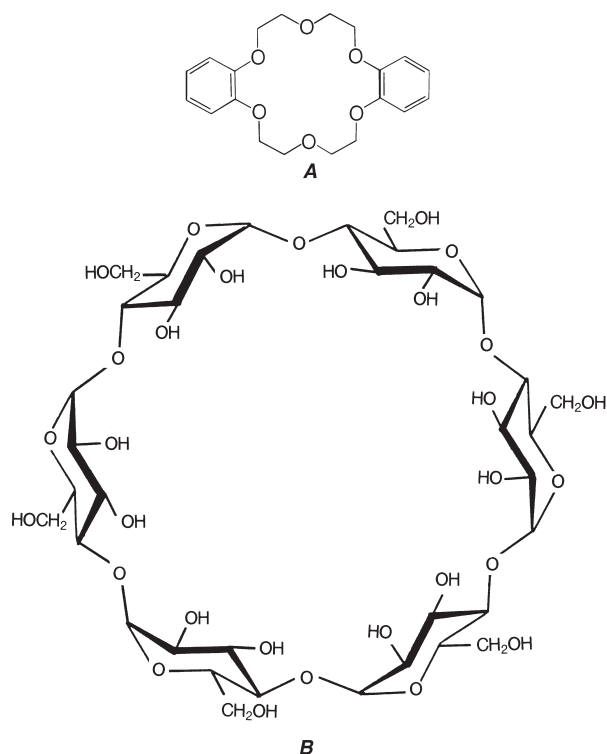


Figure 14-8. Structure of (A) dibenzo-18-crown-6 and (B) α -cyclodextrin.

macrocyclic hosts are known, and Figure 14-8A and B show an example of each. Crown ethers, such as the one shown in Figure 14-8A, present a nonpolar external molecular surface, but the interior of the cavity is relatively polar. As a consequence, polar guests such as ions can enter the cavity and, because their polarity now is masked by the surrounding host, exhibit unusual chemistry. For example, potassium permanganate, which is not soluble in nonpolar solvents, can be extracted into organic solvents from water in the presence of a crown ether.

The cyclodextrins are macrocyclic hosts that are formed by the action of certain bacterial enzymes on starch. They consist of α -D-glucose units joined with glycosidic (ether) linkages. The interior of the cavity is lined with these glycosidic bonds and, therefore, is relatively nonpolar (ie, relative to water), whereas the exterior of the molecule is quite polar because of the large number of hydroxy groups. The three commercially available cyclodextrins are called α -, β -, and γ -cyclodextrins (or, alternatively, cyclohexamylose, cycloheptaamylose, and cyclooctamylose), and they consist of 6, 7, and 8 glucose units, respectively. The diameters of the cavities of the cyclodextrins are approximately 5 Å (for α), 6 to 7 Å (for β), or 8 to 9 Å (for γ). Thus, small guest molecules, or parts of molecules, may enter the host cavity to form inclusion complexes, whose stabilities are in part the result of the hydrophobic effect. Many properties of a guest molecule may be altered by inclusion in a cyclodextrin;¹⁹ these include volatility, solubility, and chemical stability, so numerous practical applications have been suggested.²⁰ The stabilities of cyclodextrin complexes have been discussed.²¹

There is a special type of inclusion compound, called a *clathrate*, in which the host molecules form a crystal lattice containing spaces into which guest molecules can fit.²² In cage clathrates, the cavity is a space completely surrounded by a network of host molecules. Some “gas” hydrates are examples; in these structures, a hydrogen-bonded network of water molecules, analogous to ice, encloses gaseous small molecules such as argon, methane, or nitrogen. The stoichiometry is not integral, but it can be explained on the basis of the hydrate crystal structure and size of the cages.²³

Channel clathrates form when the host crystal contains continuous channels in which the guest can be included. Urea, $(\text{H}_2\text{N})_2\text{C}=\text{O}$, forms channel clathrates with many long-chain molecules as the guests. Such clathrates have been used to isolate guest molecules from mixtures by crystallization in the clathrate form.

The literature on molecular complexes frequently now uses the term *molecular recognition*, which can be taken to mean a noncovalent interaction in which complementary features of the two interactants (exemplified by the hydrogen-bonding sites in Structure 8) result in significant specificity in the complex-formation process.

COMPLEX STABILITY

Binding Constants and Stoichiometric Models

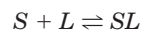
For the general complex-formation equilibrium



the overall binding constant, β_{mn} , is defined

$$\beta_{mn} = \frac{[S_mL_n]}{[S]^m[L]^n} \quad (3)$$

where brackets denote molar concentrations. Actually, complexes probably form in a stepwise fashion by the coming together of two interactant species at a time. For example, the 1:2 complex SL_2 is formed in these two consecutive steps.



Therefore, defining the stepwise binding constants as

$$K_{11} = \frac{[S]}{[S][L]} \quad (4)$$

$$K_{12} = \frac{[SL_2]}{[SL][L]} \quad (5)$$

Algebraic substitution shows that $\beta_{12} = K_{11}K_{12}$. Binding constants also are known as stability constants, formation constants, or association constants. The reciprocal quantity is a dissociation constant or an instability constant. These constants obviously depend on the identities of the substrate, S , and ligand, L ; they also depend on the solvent and the temperature. Throughout most of this discussion the simplest example will be used, that of 1:1 complex formation, to illustrate concepts and methods, but in many situations it also may be necessary to consider the possibility of other stoichiometric ratios.²⁴

The binding constant is an important measure of complex stability, and it is related to the standard free energy of complex formation by

$$\Delta G_{11}^0 = -RT \ln K_{11} \quad (6)$$

where R is the gas constant and T is the absolute temperature. The standard free-energy change is related to the standard enthalpy change ΔH_{11}^0 and the standard entropy change ΔS_{11}^0 by

$$\Delta G_{11}^0 = \Delta H_{11}^0 - T\Delta S_{11}^0 \quad (7)$$

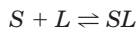
ΔH_{11}^0 can be determined from measurements of K_{11} at several temperatures. From

$$\log K_{11} = \frac{\Delta H_{11}^0}{2.303RT} + \text{constant} \quad (8)$$

a linear plot of $\log K_{11}$ against $1/T$ (a van't Hoff plot) yields ΔH_{11}^0 from the slope. From Equation 7, ΔS_{11}^0 then can be calculated.

Complex stability commonly is discussed in terms of K_{11} , $\log K_{11}$, ΔG_{11}^0 , or (less often) ΔH_{11}^0 .

Before an experimentally measured binding constant can be accepted as a valid measure of complex stability, there must be a firm basis for believing that the stoichiometry has been identified correctly. This is achieved by formulating and then testing a hypothesis. This hypothesis is simply a statement or an equation giving the assumed stoichiometry; for example,



which expresses the assumption of 1:1 stoichiometry. The test of this model consists of showing that K_{11} is a constant over all concentration ranges.

This procedure is oversimplified above, because it ignores nonideality effects that lead to differences between concentrations and activities. A more rigorous discussion is given elsewhere.²⁴

This may be illustrated by building and testing a 1:1 model. The first step is to define K_{11} as in Equation 4. The second step is to write the material-balance relationship for the substrate.

$$S_t = [S] + [SL] \quad (9)$$

Here, S_t is the total substrate concentration. Also f_{11} is defined as the fraction of substrate in the complexed form.

$$f_{11} = [SL]/S_t \quad (10)$$

Algebraic combination of Equations 4, 9, and 10 yields

$$f_{11} = \frac{K_{11}[L]}{1 + K_{11}[L]} \quad (11)$$

Equation 11 is the *binding isotherm* for this model; it shows how f_{11} depends on the free-ligand concentration. The mathematical form of Equation 11 is very important in all 1:1 equilibria. The model is tested by measuring f_{11} (or some experimental variable that is proportional to f_{11}) and showing that it is quantitatively related to $[L]$ by Equation 11.

This procedure is of sufficient importance to be illustrated with a hypothetical example. Suppose $K_{11} = 10 M^{-1}$; by assigning reasonable values to $[L]$ the corresponding values of f_{11} can be calculated with Equation 11. The result is plotted in Figure 14-9. Several features are of interest. The binding isotherm is nonlinear; in fact, it is a rectangular hyperbola. At very low values of the free-ligand concentration the fraction bound rises sharply (the slope is relatively steep), but at high values of $[L]$ the curve flattens out and approaches the value $f_{11} = 1$, asymptotically. This change to a very small slope value at high $[L]$ is called a *saturation effect*; the physical interpretation is that in this region most of the substrate molecules are already bound (complexed) to ligand, so addition of more ligand cannot create additional complex as efficiently as at lower values of f_{11} . Notice, also, that when $f_{11} = 1/2$, $[L] = 1/K_{11}$, as can be seen from

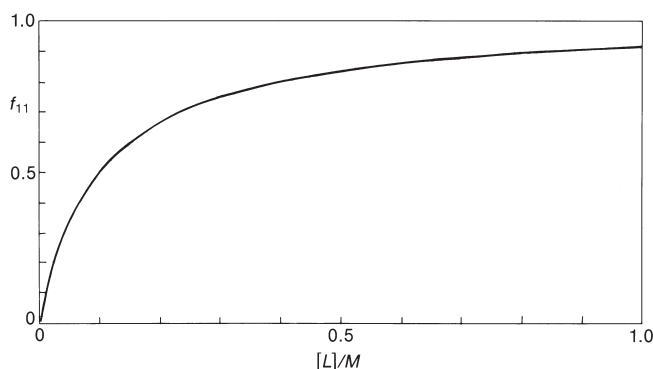


Figure 14-9. Plot of the 1:1 binding isotherm, Equation 11, with $K_{11} = 10 M^{-1}$.

Equation 11. This condition is familiar in the context of acid–base chemistry, for at the condition of half-neutralization, $[H^+] = K_a$ or $pH = pK_a$, where K_a is defined to be a dissociation constant.

One way to test the assumed model against the experimental data is to perform a nonlinear least-squares regression analysis of f_{11} on $[L]$ according to Equation 11, observing the goodness of fit of the regression line to the data points. Another way is to rearrange Equation 11 into a linear form and plot the data accordingly. For example, Equation 11 is transformed easily to the “double-reciprocal” form

$$\frac{1}{f_{11}} = \frac{1}{K_{11}[L]} + 1 \quad (12)$$

This predicts that a plot of $1/f_{11}$ against $1/[L]$ will be linear if the model is valid. Other linear transformations also are possible, as shown below. Note that K_{11} can be evaluated from the slope of the double-reciprocal plot.

The *Michaelis-Menten equation* of enzyme kinetics has the same mathematical form as Equation 11, because it is based on the formation of a 1:1 enzyme–substrate complex. Another important example arises in the study of the binding of drugs to proteins. The simplest model of this process supposes that the protein possesses n identical, independent binding sites for drug L , each site having a site-binding constant of k . Mathematical treatment of this model leads to Equation 13 as the isotherm.

$$\bar{i} = \frac{nk[L]}{1 + k[L]} \quad (13)$$

where \bar{i} is defined to be the average number of drug molecules bound per protein molecule at free-drug concentration $[L]$; \bar{i} is defined by

$$\bar{i} = \frac{L_t - [L]}{S_t} \quad (14)$$

where L_t is the total drug concentration and S_t is the total protein concentration. The quotient $\bar{i}/n = \theta$ is called the degree of saturation.

Once again, in Equation 13, can be seen the characteristic hyperbolic dependence on ligand concentration. Drug–protein-binding often is analyzed with the aid of another linear transformation, Equation 15, according to which a plot of $\bar{i}/[L]$ against \bar{i} will be linear.

$$\frac{\bar{i}}{[L]} = -k \cdot \bar{i} + n/k \quad (15)$$

From the slope and either intercept, the parameters n and k can be estimated. A plot according to Equation 15 is called a Scatchard plot. If the Scatchard plot is curved, evidently the simple model leading to Equation 13 is not valid.

Measurement of Complex Stability

If a property of the substrate is altered upon its complexation with a ligand, measurement of the property as a function of ligand concentration provides a means for estimating the binding constant. Many properties are suitable for this purpose. To demonstrate the method, a 1:1 complex formation will be used as a model, for just a few of these.

SPECTROMETRY—Suppose the absorption spectrum of the substrate is changed significantly upon binding. Figure 14-10 shows a typical example in which the ultraviolet spectrum of *p*-nitrophenol changes upon complexation with α -cyclodextrin. The presence of well-defined isosbestic points is consistent with the assumption of 1:1 stoichiometry. Selecting a wavelength at which a substantial change in absorption occurs and assuming that Beer’s law is obeyed by all species, then, at total substrate concentration S_t in the absence of a ligand, the solution

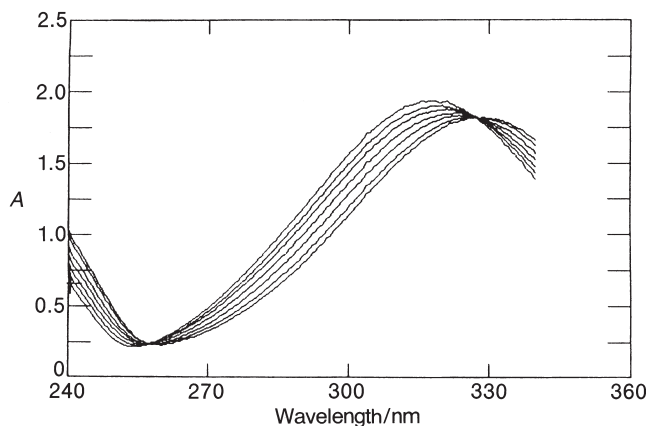


Figure 14-10. Ultraviolet absorption spectrum of *p*-nitrophenol in the presence of varying concentrations of α -cyclodextrin. The *p*-nitrophenol concentration is $1.99 \times 10^{-4} M$, and the cyclodextrin concentration ranges from zero (topmost spectrum) to $0.01 M$.

absorbance is

$$A_0 = \epsilon_S b S_t \quad (16)$$

where b is the path length and ϵ_S is the molar absorptivity. In the presence of a ligand the absorbance is

$$A_L = \epsilon_S b [S] + \epsilon_L b [L] + \epsilon_{11} b [SL] \quad (17)$$

where ϵ_{11} is the absorptivity of the complex. Combining Equation 17 with the mass balances $S_t = [S] + [SL]$ and $L_t = [L] + [SL]$ gives

$$A_L = \epsilon_S b S_t + \epsilon_L b L_t + \Delta \epsilon_{11} b [SL] \quad (18)$$

where $\Delta \epsilon_{11} = \epsilon_{11} - \epsilon_S - \epsilon_L$. If the solution absorbance is measured against a reference solution containing the same total ligand concentration, L_t , the measured absorbance is

$$A = \epsilon_S b S_t + \Delta \epsilon_{11} b [SL] \quad (19)$$

Equation 19 is combined with Equation 4 to give Equation 20, the binding isotherm, where $\Delta A = A - A_0$.

$$\frac{\Delta A}{b} = \frac{S_t K_{11} \Delta \epsilon_{11} [L]}{1 + K_{11} [L]} \quad (20)$$

Equation 20 is identical in form with Equation 11, and it can be analyzed in the same way. The two unknown parameters K_{11} and $\Delta \epsilon_{11}$ are obtained from this analysis. From the data shown in Figure 14-10 the values $K_{11} = 256 M^{-1}$ and $\Delta \epsilon_{11} = -1726 M^{-1} \text{ cm}^{-1}$ (at 317 nm) were obtained.

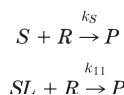
There is a further matter to consider in this treatment of the data. Equation 20 is expressed in terms of free-ligand concentration $[L]$, but only the total ligand concentration L_t is known. From the relationship $L_t = [L] + [SL]$ is found

$$L_t = [L] + \frac{S_t K_{11} [L]}{1 + K_{11} [L]} \quad (21)$$

The assumption that $[L] = L_t$, is used widely, but when this approximation is not justified, $[L]$ must be estimated with the aid of Equation 21. Methods have been devised to solve this problem.²⁵

This spectrometric method is applicable in the ultraviolet, visible, and infrared regions. Nuclear magnetic resonance (NMR) spectrometry can be applied in a similar manner, but with NMR a change in the *chemical shift* is measured.

CHEMICAL REACTIVITY—If the rate of a chemical reaction (such as hydrolysis) undergone by the substrate is either increased or decreased by binding to the ligand, the stability constant can be measured. Consider this kinetic scheme:



Here, R is a reagent that reacts with S and SL , but does not form complexes, P is the product of the reaction, and k_S , k_{11} are second-order rate constants. The mathematical development is similar to that in the spectrometric treatment, and the result is

$$\frac{k_S - k'_S}{k_S} = \frac{q_{11} K_{11} [L]}{1 + K_{11} [L]} \quad (22)$$

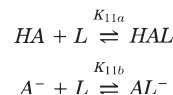
where q_{11} is given by $q_{11} = 1 - k_{11}/k'_S$ and k'_S is the measured second-order rate constant in a solution having ligand concentration, $[L]$. If the reaction rate is decreased upon binding, then $k'_S < k_S$, and q_{11} will lie between 0 and 1. Equation 22 has the usual hyperbolic form, and is treated as described earlier for similar functions.

POTENTIOMETRY—If the activity of an ion is changed upon complex formation, it may be possible to make use of the measurement of electrical potential, E , according to the *Nernst equation*:

$$E = \text{constant} + \frac{RT}{nF} \ln a$$

where a is the ion activity, n is the number of electrons in the redox process, and F is the Faraday. Potentiometry is the most widely used method for the study of metal-ion coordination complexes, for which the activity of the metal ion, the ligand, or the hydrogen ion may be measured.²⁶

Potentiometry also is applicable to structures that are weak acid-bases. If HA and A^- are the conjugate acid and base of such a substrate, with L being the ligand, two possible complexes can form:



The experiment consists of measuring the apparent acid dissociation constant K'_a of HA in the presence of the ligand. Mathematical treatment gives Equation 23 as the binding isotherm, where $'pK'_a = pK'_a - pK_a$, and pK_a is the value when $L_t = 0$.

$$\Delta pK'_a = \log \frac{(1 + K_{11a} [L])}{(1 + K_{11b} [L])} \quad (23)$$

Thus, if $\Delta pK'_a \neq 0$, $K_{11a} \neq K_{11b}$; that is, the conjugate acid and base forms of the substrate have different affinities for the ligand. The sign of $\Delta pK'_a$ indicates which form of the substrate forms the stronger complex, and K_{11a} and K_{11b} can be evaluated from the dependence of $\Delta pK'_a$ on $[L]$.

SOLUBILITY—In this technique the total apparent solubility, S_t , of the substrate is measured as a function of total ligand concentration, L_t . Because the system is prepared to contain excess (solid) substrate, the free-substrate concentration is maintained constant at its intrinsic molar solubility, s_0 . Therefore, the mass balance on substrate can be written

$$S_t = s_0 + [SL]$$

which, combined with Equation 4 and $L_t = [L] + [SL]$, yields

$$S_t = s_0 + \frac{K_{11} s_0 L_t}{1 + K_{11} s_0} \quad (24)$$

Equation 24 predicts that S_t is a linear function of L_t . The binding constant is obtained with

$$K_{11} = \frac{\text{slope}}{s_0(1 - \text{slope})} \quad (25)$$

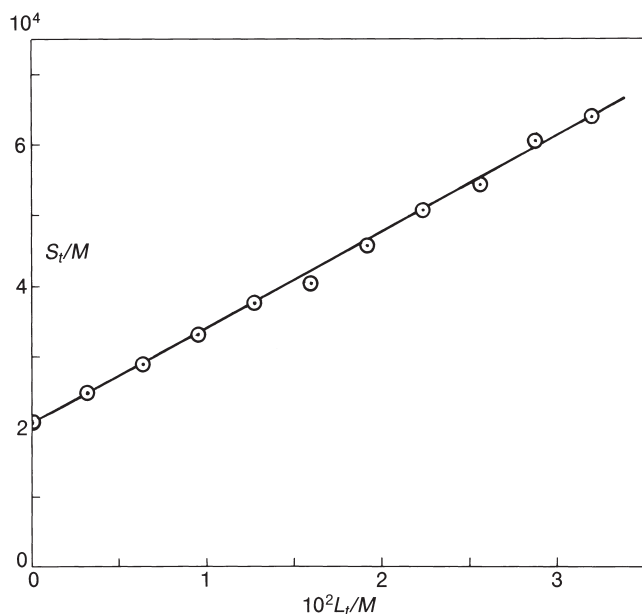


Figure 14-11. Solubility (S_t) of naphthalene as a function of concentration (L_t) of theophylline, in water at 25°.

Figure 14-11 is a plot according to Equation 24 for the system naphthalene (substrate)–theophylline (ligand). The equilibrium constant evaluated with Equation 25 is $K_{11} = 64 M^{-1}$.

It will be noted that in the solubility method, the isotherm is linear, rather than hyperbolic. This is because $[S]$ is held constant in this method, whereas in the methods discussed earlier, S_t is constant and $[S]$ varies.

There are other methods that, like the solubility method, involve a distribution between two phases. The apparent partition coefficient of a solute between two immiscible solvents can be a measure of complex formation. Several chromatographic methods are based on a similar principle, the retention volume, or time, of a substrate being measured as a function of ligand concentration.

DIALYSIS—This is a technique applicable when one inter-actant, such as the substrate, is a very large molecule, and the other, the ligand, is a small molecule. Therefore, it is used widely to study the binding of drugs to proteins.

In dialysis, two compartments containing solvent are separated by a semipermeable membrane, i.e., a membrane whose pores permit the free transport of the small ligand molecules but do not permit the passage of the large substrate molecule. In one compartment (No 1) this nondiffusible substrate is placed, and in the other (No 2) the diffusible ligand is placed. The system then is allowed to come to equilibrium.

At equilibrium the free-ligand concentration $[L]$ is equal in the two compartments. The solutions in the two compartments are analyzed for their total ligand concentrations.

With the above designations of compartment numbers, we can write

$$(L_t)_1 = [L]_1 + [\text{bound } L]_1$$

$$(L_t)_2 = [L]_2$$

and the equilibrium condition is $[L]_1 = [L]_2$. Therefore, \bar{i} can be calculated for compartment No 1 using Equation 14, because S_t , the total protein or macromolecule concentration, is known. The experiment is repeated at different ligand concentrations to obtain \bar{i} as a function of $[L]$. The data then are analyzed in terms of the model equation.²⁵

Factors Affecting Complex Stability

This is a large and poorly understood subject so any treatment must be cursory. Much of the earlier discussion on bonding and intermolecular forces is pertinent here.

Consider a general effect that operates in all systems having multiple equilibria. In the simplest example there exists a substrate, S , with n identical independent binding sites, so that complexes $SL, SL_2, SL_3, \dots, SL_n$ may form, with corresponding binding constants $K_{11}, K_{12}, K_{13}, \dots, K_{1n}$. Even though the binding sites are identical, it will be found that $K_{11} > K_{12} > K_{13}, \dots, > K_{1n}$. This is a result of a *statistical effect*. The origin of the statistical effect can be demonstrated readily for the case $n = 2$. The formation of the 1:1 complex is favored over the formation of the 1:2 complex by a factor of 2, because there are two available sites for binding in reactant S , whereas there is only one available site in reactant SL . Moreover, dissociation of SL_2 is favored over dissociation of SL by a factor of 2 because SL_2 has twice as many ligands to surrender. The combination of these statistical factors leads to the result $K_{11} = 4K_{12}$, solely as a consequence of the statistical effect. This argument was generalized by Jones.²⁷

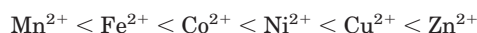
Considering the stability of metal-ion coordination complexes, when successive complexes form, two additional factors may operate in addition to the statistical effect.

One of these is the *steric effect*, which is a result of the bulky nature of the ligand (relative to the H_2O that it replaces). As successive ligands are added to the metal ion, crowding inhibits the addition of the next ligand, resulting in a decrease in the value of the binding constants.

A second factor is the *electrostatic effect*, which plays a role when the central cation complexes with an anionic ligand. Then, as successive ligands approach the central ion they experience different fields, because the net charge on the central ion changes with the addition of each ligand.²⁶

The *chelate effect* was mentioned earlier in this chapter. The formation of a cyclic complex upon binding of a metal ion to a multidentate ligand leads to greater complex stability than when the same metal ion complexes with an analogous unidentate ligand. Complex stability is favored especially by the formation of 5- and 6-membered rings. A multidentate ligand that is also a macrocycle (such as a crown ether) can form particularly strong complexes; this is called the *macrocyclic effect*.²⁸

A useful approach in understanding complex stability is to seek correlations of stability with other properties of the inter-actants. For example, the *Irving-Williams order* of stability of complexes of divalent cations with a common ligand,



can be correlated with the ionization potentials (corresponding to the last electron lost) of the ions. Similarly, for complexes of a common metal ion, with a series of structurally related ligands of measurable Brønsted basicity, the complex stabilities (expressed as the logarithms of the binding constants) often are correlated linearly with the pK_a values of the bases.²⁹ Bases of different structural classes (eg, aliphatic primary amines or substituted pyridines) usually give rise to different lines, showing that basicity is not the only controlling feature. The hard–soft acid–base concept described earlier provides additional insight into the effects that properties such as polarizability, electronegativity, ionization potential, electron affinity, and basicity can have in affecting complex stability.

In molecular complexes, it is useful to start with Equation 2, in which ΔG_{net} corresponds to ΔG_{11}^0 in Equation 6. The value of ΔG_{net} is determined by the three terms ΔG_{MM} , ΔG_{MS} , and ΔG_{SS} . If one of these terms greatly predominates over the others, then fairly simple correlations between ΔG_{11}^0 and a molecular property related to the dominant term might be expected. If, however, two or three terms contribute significantly to ΔG_{net} , they may combine in complicated ways, perhaps even opposing each other, so clear relationships may not be observed. Often the

most fruitful experiments are those in which one interactant is held as a constant feature, and changes in the structure of the other interactant are made.

Table 14-5 provides some theoretical guidance. If solute–solute interactions of the dipole or induced-dipole type are important, one might anticipate correlations with interactant dipole moment or polarizability. In charge-transfer complexing, substituent effects that increase electron density in the donor or decrease it in the acceptor (Structures 5, 6, and 7 are examples of the latter type) may be expected to increase complex stability. Such effects have been observed.^{30,31}

If the hydrophobic interaction makes an important contribution to complex stability, the incorporation of organic solvents will reduce the stability. According to the cavity theory of the hydrophobic effect, complex stability is related to the change in surface area upon complex formation, so it may be anticipated that, for such systems, complex stability is related to the size of the interactants. Such a dependence has been seen, but it is complicated by the presence of additional effects.³² Another prediction of the cavity model is that, for a given complex, stability should be determined primarily by the solvent surface tension, and there is some experimental support for this prediction.^{17,21,33}

COMPLEXES IN PHARMACY

APPLICATION TO DRUG DELIVERY—Some of the properties of a drug are so pertinent to dosage forms and drug delivery that it is reasonable to identify them as pharmaceutical or biopharmaceutical properties. Complex formation may affect these properties, sometimes to advantage and sometimes adversely. Many of these properties, with corresponding examples of drug complexes, are given in Table 14-9.³⁴

A dosage form might be prepared either with the separate components *S* (the substrate or drug) and *L* (the ligand or complexing agent), or with the preformed solid complex.

In a solution dosage form the method of preparation makes no difference, because the complexation equilibrium immediately establishes the equilibrium composition. It must be remembered that the fraction of drug in the complexed form is given by Equation 11, so that the free-ligand concentration is a critical variable, and excess ligand may have to be added in order to “drive the equilibrium” in favor of the bound (complexed) form.

In a solid dosage form it may be preferable to incorporate the solid complex rather than a physical mixture of the drug and complexing agent. For many systems it has been shown that the complex provides faster dissolution and greater bioavailability than does the physical mixture. The processing characteristics (physical state, stability, flowability, etc) of the complex also may be better than those of the free drug.

Not all complexation is intentional or desirable, and some dosage-form *incompatibilities* may be the result of unwanted complexation reactions. For example, some widely used polyethers (Tweens, Carbowaxes, or PEGs) can form precipitates with H-bond donors such as phenols and carboxylic acids.

A substance used widely in liquid dosage forms as a complexer of metal ions is EDTA (ethylenediaminetetraacetic acid). The purpose of this application of complexation is to improve drug stability by inhibiting reactions (usually oxidations) that are catalyzed by metal ions, the complexed form of the metal ion being catalytically inactive. Citric acid (in the form of the citrate anion) also is used for this purpose.³⁵

The cyclodextrins have been shown to have effects on all of the properties listed in Table 14-9, and many pharmaceutical applications have been proposed.^{19,20,36,37}

COMPLEXES IN PHARMACEUTICAL ANALYSIS—The formation of metal-ion coordination complexes provides the basis of many analytical methods for the determination of metals. Titration of divalent and trivalent metal ions with a solution of EDTA is a standard procedure called complexometric or

Table 14-9. Pharmaceutical Properties Affected by Complexation

PROPERTY	EXAMPLE ^{a,b}
Physical state	Nitroglycerin-cyclodextrin
Volatility	Iodine-PVP
Solid-state stability	Vitamin A-cyclodextrin
Chemical stability	Benzocaine-caffeine
Solubility	Aspirin-caffeine
Dissolution rate	Phenobarbital-cyclodextrin
Partition coefficient	Benzoic acid-caffeine
Permeability	Prednisone-dialkylamides
Absorption rate	Salicylamide-caffeine
Bioavailability	Digoxin-cyclodextrin
Biological activity	Indomethacin-cyclodextrin

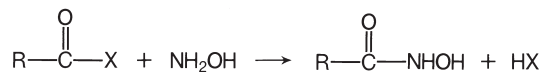
^a Listed in order of drug-complexing agent.

^b Citations of the original literature will be found in Ref 34.

chelometric titration.³⁸ The theoretical titration curve is calculated readily, and it can be shown that the very large endpoint “break” is the result of the 1:1 stoichiometry between the metal ion and the multidentate EDTA tetraanion. The endpoint can be detected visually with metallochromic indicators or, potentiometrically, with ion-selective membrane electrodes.

Very low concentrations of metal ions can be determined spectrometrically by complexation with a ligand that produces a spectral change. If the complex absorbs in the visible region of the spectrum, this is called colorimetric analysis. Thousands of such methods have been developed.³⁹ Two examples are the determination of Fe(III) by complexation with 1,10-phenanthroline (see Table 14-1), and of Hg(II) by complexation with dithizone (diphenylthiocarbazone), S=C(NHNHC₆H₅)₂. Gravimetric analysis of metal ions can be accomplished via their precipitation as insoluble coordination complexes. For example, Ni(II) forms an insoluble square planar bis(dimethylglyoxime) complex, and many metal ions yield insoluble complexes with 8-hydroxyquinoline (see Table 14-1 for the structures of these ligands).

In some instances the analytical situation can be reversed to make the metal ion serve as the analytical reagent and the organic ligand as the sample. The *ferric hydroxamate* method for the detection and determination of carboxylic acid derivatives is a good example, in which a carboxylic acid derivative such as an ester, amide, or anhydride is reacted with hydroxylamine to form the corresponding hydroxamic acid.



An excess of Fe(III) is added, and this forms a red-violet coordination complex with the hydroxamic acid; the concentration of the complex is determined spectrometrically.

Colorimetric analyses also can be based on molecular complex formation. Recall that charge-transfer complexation often is accompanied by the development of an intense charge-transfer absorption band, and this can be put to analytical use. For example, tertiary amines can be determined spectrometrically by complexation with tetracyanoethylene (Structure 5).

Many complex formation reactions are used in conjunction with, or as the basis for, a separation, either by liquid–liquid extraction or chromatography. A classical method for amines, the *acid-dye method*, is based upon complex formation between an amine and a dye molecule. The complex is extracted from the aqueous phase in which it is formed into an organic solvent, where the dye concentration is measured spectrometrically.

The success of the method is based on the condition that only the complexed form of the dye is extractable, so each molecule of amine results in the complexation of one molecule of dye, and this is extracted into the organic phase, where its concentration is an indirect measure of the amount of amine. In order to ensure the nonextractability of the excess (uncomplexed) dye, a

dye is used that is a neutral weak acid, and the aqueous pH is controlled at a level above the pK_a of the dye, thus converting it to its anionic form.⁴⁰ The principle can be reversed to determine acidic compounds with basic dyes.⁴¹ In a similar way metal ions may be extracted into organic solvents upon complexation with hydrophobic ligands.

Chromatographic separations can make use of the same principle, most notably in a technique called *ion-pair chromatography*. In an application of great pharmaceutical importance, an amine sample in its cation form is complexed with a hydrophobic anion (eg, an alkyl sulfonate, RSO_3^-), and reverse-phase liquid chromatography is performed. The mobile phase is polar (often aqueous), and the stationary phase is nonpolar (eg, a C-18-bonded packing). Although the protonated amine has little affinity for the nonpolar stationary phase, its complex (called an *ion-pair*) with the hydrophobic counterion masks its polar nature, and the ion-pair can partition between the two chromatographic phases.

Several other forms of chromatography take advantage of complex formation between a sample solute and a molecular entity in the stationary phase to generate selective chromatographic retention behavior.

In *hydrophobic chromatography* the hydrophobic interaction provides the driving force for association.

Affinity chromatography is based on fairly specific interactions between the migrating solute and a ligand that is chemically bonded to the stationary phase. For example, an enzyme can be isolated by affinity chromatography on a column prepared with an inhibitor of the enzyme; formation of the enzyme-inhibitor complex on the column removes the enzyme from the sample mixture. In a similar way the very specific antigen-antibody interaction can be applied to isolate antibodies.

Another type of chromatography based on complex formation is *chiral chromatography*, used to separate optical isomers based on interactions between the isomers and a stationary phase that possesses chiral binding sites. For example, stationary phases have been prepared with covalently bound cyclodextrins, which are capable of effecting chiral separations.

PROTEIN-BINDING OF DRUGS—Systemically delivered drugs are made available to the tissues and organs of the body by means of the blood, which is a complicated mixture of substances, some of which are capable of forming complexes with drugs. Because it is widely accepted that the pharmacological response to a drug is determined by the concentration of the “free” (ie, unbound, uncomplexed) drug rather than the total drug concentration, drug-binding by constituents of the blood has important practical implications.

Of all the constituents of blood that might take part in complex formation, the most important and most studied is the protein serum albumin (HSA for human serum albumin, BSA for the closely related bovine serum albumin). The normal HSA concentration in the blood is remarkably high, being 3.5 to 4.5 g/100 mL, and the concentration can vary with age, exercise, stress, and disease.⁴² It is a very soluble, very stable protein and consists of 585 amino acid residues, having a calculated molecular weight of 66,439 and a net charge of -15 units at pH 7. The amino acid sequence is known.⁴³

Serum albumin is a strikingly indiscriminate complexing agent, having a significant affinity for very many compounds, including drugs. The molecule appears to be appreciably flexible and able to adapt its shape to fit the molecular shape of the ligand binding to it. There are multiple binding sites, but the number that are accessible appears to depend upon the particular ligand; moreover, not all the sites are equivalent.⁴² The principal driving force for complexing is the hydrophobic interaction, and hydrophobic compounds, such as long-chain fatty acids (actually as their anions at physiological pH) are bound avidly to HSA. Typical site-binding constants are 10^4 to $10^8 M^{-1}$. Certain metal ions also can bind to HSA, and the complex with Cu(II) is particularly stable.

Since the binding sites of HSA are evidently not all identical, the simple binding model exemplified by Equation 13 is not

applicable precisely, but this equation often forms the basis for discussions of the binding equilibria. Provided that this oversimplification is recognized, some useful insights can be gained. The symbolism is recast as follows: let P = protein, P_t = total protein concentration, D = drug (ligand), D_t = total drug concentration, and $[D]$ = free (unbound) drug concentration. Then $\bar{i} = (D_t - [D])/P_t$, is the average number of drug molecules bound per molecule of protein at free-drug concentration $[D]$. Equation 13 now is written

$$\bar{i} = \frac{nk[D]}{1 + k[D]} \quad (26)$$

In the context of drug-protein binding, workers often make use of the concepts *fraction of drug bound* (f_b) and *fraction of drug unbound* (f_u). Obviously, $f_b + f_u = 1$. One can write the definitions $f_u = [D]/D_t$ and $f_b = (D_t - [D])/D_t$. Algebraic combination of these expressions leads to

$$f_b = \frac{nkP_t}{1 + k[D] + nkP_t} \quad (27)$$

and

$$f_u = \frac{1 + k[D]}{1 + k[D] + nkP_t} \quad (28)$$

Equations 27 and 28 show that f_b and f_u depend upon the concentrations of both the protein and the drug. Clearly, however, when $k[D] \ll 1$ (ie, at very low free-drug concentrations), f_b and f_u essentially become independent of drug concentration, but this condition may not always hold in a therapeutic situation. Moreover, because f_b increases as P_t increases, changes in serum protein concentration as a result of physiological or pathological states may result in significant alterations in free-drug levels. Another implication is that for a strongly bound drug (high k) the protein-binding sites may become saturated with drug, so at higher doses a larger fraction of drug is in the free form.

There are several pharmacological or pharmacokinetic consequences of drug-protein binding.⁴⁴ Besides binding to proteins in the blood, drugs also may bind to constituents of the tissues in the organs perfused by the blood supply. If the binding ability of the blood (which may be roughly measured by the product nk) is greater than that of the tissues, the drug will tend to be retained in the blood, whereas tissue retention can occur for the opposite situation. Thus, the distribution of the drug can be affected by its binding characteristics.

The drug clearance also can be affected. If the extraction ratio for a tissue is high, the clearance is determined primarily by blood flow, and blood-protein binding has little effect on the clearance; if the extraction ratio is small, the clearance depends upon binding, and only the free drug is cleared.⁴⁴

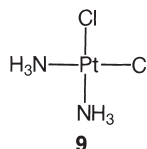
The pharmacokinetic parameters, volume of distribution, and elimination-rate constant may be dose-dependent if the protein can be saturated by the drug. A high loading dose may be appropriate in such a case to saturate the protein, followed by lower maintenance doses. It generally is advisable to perform experimental studies (eg, by dialysis) that allow free-drug concentration, as well as total-drug concentration, to be determined. Such studies can detect nonlinear dependencies of $[D]$ on D_t (ie, nonconstancy of f_b and f_u) and, therefore, can be helpful in developing dosage regimens to optimize therapeutic response and minimize undesirable side effects.

COMPLEXES IN THERAPEUTICS—Complexes occur widely in biological systems, so the application of complex formation processes in therapy is a reasonable approach to drug design. Among the most obvious and important biological manifestations of complexation are many metal-ion coordination complexes, whose study in this context constitutes a large part of bioinorganic chemistry. Examples of these complexes, with the metals involved, are hemoglobin (iron), cytochrome (iron), carboxypeptidase A (zinc), carbonic anhydrase (zinc), superox-

ide dismutase (zinc and copper), vitamin B₁₂ (cobalt), chlorophyll (magnesium), and urease (nickel). Molecular complexation in biological systems also occurs, as noted earlier for DNA base-pairing and stacking interactions. The folding of proteins is a consequence of intramolecular noncovalent interactions. Charge-transfer interactions may play a role in physiological processes, and some membrane-transport processes may involve inclusion phenomena.

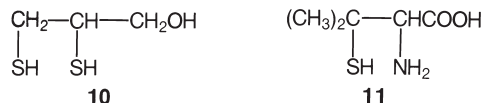
Numerous antimicrobial and antineoplastic agents are believed to exert their action by means of complex formation with DNA base-pairs. These drug molecules are large planar aromatic compounds, and they can be inserted between the planar base-pair assemblies in the DNA double helix; this type of inserted molecular interaction is called *intercalation*. Intercalating drugs include ethidium, quinacrine, proflavine, daunorubicin, adriamycin, and actinomycin D.⁴⁵

The structure of *cis*-dichlorodiammineplatinum(II) (*cis*-platin, Structure 9) is unusual for a drug.

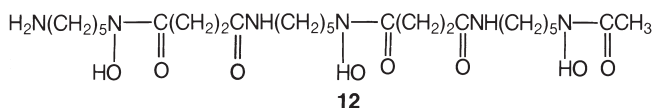


This antineoplastic drug is a square planar complex of Pt(II). Its biological activity probably arises from the *cis* geometry.

Many toxic effects of excessive metal-ion concentrations can be treated by agents that form strong coordination complexes, via chelation, thus aiding the excretion of the metal. Among the metals whose toxicity can be treated by *chelation therapy* are iron, lead, copper, cobalt, nickel, mercury, and zinc. The standard chelating agents for this purpose are the monocalcium disodium salt of EDTA, dimercaprol (BAL, Structure 10), and D-penicillamine (Structure 11).



Iron poisoning is treated with the chelator, deferoxamine, Structure 12.



REFERENCES

- Basolo F, Pearson RG. *Mechanisms of Inorganic Reactions*, 2nd ed. New York: Wiley, 1967, Chap 1.
- Basolo F, Pearson RG. *Mechanisms of Inorganic Reactions*, 2nd ed. New York: Wiley, 1967, p 141.
- Basolo F, Pearson RG. *Mechanisms of Inorganic Reactions*, 2nd ed. New York: Wiley, 1967, p 375.
- Cotton FA, Wilkinson G. *Advanced Inorganic Chemistry*, 4th ed. New York: Wiley-Interscience, 1980, p 619.
- Basolo F, Pearson RG. *Mechanisms of Inorganic Reactions*, 2nd ed. New York: Wiley, 1967, p 60.
- Pauling L. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, NY: Cornell University Press, 1960, Chap 5.
- Jones MM. *Elementary Coordination Chemistry*. Englewood Cliffs, NJ: Prentice-Hall, 1964, p 133.
- Jones MM. *Elementary Coordination Chemistry*. Englewood Cliffs, NJ: Prentice-Hall, 1964, p 144.
- Hanzlik RP. *Inorganic Aspects of Biological and Organic Chemistry*. New York: Academic Press, 1976, p 97.
- Basolo, Pearson, *Mechanisms of Inorganic Reactions*, p 104.
- Pearson RG. *J Chem Educ* 1987; 64: 561.
- Israelachvili JN. *Intermolecular and Surface Forces*. New York: Academic Press, 1985, Chap 2.
- Israelachvili JN. *Intermolecular and Surface Forces*. New York: Academic Press, 1985, p 98.
- Mulliken RS, Person WB. *Molecular Complexes*. New York: Wiley-Interscience, 1969, Chap 1.
- Tanford C. *The Hydrophobic Effect*, 2nd ed. New York: Wiley-Interscience, 1980.
- Jencks WP. *Catalysis in Chemistry and Enzymology*. New York: McGraw-Hill, 1969, p 417.
- Connors KA, Mulski MJ, Paulson A. *J Org Chem* 1992; 57: 1794.
- Watson JD. *Molecular Biology of the Gene*, 2nd ed. New York: WA Benjamin, 1970, p 132.
- Szejtli J. *Cyclodextrins and Their Inclusion Complexes*. Budapest: Akademiai Kiado, 1982.
- Duchene D, ed. *Cyclodextrins and Their Industrial Uses*. Paris: Editions de Santé, 1987.
- Connors KA. *Chem Rev* 1997; 97: 1325.
- Hagan M. *Clathrate Inclusion Compounds*. New York: Reinhold, 1962.
- Tsoucaris G. In: Duchene D, ed. *Cyclodextrins and Their Industrial Uses*. Paris: Editions de Santé, 1981, Chap 1.
- Connors KA. *Binding Constants: The Measurement of Molecular Complex Stability*. New York: Wiley-Interscience, 1987.
- Connors KA. *Binding Constants: The Measurement of Molecular Complex Stability*. New York: Wiley-Interscience, Chap 2.
- Hartley FR, Burgess C, Alcock RM. *Solution Equilibria*. Chichester: Ellis Horwood/Halsted Press, 1980.
- Jones MM. *Elementary Coordination Chemistry*. Englewood Cliffs, NJ: Prentice-Hall, 1964, p 333.
- Cotton FA, Wilkinson G. *Advanced Inorganic Chemistry*, 4th ed. New York: Wiley-Interscience, 1980, p 73.
- Hanzlik RP. *Inorganic Aspects of Biological and Organic Chemistry*. New York: Academic Press, 1976, p 118.
- Andrews LJ, Keefer RM. *Molecular Complexes in Organic Chemistry*. San Francisco: Holden-Day, 1964, Chap 4.
- Gur'yanova EN, Gol'dshtein IP, Romm IP. *Donor-Acceptor Bond*. New York: Wiley, 1975, Chap 5.
- Cohen JL, Connors KA. *J Pharm Sci* 1970; 59:1271.
- Connors KA, Sun S. *J Am Chem Soc* 1971; 93:7239.
- Connors KA. *Pharm Mfg* 1985; 2(9):23.
- Connors KA, Amidon GL, Stella VJ. *Chemical Stability of Pharmaceuticals*, 2nd ed. New York: Wiley-Interscience, 1987, p 100.
- Pitha J, Szente L, Szejtli J. In: Bruck SD, ed. *Controlled Drug Delivery*, Vol I. Boca Raton, FL: CRC Press, 1983, Chap 5.
- Duchene D, Vaution C, Glomot F. *Drug Develop Ind Pharm* 1986; 12:2193.
- Connors KA. *A Textbook of Pharmaceutical Analysis*, 3rd ed. New York: Wiley-Interscience, 1982, Chap 4.
- Sandell EB. *Colorimetric Determination of Traces of Metals*, 3rd ed. New York: Interscience, 1959.
- Higuchi T, Bodin JI. In: Higuchi T, Brochmann-Hansen E, eds. *Pharmaceutical Analysis*. New York: Interscience, 1961.
- Pezes M, Bartos J. *Colorimetric and Fluorometric Analysis of Organic Compounds and Drugs*. New York: Dekker, 1974, p 139.
- Bridges JW, Wilson AGE. *Prog Drug Metab* 1978; 1:193.
- Peters T Jr. *Adv Protein Chem* 1985; 37:161.
- Tillement J-P, et al. *Adv Drug Res* 1984; 13:59.
- Wilson WD, Jones RL. *Adv Pharmacol Chemother* 1981; 18:177.

ACKNOWLEDGMENTS—Kenneth A. Connors, PhD is acknowledged for his efforts in previous editions of this work.

Thermodynamics

Timothy S Wiedmann, PhD



Thermodynamics rests upon three basic laws that took over 500 years to establish. Although quantum mechanics has defined the limits of its scope, the concepts laid out in this chapter have remained unchanged for over a century. The reader is therefore encouraged to appreciate not only the many years of effort spent in defining the laws of thermodynamics but also the likely fact that the contents will be relevant for a lifetime of applications.

The approach will involve the development of concepts within the framework of very specific examples, as the great value of thermodynamics lies in its general applicability. Simple examples will be used to introduce the concepts that form the basis of a thermodynamic description.

A *system* is that part of the universe under consideration and, as such, is separated from the *surroundings* or, equivalently, the rest of the universe. The focus of the analysis will center on how the properties of a system are altered through an interaction with the surroundings. The interaction occurs at the boundary that separates the system from the surroundings.

When a sufficient number of properties of the system have been specified as fixed values, then the system is at equilibrium. Certain systems at equilibrium have a simple equation that provides a relationship among the values of the properties. For example, a system containing an ideal gas has the properties of pressure, P , volume, V , number of mols, n and temperature (K), T related by

$$PV = nRT \quad (1)$$

where R is the gas law constant. Such a relationship is referred to as an *equation of state* because it specifies the relationship among the properties of a system in a definite state. Furthermore, if the system is at equilibrium, only three of the above values for the properties need to be specified, as the fourth may be calculated from the equation of state.

THE FIRST LAW

The first law of thermodynamics is a statement of the principle of conservation of energy; energy may neither be created nor be destroyed. It is mathematically written as

$$dE = \delta q - \delta W \quad (2)$$

where dE is the differential change in the internal energy, δq is the differential change in the absorbed heat, and δW is the differential change in the expended work.

The change in internal energy of a system in going from state A to state B is given by

$$\Delta E = \int_A^B dE \quad (3)$$

From this equation, it is observed that the internal energy is a state function since d represents an exact differential.

The implication is that the change in the internal energy depends only on the initial and final state and does not depend on how the change in state was achieved. Because the change in energy does not depend on the path, there is no net change in the energy for any system that undergoes a cyclic change. The expression is

$$\oint dE = 0 \quad (4)$$

This fact will be useful when a system undergoing a cyclic change is considered, as will be encountered with the discussion of a heat engine. And, finally, the equation provides only a relation for the change in the internal energy and does not provide an absolute value of the internal energy of the system in a particular state.

In the first law, the change in internal energy is related to heat flow and work done. The concepts of work and heat now will be defined precisely, thereby providing the framework for the use of the first law, as well as the other laws. In contrast to the internal energy, the differential change in heat and work are inexact differentials. This means that neither the heat nor the work are state functions of the system; thus, the integral of the differential depends on the path taken. To elaborate on this point, consider the change in going from state A to state B . The heat and work are given as

$$q = \int_A^B \delta q \text{ and } W = \int_A^B \delta W \quad (5)$$

which will depend on what path was taken in going from state A to state B . The work and heat may be determined only if more information is provided concerning how the change in the state of the system was achieved.

WORK—The concept of *work* in thermodynamics may be expressed as a product of an *intensity factor* and a *capacity factor*; for example, mechanical work is given as

$$\delta W = Fdl \quad (6)$$

where the differential quantity of work done, δW , is the product of the force, F (intensity factor), and a differential distance, dl (capacity factor). Other types of work include *gravitational* (gravitational potential and mass), *electrical* (potential difference and quantity of electricity or charge), *surface increase* (surface tension and area) and, most important for our purposes, *volume expansion* or *PV* work (pressure and volume).

Some peculiarities of the work are that it appears only at the boundary and thus may be thought of as flowing into or out of the system. Work may be generated only through a change in the state of the system, and it is an algebraic quantity that may be positive or negative. The convention chosen is that if the

system does work on the surroundings, the work is a positive quantity; conversely, if the surroundings does work on the system, the work is a negative quantity.

As alluded to above, PV work is given by

$$\delta W = PdV \quad (7)$$

Under the condition that the system is kept under a constant, external pressure, P_{ext} , the pressure may be brought out from under the integral

$$\delta W = P_{\text{ext}} \int dV \quad (8)$$

with the integrated expression being

$$W = P_{\text{ext}}(V_f - V_i) \quad (9)$$

where V_f and V_i are the final and initial volumes of the system. Therefore, it becomes clear that if the system expands ($V_f > V_i$) against a constant external pressure, the system does work on the surroundings, and $W > 0$. In dealing with PV work, a distinction concerning the nature of the boundary is made. The boundary can be rigid and not allow PV work or it may be movable, thereby permitting changes in the volume of the system. If there is no heat flow into or out of the system, $\delta q = 0$, the change in the internal energy may be calculated from the work,

$$\int dE = -W \quad (10)$$

or

$$\Delta E = -P_{\text{ext}}(V_f - V_i) \quad (11)$$

which indicates that, with the expansion of a system against a constant pressure and without heat flow, there is a decrease in the internal energy of the system. This convention agrees with intuition: work is done by the system at the expense of its internal energy.

HEAT—The other quantity appearing in the first law is *heat*. It shares many properties with work. Specifically, it also appears at the boundary and only with a change in the state of the system. By definition, heat flow into the system, which is taken to be a positive quantity, results in an increase in the internal energy of the system. For a system where there is no work done, $\delta W = 0$, the change in internal energy is given by

$$\int dE = \int \delta q \quad (12)$$

$$\Delta E = q \quad (13)$$

Thus, q often is referred to as a transfer of thermal energy.

Boundaries are classified as either *diathermal*, thereby allowing free exchange of heat, or conversely, *adiabatic*, where no heat flow is allowed. As an example, consider the change in internal energy for a system where only PV work is possible. The first law is written

$$\int dE = \int \delta q - \int PdV \quad (14)$$

where $\int PdV$ has been substituted for the work term.

Further, stipulating that the boundary be rigid, or $dV = 0$, the change in the internal energy is equal to the heat flow, or equivalently

$$\Delta E = q_v \quad (15)$$

where the subscript v has been added to the heat term to reflect the constraint of constant volume. Thus, adding a quantity of heat to the system increases the internal energy.

Q—What is the change in the internal energy and heat for a system that does 1.0 kcal of work on the surroundings? Assume a closed system where there is no exchange of matter across the boundary.

A—The adiabatic boundary prevents the transfer of heat, $q = 0$ and $\Delta E = 0 - 1.0$ or $\Delta E = -1.0$ kcal.

Q—What is the work done by a system expanding from 2 L to 8 L against a constant external pressure of 2 atm?

$$\begin{aligned} \text{A—} \quad W &= \int PdV = P_{\text{ext}} \int dV = P_{\text{ext}}(V_f - V_i) \\ &= 2 \text{ atm} (8 \text{ L} - 2 \text{ L}) = 12 \text{ L atm.} \end{aligned}$$

It is desirable to quantify the change in thermal energy for the purposes of determining the heat flow associated with chemical reactions or physical changes. The pressure would be constant, as determined by the atmosphere. As reactions frequently are carried out under such conditions, it is expedient to define another state function, the *enthalpy*, H .

$$H \equiv E + PV \quad (16)$$

The definition is given in terms of state properties of the system; however, to determine the change in the system, the differential is taken, resulting in

$$dH = dE + d(PV) \quad (17)$$

Expansion of the latter term yields

$$dH = dE + PdV + VdP \quad (18)$$

but because the pressure is constant, $VdP = 0$, the change in enthalpy becomes

$$dH = dE + PdV \quad (19)$$

However, from the first law the change in energy of systems restricted to PV work is

$$dE = \delta q_p - PdV \quad (20)$$

where δq_p is the heat absorbed under constant pressure. Combining the latter two equations reveals

$$dH = \delta q_p \quad (21)$$

or with integration

$$\Delta H = q_p \quad (22)$$

which implies that the enthalpy is no more than the heat absorbed by the system under the condition of constant pressure.

The heat capacity at constant pressure may be defined as

$$C_p \equiv (dq_p / dT) \quad (23)$$

or assuming the heat capacity is constant over the range of ΔT , given as

$$\int dq_p = \int C_p dT \quad (24)$$

and the integrated form

$$q_p = C_p \Delta T = \Delta H \quad (25)$$

Thus, knowing the heat capacity, the change in enthalpy of the system with a change in temperature may be determined.

In an analogous fashion, the heat capacity at constant volume may be defined as

$$C_v \equiv dq_v / dT \quad (26)$$

with the integrated form being

$$q_v = C_v \Delta T \quad (27)$$

assuming the heat capacity is constant over the range of ΔT .

Before proceeding to the application of these concepts to specific problems, the two types of processes involved with a change in the state of the system require elaboration. A process is *reversible* if, and only if, the difference between the driving and opposing force is infinitesimal; a process is *irreversible* if the forces are not infinitesimally different. The important point is that if a system is displaced from equilibrium by some vanishing small force, equilibrium may be restored by application

of an equal force in the opposite direction. That is, the restoring force applied is equal in magnitude but is in the reverse direction of the force originally applied. The process then is said to be reversible. Any process that occurs in a different manner is irreversible, and the original state of the system may not be restored without some change in the surroundings. Because vanishing small forces cannot be achieved experimentally, all real processes are irreversible, although a reversible process can be approximated closely. Nevertheless, reversibility is an important concept to establish the maximum value of work associated with a process. If a specific transformation occurs that produces work, carrying out the process reversibly yields the maximum work that can be obtained. The actual process is carried out irreversibly and thereby must yield less work.

Q—Suppose a system containing 1 mole of an ideal gas at 300 K undergoes a reversible, isothermal compression from 4 to 2 liters. What are the values of ΔE , q , W , ΔH and the final pressure if, from the kinetic molecular theory, the internal energy of an ideal gas is known to be a function solely of the temperature?

A—Although $W = \int PdV$ is correct, the expression no longer may be integrated directly, because the pressure is not constant. However, the equation of state for an ideal gas provides the pressure in terms of the volume, that is, $P = nRT/V$; thus, upon substitution

$$\int \delta W = \int nRTdV/V$$

and with integration

$$W = nRT \ln (V_f/V_i)$$

$$= (1 \text{ mol})(0.082 \text{ L atm/mol K})(300 \text{ K}) \ln (2/4) = -17.1 \text{ L atm}$$

For all isothermal processes involving ideal gases, $\Delta E = 0$. Using this fact, the heat flow out of the system is

$$\Delta E = 0 = q - W$$

or

$$q = W = -17.1 \text{ L atm}$$

The change in enthalpy is found from the relationship

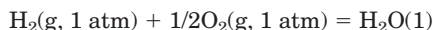
$$\Delta H = \Delta E + \Delta(PV)$$

and, noting that $\Delta(PV) = P_fV_f - P_iV_i = nRT - nRT = 0$, and from above, $\Delta E = 0$, thus, $\Delta H = 0$.

For the final part of the problem, calculation of the final pressure may be accomplished, again, by using the equation of state

$$P_f = nRT/V_f = (1 \text{ mol})(0.082 \text{ L atm/mol K})(300 \text{ K})/(2) \\ = 12.3 \text{ atm}$$

HEAT OF REACTION—The subject that deals with heat effects, associated with chemical reactions and certain physical processes, is known as *thermochemistry*. The heat of reaction is an important measure in chemical reactions. Consider the following reaction, which represents both a mass and energy balanced equation,



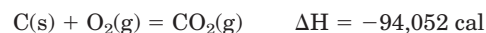
$$\Delta H_{298} = -68,300 \text{ cal}$$

where ΔH_{298} is defined as the heat of reaction. This quantity specifies the change in enthalpy of the above reaction as written at the specified temperature of 298 K. The implication is that when a mole of H_2 combines with 1/2 mol of O_2 at 298 K, 68,300 cal of heat are released in this exothermic reaction, and the enthalpy of the system is reduced by the same value. Alternatively, a reaction is endothermic if heat is absorbed by the system from the surroundings, which would result in an increase in enthalpy of the system. Because this is a balanced energy equation, the reverse reaction of the breakdown of liquid

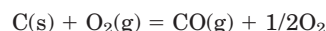
water into the respective components of H_2 and O_2 is an endothermic process requiring 68,300 cal of heat. The choice of the temperature of 298 K is arbitrary, although by convention it is taken as typical room temperature of 25°. It is important to note that, in general, the enthalpy change would be different if the reaction were carried out at another temperature.

The thermodynamic concept of the enthalpy change of the system may be extended to include coupled reactions. Because the enthalpy is a state function, only the difference between the initial and final state is important for determining the change in enthalpy for the entire process. This is simply Hess's law, which states that the enthalpy change of a reaction is the same, whether it occurs in one or several steps. Therefore, energy equations may be manipulated algebraically just like the corresponding mass balanced equations.

Consider the following two equations:



From these equations, the enthalpy change associated with the reaction



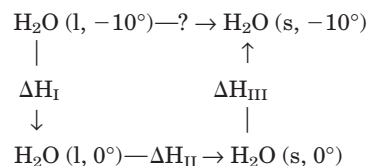
may be calculated as

$$\Delta H = (-94,052) - (-67,636 \text{ cal}) = -26,416 \text{ cal}$$

HEAT OF REACTION AS A FUNCTION OF TEMPERATURE—At this point the information presented is useful if every reaction was carried out under standard conditions of 298 K and 1 atm pressure; however, most reactions are not. Nevertheless, the heat of the reaction may be calculated at any other temperature by the temperature dependence of the heat capacity.

Consider the problem of the heat of reaction for the freezing of water at -10° (263 K). One should note that a simple phase change may be treated in a manner analogous to a chemical reaction for the purposes of calculating the enthalpy change.

The approach to the problem is to calculate the enthalpy change associated with the temperature change of the reactant (water) from 0° to -10° and the product (ice) from -10° to 0° , and then use the value for the enthalpy change at the melting point. The following schematic illustrates the approach and also provides insight into the concept of path independence of state functions.



Specifically, the molar enthalpy for the reaction at -10° is given by

$$\Delta H(-10^\circ) = \Delta H_{\text{I}} + \Delta H_{\text{II}} + \Delta H_{\text{III}} \quad (28)$$

with

$$\Delta H_{\text{I}} = \int c_p(\text{H}_2\text{O}, \text{l})dT = c_p(\text{l})\Delta T$$

$$\Delta H_{\text{I}} = (8.7)(263 - 273) = \underline{-87 \text{ cal/mol}}$$

$$\Delta H_{\text{II}} = \underline{-1436 \text{ cal/mol}} \text{ (found in tables)}$$

$$\Delta H_{\text{III}} = \int c_p(\text{H}_2\text{O}, \text{s})dT = c_p(\text{s})\Delta T$$

$$\Delta H_{\text{III}} = (18)(273 - 263) = \underline{180 \text{ cal/mol}}$$

Thus, the enthalpy for the conversion of water to ice at -10° is given by the sum or

$$\Delta H(-10^\circ) = -87 - 1436 + 180 = \underline{-1343 \text{ cal/mol}}$$

As shown, the direction around the circle dictates the limits of integration (final less initial), thus care must be taken not to confuse them. The heat capacity was assumed to be independent of temperature, which in this case is reasonable for such a small temperature change (10°). However, in general, heat capacity is a function of temperature, which will not present severe difficulties if the functional relationship is known. In this case the temperature dependence may be substituted into the equation and then integrated with the appropriate limits. For example, the molar heat capacity of oxygen over the range of 300 to 1500 K has been determined experimentally and is approximated closely by

$$c_p = 6.0954 + 3.2533 \times 10^{-3}T - 10.171 \times 10^{-7}T^2 \quad (29)$$

The change in the molar heat of formation for oxygen from temperature T_1 to T_2 would then be given by

$$\int dH = \int (6.0954 + 3.2533 \times 10^{-3}T - 10.171 \times 10^{-7}T^2) dT \quad (30)$$

$$\Delta H = [6.0954(T_2 - T_1) + (1/2)(3.2533 \times 10^{-3})[T_2^2 - T_1^2] + (1/3)(-10.171 \times 10^{-7})(T_2^3 - T_1^3)] \quad (31)$$

HEAT OF SOLUTION—When a compound is dissolved in a solvent, the resulting enthalpy change of the system is referred to as the *heat of solution*. The heat evolved or absorbed reflects the energy required to disrupt the cohesive forces of the solid and the energy generated from interaction of the solute molecules with solvent molecules. There are two ways of expressing the enthalpy change per mol of material dissolved: the integral and differential heats of solution. The two conventions arise from the dependency of the heat of solution on the amount of solvent used to dissolve the solute. Thus, the *integral heat of solution* describes the enthalpy change when 1 mol of solute is dissolved to yield a specified concentration, perhaps a 1 molar solution, whereas the *differential heat of solution* provides a value of the enthalpy change when the amount of solute dissolved is negligible.

One way of understanding the difference between these representations, and also a way of remembering, is as follows:

- The integral heat of solution gives the enthalpy change for a discrete or integral change in concentration of the solution.
- The differential heat of solution provides the enthalpy change for an infinitesimally or differential (dC) change in concentration.

Q—If the differential heat of solution of two polymorphic forms of a drug were measured in water at standard temperature and pressure (STP) and form A had a larger heat than form B, which is more stable at STP?

A—More energy is required to dissolve form A; therefore, it must be the more stable polymorph at STP.

ENTROPY AND THE SECOND LAW

Although the first law provides the framework for calculating the change in energy associated with chemical reactions or physical changes in state, there is insufficient information to allow prediction of the likelihood of whether the change will occur. Consider a system composed of two parts that are at different temperatures, T_1 and T_2 , separated by an impermeable, adiabatic partition. When the partition is removed, heat will flow from the part at a higher temperature to the part at a lower temperature. According to the first law, the energy of the whole system, the sum of parts one and two, has not changed.

Intuitively it is known that the above change will occur regardless of the fact that the first law does not provide a method of predicting the occurrence. Such changes are described as *spontaneous*, for the obvious reason that they occur without additional stimulation. It should be noted that this spontaneous change involved an increase in the disorder or, if

you will, the randomness of the system. Thus, the system initially was separated into two parts at different temperatures, but after thermal contact, a uniform temperature was reached. The *entropy*, S , is the function that provides a quantitative description of the randomness or disorder of the system and is fundamental for predicting the spontaneity of chemical reactions and physical changes. The entropy is a state function that depends only on the initial and final state of the system.

The definition of the entropy change is given by the seemingly surprising form

$$dS = \delta q_{rev}/T \quad (32)$$

where the subscript *rev* denotes that the heat flow occurs in a reversible manner. By carrying out the integration, the change in entropy for a reversible, isothermal change from state 1 to state 2 is given by

$$\Delta S = \int \delta q_{rev}/T = q_{rev}/T \quad (33)$$

With the introduction of entropy, the second law may be stated as follows:

For any spontaneous process in an isolated system, there is an increase in the value of entropy. Alternatively, the first and second laws may be combined with the classic thermodynamic statement, “the energy of the universe is constant; the entropy is increasing.”

CARNOT CYCLE—Before giving specific examples for the calculation of the entropy, it is instructive to provide the background leading to the above definition. The concepts of heat and work have been developed already and thus can be used to show the origin of the entropy function. By permitting the flow of heat into a system, work may be done by the system. The hypothetical instrument that is capable of converting heat to work is referred to as a *heat engine* (Fig 15-1). The second law dictates that not all of the heat may be converted into work, even if all changes occur in a reversible manner. In fact, the maximum work, W_{max} , that may be obtained is specified by the heat flow into the system and the temperature difference over which the heat engine is operating; that is,

$$W_{max} = q_1(T_1 - T_2)/T_1 \quad (34)$$

where $T_1 > T_2$.

In 1824 Carnot established this equation, which perhaps may be understood best by introducing the *Carnot cycle*. Consider the system, as shown in Fig 15-1, containing an ideal

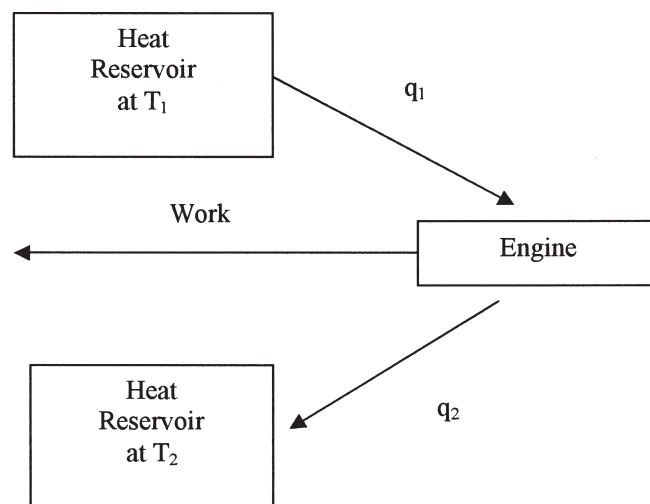


Figure 15-1. A schematic of one possible heat engine.

gas that can perform PV work due to its connections with two heat reservoirs. A *heat reservoir* is a system that has a constant temperature throughout, and the temperature is not affected by the transfer of heat into or out of the reservoir. Take a starting point at some pressure and volume and let the system undergo a cyclic, reversible change involving three other points on a pressure and volume diagram as shown in Figure 15-2. The first step is an isothermal expansion; the second is an adiabatic expansion; the third is an isothermal compression; and finally, the fourth is an adiabatic compression. The proof of Equation 32 is provided by calculating the heat and work for each step and noting that not all of the heat energy may be converted into work.

As this is a cyclic change, the total energy change in one complete cycle, ΔE_{tot} , is zero. The change in energy, along with the heat and work with each step, may be determined using the first law. For the first step of the cycle, consisting of an isothermal expansion, the energy is given as

$$\Delta E_1 = q_1 - W_1 \quad (35)$$

Since the energy change for an isothermal process involving an ideal gas is zero, the heat is equal to the work:

$$q_1 = W_1 = \int PdV \quad (36)$$

Substituting for the pressure, to allow integration between the limits of the initial and final volumes, V_1 and V_2 :

$$q_1 = \int nRTdV/V \quad (37)$$

$$q_1 = nRT \ln (V_2/V_1) \quad (36)$$

The second step is an adiabatic expansion. The heat flow is zero, and thus,

$$\Delta E_2 = -W_2 \quad (38)$$

The energy change may be obtained from the definition of the heat capacity at constant volume, or mathematically,

$$\int C_v dT = -W_2 \quad (39)$$

which is equal to the following, for an ideal gas.

$$C_v(T_2 - T_1) = -W_2 \quad (40)$$

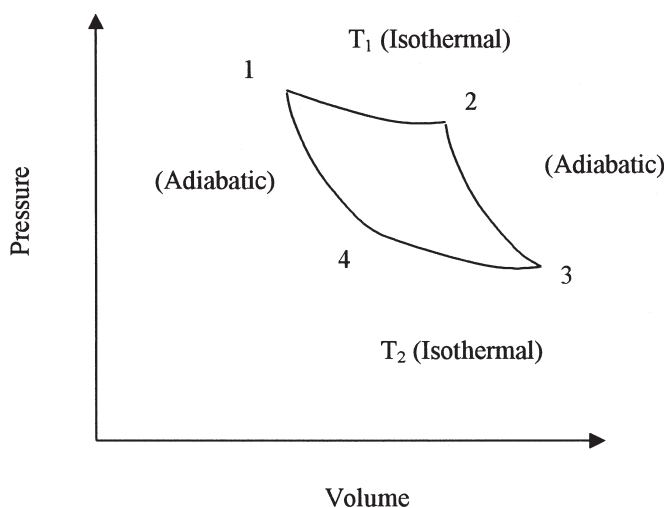


Figure 15-2. The four steps involved in the completion of the Carnot cycle, indicating the relationship among the pressure, volume, and temperature.

The third step is an isothermal compression, which is essentially the reverse of the first step, with the volume limits altered accordingly:

$$q_3 = nRT \ln (V_4/V_3) \quad (41)$$

The final step is an adiabatic compression, which is dealt with in a manner similar to the second step by noting that $q_4 = 0$, yielding

$$C_v(T_1 - T_2) = -W_4 \quad (42)$$

Summing up the results for each individual step to obtain the total work for the cycle, W_{tot} ,

$$W_{\text{tot}} = nRT \ln (V_2/V_1) - C_v(T_2 - T_1) + nRT \ln (V_4/V_3) - C_v(T_1 - T_2) \quad (43)$$

which may be simplified by canceling terms, yielding

$$W_{\text{tot}} = nRT_1 \ln (V_2/V_1) + nRT_2 \ln (V_4/V_3) \quad (44)$$

The heat flow into the system occurs with the first step, thus

$$q_1 = nRT_1 \ln (V_2/V_1) \quad (45)$$

The efficiency of an engine, ε , is given by the amount of work extracted divided by the heat flow into the system,

$$\varepsilon = W_{\text{tot}}/q_1 \quad (46)$$

where $W_{\text{tot}} = W_{\text{max}}$ for reversible changes or, equivalently,

$$\varepsilon = [nRT_1 \ln (V_2/V_1) + nRT_2 \ln (V_4/V_3)]/[nRT_1 \ln (V_2/V_1)] \quad (47)$$

Although it is not obvious, it may be shown that $V_3/V_4 = V_2/V_1$, thus

$$\varepsilon = (T_1 - T_2)/T_1 \quad (48)$$

The efficiency is proportional to the difference in temperature between the heat reservoirs. In addition, there is no work done unless there is a difference in temperature between the heat reservoirs. Finally, 100% efficiency is obtained only when $T_2 \rightarrow 0$, which as will be noted from the third law, is impossible.

The above analysis elucidates the connection between entropy and heat flow. The second law provides the quantitative limit on the amount of work that can be done with a cyclic operation performed in a reversible manner. This result has the powerful implication that it is impossible to construct a perpetual motion machine. The latter is a hypothetical device that once set into motion continues to interconvert work and heat without exhausting its finite source of energy. No machine is 100% efficient; therefore, there can be no perpetual motion machine!

ENTROPY CHANGES FOR REVERSIBLE AND IRREVERSIBLE PROCESSES—Given the above background, the entropy changes associated with several reversible processes may be determined. Consider the melting of ice at the melting point under 1 atm pressure where $\Delta H_{\text{fus}} = 1436$ cal/mol (heat of fusion). The entropy of fusion, ΔS_{fus} , is given by

$$\int dS = \int \delta q/T \quad (49)$$

which may be determined from $\int \delta q = q_p = \Delta H$, because the pressure is constant and the temperature is the melting point. Thus,

$$\int dS = \Delta H_{\text{fus}}/T_m \quad (50)$$

$$\Delta S = 1436/273 = 5.275 \text{ cal/mol K}$$

The positive change in entropy also confirms intuition concerning such an event; liquids are in a state of more disorder than solids, thus the entropy also is greater in the liquid state.

Q—What is the entropy change for a reversible, adiabatic expansion of an ideal gas?

A—As $q = 0$ for an adiabatic change, $\Delta S = 0$.

Q—What is the relationship for the entropy change of a reversible expansion of an ideal gas, given that the pressure (isobaric condition) and heat capacity are constant over the range of T_1 to T_2 ?

A—Although $\Delta S = \int \delta q/T$, T is no longer a constant and, therefore, may not be taken outside the integral. However, note that

$$\Delta S = \int dH/T = \int (C_p/T)dT \quad (51)$$

$$\Delta S = C_p \ln (T_2/T_1) \quad (52)$$

The discussion, thus far, has been limited to reversible processes that are strictly impossible to achieve in the laboratory (even though such a process may be approximated very closely). The question is, what is the entropy change for an irreversible change? Here the entropy change is given by

$$dS > \delta q_{\text{irr}}/T \quad (53)$$

Thus, all real processes may be written as

$$dS \geq \delta q/T \quad (54)$$

This concept may be extended to determine the condition of spontaneity. Consider a system that is transformed irreversibly from state 1 to state 2 and then reversibly from state 2 back to state 1. The overall change is given by

$$\int_{\text{State 1}}^{\text{State 2}} \delta q_{\text{irr}}/T + \int_{\text{State 2}}^{\text{State 1}} \delta q_{\text{rev}}/T < 0 \quad (55)$$

$$\int_{\text{State 1}}^{\text{State 2}} \delta q_{\text{irr}}/T + \int_{\text{State 1}}^{\text{State 2}} dS < 0 \quad (56)$$

which may be rearranged, with changing the limits of integration, to yield

$$\int_{\text{State 1}}^{\text{State 2}} \delta q_{\text{irr}}/T < \int_{\text{State 1}}^{\text{State 2}} dS \quad (57)$$

or, equivalently, for infinitesimal changes

$$\delta q_{\text{irr}}/T < dS \quad (58)$$

This is known as the *Clausius inequality*. For isolated systems, where boundaries do not permit the passage of energy or matter, $\delta q_{\text{irr}} = 0$, the result is given

$$dS > 0 \quad (59)$$

That is, for every spontaneous change in an isolated system there is an increase in the entropy.

The second law may be generalized in another way. The total entropy for any process is given by the sum of the entropy of the system and the surroundings; that is,

$$dS_{\text{tot}} = dS_{\text{sys}} + dS_{\text{surr}} \quad (60)$$

For reversible processes, the entropy change in a system is the negative of the entropy change produced in the surroundings. The total entropy, therefore, is zero. For irreversible processes the total entropy, system plus surroundings, increases. The mathematical statement of this relationship is

$$\sum \Delta S_{\text{tot}} = 0 \text{ reversible process} \quad (61)$$

$$\sum \Delta S_{\text{tot}} > 0 \text{ irreversible process} \quad (62)$$

THE THIRD LAW

The third law of thermodynamics simply defines the zero point of the entropy scale. The entropy of a pure, perfectly crystalline substance is zero at absolute zero. Intuitively, at the lowest possible temperature a system that has perfect three-dimensional order should have no entropy. The defining of a zero for the

entropy is unlike the other state functions introduced previously. Thus, the value of the entropy, S , of a system in any state, in principle, may be calculated.

What would be the entropy of a crystalline solid at 150 K, S_{150} ? This may be calculated as

$$\Delta S = S_{150} - S_0 \quad (63)$$

$$\Delta S = \int (c_p/T)dT - 0 \quad (64)$$

or

$$\Delta S = S_{150} \quad (65)$$

If the heat capacity over the range of 0 to 150 K is known, the value of the entropy may be calculated.

Free Energy

The concept of *free energy* is probably the most useful aspect of thermodynamics. The criteria for determining the spontaneity of a chemical reaction or phase change were presented above; however, it involved carrying out the change in an isolated system. One can imagine how inconvenient and often impossible it would be to apply such a constraint to the laboratory setting. For this sake, additional state functions have been defined to allow prediction of the spontaneity of a change in state. The rationale for the development of other functions was to allow maximum flexibility in their application. The two functions introduced are *Helmholtz free energy*, A , and *Gibbs free energy*, G . The functions for predicting spontaneity are

1. Isolated system: $dS > 0$
2. Isothermal and isochoric system: $dA < 0$
3. Isothermal and isobaric: $dG < 0$
4. Constant volume and entropy: $dE < 0$

Helmholtz free energy is defined as

$$A \equiv E - TS \quad (66)$$

Helmholtz free energy is the energy available to do pressure-volume work for reversible isothermal processes; a decrease in the Helmholtz free energy is equal to the capacity of the system to do work. An alternative view is that, for systems at constant volume and temperature, a change in state is spontaneous if, and only if, there is a decrease in the Helmholtz free energy. Thus, with the introduction of ΔA , the spontaneity of changes occurring at constant volume and temperature may be predicted.

As most reactions carried out in the laboratory are under conditions of constant pressure and temperature, Gibbs free energy is the most useful function and is defined as

$$G \equiv E + PV - TS \quad (67)$$

which can be converted into a more usable form by an analogous method used with the Helmholtz function. Taking the differential and applying the constraints of constant pressure and temperature yields

$$dG = dE + PdV - TdS \quad (68)$$

but $dE = \delta q - \delta W = TdS - \delta W$; thus, upon substitution,

$$-dG = \delta W - PdV \quad (69)$$

A decrease in Gibbs free energy is equal to the non-PV work done by the system or, equivalently,

$$dG = -\delta W_{(\text{non-PV})} \quad (70)$$

which also provides the conditions of a spontaneous change under the constraints of constant temperature and pressure. A direct application of the relationship between Gibbs free energy and non-PV work is used in potentiometry.

These relationships for predicting spontaneity often are expressed in a differential form, which presents the state functions in a concise manner as well as facilitating their use to specific problems. The four differential equations are

$$dE = TdS - PdV \quad (71)$$

$$dH = TdS + VdP \quad (72)$$

$$dA = -SdT - PdV \quad (73)$$

$$dG = -SdT + VdP \quad (74)$$

These expressions represent the four fundamental equations of thermodynamics, which in reality are four ways of looking at one fundamental equation describing the conditions of spontaneity.

Q—One mol of liquid water is vaporized reversibly at 100° and 1 atm pressure. The molar heat of vaporization is 9.725 kcal/mol; what are q_p , ΔH , ΔE , ΔA , ΔG , and ΔS ?

A—The value of q_p actually is given in the question, as the heat required to vaporize 1 mol of liquid is the definition of the molar heat of vaporization; thus, $q_p = 9.725$ kcal. Recognizing that the pressure is constant, $\Delta H = q_p = 9.725$ kcal. To calculate ΔE , the work first must be determined. The work is given by

$$W = \int PdV = P\Delta V \quad (75)$$

$$W = P(V_g - V_l) \quad (76)$$

However, the volume of the gas, V_g , is much larger than the volume of the liquid, V_l , which implies that the work is given by

$$W \approx PV_g \quad (77)$$

Assuming the gas is ideal, the work is

$$W = nRT = (1 \text{ mol})(1.987 \text{ cal/mol K})(373 \text{ K}) = 741 \text{ cal}$$

From the above, ΔE may be calculated from

$$\Delta E = q - w = 9725 - 741 = 8984 \text{ cal}$$

The change in entropy is a straightforward calculation, once the enthalpy is known:

$$\Delta S = \Delta H/T_m = 9725/373 = 26 \text{ cal/K}$$

Helmholtz free energy is given by

$$\Delta A = \Delta E - T\Delta S = 8984 - (373)(26.0) = -741 \text{ cal}$$

which also may have been obtained by recognizing that

$$\Delta A = -W_{\text{rev}} = -741 \text{ cal}$$

Finally, the change in Gibbs free energy is determined from

$$\Delta G = \Delta E + P\Delta V - T\Delta S = 8984 + 741 - (373)(26.0) = 0 \text{ cal}$$

which, too, may have been obtained by recognizing the absence of non-PV work.

STANDARD MOLAR GIBBS FREE ENERGY—The fundamental equation for Gibbs free energy has been given as

$$dG = -SdT + VdP \quad (78)$$

Consider the change in free energy with pressure at constant temperature. One may begin by defining a standard free energy, $G^\circ(T)$, which corresponds to the free energy of the ideal gas under a pressure of 1 atm. Because the temperature is constant, the change in free energy is given by

$$\int dG = \int VdP \quad (79)$$

between the limits of 1 atm and the pressure, P . For an ideal gas, the volume is a strong function of pressure; thus, with substitution, and after integration, the result is

$$G(T, P) - G^\circ(T) = \int (nRT/P)dP \quad (80)$$

Simplifying yields

$$G = G^\circ + nRT \ln P \quad (81)$$

Dividing through by the number of mols gives

$$G/n = G^\circ/n + RT \ln P \quad (82)$$

Molar free energy, G/n , is encountered so frequently it is given a special symbol, μ , and Equation 82 is written as

$$\mu = \mu^\circ + RT \ln P \quad (83)$$

The molar free energy also is referred to as the *chemical potential*.

NONIDEALITY—Equation 83 describes the molar free energy of an ideal gas, but for real gases the molar free energy is not related directly to the pressure. Thus, a function, the *fugacity*, f , has been introduced, which provides the same functional form of equation for a real gas:

$$\mu = \mu^\circ + RT \ln f \quad (84)$$

The fugacity is related to the pressure by the following equation, which is provided without derivation:

$$\ln f = \ln P + (1/nRT) \int (V - V_{\text{id}})dP \quad (85)$$

where V_{id} represents the volume of an ideal gas. This equation may be justified by considering the assumptions of an ideal gas, which are that the molecules are point particles without volume, and no intermolecular attractive or repulsive forces. Both of these effects have a direct impact on the measured volume; thus, the fugacity may be considered as a function that corrects for inaccuracies of these assumptions. Clearly, the fugacity approaches the pressure as the real volume approaches the ideal volume.

A similar approach is applied when dealing with mixtures. Consider the molar free energy of a mixture of gases. From Raoult's law, the partial pressure, P_i , of a gas is given by

$$P_i = x_i P \quad (86)$$

where x_i is the mol fraction of the i th component and P is the total pressure. Molar free energy is

$$\mu = \mu^\circ(T) + RT(\ln P + \ln x_i) \quad (87)$$

For the purposes of evaluating mixtures, it generally is more convenient to define a new standard state, $\mu^\circ(T, P)$, which consists simply of the pure gas at 1 atm, thereby yielding

$$\mu_i = \mu_i^\circ(\text{pure})(T, P) + RT \ln x_i \quad (88)$$

In fact, this equation is applicable not only to the gas state but any ideal state of aggregation. This becomes more apparent by letting the mole fraction go to unity (that is, a pure substance), whence the logarithmic term goes to zero and the molar free energy is equal to the standard-state molar free energy.

For solutions, there is a corresponding term that describes the departure for an ideal mixture, the *activity*, a . Equation 88 is applicable only for ideal mixtures. However, with the introduction of the activity, a , the above expression may be written as follows, which is general for all mixtures:

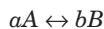
$$\mu_i = \mu_i^\circ(T, P) + RT \ln a_i \quad (89)$$

For solutions, $\mu_i^\circ(T, P)$ is the molar free energy of the liquid in the pure state.

Equilibria

Equilibrium is related intimately to spontaneity, thus the functions above used to predict the spontaneity also may be used for establishing conditions of equilibrium. In essence, if no spontaneous change is predicted, the system is at equilibrium.

Consider the following chemical reaction for an ideal gas:



For this reaction, the equilibrium constant is written as

$$K = P_B^b / P_A^a \quad (90)$$

Let the molar free energy of each component for the condition of equilibrium be defined as G_A and G_B , and as G'_A and G'_B for the nonequilibrium state. The changes in free energy at equilibrium and in a nonequilibrium state are given as

$$\Delta G = bG_B - aG_A \quad (91)$$

$$\Delta G' = bG'_B - aG'_A \quad (92)$$

The change in free energy between these two states is given by the difference between the changes in free energy; that is,

$$\Delta G' - \Delta G = b(G'_B - G_B) - a(G'_A - G_A) \quad (93)$$

The difference between G'_B and G_B may be calculated for an ideal gas with the use of the fundamental equation given above:

$$dG = VdP - SdT \quad (94)$$

which is related, under the condition of constant temperature, as

$$dG = VdP \quad (95)$$

$$\Delta G = \int VdP \quad (96)$$

$$\Delta G = \int (nRT/P)dP \quad (97)$$

After integration between limits of the two states, it yields

$$\Delta G = nRT \ln (P'_B/P_B) \quad (98)$$

Substituting into Equation 93, to find the overall change in ΔG :

$$\Delta G' - \Delta G = bRT \ln (P'_B/P_B) - aRT \ln (P'_A/P_A) \quad (99)$$

The quantity under the logarithm is given a special definition because it may be generalized to other cases not involving ideal gases; thus, the reaction quotient is defined as

$$Q = [B']^b/[A']^a \quad (100)$$

where including the equilibrium constant yields

$$\Delta G' - \Delta G = RT \ln Q - RT \ln K \quad (101)$$

Under conditions of both constant pressure and temperature, dG and $\Delta G = 0$; thus,

$$\Delta G' = RT \ln Q - RT \ln K \quad (102)$$

In a similar fashion to the standard enthalpies of formation of specific compounds, a standard Gibbs free energy, ΔG° , for the above reaction, may be defined as the free energy associated with the conversion of a mols of reactants to b mols of products when the pressure and temperature are held constant, or $\Delta G' = \Delta G^\circ$ and $\ln Q = 0$. The concentrations (or in this example, the pressures) are equal to unity, thus

$$\Delta G^\circ = -RT \ln K \quad (103)$$

This equation is of great importance because it provides the energy per mole of any chemical reaction, provided the equilibrium constant is known under standard conditions. Alternatively, the equilibrium constant may be calculated if the free energy is known. The universal applicability of thermodynamics also is displayed. Although it was derived for an ideal gas, it is equally applicable to reactions conducted in solution or even in the solid state.

TEMPERATURE DEPENDENCE OF THE EQUILIBRIUM CONSTANT—A related aspect is the question of the temperature dependence of the equilibrium constant or, from a different perspective, the temperature dependence of the change in free energy. Using the fundamental equations, it can

be shown that Gibbs free energy is related to the following state functions:

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \quad (104)$$

The determination of the temperature dependence is obtained through the Gibbs–Helmholtz equation, which is derived as follows. First, both sides are divided by the temperature:

$$\Delta G^\circ/T = \Delta H^\circ/T - \Delta S^\circ \quad (105)$$

then, the derivative with respect to temperature is taken:

$$\partial(\Delta G^\circ/T)/\partial T = \partial(\Delta H^\circ/T)/\partial T \quad (106)$$

$$\partial(\Delta G^\circ/T)/\partial T = -\Delta H^\circ/T^2 \quad (107)$$

This is referred to as the *Gibbs–Helmholtz equation*, which provides the relationship between the change in Gibbs free energy with temperature and the enthalpy change. However, the standard free-energy change also is related to the equilibrium constant, which, in general, also is temperature dependent, as

$$\partial(\Delta G^\circ/T)\partial T = -\partial(R \ln K)\partial T \quad (108)$$

Combining Equations 105 and 107 yields

$$-\partial(R \ln K)/\partial T = -\Delta H^\circ/T^2 \quad (109)$$

which may be rearranged and integrated over the limits of T_1 and T_2 , assuming that ΔH° is not a function of temperature, which is a reasonable approximation for a small temperature range:

$$\int \partial(\ln K) = -\int (\Delta H^\circ/RT^2)\partial T \quad (110)$$

Thus, the temperature dependence of the equilibrium constant is given as

$$\ln [K_2/K_1] = \Delta H^\circ/R[(1/T_1) - (1/T_2)] \quad (111)$$

where K_1 and K_2 are the equilibrium constants at temperature T_1 and T_2 , respectively. This is known as the *van't Hoff equation*; this equation is extremely important because of its wide applicability, to not only equilibrium constants of chemical reactions, but also other phenomena such as solubility, complexation, dissociation, and vapor pressure.

Q—If the equilibrium constant is 13.6 at STP, at what temperature will it be 20 if the standard enthalpy for the reaction is 8.3 kcal/mol?

A—Using the following equation:

$$(-R/\Delta H^\circ) \ln (K_2/K_1) + 1/T_1 = 1/T_2$$

$$1/T_2 = (-1.987/8300) \ln (20/13.6) + 1/298$$

$$T_2 = 3.26 \times 10^{-3} = 306 \text{ K} = \underline{33^\circ}$$

CLAPEYRON EQUATION—A special case of the van't Hoff equation is known as the *Clapeyron equation*. The interesting feature is that the free-energy change with temperature, in connection with phase changes, is approached in a different manner, with the same result. Consider a liquid in equilibrium with a vapor, such that the free energy associated with each phase, liquid and vapor, may be written from the fundamental equations as

$$dG_l = -S_l dT + V_l dP \quad (112)$$

$$dG_v = -S_v dT + V_v dP \quad (113)$$

However, because the phases are in equilibrium, $dG_l = dG_v$, or equating the above relationships,

$$-S_l dT + V_l dP = -S_v dT + V_v dP \quad (114)$$

these can be rearranged to yield

$$(S_v - S_l)dT = (V_v - V_l)dP \quad (115)$$

or with separation of the differential with the incremental changes to give

$$dP/dT = (S_v - S_l)/(V_v - V_l) \quad (116)$$

or equivalent

$$dP/dT = \Delta S/\Delta V \quad (117)$$

The interesting aspect of this equation is that the derivative, dP/dT , is related to the discontinuous changes that occur with a phase change. Although this relation was derived for a liquid–vapor equilibrium, it is general and may be applied to any phase change.

The relationship may be manipulated further by recalling that $\Delta S = \Delta H/T$, where T is the temperature at the point of equilibrium. Thus, by substitution, the following is obtained:

$$dP/dT = \Delta H/T\Delta V \quad (118)$$

An approximation may be made, as before, by noting that $\Delta V = V_v - V_l \approx V_v$, which for 1 mol of an ideal gas is $V_v = RT/P$ and affords

$$dP/dT = P\Delta H/RT^2 \quad (119)$$

The expression may be rearranged, thereby providing a means for measuring the change in enthalpy and entropy:

$$dP/P = (\Delta H/RT^2)dT \quad (120)$$

This is known as the *Clausius–Clapeyron equation*. Assuming ΔH is constant over the small temperature range between T_2 and T_1 , this expression may be integrated yielding

$$\ln(P_2/P_1) = (\Delta H/R)[(1/T_1) - (1/T_2)] \quad (121)$$

Q—If the equilibrium constant is 1.3×10^{-2} at 25° and 1.7×10^{-1} at 150° , what are ΔH° , ΔG° , and ΔS° ?

A—The temperature of 25° is taken as the standard temperature; Gibbs free energy is given by

$$\begin{aligned} \Delta G^\circ &= -RT \ln K(298) = -(1.987)(298) \ln (1.2 \times 10^{-2}) \\ &= 2.62 \text{ kcal/mol} \end{aligned}$$

The standard enthalpy change may be calculated from the temperature dependence as follows:

$$\Delta H^\circ = R \ln [K(423)/K(298)] / [(1/T_1) - (1/T_2)] = 5.15 \text{ kcal/mol}$$

Finally, ΔS° may be calculated, knowing the free energy and enthalpy changes, from

$$\begin{aligned} \Delta S^\circ &= -(1/T)(\Delta G^\circ - \Delta H^\circ) = -(1/298)(2620 - 5150) \\ &= 8.48 \text{ eu (entropy units)} \end{aligned}$$

Q—Justify the following expression of the temperature dependence of the vapor pressure of a liquified gas:

$$\ln P = -\Delta H_{\text{vap}}/RT + C$$

where C is a constant.

A—Taking the indefinite integral of Equation 120,

$$\int dP/P = \int (\Delta H_{\text{vap}}/RT^2) dT$$

the result is obtained directly:

$$\ln P = (-\Delta H_{\text{vap}}/RT) + C$$

Solubility and Partitioning Behavior

Consider a system consisting of solid drug in equilibrium with a saturated solution. The molar free energy of the solute is the same everywhere, thereby permitting the following

$$\mu_{2,\text{solution}}(T,P,a_2) = \mu_{2,\text{solid}}(T,P) \quad (122)$$

from which

$$\mu_{2,\text{liquid}}(T,P) + RT \ln a = \mu_{2,\text{solid}}(T,P) \quad (123)$$

where $\mu_{2,\text{liquid}}$ is the chemical potential of the pure liquid solute. Solving for the activity of the pure liquid solute yields

$$\ln a_2 = [\mu_{2,\text{solid}}(T,P) - \mu_{2,\text{liquid}}(T,P)]/RT \quad (124)$$

The relationship between activity and solubility is given as follows:

$$a_2 = \gamma_2 X_2 \quad (125)$$

where γ_2 is the activity coefficient of the solute in water. From this analysis, the mole fraction solubility is seen to depend only on the chemical potential of the solute. Thus, the ideal solubility of a drug is the same in every solvent. Furthermore, the difference in solubility observed between solvents is related to the nonideality of the solute, which is quantitatively determined by the activity coefficient.

The cause of nonideality arises from intermolecular interactions, which may be favorable leading to high solubility and a low activity coefficient. Alternatively, the interactions may be unfavorable, which lead to low solubility and high activity coefficient. For this reason, the activity coefficient quantitatively describes the escaping tendency of the solute from the solution.

In an analogous manner, the activity coefficients may be related to the partition coefficient. Consider a solute distributed between two immiscible solvents, A and B. At equilibrium, the chemical potential of the solute is the same in each phase, thus

$$\mu_{2A} = \mu_{2B} \quad (126)$$

and

$$\mu_2^\circ + RT \ln a_A = \mu_2^\circ + RT \ln a_B \quad (127)$$

The standard state for the solute in both solvents A and B, μ_2° , must be the same since it is based on the pure liquid solute, the activities must be equal.

$$a_A = a_B \quad (128)$$

The partition coefficient is defined in terms of the mole fractions of infinitely dilute solutions,

$$P_{o/w} = (X_o)^\infty / (X_w)^\infty = \gamma_w / \gamma_o \quad (129)$$

where the activities of the solute in each phase have been cancelled. Noteworthy, in using this system of standard states is that for ideal oil and water solutions of solutes, the partition coefficient is unity, since the activity coefficients, γ_o and γ_w , are equal to one.

These two concepts have been combined by Yalkowski and his coworkers to predict the aqueous solubility of drugs. They proposed the following:

$$\ln(X_{2w}) = \ln(X_{2oct}) - \ln(P_{o/w}) \quad (130)$$

where X_{2w} is the mole fraction in water, $P_{o/w}$ is the octanol/water partition coefficient, and X_{2oct} is the mole fraction solubility of the solute in octanol. From above, the $P_{o/w}$ is given by the ratio of the activity coefficients, γ_w/γ_o . However, the drug in octanol generally forms an ideal solution, i.e. $\gamma_o = 1$. As such, the octanol/water partition coefficient is a measure of the activity coefficient of the drug in water. Finally, the mole fraction of the solute in an ideal solution of octanol may be determined from the heat of fusion and the melting point. The resulting predictive equation for the water solubility, which has yielded highly correlated data, is:

$$\ln X_2 = -(\Delta H_{\text{fus}}/R)(1/T - 1/T_m) - \ln P_{o/w}^x \quad (131)$$

The melting point and heat of fusion may be readily measured, and the partition coefficient can be estimated by group contribution methods. Thus, a conceptually elegant foundation has been developed into a practically useful scheme to estimate the water solubility of drugs from the thermal properties and the octanol/water partition coefficient of drugs.

PROTEIN BINDING—As a final example of equilibria, the protein-binding of drugs should be mentioned. Consider the case where a protein has a single binding site for a drug. A mass-balanced equation may be written as

$$[P] + [D] = [PD] \quad (132)$$

where $[P]$ is the concentration of unbound protein, $[D]$ is the concentration of unbound drug, and $[PD]$ is the concentration of the drug–protein complex. The equilibrium constant may be written for this reaction as

$$K_a = [PD]/[P][D] \quad (133)$$

where K_a is the association constant. Assuming ideality, the standard free energy for the above equilibrium may be immediately identified as

$$\Delta G^\circ = -RT \ln (K_a) \quad (134)$$

This concept often is taken a step farther in order to characterize the nature of the binding site of the drug. This has application in structure–activity relationships used for predicting pharmacological activity.

Suppose the association constants of two structurally related drugs, K'_a and K''_a , were determined experimentally. The standard free energy of each association is given as $\Delta G^{\circ'}$ and $\Delta G^{\circ''}$. The effect of the change in the chemical structure on the energetics of the association then can be calculated from the association constants as

$$\Delta \Delta G^\circ = \Delta G^{\circ''} - \Delta G^{\circ'} = RT \ln (K'_a/K''_a) \quad (135)$$

where $\Delta \Delta G^\circ$ is the standard free–energy change of protein–binding associated with the specific chemical modification.

Q—At room temperature and a protein concentration of 2 μM , the fraction of penicillin G and penicillin V bound was found to be 0.65 and 0.80, respectively. Calculate the change in the standard free energy of binding associated with the replacement of the benzyl group in penicillin G by the phenoxy moiety in penicillin V.

A—The fraction bound may be related to the equilibrium constant by assuming there is only one binding site on each protein molecule. The fraction of drug bound, F , is defined as

$$F = [DP]/([D] + [DP]) \quad (136)$$

and since the concentration of the drug–protein complex is given by

$$[PD] = K_a[P][D] \quad (137)$$

this may be substituted into the above equation yielding

$$F = K_a[P][D]/([D] + K_a[P][D]) \quad (138)$$

After canceling terms and solving for K_a , the desired expression is obtained:

$$K_a = F/[P](1 - F) \quad (139)$$

The association constants for each drug then are calculated:

$$K_a(G) = (0.65)/[2 \mu\text{M}](1 - 0.65) = 0.93 \mu\text{M}^{-1}$$

$$K_a(V) = (0.80)/[2 \mu\text{M}](1 - 0.80) = 2.0 \mu\text{M}^{-1}$$

From the association constants, the change in standard free energy associated with replacing the benzyl group with a phenoxy moiety is

$$\Delta \Delta G^\circ = \Delta G^\circ(V) - \Delta G^\circ(G) = RT \ln (0.93/2.0)$$

$$\Delta \Delta G^\circ = \underline{453 \text{ cal/mol K}}$$

The change in the standard free energy is negative, in agreement with the concept that binding of the phenoxy group is more favorable than the benzyl group.

BIBLIOGRAPHY

Introductory

- Alberty RA, Silbey RJ. *Physical Chemistry A Basic Theory and Methods*, 2nd ed. New York: Wiley, 1996.
- Connors KA. *Thermodynamics of Pharmaceutical Systems, An Introduction for Students of Pharmacy*. New York: Wiley Interscience, 2002.
- Levine IN. *Physical Chemistry*, 4th ed. New York: McGraw-Hill, 1995.
- Reiss H. *Methods of Thermodynamics*. New York: Dover, 1997.

Comprehensive

- Glasstone S. *Thermodynamics for Chemists*, New York: Van Nostrand, 1946.
- Lewis GN, Randall M. *Thermodynamics*. Revised by Pitzer KS, Brewer L. New York: McGraw-Hill, 1961.
- Kondepudi DK, Prigogine I. *Modern Thermodynamics: From Heat Engines to Dissipative Structures*. New York: Wiley, 1996.

Solutions and Phase Equilibria

Pardeep K Gupta, PhD



SOLUTIONS AND SOLUBILITY

A solution is a chemically and physically homogeneous mixture of two or more substances. The term *solution* generally denotes a homogeneous mixture that is liquid, even though it is possible to have homogeneous mixtures that are solid or gaseous. Thus, it is possible to have solutions of solids in liquids, liquids in liquids, gases in liquids, gases in gases, and solids in solids. The first three of these are most important in pharmacy, and ensuing discussions will be concerned primarily with them.

In pharmacy different kinds of liquid dosage forms are used and all consist of the dispersion of some substance or substances in a liquid phase. Depending on the size of the dispersed particle, they are classified as *true solutions*, *colloidal solutions*, or *disperse systems*. If sugar is dissolved in water, it is supposed that the ultimate sugar particle is of molecular dimensions and that a true solution is formed. On the other hand, if very fine sand is mixed with water, a suspension of comparatively large particles, each consisting of many molecules, is obtained. Between these two extremes lie colloidal solutions, the dispersed particles of which are larger than those of true solutions but smaller than the particles present in suspensions. In this chapter only true solutions will be discussed.

It is possible to classify broadly all solutions as one of two types. In the first type, although there may be lesser or greater interaction between the dispersed substance (the solute) and the dispersing medium (the solvent), the solution phase contains the same chemical entity as found in the solid phase; thus, upon removal of the solvent, the solute is recovered unchanged. One example would be sugar dissolved in water where, in the presence of sugar in excess of its solubility, there is an equilibrium between sugar molecules in the solid phase with sugar molecules in the solution phase. A second example would be dissolving silver chloride in water. Admittedly, the solubility of this salt in water is low, but it is finite. In this case the solvent contains silver and chloride ions and the solid phase contains the same material. The removal of the solvent yields initial solute.

In the second type the solvent contains a compound that is different from the one in the solid phase. The difference between the compound in the solid phase and solution is due generally to some chemical reaction that has occurred in the solvent. An example would be dissolving aspirin in an aqueous solvent containing some basic material capable of reacting with the acid aspirin. Now the species in solution would not only be undissociated aspirin, but aspirin also as its anion, whereas the species in the solid phase is aspirin in only its undissociated acid form. In this situation, if the solvent were removed, part of the substance obtained (the salt of aspirin) would be different from what was present initially in the solid.

Solutions of Solids in Liquids

REVERSIBLE SOLUBILITY WITHOUT CHEMICAL REACTION—From a pharmaceutical standpoint, solutions of solids in liquids, with or without accompanying chemical reaction in the solvent, are of the greatest importance, and many quantitative data on the behavior and properties of such solutions are available. This discussion will be concerned with definitions of solubility, with the rate at which substances go into solution, and with temperature and other factors that control solubility.

SOLUBILITY—When an excess of a solid is brought into contact with a liquid, molecules of the former are removed from its surface until equilibrium is established between the molecules leaving the solid and those returning to it. The resulting solution is said to be saturated at the temperature of the experiment, and the extent to which the solute dissolves is referred to as its *solubility*. The extent of solubility of different substances varies from almost imperceptible amounts to relatively large quantities, but for any given solute the solubility has a constant value at a given constant temperature.

Under certain conditions it is possible to prepare a solution containing a larger amount of solute than is necessary to form a saturated solution. This may occur when a solution is saturated at one temperature, the excess of solid solute is then removed, and the solution cooled. The solute present in solution, even though it may be less soluble at the lower temperature, does not always separate from the solution and there is produced a supersaturated solution. Such solutions, formed by sodium thiosulfate or potassium acetate, for example, may be made to deposit their excess of solute by vigorous shaking, scratching the side of the vessel in contact with the solution, or introducing into the solution a small crystal of the solute.

METHODS OF EXPRESSING SOLUBILITY—When quantitative data are available, solubilities may be expressed in many ways. For example, the solubility of sodium chloride in water at 25° may be stated as

- 1 g of sodium chloride dissolves in 2.786 mL of water. (An approximation of this method is used by the USP.)
- 35.89 g of sodium chloride dissolves in 100 mL of water.
- 100 mL of a saturated solution of sodium chloride in water contains 31.71 g of solute.
- 100 g of a saturated solution of sodium chloride in water contains 26.47 g of solute.
- 1 L of a saturated solution of sodium chloride in water contains 5.425 mols of solute. This also may be stated as a saturated solution of sodium chloride in water is 5.425 molar with respect to the solute.

In order to calculate item 3 above from items 1 or 2, it is necessary to know the density of the solution, in this case 1.198 g/mL.

To calculate item 5, the number of grams of solute in 1000 mL of solution (obtained by multiplying the data in item 3 by 10) is divided by the molecular weight of sodium chloride, namely 58.45.

Several other concentration expressions are used. Molality is the number of mols of solute in 1000 g of solvent and could be calculated from the data in item 4 by subtracting grams of solute from grams of solution to obtain grams of solvent, relating this to 1000 g of solvent and dividing by molecular weight to obtain mols.

Mol fraction is the number of mols of a component divided by total number of mols in that solution. Mol % may be obtained by multiplying mol fraction by 100. Normality refers to the number of gram equivalent weights of solute dissolved in 1000 mL of solution.

In pharmacy, use also is made of three other concentration expressions. Percent by weight (% w/w) is the number of grams of solute per 100 g of solution and is exemplified by item 4 above. Percent weight in volume (% w/v) is the number of grams of solute per 100 mL of solution and is exemplified by item 3 above. Percent by volume (% v/v) is the number of milliliters of solute in 100 mL of solution, referring to solutions of liquids in liquids. The USP indicates that the term *percent*, when unqualified, means percent weight in volume for solutions of solids in liquids and percent by volume for solutions of liquids in liquids.

In pharmacopeial texts, when it has not been possible, or in some instances not desirable, to indicate exact solubility, a descriptive term is used. Table 16-1 indicates the meaning of such terms.

RATE OF SOLUTION—It is possible to define quantitatively the rate at which a solute goes into solution. The simplest treatment is based on a model depicted in Figure 16-1. A solid particle dispersed in a solvent is surrounded by a thin layer of solvent having a finite thickness, l , in centimeters. The layer is an integral part of the solid, and thus is referred to characteristically as the *stagnant layer*. This means that, regardless of how fast the bulk solution is stirred, the stagnant layer remains a part of the surface of the solid, moving wherever the particles move. The thickness of this layer may get smaller as the stirring of the bulk solution increases, but it is important to recognize that this layer will always have a finite thickness however small it may get.

Using Fick's First Law of Diffusion, the rate of solution of the solid can be explained, in the simplest case, as the rate at which a dissolved solute particle diffuses through the stagnant layer to the bulk solution. The driving force behind the movement of the solute molecule through the stagnant layer is the difference in concentration that exists between the concentration of the solute, C_1 , in the stagnant layer at the surface of the solid and its concentration, C_2 , on the farthest side of the stagnant layer. The greater this difference in concentration ($C_1 - C_2$), the faster the rate of dissolution.

According to Fick's Law, the rate of solution also is directly proportional to surface area of the solid, A in cm^2 , exposed to solvent and inversely proportional to the length of the path through which the dissolved solute molecule must diffuse.

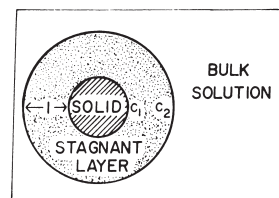


Figure 16-1. Physical model representing the dissolution process.

Mathematically, then, the rate of solution of the solid is given by

$$\text{Rate of solution} = \frac{DA}{l} (C_1 - C_2) \quad (1)$$

where D is a proportionality constant called the *diffusion coefficient* in cm^2/sec . In measuring the rate of solution experimentally, the concentration C_2 is maintained at a low value compared to C_1 and hence is considered to have a negligible effect on the rate. Furthermore, C_1 most often is the saturation solubility of the solute. Hence Equation 1 is simplified to

$$\text{Rate of solution} = \frac{DA}{l} (\text{saturation solubility}) \quad (2)$$

Equation 2 quantitatively explains many of the phenomena commonly observed that affect the rate at which materials dissolve.

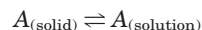
1. Small particles go into solution faster than large particles. For a given mass of solute, as the particle size becomes smaller, the surface area per unit of mass of solid increases; Equation 2 shows that as area increases, the rate must increase proportionately. Hence, if a pharmacist wishes to increase the rate of solution of a drug, its particle size should be decreased.
2. Stirring a solution increases the rate at which a solid dissolves. This is because the thickness of the stagnant layer depends on how fast the bulk solution is stirred; as stirring rate increases, the length of the diffusional path decreases. Because the rate of solution is proportional inversely to the length of the diffusional path, the faster the solution is stirred, the faster the solute will go into solution.
3. The more soluble the solute, the faster is its rate of solution. Again, Equation 2 predicts that the larger the saturation solubility, the faster the rate.
4. With a viscous liquid the rate of solution is decreased. This is because the diffusion coefficient is proportional inversely to the viscosity of the medium; the more viscous the solvent, the slower the rate of solution.

HEAT OF SOLUTION AND TEMPERATURE DEPENDENCY—Turning from the kinetic aspects of dissolution, this discussion will be concerned with the situation where there is thermodynamic equilibrium between solute in its solid phase and the solute in solution. (It is assumed that there is an amount of solid material in excess of the amount that can go into solution; hence, a solid phase is always present.) As defined earlier, the concentration of solute in solution at equilibrium is the saturation solubility of the substance.

When a solid (Solute A) dissolves in some solvent, two steps may be considered as occurring: the solid absorbs energy to become a liquid, then the liquid dissolves.



For the overall dissolution, the equilibrium existing between solute molecules in the solid and solute molecules in solution may be treated as an equilibrium. Thus, for Solute A in equilibrium with its solution,



Using the Law of Mass Action, an equilibrium constant for this system can be defined, just as any equilibrium constant may be written, as

Table 16-1. Descriptive Terms for Solubility

DESCRIPTIVE TERMS	PARTS OF SOLVENT FOR 1 PART OF SOLUTE
Very soluble	Less than 1
Freely soluble	From 1 to 10
Soluble	From 10 to 30
Sparingly soluble	From 30 to 100
Slightly soluble	From 100 to 1000
Very slightly soluble	From 1000 to 10,000
Practically insoluble, or insoluble	More than 10,000

$$K_{\text{eq}} = \frac{a_{(\text{solution})}}{a_{(\text{solid})}}$$

where a denotes the activity of the solute in each phase. Because the activity of a solid is defined as unity,

$$K_{\text{eq}} = a_{(\text{solution})}$$

Because the activity of a compound in dilute solution is approximated by its concentration, and because this concentration is the saturation solubility, K_S , the van't Hoff equation (for a more complete treatment, see Martin et al¹) may be used, which defines the relationship between an equilibrium constant (here, solubility) and absolute temperature.

$$\frac{d \log K_S}{dT} = \frac{\Delta H}{2.3RT^2} \quad (3)$$

where $d \log K_S/dT$ is the change of $\log K_S$ with a unit change of absolute temperature, T ; ΔH is a constant that, in this situation, is the heat of solution for the overall process (solid \rightleftharpoons liquid \rightleftharpoons solution); and R is the gas constant, 1.99 cal/mol/deg. Equation 3, a differential, may be solved to give

$$\log K_S = -\frac{\Delta H}{2.3RT} + J \quad (4)$$

where J is a constant. A more useful form of this equation is

$$\log \frac{K_{S,T_2}}{K_{S,T_1}} = \frac{\Delta H(T_2 - T_1)}{2.3RT_1T_2} \quad (5)$$

where K_{S,T_1} is the saturation solubility at absolute temperature T_1 , and K_{S,T_2} is the solubility at temperature T_2 . Through the use of Equation 5, if ΔH and the solubility at one temperature are known, the solubility at any other temperature can be calculated.

EFFECT OF TEMPERATURE—As is evident from Equation 4, the solubility of a solid in a liquid depends on the temperature. In the process of solution, if heat is absorbed (as evidenced by a reduction in temperature), ΔH is by convention positive and the solubility of the solute will increase with increasing temperature. Such is the case for most salts, as is shown in Figure 16-2 in which the solubility of the solute is plotted as the ordinate and the temperature as the abscissa, and the line joining the experimental points represents the solubility curve for that solute.

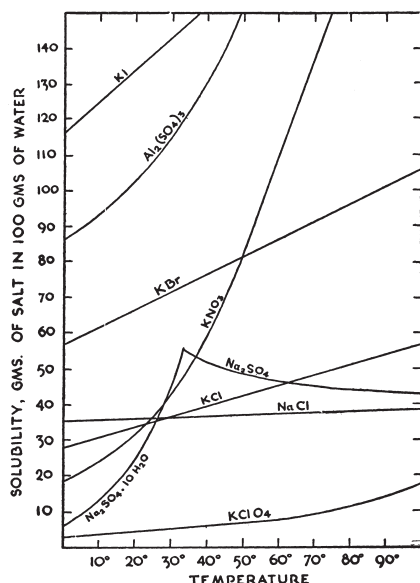


Figure 16-2. Effect of heat on solubility.

If a solute gives off heat during the process of solution (as evidenced by an increase in temperature), by convention ΔH is negative and solubility decreases with an increase in temperature. This is the case with calcium hydroxide and, at higher temperatures, with calcium sulfate. (Because of the slight solubility of these substances, their solubility curves are not included.) When heat is neither absorbed nor given off, the solubility is not affected by variation of temperature as is nearly the case with sodium chloride.

Solubility curves usually are continuous as long as the chemical composition of the solid phase in contact with the solution remains unchanged, but if there is a transition of the solid phase from one form to another, a break will be found in the curve. Such is the case with $\text{Na}_2\text{SO}_4 \cdot 10\text{H}_2\text{O}$, which dissolves with absorption of heat up to a temperature of 32.4°, at which point there is a transition of the solid phase to anhydrous sodium sulfate, Na_2SO_4 , which dissolves with evolution of heat. This change is evidenced by increased solubility of the hydrated salt up to 32.4°, but above this temperature the solubility decreases.

These temperature effects are what would be predicted from Equation 4. When the heat of solution is negative, signifying that energy is released during dissolution, the relation between $\log K_S$ and $1/T$ is typified in Figure 16-3 (Curve A), where as $1/T$ increases, $\log K_S$ increases. It can be seen that with increasing temperature (T itself actually increases proceeding left in Fig 16-3A), there is a decrease in solubility. On the other hand, when the heat of solution is positive—that is, when heat is absorbed in the solution process—the relation between $\log K_S$ and $1/T$ is typified in Figure 16-3B. Hence, as temperature increases ($1/T$ decreases), the solubility increases.

EFFECT OF SALTS—The solubility of a nonelectrolyte in water either is decreased or increased generally by the addition of an electrolyte; it is only rarely that the solubility is not altered. When the solubility of a nonelectrolyte is decreased, the effect is referred to as *salting-out*; if it is increased, it is described as *salting-in*. Inorganic electrolytes commonly decrease solubility, though there are some exceptions to the generalization.

Salting-out occurs because the ions of the added electrolyte interact with water molecules, and thus, in a sense, reduce the amount of water available for dissolution of the nonelectrolyte. (Refer to the section on *Thermodynamics of the Solution Process* for another view.) The greater the degree of hydration of the ions, the more the solubility of the nonelectrolyte is decreased. If, for example, one compares the effect of equivalent amounts of lithium chloride, sodium chloride, potassium chloride, rubidium chloride, and cesium chloride (all of which belong to the family of alkali metals and are of the same valence type),

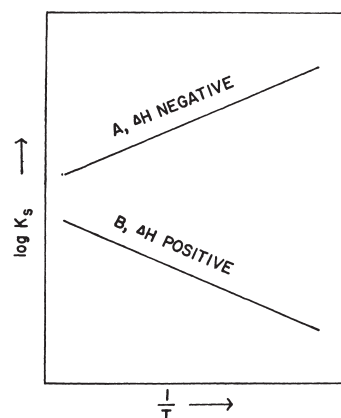


Figure 16-3. Typified relationship between the logarithm of the saturation solubility and the reciprocal of the absolute temperature.

lithium chloride decreases the solubility of a nonelectrolyte to the greatest extent and the salting-out effect decreases in the order given.

This is also the order of the degree of hydration of the cations; lithium ion—being the smallest ion and, therefore, having the greatest density of positive charge per unit of surface area (see Chapter 13 under *Electronegativity Values*)—is the most extensively hydrated of the cations, whereas cesium ion is hydrated the least. Salting-out is encountered frequently in pharmaceutical operations.

Salting-in commonly occurs when either the salts of various organic acids or organic-substituted ammonium salts are added to aqueous solutions of nonelectrolytes. In the first case, the solubilizing effect is associated with the anion; in the second, it is associated with the cation. In both cases the solubility increases as the concentration of added salt is increased. The solubility increase may be relatively great, sometimes amounting to several times the solubility of the nonelectrolyte in water.

SOLUBILITY OF SOLUTES CONTAINING TWO OR MORE SPECIES—In cases where the solute phase consists of two or more species (as in an ionizable inorganic salt), when the solute goes into solution, the solution phase often contains each of these species as discrete entities. For some such substance, AB , the following relationship for the solution process may be written.



As there is an equilibrium between the solute and saturated solution phases, the Law of Mass Action defines an equilibrium constant, K_{eq}

$$K_{\text{eq}} = \frac{\alpha_{A(\text{solution})} \cdot \alpha_{B(\text{solution})}}{\alpha_{AB(\text{solid})}} \quad (6)$$

where $\alpha_{A(\text{solution})}$, $\alpha_{B(\text{solution})}$, and $\alpha_{AB(\text{solid})}$ are the activities of A and B in solution and of AB in the solid phase. Recall from the earlier discussion that the activity of a solid is defined as unity, and that in a very dilute solution (eg, for a slightly soluble salt) concentrations may be substituted for activities. Equation 6 then becomes

$$K_{\text{eq}} = C_A C_B$$

where C_A and C_B are the concentrations of A and B in solution. In this situation K_{eq} has a special name, the *solubility product*, K_{SP} . Thus,

$$K_{\text{SP}} = C_A C_B \quad (7)$$

This equation will hold true theoretically only for slightly soluble salts.

As an example of this type of solution, consider the solubility of silver chloride,

$$K_{\text{SP}} = [\text{Ag}^+][\text{Cl}^-]$$

where the brackets [] designate molar concentrations.

At 25° the solubility product has a value of 1.56×10^{-10} , the concentration of silver and chloride ions being expressed in mol/liter. The same numerical value applies also to solutions of silver chloride containing an excess of either silver or chloride ions. If the silver-ion concentration is increased by the addition of a soluble silver salt, the chloride-ion concentration must decrease until the product of the two concentrations again is equal numerically to the solubility product. To effect the decrease in chloride-ion concentration, silver chloride is precipitated, and hence its solubility is decreased. In a similar manner, an increase in chloride-ion concentration by the addition of a soluble chloride effects a decrease in the silver-ion concentration until the numerical value of the solubility product is attained. Again, this decrease in silver-ion concentration is brought about by the precipitation of silver chloride. This phenomenon of decrease in solubility due to the presence of one of the ions in solution is known as the *common-ion effect*.

The solubility of silver chloride in a saturated aqueous solution of the salt may be calculated by assuming that the concentration of silver ion is the same as the concentration of chloride ion, both expressed in mol/liter, and that the concentration of dissolved silver chloride is numerically the same as each silver chloride molecule gives rise to one silver ion and one chloride ion, because

$$[\text{dissolved AgCl}] = [\text{Ag}^+] = [\text{Cl}^-]$$

the solubility of AgCl is equal to $\sqrt{1.56 \times 10^{-10}}$, which is 1.25×10^{-5} mol/liter. Multiplying this by the molecular weight of silver chloride (143), we obtain a solubility of approximately 1.8 mg/liter.

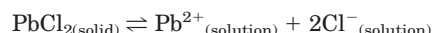
For a salt of the type PbCl_2 the solubility product expression takes the form

$$[\text{Pb}^{2+}][\text{Cl}^-]^2 = K_{\text{SP}}$$

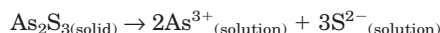
while for As_2S_3 it would be

$$[\text{As}^{3+}]^2[\text{S}^{2-}]^3 = K_{\text{SP}}$$

because from the Law of Mass Action



and



For further details of methods of using solubility-product calculations, see textbooks on qualitative or quantitative analyses or physical chemistry.

Recall that the solubility-product principle is valid for aqueous solutions of slightly soluble salts, provided that the concentration of added salt is not too great. Where the concentrations are high, deviations from the theory occur and these have been explained by assuming that in such solutions the nature of the solvent has been changed. Frequently, deviations also may occur as the result of the formation of complexes between the two salts. An example of increased solubility, by virtue of complex-ion formation, is seen in the effect of solutions of soluble iodides on mercuric iodide. According to the solubility-product principle, it might be expected that soluble iodides would decrease the solubility of mercuric iodide, but because of the formation of the more soluble complex salt K_2HgI_4 , which dissociates as



the iodide ion no longer functions as a common ion.

It is possible to formulate some general rules regarding the effect of the addition of soluble salts to slightly soluble salts where the added salt does not have an ion common to the slightly soluble salt. If the ions of the added soluble salt are not highly hydrated (see the previous section, *Effect of Salts*), the solubility product of the slightly soluble salt will increase because the ions of the added salt tend to decrease the interionic attraction between the ions of the slightly soluble salt. On the other hand, if the ions of the added soluble salt are hydrated, water molecules become less available and the interionic attraction between the ions of the slightly soluble salt increases with a resultant decrease in solubility product. Another way of considering this effect is discussed later (see *Thermodynamics of the Solution Process*).

In general, the effect of temperature is what would be expected: increasing the temperature of the solution results in an increase of the solubility product.

SOLUBILITY FOLLOWING A CHEMICAL REACTION—Thus far the discussion has been concerned with solubility that comes about because of interplay of entirely physical forces. The dissolution of some substance resulted from overcoming the physical interactions between solute molecules and solvent molecules by the energy produced when a solute molecule interacted physically with a solvent molecule. The solution process, however, can be facilitated also by a chemical re-

action. Almost always the chemical enhancement of solubility in aqueous systems is due to the formation of a salt following an acid–base reaction.

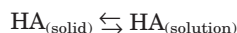
An alkaloidal base, or any other nitrogenous base of relatively high molecular weight, generally is slightly soluble in water, but if the pH of the medium is reduced by addition of acid, the solubility of the base is considerably increased as the pH continues to be reduced. The reason for this increase in solubility is that the base is converted to a salt, which is relatively soluble in water. Conversely, the solubility of a salt of an alkaloid or other nitrogenous base is reduced as pH is increased by addition of alkali.

The solubility of slightly soluble acid substances is, on the other hand, increased as the pH is increased by addition of alkali, the reason again being that a salt, relatively soluble in water, is formed. Examples of acid substances whose solubility is thus increased are aspirin, theophylline and the penicillins, cephalosporins, and barbiturates. Conversely, the solubility of salts of the same substances is decreased as the pH decreases.

Among some inorganic compounds a somewhat similar behavior is observed. Tribasic calcium phosphate, $\text{Ca}_3(\text{PO}_4)_2$, for example, is almost insoluble in water, but if an acid is added its solubility increases rapidly with a decrease in pH. This is because hydrogen ions have such a strong affinity for phosphate ions forming nonionized phosphoric acid that the calcium phosphate is dissolved in order to release phosphate ions. Or, stated in another way, the solubilization is an example of a reaction in which a strong acid (the source of the hydrogen ions) displaces a weak acid.

In all of these examples solubilization occurs as the result of an interaction of the solute with an acid or a base, and thus the species in solution is not the same as the undissolved solute. Compounds that do not react with either acids or bases are slightly, or not at all, influenced in their aqueous solubility by variations of pH. Such effects if observed are generally due to ionic *salt effects*.

It is possible to analyze quantitatively the solubility following an acid–base reaction by considering it as a two-step process. The first example is an organic acid, designated as HA , that is relatively insoluble in water. Its two-step dissolution can be represented as



followed by



The equilibrium constant for the first step is the solubility of HA ($K_S = [HA]_{(\text{solution})}$), just as was developed earlier when no chemical reaction took place, and the equilibrium constant for the second step is the dissociation constant of the acid is

$$K_a = \frac{[H^+][A^-]}{[HA]}$$

Since the total amount of compound *in solution* is the sum of nonionized and ionized forms of the acid, the total solubility may be designated as $S_{i(HA)}$, or

$$S_{i(HA)} = [HA] + [A^-] = [HA] + K_a \frac{[HA]}{[H^+]} \quad (8)$$

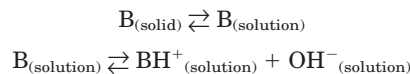
and because $K_S = [HA]$, Equation 8 becomes

$$S_{i(HA)} = K_S \left(1 + \frac{K_a}{[H^+]} \right) \quad (9)$$

Equation 9 is very useful because it equates the total solubility of an acid drug with the hydrogen-ion concentration of the solvent. If the water solubility, K_S , and the dissociation constant, K_a , are known, the total solubility of the acid can be calculated at various hydrogen-ion concentrations.

Equation 9 demonstrates quantitatively how the total solubility of the acid increases as the hydrogen-ion concentration decreases (ie, as the pH increases).

It is possible to develop an equation similar to Equation 9 for the solubility of a basic drug B , such as a relatively insoluble nitrogenous base (eg, an alkaloid), at various hydrogen-ion concentrations. The solubility of the base in water may be represented in two steps as



Again, if K_S is the solubility of the free base in water and K_b is its dissociation constant,

$$K_b = \frac{[BH^+][OH^-]}{[B]}$$

the total solubility of the base in water $S_{i(B)}$ is given by

$$S_{i(B)} = [B] + [BH^+] = [B] + \frac{K_b[B]}{[OH^-]} = K_S \left(1 + \frac{K_b}{[OH^-]} \right) \quad (10)$$

It is convenient to rewrite Equation 10 in terms of hydrogen-ion concentration by making use of the dissociation constant for water

$$K_W = [H^+][OH^-] = 1 \times 10^{-14}$$

Equation 10 then becomes

$$S_{i(B)} = K_S \left(1 + \frac{K_b}{K_W/[H^+]} \right) = K_S \left(1 + \frac{K_b[H^+]}{K_W} \right) \quad (11)$$

Equation 11 quantitatively shows how the total solubility of the base increases as the hydrogen-ion concentration of the solvent increases. If K_S and K_b are known, it is possible to calculate the total solubility of a basic drug at various hydrogen-ion concentrations using this equation.

Equations 9 and 11 have assumed that the salt formed following a chemical reaction is infinitely soluble. This, of course, is not an acceptable assumption, as suggested and demonstrated by Kramer and Flynn.² Rather, for an acidic or basic drug there should be a pH at which maximum solubility occurs where this solubility remains the sum of the solution concentrations of the free and salt forms of the drug at that pH. Using a basic drug B as the example, this would mean that a solution of B , at pH values greater than the pH of maximum solubility, would be saturated with free-base form but not with the salt form, and the use of Equation 11 would be valid for the prediction of solubility. On the other hand, at pH values less than the pH of maximum solubility, the solution would be saturated with salt form and Equation 11 is no longer really valid. Because in this situation the total solubility of the base, $S_{i(B)}$, is

$$S_{i(B)} = [B] + [BH^+]_s$$

where the subscript s designates a solution saturated with salt, the correct equation to use at pH values less than the pH maximum would be

$$S_{i(B)} = [BH^+]_s \left(1 + \frac{[OH^-]}{K_b} \right) = [BH^+]_s \left(1 + \frac{K_W}{K_b[H^+]} \right) \quad (12)$$

A relationship similar to Equation 12 likewise can be developed for an acidic drug at a pH greater than its pH of maximum solubility.

EFFECTING SOLUTION OF SOLIDS IN THE PRESCRIPTION LABORATORY—The method usually employed by the pharmacist when soluble compounds are to be dissolved in water in compounding a prescription requires the use of the mortar and pestle. The ordinary practice is to crush the substance into fragments in the mortar with the pestle and pour the solvent on it, meanwhile stirring with the pestle until solution is effected. If definite quantities are used and the whole of

the solvent is required to dissolve the given weight of the salt, only a portion of the solvent should be added first, and, when this is saturated, the solution is poured off and a fresh portion of solvent added. This operation is repeated until the solid is dissolved entirely and all the portions combined. Other methods of affecting solution are to shake the solid with the liquid in a bottle or flask or to apply heat to the substances in a suitable vessel.

Substances vary greatly in the rate at which they dissolve; some are capable of producing a saturated solution quickly, others require several hours to attain saturation.

With hygroscopic substances like pepsin, silver protein compounds, and some others, the best method of effecting solution in water is to place the substance directly upon the surface of the water and then stir vigorously with a glass rod. If the ordinary procedure, such as using a mortar and pestle, is employed with these substances, gummy lumps form that are exceedingly difficult to dissolve.

The *solubility* of chemicals and the *miscibility* of liquids are important physical factors for the pharmacist to know, as they often have a bearing on intelligently and properly filling prescriptions. For the information of the pharmacist, the USP provides tabular data indicating the degree of solubility or miscibility of many official substances.

DETERMINATION OF SOLUBILITY—For the pharmacist and pharmaceutical chemist, the question of solubility is of paramount importance. Not only is it necessary to know solubilities when preparing and dispensing medicines, but such information is also necessary to effect separation of substances in qualitative and quantitative analysis. Furthermore, the accurate determination of the solubility of a substance is one of the best methods for determining its purity.

The details of the determination of the solubility are affected markedly by the physical and chemical characteristics of the solute and solvent and also by the temperature at which the solubility is to be determined. Accordingly, it is not possible to describe a universally applicable method, but in general the following rules must be observed in solubility determinations.

1. The purity of both the dissolved substance and the solvent is essential, because impurities in either affect the solubility.
2. A constancy of temperature must be maintained accurately during the course of the determination.
3. Complete saturation must be attained.
4. Accurate analysis of the saturated solution and correct expression of the results are imperative.

Consideration should be given also to the varying rates of dissolution of different compounds and to the marked effect of the degrees of fineness of the particles on the time required for the saturation of the solution.

THE PHASE RULE AND PHASE-SOLUBILITY ANALYSIS—Phase-solubility analysis is a useful and accurate method for the determination of the purity of a substance. It involves the application of precise solubility methods to the principle that constancy of solubility, in the same manner as constancy of melting point, indicates that a material is pure or free from foreign admixture. It is important to recognize that the technique can be used to obtain the exact solubility of the pure substance without the necessity of the experimental material itself being pure.

The method is based on the thermodynamic principles of heterogeneous equilibria that are among the soundest of theoretical concepts of chemistry. Thus, it does not depend on any assumptions regarding kinetics or structure of matter, but is applicable to all species of molecules, and is sufficiently sensitive to distinguish between optical isomers. The requirements for an analysis are simple, as the equipment needed is basic to most laboratories and the quantities of substances required are small.

The standard solubility method consists of five steps:

1. Mixing, in separate systems, increasing amounts of a substance with measured amounts of a solvent.

2. Establishment of equilibrium for each system at identical constant temperature and pressure.
3. Separation of the solid phase from the solutions.
4. Determination of the concentration of the material dissolved in the various solutions.
5. Plotting the concentration of the dissolved material of interest per unit of solvent (*y*-axis, or solution concentration) against the mass of total material per unit of solvent (*x*-axis or system concentration).

The solubility method has been established on the sound theoretical principles of the Gibbs phase rule: $F = C - P + 2$, which relates *C*, the number of components; *F*, the degrees of freedom (pressure, temperature, and concentration); and *P*, the number of phases for a heterogeneous equilibrium.

Solubility analyses are carried out at constant temperature and pressure, so a pure solid in solution would show only one degree of freedom, because only one phase is present at concentrations below saturation. This is represented by section *AB* in Figure 16-4. For a pure solid in a saturated solution at equilibrium (Fig 16-4, *BC*), two phases are present, solid and solution; there is no variation in concentration, and thus, at constant temperature and pressure, no degrees of freedom.

The curve *ABC* of Figure 16-4 represents the type of solubility diagram obtained for: (1) a pure material, (2) equal amounts of two or more materials having identical solubilities, or (3) a mixture of two or more materials present in the unique ratio of their solubilities. These latter two cases are rare and often may be detected by a change in solvent system.

Line segment *BC* of Figure 16-4 indicates purity because it has no slope. If, however, this section does exhibit a slope, its numerical value indicates the fraction of impurity present. Line segment *BC*, extrapolated to the *y*-axis at *D*, is the actual solubility of the pure substance.

A representative type of solubility curve, which is obtained when a substance contains one impurity, is illustrated in Figure 16-5. Here, at *B* the solution becomes saturated with one component. From *B* to *C* there are two phases present: a solution saturated with Component I (usually the major component) containing also some Component II (usually the minor component), and a solid phase of Component I. The one degree of freedom revealed by the slope of the line segment *BC* is the concentration of Component II, which is the impurity (usually the minor component). A mixture of *d* and *l* isomers could have

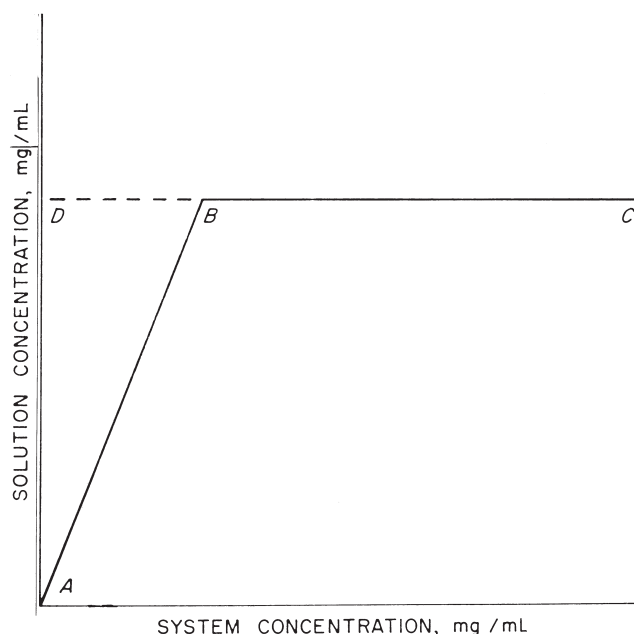


Figure 16-4. Phase-solubility diagram for a pure substance.

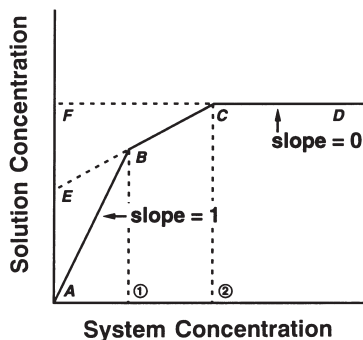


Figure 16-5. Type of solubility curve obtained when a substance contains one impurity.

such a curve, as would any simple mixtures in which the solubilities are independent of each other.

The section CD indicates that the solvent is saturated with both components of the two-component mixture. Here, three phases are present: a solution saturated with both components and the two solid phases. No variation of concentration is possible; hence, no degree of freedom is possible (indicated by the lack of slope of section CD). The distance AE on the ordinate represents the solubility of the major component, and the distance EF represents the solubility of the minor component.

The equilibration process is time consuming, requiring as long as 3 weeks in certain cases, but this is offset by the fact that all of the sample can be recovered after a determination. This adds to the general usefulness of the method, particularly in cases where the substance is expensive or difficult to obtain. A use for the method other than the determination of purity or of solubility is to obtain especially pure samples by recovering the solid residues at system concentration, corresponding to points on section BC in Figure 16-5. Thus, the method is useful not only as a quantitative analytical tool, but also for purification.

Solutions of Liquids in Liquids

BINARY SYSTEMS—The following types of liquid-pairs may be recognized as binary systems.

1. Those that are soluble completely in each other in all proportions. Examples: alcohol and water, glycerin and water, alcohol and glycerin.
2. Those that are soluble in each other in definite proportions. Examples: phenol and water, ether and water, nicotine and water.
3. Those that are imperceptibly soluble in each other in any proportion. Examples: castor oil and water, liquid petrolatum and water.

The mutual solubility of liquid pairs of Type 2 has been studied extensively and found to show interesting regularities. If a series of tubes containing varying, but known, percentages of phenol and water are heated (or cooled, if necessary) just to the point of formation of a homogeneous solution, and the temperatures at such points is noted, upon plotting the results a curve is obtained similar to that in Figure 16-6. On this graph the area inside the curve represents the region where mixtures of phenol and water will separate into two layers, while in the region outside of the curve homogeneous solutions will be obtained. The maximum temperature on this curve is called the *critical solution temperature*, that is, the temperature above which a homogeneous solution occurs regardless of the composition of the mixture. For phenol and water the critical solution temperature occurs at a composition of 34.5% phenol in water.

Temperature versus composition curves, as depicted in Figure 16-6, provide much useful information in the preparation of

homogeneous mixtures of substances showing mutual-solubility behavior. At room temperature (here assumed to be 25°), by drawing a line parallel to the abscissa at 25° , we find that we actually can prepare two sets of homogeneous solutions, one containing from 0% to about 7.5% phenol and the other containing phenol from 72% to about 95% (its limit of solubility). At compositions between 7.5% and 72% phenol at 25° two liquid layers or phases will separate. In sample tubes containing a concentration of phenol in this two-layer region at 25° one layer always will be phenol-rich and always contain 72% phenol while the other layer will be water-rich and always contain 7.5% phenol. These values are obtained by interpolation of the two points of intersection of the line drawn at 25° with the experimental curve.

As it may be deduced, at other temperatures, the composition of the two layers in the two-layer region is determined by the points of intersection of the curve with a line (called the *tie line*) drawn parallel to the abscissa at that temperature. The relative amounts of the two layers or phases, phenol-rich and water-rich in this example, will depend on the concentration of phenol added. As expected, the proportion of phenol-rich layer relative to the water-rich layer increases as the concentration of phenol added increases. For example, at 20% phenol in water at 25° , there would be more of the water-rich layer than of the phenol-rich layer, whereas at 50% phenol in water there would be more of the phenol-rich layer. The relative portion of each layer may be calculated from such tie lines at any temperature and compositions as well as the amount of phenol present in each of the two phases. To determine how these calculations are made and for further discussion of this topic the student should consult Martin et al.¹

A simple and practical advantage in the use of phase diagrams is pointed Martin et al.¹ Based on diagrams such as Figure 16-6, they point out that the most concentrated stock solution of phenol that should perhaps be used by pharmacists is one containing 76% *w/w* phenol in water (equivalent to 80% *w/v*). At room temperature this mixture is a homogeneous solution and will remain homogeneous to around 3.5° , at which temperature freezing occurs. It should be noted that Liquefied Phenol USP contains 90% *w/w* phenol and freezes at 17° . This means that if the storage area in the pharmacy falls to about 63°F , the preparation will freeze, resulting in a stock solution no longer convenient to use.

In the case of phenol and water, the mutual solubility increases with an increase in temperature and the critical solu-

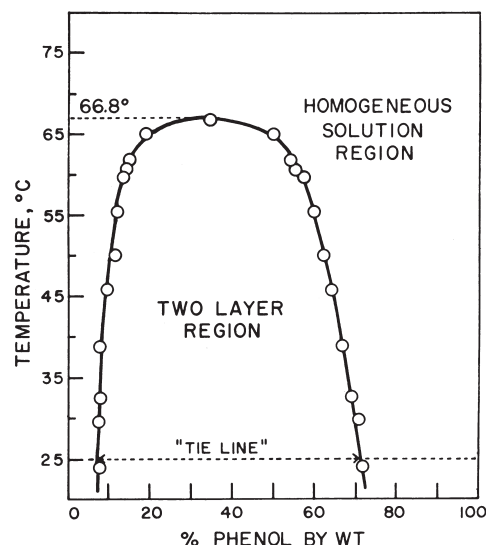


Figure 16-6. Phenol–water solubility. (From Campbell AN, Campbell AJR. *J Am Chem Soc* 1937; 59:2481.)

tion temperature occurs at a relatively high point. In a certain number of cases, however, the mutual solubility increases with decrease in temperature and the critical solution temperature occurs at a relatively low value. Most of the substances that show lower critical solution temperatures are amines as, for example, triethylamine with water.

In addition to pairs of liquids that show *either* upper or lower critical solution temperatures, there are other pairs that show *both* upper and lower critical solution temperatures and the mutual solubility curve is of the closed type. An example of this type of liquid pair is found in the case of nicotine and water (Fig 16-7). Mixtures of nicotine and water represented by points within the curve will separate into two layers, but mixtures represented by points outside of the curve are perfectly miscible with each other.

In a discussion of solutions of liquids in liquids it is evident that the distinction between the terms solute and solvent loses its significance. For example, in a solution of water and glycerin, which shall be considered to be the soluble and which the solvent? Again, when two liquids are only partially soluble in each other, the distinction between solute and solvent might be reversed easily. In such cases the term solvent usually is given to the constituent present in larger quantity.

TERNARY SYSTEMS—The addition of a third liquid to a binary liquid system to produce a ternary or three-component system can result in several possible combinations.

If the third liquid is soluble in only one of the two original liquids or if its solubility in the two original liquids is markedly different, the mutual solubility of the original pair will be decreased. An upper critical solution temperature will be elevated and a lower critical solution temperature lowered. On the other hand, the addition of a liquid having roughly the same solubility in both components of the original pair will result in an increase in their mutual solubility. An upper critical solution temperature then will be lowered and a lower critical solution temperature elevated.

An equilateral-triangle graph may be used to represent ternary systems. In this type of graph, each side of the triangle represents 0% of one of the components and the apex opposite that side represents 100% of that component. This is illustrated using a particularly common ternary system involving two solvents that are completely miscible and a third that is miscible with only one of the two. In Figure 16-8, water and alcohol are the miscible solvents and castor oil is the third solvent that is soluble in alcohol but not in water. Such diagrams could be ap-

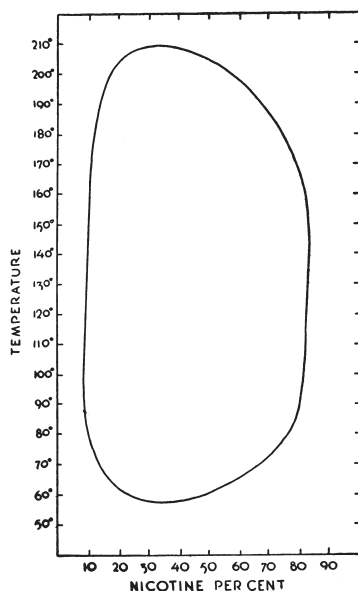


Figure 16-7. Nicotine–water solubility.

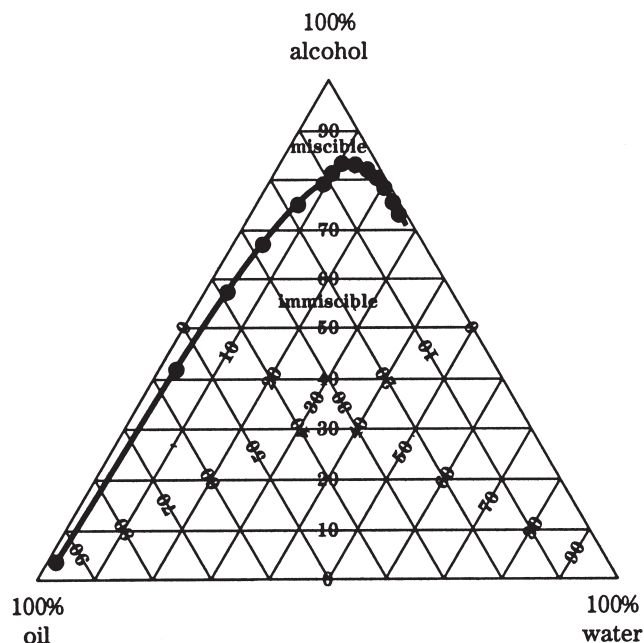


Figure 16-8. Phase diagram at constant temperature for a ternary system: two liquids completely miscible in one another with a third liquid soluble in only one of the two. (Data from Lorán MR, Guth EP. *J APhA Sci Ed* 1951; 40:465.)

plied, for example, to surfactant/oil/water systems, flavor/water/alcohol systems, drug/propellant mixture systems, drug/water/propylene glycol systems or any other such system you might think of that would fit into this category.

The data in Figure 16-8 were obtained by determining the amount of water needed to just cloud solutions of oil in alcohol at different concentrations and at room temperature. The percentage of each solvent just clouding the system was then calculated and plotted as shown in the figure. For example, a cloudy solution developed at a mixture of about 67% alcohol, 27% oil, and 6% water. Note that the percentages of the three components must always equal 100%. In the region labeled *miscible*, any combination of the three components will result in a solution. The pharmacist can pick any combination in this region for reasons of taste, safety, stability, or cost. Figure 16-8 is constructed for room temperature; any other temperature would have its own phase diagram. Including temperature as a variable would create a three-dimensional relationship with ternary diagrams such as Figure 16-8 stacked in the x - y plane as a function of temperature on the z -axis.

Other possibilities exist in ternary liquid systems—for example, those in which two components are completely miscible and the third is partially miscible with each, and that in which all combinations of two of the three components are only partially miscible.

Solutions of Gases in Liquids

Nearly all gases are more or less soluble in liquids. One has but to recall the solubility of carbon dioxide, hydrogen sulfide, or air in water as common examples.

The amount of gas dissolved in a liquid in general follows *Henry's law*, which states that the weight of gas dissolved by a given amount of a liquid at a given temperature is proportional to its pressure. Thus, if the pressure is doubled, twice as much gas will dissolve as at the initial pressure. The extent to which a gas is dissolved in a liquid, at a given temperature, may be expressed in terms of the solubility coefficient, which is the volume of gas measured under the conditions of the experiment

that is absorbed by one volume of the liquid. The degree of solubility also is expressed sometimes in terms of the *absorption coefficient*, which is the volume of gas, reduced to standard conditions, dissolved by one volume of liquid under a pressure of one atmosphere.

Although Henry's law expresses fairly accurately the solubility of slightly soluble gases, it deviates considerably in the case of very soluble gases such as hydrogen chloride and ammonia. Such deviations most frequently are due to chemical interaction of solute and solvent.

The solubility of gases in liquids decreases with a rise in temperature and, in general, also when salts are added to the solvent, the latter effect being referred to as the *salting-out* of the gas.

Solutions of gases potentially are dangerous when exposed to warm temperatures because of the liberation and expansion of the dissolved gas, which may cause the container to burst. Bottles containing such solutions (eg, strong ammonia solution) should be cooled before opening, if practical, and the stopper should be covered with a cloth before attempting its removal.

Solutions of Solids in Solids

Various mixtures of one solid in another are being considered in the pharmaceutical sciences primarily as a means to increase bioavailability. For example, melts of solid mixtures of drugs with excipients and eutectic mixtures are being investigated (see Chapter 13). It is possible to have a true solution of one solid in another to give rise to a continuum of one solid dispersed in another as depicted in Figure 16-9. Such a system is referred to as a *continuous* dispersion; it is very rarely found. To achieve this would mean that two materials would have to be of similar size, structure, and interaction energy so that they might enter and occupy a mutual crystalline structure at the molecular level. Hence, such solid solutions may occur only among racemic mixtures of chiral compounds. If it were possible to form a solid solution of a drug in a water-soluble excipient, bioavailability could increase dramatically because the drug would transfer into water as individual molecules.

There are three types of continuous dispersions depicted in Figure 16-9: [1] shows an *ideal* dispersion of constant melting point, while [2] and [3] (*nonideal* solutions) show dispersions having a maximum or minimum, respectively. Each of the latter dispersions show an upper or *liquidus line* and a lower or *solidus line* that might be viewed as representing the direction (cooling or heating) used to arrive at the temperature of melting or solidification in each mixture. The composition of the liquid and solid phases in the region between the two lines can be quantified in a way that is related to the tie-line treatment for phenol-water systems, although more complicated.

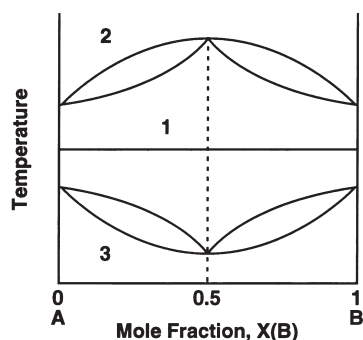


Figure 16-9. Phase diagram for a continuous solution of a Solid A in a Solid B (or of B in A: [1] is an ideal solution, [2] and [3], nonideal solutions. (Data from Duddu S. PhD thesis. University of Minnesota, 1993.)

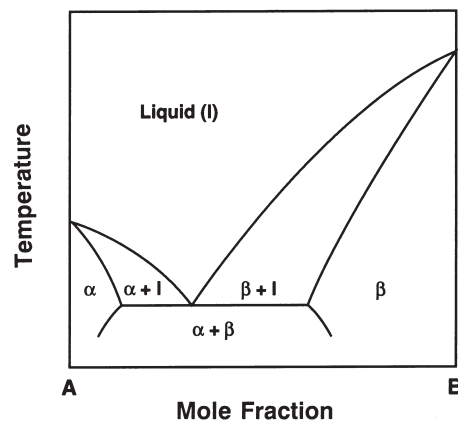


Figure 16-10. Phase diagram for a discontinuous solid solution for Solid A and Solid B: true solid solutions α and β are separated by a eutectic phase. (Adapted from Grant DJW, Abougela IKA. *Analytical Proceedings: Proceedings of the Analytical Division of the Royal Society of Chemistry*. Dec 1992.)

More common are the *discontinuous* solid dispersions illustrated in Figure 16-10 where two true solid solutions α and β are separated by a eutectic phase. Such a system is found for urea/acetaminophen; which exists as solid solutions in very small regions at very high urea concentration and very high acetaminophen concentration.

At this point it is worthwhile to briefly consider solid complexes. The interaction of a drug with an excipient to form a new solid phase through strong hydrogen-bond formation can give a solid phase that is not precisely a solid solution, but nonetheless potentially important in its effects on bioavailability—both in a positive and negative sense. The phase diagram in Figure 16-11 was obtained by fusing and cooling mixtures of griseofulvin (G) and phenobarbitone (P). When a complex is stable up to its melting point, the liquidus curve shows a peak referred to as a *congruent melting point*. Two congruently melting complexes, PG_3 and P_3G (at $x = 0.25$ and 0.75 in Fig 16-11), are found for the griseofulvin-phenobarbital system.

Thermodynamics of the Solution Process

In this discussion of the thermodynamics of the solution process, the solute is assumed to be in the liquid state, hence, the

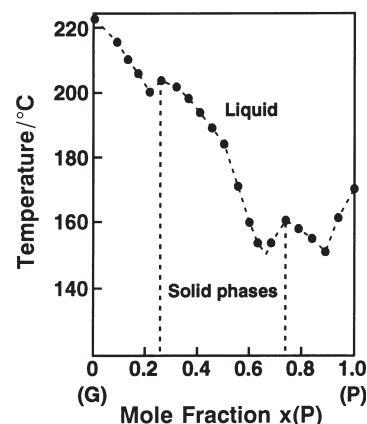
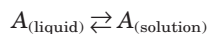


Figure 16-11. Temperature-concentration relationship for fused mixtures of griseofulvin (G) and phenobarbital (P). (Adapted from Grant DJW, Abougela IKA. *Analytical Proceedings: Proceedings of the Analytical Division of the Royal Society of Chemistry*. Dec 1992.)

heat of solution ($\Delta H'$) is a term different from that in Equation 3 (ΔH). The heat of solution for a solid solute going into solution as defined in Equation 3 is the net heat effect for the overall dissolution



Considering only the process,



and assuming that the solute is a liquid (or a super-cooled liquid in the case of a solid) at a temperature close to room temperature, where the energy needed for melting (heat of fusion) is not being considered.

For a physical or chemical reaction to occur spontaneously at a constant temperature and pressure, the net free-energy change, ΔG , for the reaction should be negative (see *Thermodynamics*, Chapter 15). Furthermore, it is known that the free-energy change depends on heat-related enthalpy ($\Delta H'$) and order-related entropy (ΔS) factors as seen in

$$\Delta G = \Delta H' - T\Delta S \quad (13)$$

where T is the temperature. Recall also that the relation between free energy and the equilibrium constant, K , for a reaction is given by

$$\Delta G = -1 RT \ln K \quad (14)$$

Equations 13 and 14 certainly apply to the solution of a drug. Because the solubility is, in reality, an equilibrium constant, Equation 14 indicates that the greater the negative value of ΔG , the greater the solubility.

The interplay of these two factors, $\Delta H'$ and ΔS in Equation 13, determines the free-energy change, and hence whether dissolution of a drug will occur spontaneously. Thus, if in the solution process $\Delta H'$ is negative and ΔS positive, dissolution is favored because ΔG will be negative.

As the heat of solution is quite significant in the dissolution process one must look at its origin. (For an excellent and more complete discussion of the interactions and driving forces underlying the dissolution process, see Higuchi.⁶) The mechanism of solubility involves severing of the bonds that hold together the ions or molecules of a solute, the separation of molecules of solvent to create a space in the solvent into which the solute can be fitted and the ultimate response of solute and solvent to whatever forces of interaction may exist between them. In order to sever the bonds between molecules or ions of solute in the liquid state, energy must be supplied, as is the case also when molecules of solvent are to be separated. If heat is the source of energy it is apparent that both processes require the absorption of heat.

Solute-solvent interaction, on the other hand, generally is accompanied by the evolution of heat as the process occurs spontaneously. In effecting solution there is, accordingly, a heat-absorbing effect and a heat-releasing effect to be considered beyond those required to melt a solid. If there is no, or very little, interaction between solute and solvent, the only effect will be that of absorption of heat to produce the necessary separations of solute and solvent molecules or ions. If there is a significant interaction between solute and solvent, the amount of heat in excess of that required to overcome the solute-solute and the solvent-solvent forces is liberated. If the opposing heat effects are equal, there will be no change of temperature.

When $\Delta H'$ is zero, and there is no volume change, an *ideal solution* is said to exist because the solute-solute, solvent-solvent, and solute-solvent interactions are the same. For such an ideal solution, the solubility of a solid can be predicted from its heat of fusion (the energy needed to melt the solid) at temperatures below its melting point. The student is referred to Martin et al.¹ to see how this calculation is made.

When the heat of solution has a positive (energy absorbed) or negative (energy liberated) value, the solution is said to be a *nonideal solution*. A negative heat of solution favors solubility while a positive heat works against dissolution.

The magnitude of the various attractive forces involved between solute, solvent, and solute-solvent molecules may vary greatly and thus could lead to varying degrees of positive or negative enthalpy changes in the solution process. The reason for this is that the molecular structure of the various solutes and solvents determining the interactions can themselves vary greatly. For a discussion of these effects, see Martin et al.¹

The solute-solute interaction that must be overcome can vary from the strong ion-ion interaction (as in a salt), to the weaker dipole-dipole interaction (as in nearly all organic medicinals that are not salts), to the weakest induced dipole-induced dipole interaction (as with naphthalene).

The attractive forces in the solvent that must be overcome are, most frequently, the dipole-dipole interaction (as found in water or acetone) and the induced dipole-induced dipole interaction (as in liquid petrolatum).

The energy-releasing solute-solvent interactions that must be taken into account may be one of four types. In decreasing energy of interaction these are ion-dipole interactions (eg, a sodium ion interacting with water), dipole-dipole interactions (eg, an organic acid dissolved in water), dipole-induced dipole interaction, to be discussed later (eg, an organic acid dissolved in carbon tetrachloride) and induced dipole-induced dipole interactions (eg, naphthalene dissolved in benzene).

Since the energy-releasing solute-solvent interaction should approximate the energy needed to overcome the solute-solute and solvent-solvent interactions, it should be apparent why it is not possible to dissolve a salt like sodium chloride in benzene. The interaction between the ions and benzene does not supply enough energy to overcome the interaction between the ions in the solute and therefore gives rise to a positive heat of solution. On the other hand, the interaction of sodium and chloride ions with water molecules does provide an amount of energy approximating the energy needed to separate the ions in the solute and the molecules in the solvent.

Consideration must next be given to entropy effects in dissolution processes. Entropy is an indicator of the disorder or randomness of a system. The more positive the entropy change (ΔS) is, the greater the degree of randomness or disorder of the reaction system and the more favorably disposed is the reaction. Unlike $\Delta H'$, the entropy change (an entropy of mixing) in an ideal solution, is not zero, but has some positive value as there is an increase in the disorderliness or entropy of the system upon dissolution. Thus, in an ideal solution with $\Delta H'$ zero and ΔS positive, ΔG would have a negative value and the process would therefore be spontaneous.

In a nonideal solution, on the other hand, where $\Delta H'$ is not zero, ΔS can be equal to, greater than, or less than the entropy of mixing found for the ideal solution. A nonideal solution with an entropy of mixing equal to that of the ideal solution is called a *regular solution*. These solutions usually occur with nonpolar or weakly polar solutes and solvents. Such solutions are accompanied by a positive enthalpy change, implying that the solute-solvent molecular interaction is less than the solute-solute and solvent-solvent molecular interactions. Regular solutions are amenable to rigorous physical chemical analysis, which will not be covered in this chapter but which can be found in outline form in Martin et al.¹

The possibility exists in a nonideal solution that the entropy change is greater than for an ideal solution. Such a solution occurs when there is an association among solute or solvent molecules. In essence, the dissolution process occurs when starting at a relatively ordered (low entropy) state and progressing to a disorderly (high entropy) state.

The overall entropy change is positive, greater than that of the ideal case, and favorable to dissolution. As may be expected, the enthalpy change in such a solution is positive because association in a solute or solvent must be overcome. The facilitated solubility of citric acid (an unsymmetrical

molecule), as compared to inositol (a symmetrical molecule), may be explained on the basis of such a favorable entropy change.⁶

The solubility of citric acid is greater than that of inositol, yet on the basis of their heats of solution, inositol should be more soluble. One may regard this phenomenon in another way. The reason for the higher solubility of citric acid is that, although there is no hindrance in the transfer of a citric acid molecule as it goes from the solute to the solution phase, when the structurally unsymmetrical citric acid attempts to return to the solute phase from solution, it must assume an orientation that will allow ready interaction with polar groups already oriented. If it does not have the required orientation, it will not return readily to the solute, but rather will remain in solution, thus bringing about a solubility larger than expected on the basis of heat of solution.

On the other hand, the structurally symmetrical inositol, as it leaves the solution phase, can interact with the solute phase without requiring a definite orientation; all orientations are equivalent. Hence, inositol can enter the solute phase without hindrance, and therefore no facilitation of its solubility is observed. In general, unsymmetrical molecules tend to be more soluble than symmetrical molecules.

Another type of nonideal solution occurs when there is an entropy change less than that expected of an ideal solution.

Such nonideal behavior can occur with polar solutes and solvents. In a nonideal solution of this type there is significant interaction between solute and solvent. As may be expected, the enthalpy change ($\Delta H'$) in such a solution is negative and favors dissolution, but this effect is tempered by the unfavorable entropy change occurring at the same time. The reason for the lower-than-ideal entropy change can be visualized where the equilibrium system is more orderly and has a lower entropy than that expected for an ideal solution. The overall entropy change of solution thus would be less and not favorable to dissolution.

One may rationalize the lower-than-expected solubility of lithium fluoride on the basis of this phenomenon. Compared with other alkali halides, it has a solubility lower than would be expected based solely on enthalpy changes. Because of the small size of ions in this salt there may be considerable ordering of water molecules in the solution. This effect must, of course, lead to a lowered entropy and an unfavorable effect on solubility. The effect of soluble salts on the solubility of nonelectrolytes may be considered as a result of an unfavorable entropy effect (see *Solubility of Solute Containing Two or More Species*, above).

PHARMACEUTICAL SOLVENTS

The discussion will focus now on solvents available to pharmacists and on the properties of these solvents. Pharmacists must obtain an understanding of the possible differences in solubility of a given solute in various solvents because they are often called on to select a solvent that will dissolve the solute. A knowledge of the properties of solvents will allow the intelligent selection of suitable solvents.

On the basis of the forces of interaction occurring in solvents one may broadly classify solvents as one of three types:

1. *Polar solvents*—those made up of strong dipolar molecules having hydrogen bonding (water or hydrogen peroxide).
2. *Semipolar solvents*—those also made up of strong dipolar molecules but that do not form hydrogen bonds (acetone or pentyl alcohol).
3. *Nonpolar solvents*—those made up of molecules having a small or no dipolar character (benzene, vegetable oil, or mineral oil).

Naturally, there are many solvents that may fit into more than one of these broad classes; for example, chloroform is a weak dipolar compound but generally is considered nonpolar in char-

acter, and glycerin could be considered a polar or semipolar solvent even though it is capable of forming hydrogen bonds.

Solvent Types

WATER—Water is a unique solvent. Besides being a highly associated liquid, giving rise to its high boiling point, it has another very important property, a high dielectric constant. The *dielectric constant* (ϵ) indicates the effect that a substance has, when it acts as a medium, on the ease with which two oppositely charged ions may be separated. The ease of solubilizing salts in solvents like water and glycerin can be explained on the basis of their high dielectric constant. Also, in general, the more polar the solvent, the greater its dielectric constant.

An important concept has been introduced to pharmaceutical systems: pharmacists frequently are concerned with dissolving relatively nonpolar drugs in aqueous or mixed polar aqueous solvents.¹¹ To understand what may be happening in such cases, factors concerned with the entropic effects arising from interactions originating with the nonpolar solutes must be considered. Previously it had been noted that the favorable entropic effect on dissolution was due to the disruption of associations occurring among solute or solvent molecules. Now consider the effects on solubility due to solute interactions in the solution phase—because the solutes under discussion are relatively nonpolar, the interactions are of the London Force type or a *hydrophobic association*.

This hydrophobic association in aqueous solutions may cause significant structuring of water with a resultant ordered or low-entropy system that is unfavorable to solution. Therefore, the solution of an essentially nonpolar molecular in water is not a favorable process. It should be stressed that this is due to not only an unfavorable enthalpy change but also an unfavorable entropy change generated by water structuring.

Such an unfavorable entropy change, known as the *hydrophobic effect*, is quite significant in the solution process. As an example of this effect, the aqueous solubility of a series of alkyl *p*-aminobenzoates shows a 10-million-fold decrease in solubility in going from the 1-carbon analog to the 12-carbon analog. These findings demonstrate clearly the considerable effect that hydrophobic associations can have.

ALCOHOLS—*Ethanol*, as a solvent, is next in importance to water. An advantage of ethanol is that growth of microorganisms does not occur in solutions containing alcohol in a reasonable concentration.

Resins, volatile oils, alkaloids, glycosides, etc are dissolved by alcohol, but many therapeutically inert principles, such as gums, albumin, and starch, are insoluble, which makes it more useful as a *selective* solvent. Mixtures of water and alcohol, in proportions varying to suit specific cases, are used extensively. They are often referred to as *hydroalcoholic solvents*.

Glycerin is an excellent solvent, although its range is not as extensive as that of water or alcohol. In higher concentrations it has preservative action. It dissolves the fixed alkalies, a large number of salts, vegetable acids, pepsin, tannin, and some active principles of plants, but it also dissolves gums, soluble carbohydrates, and starch. It also is of special value as a simple solvent (as in phenol glycerite), or where the major portion of the glycerin simply is added as a preservative and stabilizer of solutions that have been prepared with other solvents (see *Glycerines*, Chapter 41).

Propylene glycol, which has been used widely as a substitute for glycerin, is miscible with water, acetone, or chloroform in all proportions. It is soluble in ether and will dissolve many essential oils but is immiscible with fixed oils. It is claimed to be as effective as ethyl alcohol in its power of inhibiting mold growth and fermentation.

Isopropyl alcohol possesses solvent properties similar to those of ethyl alcohol and is used instead of the latter in a number of pharmaceutical manufacturing operations. It has the advantage in that the commonly available product contains not

over 1% of water, whereas ethyl alcohol contains about 5% water, often a disadvantage. Isopropyl alcohol is employed in some liniment and lotion formulations. It cannot be taken internally.

General Properties—Low-molecule-weight and polyhydroxy alcohol forms associated structures through hydrogen bonds just as in water. When the carbon-atom content of an alcohol rises above five, generally only monomers then are present in the pure solvent. Although alcohols have high dielectric constants compared to other types of solvents, they are small compared to water. As has been discussed, the solubility of salts in a solvent should be paralleled by its dielectric constant. That is, as the dielectric constant of a series of solvents increases, the probability of dissolving a salt in the solvent increases. This behavior is observed for the alcohols. Table 16-2, taken from Higuchi,⁶ shows how the solubility of salts follows the dielectric constant of the alcohols.

As mentioned earlier, absolute alcohol rarely is used pharmaceutically. However, hydroalcoholic mixtures such as elixirs and spirits frequently are encountered. A very useful generalization is that the dielectric properties of a mixed solvent, such as water and alcohol, can be approximated as the weighted average of the properties of the pure components. Thus, a mixture of 60% alcohol (by weight) in water should have a dielectric constant approximated by

$$\epsilon_{(\text{mixture})} = 0.6(\epsilon_{(\text{alcohol})}) + 0.4(\epsilon_{(\text{water})})$$

$$\epsilon_{(\text{mixture})} = 0.6(25) + 0.4(80) = 47$$

The dielectric constant of 60% alcohol in water is found experimentally to be 43, which is in close agreement with that just calculated. The dielectric constant of glycerin is 46, close to the 60% alcohol mixture. One would therefore expect a salt like sodium chloride to have about the same solubility in glycerin as in 60% alcohol. The solubility of sodium chloride in glycerin is 8.3 g/100 g of solvent and in 60% alcohol about 6.3 g/100 g of solvent. This agreement would be even closer if comparisons were made on a volume rather than weight basis. At least qualitatively it can be said that the solubility of a salt in a solvent or a mixed solvent closely follows the dielectric constant of the medium, or conversely that the polarity of mixed solvents is paralleled by their dielectric constant, based on salt solubility.

Although the dielectric constant is useful in interpreting the effect of mixed solvents on salt solubility, it cannot be applied properly to the effect of mixed solvents on the solubility of nonelectrolytes. It was seen earlier that unfavorable entropic effects can occur upon dissolution of relatively nonpolar nonelectrolytes in water. Such an effect due to hydrophobic association considerably affects solubility. Yalkowsky¹¹ studied the ability of cosolvent systems to increase the solubility of nonelectrolytes in polar solvents where the cosolvent system essentially brings about a reduction in structuring of solvent. Thus, by increasing, in a positive sense, the entropy of solution by using cosolvents,

it was possible to increase the solubility of the nonpolar molecule. Using as an example the solubility of alkyl *p*-aminobenzoates in propylene glycol-water systems, Yalkowsky reported that it is possible to increase the solubility of the nonelectrolyte by several orders of magnitude by increasing the fraction of propylene glycol in the aqueous system.⁸ Sometimes, it is found that, as a good first approximation, the logarithm of the solubility is related linearly to the fraction of propylene glycol added by

$$\log S_f = \log S_{f=0} + \epsilon f$$

where S_f is the solubility in the mixed aqueous system containing the volume fraction f of nonaqueous cosolvent, $S_{f=0}$ is the solubility in water, and ϵ is a constant (not dielectric constant) characteristic of the system under study. Specifically, when a 50% solution of propylene glycol in water is used, there is a 1000-fold increase in solubility of dodecyl *p*-aminobenzoate, in comparison to pure water.

Another empirical equation sometimes used to estimate solubility of a poorly water-soluble substance in a mixed-solvent system is written as

$$\log S_t = \log S_w \times f_w + \log S_1 \times f_1 + \dots$$

where S_t is the total solubility, S_w and S_1 are solubilities in pure water and cosolvent 1, respectively, and f_w and f_1 are the fractions of water and cosolvent 1, respectively.

In a series of studies, Martin et al⁹ have made attempts to predict solubility in mixed solvent systems through an extension of the *regular solution* theory. The equations are logarithmic in nature and can reduce in form to the equations of Yalkowsky.⁸

ACETONE AND RELATED SEMIPOLAR MATERIALS—Even though acetone has a very high dipole moment (2.8×10^{-18} esu), as a pure solvent it does not form associated structures. This is evidenced by its low boiling point (57°) in comparison with the boiling point of the lower molecular weight water (100°) and ethanol (79°). The reason why it does not associate is because the positive charge in its dipole does not reside in a hydrogen atom, precluding the possibility of its forming a hydrogen bond. However, if some substance that is capable of forming hydrogen bonds, such as water or alcohol, is added to acetone, a very strong interaction through hydrogen bonding will occur (see *Mechanism of Solvent Action*, below). Some substances that are semipolar and similar to acetone are aldehydes, low-molecular-weight esters, other ketones, and nitro-containing compounds.

NONPOLAR SOLVENTS—The nonpolar class of solvents includes fixed oils such as vegetable oil, and petroleum ether (ligroin), carbon tetrachloride, benzene, and chloroform. On a relative basis there is a wide range of polarity among these solvents; for example, benzene has no dipole moment whereas that of chloroform is 1.05×10^{-18} esu.

It should be emphasized that when a solvent (such as chloroform) has highly electronegative halogen atoms attached to a carbon atom that also contains at least one hydrogen atom, such a solvent will be capable of forming strong hydrogen bonds with solutes which are polar in character. Thus, through the formation of hydrogen bonds such solvents will dissolve polar solutes. For example, it is possible to dissolve alkaloids in chloroform.

Table 16-2. Solubilities of Potassium Iodide and Sodium Chloride in Several Alcohols and Acetone^a

SOLVENT	g KI/100 g SOLVENT	g NaCl/100 g SOLVENT
Water	148	35.9
Glycerin	...	8.3 (20°)
Propylene glycol	50	7.1 (30°)
Methanol	17	1.4
Acetone	2.9	...
Ethanol	1.88	0.065
1-Propanol	0.44	0.0124
2-Propanol	0.18	0.003
1-Butanol	0.2	0.005
1-Pentanol	0.089	0.0018

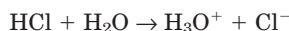
^a All measurements are at 25° unless otherwise indicated. Data from Duddu S. PhD thesis. University of Minnesota, 1993.

Mechanism of Solvent Action

A solvent may function in one or more ways. When an ionic salt is dissolved (eg, by water), the process of solution involves separation of the cations and anions of the salt with attendant orientation of molecules of the solvent about the ions. Such orientation of solvent molecules about the ions of the solute—a process called *solvation* (*hydration*, if the solvent is water)—is possible only when the solvent is highly polar, whereby the dipoles of the solvent are attracted to and held by the ions of the

solute. The solvent also must possess the ability to keep the solvated, charged ions apart with minimal energy.

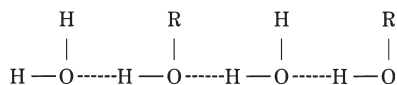
A polar liquid such as water may exhibit solvent action also by virtue of its ability to break a covalent bond in the solute and bring about ionization of the latter. For example, hydrogen chloride dissolves in water and functions as an acid as a result of



The ions formed by this preliminary reaction of breaking the covalent bond are subsequently maintained in solution by the same mechanism as ionic salts.

Still another mechanism by which a polar liquid may act as a solvent is that involved when the solvent and solute are capable of being coupled through hydrogen-bond formation.

The solubility of the low-molecular-weight alcohols in water, for example, is attributed to the ability of the alcohol molecules to become part of a water-alcohol association complex.

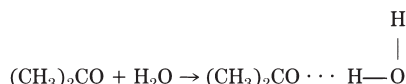


As the molecular weight of the alcohol increases, it becomes progressively less polar and less able to compete with water molecules for a place in the lattice-like arrangement formed through hydrogen bonding; high-molecular-weight alcohols are, therefore, poorly soluble or insoluble in water. When the number of carbon atoms in a normal alcohol reaches five, its solubility in water is reduced materially.

When the number of hydroxyl groups in the alcohol is increased, its solubility in water generally is increased greatly; it is principally, if not entirely, for this reason that such high-molecular-weight compounds as sugars, gums, and many glycosides, and synthetic compounds such as the polyethylene glycols, are very soluble in water.

The solubility of ethers, aldehydes, ketones, acids, and anhydrides in water and in other polar solvents also is attributable largely to the formation of an association complex between solute and solvent by means of the hydrogen bond. The molecules of ethers, aldehydes, and ketones, unlike those of alcohols, are not associated themselves, because of the absence of a hydrogen atom that is capable of forming the characteristic hydrogen bond. Notwithstanding, these substances are more or less polar because of the presence of a strongly electronegative oxygen atom, which is capable of association with water through hydrogen-bond formation.

Acetone, for example, dissolves in water, in all likelihood principally because of the following type of association:



The maximum number of carbon atoms that may be present per molecule possessing a hydrogen-bondable group, while still retaining water solubility, is approximately the same as for the alcohols.

Although nitrogen is less electronegative than oxygen, and thus tends to form weaker hydrogen bonds, amines are at least as soluble as alcohols containing an equivalent chain length. The reason for this is that alcohols form two hydrogen bonds with a net interaction of 12 kcal/mol. Primary amines can form three hydrogen bonds; two amine protons are shared with the oxygens of two water molecules, and the nitrogen accepts one water proton. The net interaction for the primary amine is between 12 and 13 kcal/mol; hence, it shows an equal or greater solubility compared with corresponding alcohols.

The solvent action of nonpolar liquids involves a somewhat different mechanism. Because they are unable to form dipoles with which to overcome the attractions between ions of an ionic salt, or to break a covalent bond to produce an ionic compound or form association complexes with a solute, nonpolar liquids

are incapable of dissolving polar compounds. They only can dissolve, in general, other nonpolar substances in which the bonds between molecules are weak. The forces involved usually are of the induced dipole-induced dipole type. Such is the case when one hydrocarbon is dissolved in another, or an oil or a fat is dissolved in petroleum ether.

Sometimes it is observed that a polar substance, such as alcohol, will dissolve in a nonpolar liquid, such as benzene. This apparent exception to the preceding generalization may be explained by the assumption that the alcohol molecule induces a temporary dipole in the benzene molecule which forms an association complex with the solvent molecules. A binding force of this kind is referred to as a *permanent dipole-induced dipole force*.

SOME USEFUL GENERALIZATIONS—The preceding discussion indicates that enough is known about the mechanism of solubility to be able to formulate some generalizations concerning this important physical property of substances. Because of the greater importance of organic substances in the field of medicinal chemistry, certain of the more useful generalizations about organic chemicals are presented here in summary form. However, it should be remembered that the phenomenon of solubility usually involves several variables, and there may be exceptions to general rules.

One general maxim that holds true in most instances is, the greater the structural similarity between solute and solvent, the greater the solubility. As often stated to the student, *like dissolves like*. Thus, phenol is almost insoluble in petroleum ether but is very soluble in glycerin.

Organic compounds containing polar groups capable of forming hydrogen bonds with water are soluble in water, provided that the molecular weight of the compound is not too great. It is demonstrated easily that the polar groups OH, CHO, COH, CHOH, CH₂OH, COOH, NO₂, CO, NH₂, and SO₃H tend to increase the solubility of an organic compound in water. On the other hand, nonpolar or very weak polar groups, such as the various hydrocarbon radicals, reduce solubility; the greater the number of carbon atoms in the radical, the greater the decrease in solubility. Introduction of halogen atoms into a molecule in general tends to decrease solubility because of an increased molecular weight without a proportionate increase in polarity.

The greater the number of polar groups contained per molecule, the greater the solubility of a compound, provided that the size of the rest of the molecule is not altered; thus, pyrogallol is much more soluble in water than phenol. The *relative positions* of the groups in the molecule also influence solubility; thus, in water, resorcinol (*m*-dihydroxybenzene) is more soluble than catechol (*o*-dihydroxybenzene), and the latter is more soluble than hydroquinone (*p*-dihydroxybenzene).

Polymers and compounds of high molecular weight can be poorly soluble.

High melting points frequently are indicative of low solubility for organic compounds. One reason for high melting points

Table 16-3. Demonstration of Solubility Rules

CHEMICAL COMPOUND	SOLUBILITY ^a
Aniline, C ₆ H ₅ NH ₂	28.6
Benzene, C ₆ H ₆	1430.0
Benzoic acid, C ₆ H ₅ COOH	275.0
Benzyl alcohol C ₆ H ₅ CH ₂ OH	25.0
1-Butanol, C ₄ H ₉ OH	12.0
<i>t</i> -Butyl alcohol, (CH ₃) ₃ COH	Miscible
Carbon tetrachloride, CCl ₄	2000.0
Chloroform, CHCl ₃	200.0
Fumaric acid (<i>trans</i> -butenedioic acid)	150.0
Hydroquinone, C ₆ H ₄ (OH) ₂	14.0
Maleic acid, <i>cis</i> -butenedioic acid	5.0
Phenol, C ₆ H ₅ OH	15.0
Pyrocatechol, C ₆ H ₄ (OH) ₂	2.3
Pyrogallol, C ₆ H ₃ (OH) ₃	1.7
Resorcinol, C ₆ H ₄ (OH) ₂	0.9

^a The number of mL of water required to dissolve 1 g of solute.

is the association of molecules, and this cohesive force tends to prevent dispersion of the solute in the solvent.

The *cis* form of an isomer is more soluble than the *trans* form (Table 16-3).

Solvation, which is evidence of the existence of a strong attractive force between solute and solvent, enhances the solubility of the solute, provided there is not a marked ordering of the solvent molecules in the solution phase.

Acids, especially strong acids, usually produce water-soluble salts when reacted with nitrogen-containing organic bases.

COLLIGATIVE PROPERTIES OF SOLUTIONS

Up to this point our concern has been with dissolving a solute in a solvent. Once the dissolution has been brought about, naturally the solution has a number of properties that are different from that of the pure solvent. Of very great importance are the colligative properties that a solution possesses.

The *colligative properties* of a solution are those that depend on the number of solute particles in solution, irrespective of whether these are molecules or ions, large or small. Ideally, the effect of a solute particle of one species is considered to be the same as that of an entirely different kind of particle, at least in dilute solution. Practically, there may be differences that may become substantial as the concentration of the solution is increased.

The colligative properties that will be considered are

1. Osmotic pressure.
2. Vapor-pressure lowering.
3. Boiling-point elevation.
4. Freezing-point depression.

Of these four, all of which are related, osmotic pressure has the greatest direct importance in the pharmaceutical sciences. It is the property that largely determines the physiological acceptability of a variety of solutions used for therapeutic purposes.

Osmotic-Pressure Elevation

OSMOSIS—The phenomenon of osmosis is based on the fact that substances tend to move or diffuse from regions of higher concentration to regions of lower concentration. When a solution is separated from the solvent by means of a membrane that is permeable to the solvent but not to the solute (such a membrane is referred to as a *semipermeable* membrane), it is possible to demonstrate visibly the diffusion of solvent into the concentrated solution, as volume changes will occur. In a similar manner, if two solutions of different concentration are separated by a membrane, the solvent will move from the solution of lower solute concentration to the solution of higher solute concentration. This diffusion of solvent through a membrane is called *osmosis*.

There is a difference between the activity or escaping tendency of the water molecules found in the solvent and salt solution separated by the semipermeable membrane. Because *activity*, which is related to water concentration, is higher on the pure solvent side, water moves from solvent to solution in order to equalize *escaping-tendency* differences. The difference in escaping-tendency gives rise to what is referred to as the *osmotic pressure* of the solution, which might be visualized as follows. A semipermeable membrane is placed over the end of a tube and a small amount of salt solution placed over the membrane in the tube. The tube then is immersed in a trough of pure water so that the upper level of the salt solution initially is at the same level as the water in the trough. With time, solvent molecules will move from solvent into the tube. The height of the solution will rise until the *hydrostatic pressure* exerted by the column of solution is equal to the *osmotic pressure*.

OSMOTIC PRESSURE OF NONELECTROLYTES—Quantitative studies using solutions of varying concentration of a solute that does not ionize have demonstrated that osmotic pressure is proportional to the concentration of the solute; that is, twice the concentration of a given nonelectrolyte will produce twice the osmotic pressure in a given solvent. (This is not strictly true in solutions of fairly high solute concentration, but does hold quite well for dilute solutions.)

Furthermore, the osmotic pressures of solutions of different nonelectrolytes are proportional to the number of molecules in each solution. Stated in another manner, the osmotic pressures of two nonelectrolyte solutions of the same molal concentration are identical. Thus, a solution containing 34.2 g of sucrose (mol wt 342) in 1000 g of water has the same osmotic pressure as a solution containing 18.0 g of anhydrous dextrose (mol wt 180) in 1000 g of water. These solutions are said to be *iso-osmotic* (*isosmotic*) with each other because they have identical osmotic pressures.

OSMOTIC PRESSURE OF ELECTROLYTES—In discussing the generalizations concerning the osmotic pressure of solutions of nonelectrolytes it was stated that the osmotic pressures of two solutions of the same molal concentration are identical. This generalization, however, cannot be made for solutions of electrolytes—acids, alkalies, and salts (see Chapter 17).

For example, sodium chloride is assumed to ionize as



It is evident that each molecule of sodium chloride that ionizes produces two ions; if sodium chloride is completely ionized, there will be twice as many particles as would be the case if it were not ionized at all. Furthermore, if each ion has the same effect on osmotic pressure as a molecule, it might be expected that the osmotic pressure of the solution would be twice that of a solution containing the same molal concentration of nonionizing substance.

For solutions that yield more than two ions—for example,



it is expected that the complete dissociation of the molecules would give rise to osmotic pressures that are three and four times, respectively, the pressure of solutions containing an equivalent quantity of a nonionized solute. Accordingly, the equation $PV = nRT$, which may be employed to calculate the osmotic pressure of a dilute solution of a nonelectrolyte, also may be applied to dilute solutions of electrolytes if it is changed to $PV = inRT$, where the value of i approaches the number of ions produced by the ionization of the stronger electrolytes cited in the preceding examples. For weak electrolytes i represents the total number of particles, ions, and molecules together in the solution, divided by the number of molecules that would be present if the solute did not ionize. The experimental evidence indicates that at least in dilute solutions the osmotic pressures approach the predicted values. It should be emphasized, however, that in more concentrated solutions of electrolytes the deviations from this simple theory are considerable, due to interionic attraction, solvation, and other factors.

BIOLOGICAL ASPECTS OF OSMOTIC PRESSURE—Osmotic pressure experiments were made as early as 1884 by the Dutch botanist Hugo de Vries in his study of *plasmolysis*, the term applied to the contraction of the contents of plant cells placed in solutions of comparatively high osmotic pressure. The phenomenon is caused by the osmosis of water out of the cell through the practically semipermeable membrane surrounding the protoplasm. If suitable cells (eg, the epidermal cells of the leaf of *Tradescantia discolor*) are placed in a solution of higher osmotic pressure than that of the cell contents, water flows out of the cell, causing the contents to draw away from the cell wall. On the other hand, if the cells are placed in solutions of lower osmotic pressure, water enters the

cell, producing an expansion that is limited by the rigid cell wall. By immersing cells in a series of solutions of varying solute concentration, a solution may be found in which plasmolysis is barely detectable or absent. The osmotic pressure of such a solution is then the same, or very nearly the same, as that of the cell contents, and it is then said that the solution is *isotonic* with the cell contents. Solutions of greater concentration than this are said to be *hypertonic*, and solutions of lower concentration are called *hypotonic*.

Red blood cells, or erythrocytes, have been studied similarly by immersion into solutions of varying concentration of different solutes.¹⁰ When introduced into water or into sodium chloride solutions containing less than 0.90 g of solute per 100 mL, human erythrocytes swell, and often burst, because of the diffusion of water into the cell and the fact that the cell wall is not sufficiently strong to resist the pressure. This phenomenon is referred to as *hemolysis*. If the cells are placed in solutions containing more than 0.90 g of sodium chloride per 100 mL, they lose water and shrink. By immersing the cells in a solution containing exactly 0.90 g of sodium chloride in 100 mL, no change in the size of the cells is observed; because in this solution the cells maintain their *tone*, the solution is said to be *isotonic* with human erythrocytes. For the reasons indicated it is desirable that solutions to be injected into the blood should be made isotonic with erythrocytes. The manner in which this may be done is described in Chapter 18.

DISTINCTION BETWEEN ISO-OSMOTIC (ISOSMOTIC) AND ISOTONIC—The terms *isosmotic* and *isotonic* are not to be considered as equivalent, although a solution often may be described as being both isosmotic and isotonic. If a plant or animal cell is in contact with a solution that has the same osmotic pressure as the cell contents, there will be no net gain or loss of water by either solution provided the cell membrane is impermeable to all the solutes present. As the volume of the cell contents remains unchanged, the *tone*, or normal state, of the cell is maintained, and the solution in contact with the cell may be described not only as being isosmotic with the solution in the cell, but also as being isotonic with it. If, however, one or more of the solutes in contact with the membrane can pass through the latter, it is evident that the volume of the cell contents will change, thus altering the tone of the cell; in this case the two solutions may be isosmotic, yet not be isotonic.

Vapor-Pressure Lowering

When a nonvolatile solute is dissolved in a liquid solvent the vapor pressure of the solvent is lowered. This easily can be described qualitatively by visualizing solvent molecules on the surface of the solvent, which normally could escape into the vapor, being replaced by solute molecules which have little if any vapor pressure of their own. For ideal solutions of nonelectrolytes the vapor pressure of the solution follows Raoult's law

$$P_A = X_A P_A^0 \quad (15)$$

where P_A is the vapor pressure of the solution, P_A^0 is the vapor pressure of the pure solvent, and X_A is the mol fraction of solvent. This relationship states that the vapor pressure of the solution is proportional to the number of molecules of solvent in the solution. Rearranging Equation 15 gives

$$\frac{P_A^0 - P_A}{P_A^0} = (1 - X_A)X_B \quad (16)$$

where X_B is the mole fraction of the solute. This equation states that the lowering of vapor pressure in the solution relative to the vapor pressure of the pure solvent—called simply the *relative vapor-pressure lowering*—is equal to the mol fraction of the solute. The *absolute lowering* of vapor pressure of the solution is defined by

$$P_A^0 - P_A = X_B P_A^0 \quad (17)$$

Example Calculate the lowering of vapor pressure and the vapor pressure at 20°, of a solution containing 50 g of anhydrous dextrose (mol wt 180.16) in 1000 g of water (mol wt 18.02). The vapor pressure of water at 20°, in absence of air, is 17.535 mm.

First calculate the lowering of vapor pressure, using Equation 17 in which X_B is the mol fraction of dextrose, defined by

$$X_B = \frac{n_B}{n_A + n_B}$$

where n_A is the number of mols of solvent and n_B is the number of mols of solute. Substituting numerical values

$$n_B = \frac{50}{180.2} = 0.278$$

$$n_A = \frac{1000}{18.02} = 55.5$$

$$X_B = \frac{0.278}{55.5 + 0.278} = 0.00498$$

the lowering of vapor pressure is

$$\begin{aligned} P_A^0 - P_A &= 0.00498 \times 17.535 \\ &= 0.0873 \text{ mm} \end{aligned}$$

The vapor pressure of the solution is

$$\begin{aligned} P_A &= 17.535 - 0.0873 \\ &= 17.448 \text{ mm} \end{aligned}$$

Boiling-Point Elevation

In consequence of the fact that the vapor pressure of any solution of a nonvolatile solute is less than that of the solvent, the *boiling point* of the solution—the temperature at which the vapor pressure is equal to the applied pressure (commonly 760 mm)—must be higher than that of the solvent. This is clearly evident in Figure 16-12.

Freezing-Point Depression

The *freezing point of a solvent* is defined as the temperature at which the solid and liquid forms of the solvent coexist in equilibrium at a fixed external pressure, commonly 1 atmosphere (1 atm = 760 mm [torr] of mercury). At this temperature the solid and liquid forms of the solvent must have the same vapor

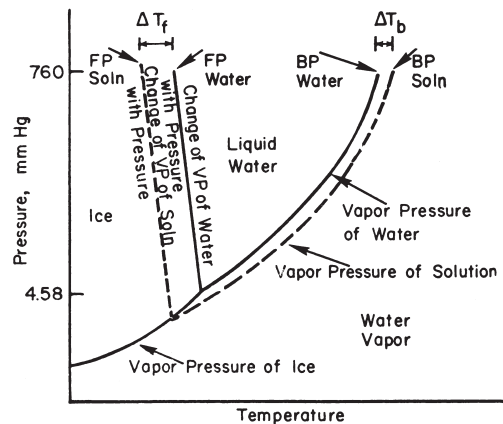


Figure 16-12. Vapor-pressure-temperature diagram for water and an aqueous solution, illustrating elevation of boiling point and lowering of freezing point of the latter.

pressure, for if this were not so, the form having the higher vapor pressure would change into that having the lower vapor pressure.

The *freezing point of a solution* is the temperature at which the solid form of the pure solvent coexists in equilibrium with the solution at a fixed external pressure, again commonly 1 atm. Because the vapor pressure of a solution is lower than that of its solvent, it is obvious that solid solvent and solution cannot coexist at the same temperature as solid solvent and liquid solvent; only at some lower temperature, where solid solvent and solution do have the same vapor pressure, is equilibrium established. A schematic pressure-temperature diagram for water and an aqueous solution, not drawn to scale and exaggerated for the purpose of more effective illustration, shows the equilibrium conditions involved in both freezing-point depression and boiling-point elevation (see Fig 16-12).

The freezing-point lowering of a solution may be quantitatively predicted for ideal solutions, or dilute solutions that obey Raoult's law, by mathematical operations similar to (though somewhat more complex than) those used in deriving the boiling-point elevation constant. The equation for the freezing-point lowering, ΔT_f , is

$$\Delta T_f = \frac{RT_0^2 M_A m}{1000 \Delta H_{\text{fus}}} = K_f m \quad (18)$$

where

$$K_f = \frac{RT_0^2 M_A}{1000 \Delta H_{\text{fus}}} \quad (19)$$

The value of K_f for water, which freezes at 273.1°K and has a heat of fusion of 79.7 cal/g, is

$$K_f = \frac{1.987 \times 273.1^2 \times 18.02}{1000 \times 18.02 \times 79.7} = 1.86^\circ \quad (20)$$

The molal freezing-point depression constant is not intended to represent the freezing-point depression for a 1-molal solution, which is too concentrated for the premise of ideal behavior to be applicable. In dilute solutions the freezing-point depression, calculated to a 1-molal basis, approaches the theoretical value—the more dilute the solution, the better the agreement between experiment and theory.

To calculate the molecular weight of the solute, the freezing point of a dilute solution of a nonelectrolyte solute may be used (as was the boiling point). The applicable equation is

$$M_B = \frac{K_f 1000 w_B}{w_A \Delta T_f} \quad (21)$$

The molecular weight of organic substances soluble in molten camphor may be determined by observing the freezing point of a mixture of the substance with camphor. This procedure, called the *Rast method*, uses camphor because it has a very large molal freezing-point-depression constant, about 40. Because the *constant* may vary with different lots of camphor and with variations of technique, the method should be standardized using a solute of known molecular weight.

Freezing-point determinations of molecular weights have the advantage over boiling-point determinations of greater accuracy and precision by virtue of the larger magnitude of the freezing-point depression compared to boiling-point elevation. Thus, in the case of water the molal freezing-point depression is approximately 3.5 times greater than the molal boiling-point elevation.

Ideal Behavior and Deviations

In setting out to derive mathematical expressions for colligative properties, such phrases as for *ideal solutions* or for *dilute solutions* were used to indicate the limitations of the expressions. Samuel Glasstone defines an ideal solution as "one which obeys

Raoult's law over the whole range of concentration and at all temperatures" and gives as specific characteristics of such solutions their formation only from constituents that mix in the liquid state without heat change and without volume change. These characteristics reflect the fact that addition of a solute to a solvent produces no change in the forces between molecules of the solvent. Thus, the molecules have the same escaping-tendency in the solution as in the pure solvent and the vapor pressure above the solution is proportional to the ratio of the number of solvent molecules in the surface of the solution to the number of the molecules in the surface of the solvent—which is the basis for Raoult's law.

Any change in intermolecular forces produced by mixing the components of a solution may result in deviation from ideality; such a deviation may be expected particularly in solutions containing both a polar and a nonpolar substance. Solutions of electrolytes, except at high dilution, are especially prone to depart from ideal behavior, even though allowance is made for the additional particles that result from ionization. When solute and solvent combine to form solvates, the escaping-tendency of the solvent may be reduced in consequence of the reduction in the number of free molecules of solvent; thus, a negative deviation from Raoult's law is introduced. On the other hand, the escaping-tendency of the solvent in a solution of nonvolatile solute may be increased, because the cohesive forces between molecules of solvent are reduced by the solute; this results in a positive deviation from Raoult's law. Chapter 17 considers deviations from ideality in more detail.

Although few solutions exhibit ideal behavior over a wide range of concentration, most solutions behave ideally at least in high dilution, where deviations from Raoult's law are negligible.

COLLIGATIVE PROPERTIES OF ELECTROLYTE SOLUTIONS (See Chapter 17)—Earlier in this chapter attention was directed to the increased osmotic pressure observed in solutions of electrolytes, the enhanced effect being attributed to the presence of ions, each of which acts, in general, in the same way as a molecule in developing osmotic pressure. Similar magnification of vapor-pressure lowering, boiling-point elevation, and freezing-point depression occurs in solutions of electrolytes. Thus, at a given constant temperature the abnormal effect of an electrolyte on osmotic pressure is paralleled by abnormal lowering of vapor pressure; the other colligative properties are (subject to variation of effect with temperature) comparably intensified. In general, the magnitude of each colligative property is proportional to the total number of particles (molecules and/or ions) in solution.

While in *very* dilute solutions the osmotic pressure, vapor-pressure lowering, boiling-point increase, and freezing-point depression of solutions of electrolytes would approach values two, three, and four times greater for NaCl, Na₂SO₄, and Na₃PO₄ than in solutions of the same molality of a nonelectrolyte, two other effects are observed as the concentration of electrolyte is increased. The first effect results in less than 2-, 3-, or 4-fold intensification of a colligative property. This reduction is ascribed to interionic attraction between the positively and negatively charged ions. Consequently, the ions are not completely dissociated from each other and do not exert their full effect in lowering vapor pressure, etc. This deviation generally increases with increasing concentration of electrolyte. The second effect intensifies the colligative properties and is attributed to the attraction of ions for solvent molecules, which holds the solvent in solution and reduces its escaping-tendency, with consequent enhancement of the vapor-pressure lowering. Solvation also may reduce interionic attraction and thereby further lower the vapor pressure.

These factors (and possibly others) combine to effect a progressive reduction in the molal values of colligative properties as the concentration of electrolyte is increased 0.5 to 1.0 molal, beyond which the molal quantities either increase (sometimes quite abruptly) or remain almost constant.

Activity and Activity Coefficient

Various mathematical expressions are employed to relate properties of chemical systems (equilibrium constants, colligative properties, pH, etc) to the stoichiometric concentration of one or more molecular, atomic, or ionic species. In deriving such expressions it is either stated or implied that they are valid only so long as intermolecular, interatomic, and/or interionic forces may be ignored or remain constant, under which restriction the system may be expected to behave ideally. But intermolecular, interatomic and/or interionic forces do exist, and not only do they change as a result of chemical reaction, but they also change with variation in the concentration or pressure of the molecules, atoms, or ions under observation. In consequence, mathematical expressions involving stoichiometric concentrations or pressures generally have limited applicability. The conventional concentration terms provide a count of molecules, atoms, or ions per unit volume, but afford no indication of the physical or chemical activity of the species measured, and it is this activity that determines the physical and chemical properties of the system.

In recognition of this, GN Lewis introduced both the quantitative concept and methods for evaluation of activity as a true measure of the physical or chemical activity of molecular, atomic, or ionic species, whether in the state of gas, liquid, or solid, or whether present as a single species or in a mixture. *Activity* may be considered loosely as a corrected concentration or pressure that takes into account not only the stoichiometric concentration or pressure but also any intermolecular attractions, repulsions, or interactions between solute and solvent in solution, association, and ionization. Thus, activity measures the net effectiveness of a chemical species.

Because only relative values of activity may be determined, a *standard state* must be chosen for quantitative comparisons to be made. Indeed, because activity measurements are needed for many different types of systems, several standard states must be selected. Because this discussion is concerned mainly with solutions, the standard state for the solvent is pure solvent, while for the solute it is a hypothetical solution with free energy corresponding to unit molality under conditions of ideal behavior of the solution. The relationship of activity to concentration is measured in terms of an activity coefficient, which is discussed in Chapter 17.

Practical Applications of Colligative Properties

One of the most important pharmaceutical applications of colligative properties is in the preparation of isotonic intravenous and isotonic lacrimal solutions, the details of which are discussed in Chapter 18.

Other applications of the colligative properties are found in experimental physiology. One such application is in the immersion of tissues in salt solutions, which are isotonic with the fluids of the tissue, in order to prevent changes or injuries that may arise from osmosis.

The colligative properties of solutions also may be used in determining the molecular weight of solutes, or in the case of electrolytes, the extent of ionization. The method of determining molecular weight depends on the fact that each of the colligative properties is altered by a constant value when a definite number of molecules of solute is added to a solvent (see Chapter 17). For example, in dilute solutions the freezing point of water is lowered at the rate of 1.855 for each mol of a nonelectrolyte dissolved in 1000 g of water.

The boiling-point elevation may be used similarly for determining molecular weights. The boiling point of water is raised at the rate of 0.52° for each mol of solute dissolved in 1000 g of water; the corresponding values for benzene, carbon tetrachloride, and phenol are 2.57°, 4.88°, and 3.60°, respectively. The

observing vapor-pressure lowering and osmotic pressure likewise may be used to calculate molecular weights.

To determine the extent to which an electrolyte is ionized, it is necessary to know its molecular weight, as determined by some other method, and then to measure one of the four colligative properties. The deviation of the results from similar values for nonelectrolytes then is used in calculating the extent of ionization.

Quantitative Treatment of Solubility

The focus of discussion so far has been on the qualitative aspects of solubility. It is, however, important to understand some quantitative relationships that can help pharmaceutical scientists predict the solubility of new-drug entities in various solvents and allow them to choose the best solvent system for a given drug. The observation that structurally similar chemical entities have better solubility in each other is based on the fact that cohesive forces operating in such molecules are of the same order of magnitude. One measure of these cohesive forces is a quantity known as internal pressure (P_i). It is given by¹

$$P_i = \left(\frac{\Delta H_v - RT}{V} \right) \quad (22)$$

where ΔH_v is the heat of vaporization of a substance and V is its molar volume at temperature T . Since Equation 22 contains ΔH_v , which depends on the amount of energy required to break intermolecular (cohesive) bonds, P_i is a measure of cohesive forces among the molecules. This value is high in polar substances; for example, water has a P_i value of 550 cal/mL. Therefore, drugs with high internal pressure show higher solubility in water. The term P_i usually is reserved for solubility of liquids in liquids.

For a quantitative estimate of solubility of solids in liquids, it is assumed that in an ideal solution the heat of solution is equal to the heat of fusion (heat required to melt one mol of solid to liquid without changing its temperature). As ideal solubility does not depend on the nature of solvent, it can be expressed by¹¹

$$-\log X_2^i = \frac{\Delta H_f}{2.303R} \left(\frac{T_0 - T}{T_0 T} \right) \quad (23)$$

where X_2^i is the mol fraction solubility in an ideal solution, ΔH_f is the molar heat of fusion of solute, T_0 is the melting point of solute, and T is the solution temperature such that $T < T_0$.

In a nonideal solution, the mol fraction solubility (X) has to be replaced by thermodynamic activity (a) of the solute. This activity can be expressed in terms of mol fraction solubility as

$$a_2 = X_2 \gamma_2 \quad (24)$$

in which γ_2 is a proportionality constant called the *activity coefficient*. The value of γ_2 in ideal solution is equal to its maximum value of 1. By taking the log of the above equation and substituting in Equation 23, one obtains the equation of non-ideal solubility as

$$-\log X_2 = \frac{\Delta H_f}{2.303R} \left(\frac{T_0 - T}{T_0 T} \right) + \log \gamma_2 \quad (25)$$

It can be seen that when $\gamma_2 = 1$, $\log \gamma_2$ is zero and the equation reduces to the ideal solubility equation.

In general, ideal solutions are rare. Solutions of nonpolar solutes in nonpolar solvents usually come close to being ideal. However, solutions involving polar solutes or solvents almost always show significant deviation from ideality. The value of γ_2 is hard to determine, and varies with concentration of solution. It can be, however, estimated by

$$\log \delta_2 = [(w_{11})^{1/2} - (w_{22})^{1/2}]^2 \frac{V_2 \Phi_1^2}{2.303RT} \quad (26)$$

where w_{11} is the amount of work involved in separating solvent molecules to create space for a solute molecule, w_{22} is the work involved in breaking a solute molecule from its bulk, V_2 is the molar volume of solute at temperature T , Φ_1 is the volume fraction of the solvent, and R is the gas constant. The terms w_{11} and w_{22} are a measure of the internal energy or cohesive forces of the solvent and solute, respectively. It can be seen from Equation 26 that deviation from ideality is high if values of w_{11} and w_{22} are different from each other, or the molar volume of the solute is high. The w terms are also known as the *solubility parameters*, denoted as δ . Thus the equation of nonideal solubility can be written as

$$-\log X_2 = \frac{\Delta H_f}{2.303R} \left(\frac{T_0 - T}{T_0 T} \right) + \frac{V_2 \Phi_1^2}{2.303RT} (\delta_1 - \delta_2)^2 \quad (27)$$

The following observations can be made from Equation 27.

1. For dilute solutions Φ_1 is approximately equal to 1 and thus may be disregarded in estimating solubility in dilute solutions
2. The closer the values of δ_1 and δ_2 , the greater the solubility for a given pair of solute and solvent. In fact, when $\delta_1 = \delta_2$, the equation reduces to the equation for ideal solution, in which case the solubility is at its maximum value and depends only on molar heat of fusion of the solute.
3. Solutions of larger solute molecules (high value of V_2) show higher deviation from ideality. It is not surprising therefore that solutions of polymers and other high-molecular-weight compounds show a very different behavior than ideal solution (see *Solutions of Polymers*, below).

The solubility parameters can be measured using property of the material that involves molecular or cohesive interactions. These include the molar heat of vaporization, surface tension, internal pressure, and several others. One method suggested by Hilderbrand *et al.*¹² is to use the expression for internal pressure to estimate the value of solubility parameter as follows.

$$\delta = \left(\frac{\Delta H_v - RT}{V} \right)^{1/2} \quad (28)$$

The meanings of the symbols are the same as defined earlier.

The values of solubility parameters are available in several references for many commonly used drugs. As intermolecular forces are composed of many kinds of forces, including polar and nonpolar forces, the individual contribution of these forces can be included in quantitative estimate of solubility parameter. Hilderbrand and Scott¹³ suggested Equation 29 for this purpose.

$$\delta^2 = \delta_D^2 + \delta_p^2 + \delta_H^2 \quad (29)$$

where δ_D is the partial solubility parameter arising from nonpolar interactions, δ_p is the partial solubility parameter from polar interactions, and δ_H is the partial solubility parameter from the hydrogen-bonding tendency among the molecules. The value of δ_D is fairly constant for all types of molecules, polar as well as nonpolar, because nonpolar forces operate in all of these molecules. This value ranges from 7 to 10 cal/cc. Because δ_p is due to polar forces, which are essentially absent in nonpolar compounds, its value range is broader, 0 to 13 cal/cc. The value of δ_H , on the other hand, has the highest contribution where present and has a range of 0 to 25 cal/cc. Therefore, for nonpolar compounds such as linear hydrocarbons, the total value of δ is comprised entirely of δ_D , and is close to about 7. For this reason most hydrocarbons show a similar behavior of solubility. In the nonhydrogen-bonding compounds that are relatively polar, δ_p has significant contribution.

Solutions of Polymers

Solubility behavior of polymers is usually significantly different from that of small molecules. Although there is no well-defined value of molecular weight cutoff point between polymers and regular molecules, polymer solutions included in the discussion

here will focus on molecules whose size approaches the colloidal range.

Depending upon the manner in which the monomers are connected to each other, polymers can be of several types. From the solubility standpoint, however, the nature of the monomers is of great significance. In general, the solubility behavior of homopolymers (consisting of monomers repeated N times) mimics the solubility behavior of the monomers. This implies that the homopolymers consisting of relatively hydrophobic monomers will be poorly soluble in water. Examples of such polymers include polystyrene and polyamines.

However, if the hydrophobic monomers form parts of the block polymers (consisting of blocks of one repeating monomer unit followed by a block of different monomer) or heteropolymers (several monomers attached in random manner), their contribution to solubility may not be as negative as one would expect from their structure. This is because polymers are long molecules and generally have the flexibility to fold themselves in a manner that allows their hydrophobic areas to be folded away from water, much the way amphiphiles aggregate to form a hydrophobic core. This arrangement allows the hydrophilic monomers to stay in contact with water, thereby allowing substantial solubility. Examples of such polymers include proteins (which may contain hydrophobic amino acid residues).

Many of the so-called *biological polymers* consist of monomers that carry a net negative or positive charge at near neutral pH. These are known as *polyelectrolytes*, and they are generally very soluble in water. Their solubility is driven by the electrostatic interactions between water and the charged monomers. Examples of such polymers include DNA, proteins, certain derivatized cellulose polymers, and carrageenans. Such polymers are of significant importance in pharmaceutical dosage forms as thickeners, additives, stabilizers, and controlled-release matrices.

Many biological polymers exist as random coil structures in aqueous solution. If the structure is treated as an approximate sphere, then its radius, known as the *radius of gyration* (R_g), is a function of its molecular weight. In polymers of very high molecular weight (typically 100 kd or higher) this radius may be so large that the polymer in solution behaves like a particle, approaching the size of the colloidal range. The volume of this particle is given by¹⁴

$$V_{\text{coil}} = \frac{4}{3} \pi R_g^3 \quad (30)$$

where V_{coil} is the volume of a single polymer chain and R_g is the radius of gyration. When the value of this volume is large, the system no longer behaves as a dilute solution even when the molar concentration is small, and polymer-polymer interactions are significant. Depending on the polymer molecular weight, significant overlapping between the polymer chains may occur at concentration as low as 0.1%.¹⁴ At higher concentration, the swollen polymer and free solvent may occupy comparable volumes in the solution.

Unlike in regular solutions, the solubility of polymers is driven primarily by the entropic changes. Upon mixing a polymer with a solvent, which is generally water in pharmaceutical solutions, two different kinds of entropic effects occur. One is the increase in entropy due to mixing of two molecular species. This effect is small in a dilute solution. The second effect is that the entropy of the polymer configuration increases due to swelling of the molecules and also due to greater flexibility in solution. Based on these entropic changes, Flory^{15,16} derived Equation 31 to describe the overall entropic change (ΔS_{mix}) in a polymer solution.

$$\Delta S_{\text{mix}} = -R (n_s \ln \Phi_s + n_p \ln \Phi_p) \quad (31)$$

where n_s and n_p are the number of molecules of solvent and polymer, respectively, and Φ_s and Φ_p represent their volume fraction, respectively. The free-energy change (ΔG_{mix}) in the process of solubility can be written as

$$\Delta G_{\text{mix}} = RT(n_s \ln \Phi_s + n_p \ln \Phi_p) + (n_s + N_p n_p) w \Phi_p \Phi_s \quad (32)$$

The first term on the right side of Equation 32 is the entropy of mixing, and the second term is the enthalpy of mixing. N_p in the second term is the degree of polymerization, and w is the effective molar interaction parameter (effectively, w is the square of the difference between solubility parameters of the polymer and solvent, multiplied by Avogadro's number). It is clear from the above equation that the value of ΔG_{mix} and therefore the polymer solubility are driven primarily by the volume fraction of the polymer in solution.

METHODS TO INCREASE SOLUBILITY OF POORLY SOLUBLE DRUGS

A large number of promising drug candidates do not make it to the market due to poor bioavailability, due primarily to their poor solubility in aqueous medium. Recently, several strategies have been used to improve solubility profile of these drugs. The strategies used to improve drug solubility include the following.

1. Use of buffers
2. Use of cosolvents
3. Surfactants
4. Complexation
5. Solid dispersions

USE OF BUFFERS—The idea behind use of buffers to improve solubility is to create and maintain pH conditions in a system that cause the drug to be in its ionized state. As discussed previously in this chapter, ionized fraction of a drug is much more soluble in water due to its increased polarity relative to the un-ionized fraction. Buffers can also help in reducing the likelihood of drug precipitation when drug solution is diluted in an aqueous medium. Consistent with the principles of solubility changes with pH, acidic drugs are formulated under relative basic conditions, while the opposite is true for the basic drugs. Some examples of drugs that are formulated with buffer systems are Amikacin sulfate (pH 3.5–5.5, citrate buffer) and Midazolam hydrochloride (pH 3).^{17–19} The drugs that make good candidates for use of pH variation or buffers are the ones that have the ability to ionize within a pH range of 2–8.

USE OF COSOLVENTS—A common way to increase drug solubility is through the use of a water miscible organic solvent. This strategy is based on the fact that poor solubility of drugs in water results due to great difference in polarity of the two components, water being of very high polarity, and the drug having low polarity. Addition of a cosolvent with a polarity value of less than that of water reduces the difference between polarity of the drug and water-cosolvent system, thereby improving solubility. Commonly used cosolvents for this purpose are the hydrogen bonding organic solvents such as ethyl alcohol, propylene glycol and glycerin.

The polarity scale of solvents is defined by a property known as dielectric constant. This value for water is 80, and for ethyl alcohol, propylene glycol and glycerin, it is 24,032 and 42, respectively. Most poorly soluble drugs have dielectric constant values of less than 20. Examples of some parenteral solution that contain cosolvents include Chlordiazepoxide (25% propylene glycol), Diazepam (10% ethyl alcohol and 40% propylene glycol), and digoxin (10% ethyl alcohol and 40% propylene glycol). Non-polar and non-ionizable drugs are good candidates for cosolvent systems.^{17–19}

SURFACTANTS—Surfactants are molecules with well defined polar and non-polar regions that allow them to aggregate in solution to form micelles. Non-polar drugs can partition into these micelles and be solubilized. Depending on the nature of the polar area, surfactants can be non-ionic (eg, polyethylene glycol), anionic (eg, sodium dodecyl sulfate), cationic (eg, tri-alkylammonium) and Zwitterionic (eg, glycine and proteins). Among these, the most commonly used ones are the anionic and non-ionic surfactants. Since the process of solubilization occurs

due to presence of micelles, generally high concentrations of surfactants are needed to significantly improve drug solubility. One example of surfactant based solution is Taxol (paclitaxel), an anti-cancer drug that is solubilized in 50% solution of Cremophor. Other examples include Valrubicin in 50% Cremophor, and Cyclosporin in 65% Cremophor.^{17–19}

COMPLEXATION—Complexation is the association between two or more molecules to form a noncovalent based complex that has higher solubility than the drug itself. From solubility standpoint, complexes can be put into two categories, stacking complexes and inclusion complexes. Stacking complexation is driven by association of nonpolar areas of the drug and complexing agent. This results in exclusion of the nonpolar areas from contact with water, thereby reducing total energy of the system. This aggregation is favored by large planar nonpolar regions on the molecules. Stacking can be homogeneous or mixed, but results in a clear solution.

Inclusion complexes are formed by insertion of drug molecule into a cavity formed by the complexing agent. In this arrangement, nonpolar area of the drug molecule is excluded from water due to its insertion in the complexing agent. One requirement for the complexing agent in such systems is that it has nonpolar core and polar exterior. The most commonly used inclusion complexing molecules are cyclodextrins. The cyclic oligomers of glucose are relatively soluble in water and have cavities large enough to accept nonpolar portions of many drug molecules. Cyclodextrins can consist of 6, 7, or 8 sugar residues and are classified as α , β , and γ , respectively. Due to geometric considerations, steroid molecules lend themselves very well for inclusion into cyclodextrin complexes.

SOLID DISPERSIONS—Solid dispersion refers to the dispersion of one or more active ingredients in an inert carrier or matrix at solid state prepared by the melting (fusion), solvent or the melting-solvent method. It has also been defined as the product formed by converting a fluid drug-carrier combination to the solid state. The term co precipitate or co evaporate has also been used frequently used when a solid dispersion is prepared by solvent method.

Classification of Solid Dispersions—Solid dispersions can be classified as follows:

- Simple eutectic mixtures
- Solid solutions
- Glass solutions of suspensions
- Compound or complex formation between the drug and the carriers
- Amorphous precipitations of drug in crystalline carrier

Simple Eutectic Mixtures—A simple eutectic mixture consists of two compounds that are completely miscible in the liquid state but only to a very limited extent in the solid state. A eutectic mixture of a sparingly water-soluble drug and a highly water-soluble carrier may be regarded thermodynamically as an intimately blended physical mixture of its two crystalline component. These components are assumed to crystallize simultaneously in very small particulate sizes. The increase in specific surface area therefore, is mainly responsible for the increased rate of dissolution of a poorly water-soluble drug.

Differential thermal analysis (DTA) of binary mixtures normally exhibits two endotherms, but a binary mixture of eutectic composition usually exhibits a single major endotherm. In the case of a simple Eutectic system, the thaw points of binary mixtures of varying compositions are equal to the eutectic temperature of the system.

Solid Solutions—Solid solution consists of a solid solute dissolved in a solid solvent. The particle size in solid solution is reduced to molecular level. Successful solubilization of Itraconazole has been achieved using solid solution techniques. Solid solutions of lower drug concentrations generally give faster dissolution rate, and drug dissolution improves considerably with an increase in molecular weight of a water-soluble polymer such as polyethylene glycol.

Glass Solutions of Suspensions—A glass solution is a homogeneous system in which a glassy or a vitreous form of the carrier solubilizes drug molecules. PVP has been used as a carrier in several formulations. In its matrix PVP dissolved an organic solvents undergoes a transition to a glassy state upon evaporation of the solvent.

Compound or Complex Formation Between the Drug and the Carriers—This system is characterized by complexation of two compo-

nents in a binary system during solid dispersion preparation. The availability of a drug from the complex is dependent on the solubility, dissociation constant, and the intrinsic absorption rate of the complex. α , β , and γ CD in combination with polyethylene glycol (PEG) 6000 have been used to formulate such systems.

Amorphous Precipitation—Amorphous precipitation occurs when the drug precipitates as an amorphous form in the inert carrier. The high-energy state of the drug in this system generally produces much greater dissolution rates than the corresponding crystalline forms of the drug.

REFERENCES

1. Martin AN et al. *Physical Pharmacy: Physical Chemical Principles in Pharmaceutical Sciences*. Philadelphia: Lea & Febiger, 1993, pp 212–237.
2. Kramer SF, Flynn GL. *J Pharm Sci* 1972; 61:1896.
3. Campbell AN, Campbell AJR. *J Am Chem Soc* 1937; 59:2481.
4. Loran MR, Guth EP. *J APhA Sci Ed* 1951; 40:465.
5. Grant DJW, Abougela IKA. *Analytical Proceedings: Proceedings of the Analytical Division of the Royal Society of Chemistry*. Dec 1992, p 545.
6. Higuchi T. In: Lyman R, ed. *Pharmaceutical Compounding and Dispensing*. Philadelphia: Lippincott, 1949, p 159.
7. Duddu S. PhD thesis. University of Minnesota, 1993.
8. Yallowky SH. *Techniques of Solubilization of Drugs*. New York: Dekker, 1981, p 91.
9. Martin A et al. *J Pharm Sci* 1982; 71: 849.
10. Setnikar I, Temelcou O. *J APhA Sci Ed* 1959; 48:628.
11. Hilderbrand JH, Wood SE. *J Chem Phys* 1933; 1:817.
12. Hilderbrand JH et al. *Regular and Related Solutions*. New York: Van Nostrand Reinhold, 1970, pp 22–23.
13. Hilderbrand JH, Scott RL. *Solubility of Nonelectrolytes*. New York: Dover, 1964, Chap 23.
14. Evans DF, Wennerstrom H. *The Colloidal Domain—Where Physics, Chemistry, Biology, and Technology Meet*. New York: VCH Publishers, 1994, pp 289–303.
15. Flory PJ. *Principles of Polymer Chemistry*. Ithaca, NY: Cornell University Press, 1953.
16. Flory PJ. *Statistical Mechanics of Chain Molecules*. New York: Interscience, 1969.
17. Strickley RG. *PDA J Pharm Sci Technol* 1999; 53:324.
18. Strickley RG. *PDA J Pharm Sci Technol* 2000; 54:69.
19. Strickley RG. *PDA J Pharm Sci Technol* 2000; 54:152.

Ionic Solutions and Electrolytic Equilibria



ELECTROLYTES

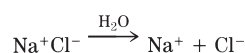
In a preceding chapter, attention was directed to the colligative properties of nonelectrolytes, or substances whose aqueous solutions do not conduct electricity. Substances whose aqueous solutions conduct electricity are known as *electrolytes* and are typified by inorganic acids, bases, and salts. In addition to the property of electrical conductivity, solutions of electrolytes exhibit anomalous colligative properties.

COLLIGATIVE PROPERTIES

In general, for nonelectrolytes, a given colligative property of two equimolar solutions will be identical. This generalization, however, cannot be made for solutions of electrolytes.

Van't Hoff pointed out that the osmotic pressure of a solution of an electrolyte is considerably greater than the osmotic pressure of a solution of a nonelectrolyte of the same molal concentration. This anomaly remained unexplained until 1887 when Arrhenius proposed a hypothesis that forms the basis for our modern theories of electrolyte solutions.

This theory postulated that when electrolytes are dissolved in water they split up into charged particles known as *ions*. Each of these ions carries one or more electrical charges, with the total charge on the positive ions (*cations*) being equal to the total charge on the negative ions (*anions*). Thus, although a solution may contain charged particles, it remains neutral. The increased osmotic pressure of such solutions is due to the increased number of particles formed in the process of ionization. For example, sodium chloride is assumed to dissociate as



It is evident that each molecule of sodium chloride that is dissociated produces two ions, and if dissociation is complete, there will be twice as many particles as would be the case if it were not dissociated at all. Furthermore, if each ion has the same effect on osmotic pressure as a molecule, it might be expected that the osmotic pressure of the solution would be twice that of a solution containing the same molal concentration of a nonionizing solute.

Osmotic-pressure data indicate that, in very dilute solutions of salts that yield two ions, the pressure is very nearly double that of solutions of equimolar concentrations of nonelectrolytes. Similar magnification of vapor-pressure lowering, boiling-point elevation, and freezing-point depression occurs in dilute solutions of electrolytes.

Van't Hoff defined a factor, i , as the ratio of the colligative effect produced by a concentration, m , of electrolyte, divided by the effect observed for the same concentration of nonelectrolyte, or

$$i = \frac{\pi}{(\pi)_0} = \frac{\Delta P}{(\Delta P)_0} = \frac{\Delta T_b}{(\Delta T_b)_0} = \frac{\Delta T_f}{(\Delta T_f)_0} \quad (1)$$

in which π , ΔP , ΔT_b , ΔT_f refer to the osmotic pressure, vapor-pressure lowering, boiling-point elevation, and freezing-point depression, respectively, of the electrolyte. The terms $(\pi)_0$ and so on refer to the nonelectrolyte of the same concentration. In general, with strong electrolytes (those assumed to be 100% ionized), the van't Hoff factor is equal to the number of ions produced when the electrolyte goes into solution (2 for NaCl and MgSO_4 , 3 for CaCl_2 and Na_2SO_4 , 4 for FeCl_3 and Na_3PO_4 , etc.).

In very dilute solutions the osmotic pressure, vapor-pressure lowering, boiling-point elevation, and freezing-point depression of solutions of electrolytes approach values two, three, four, or more times greater (depending on the type of strong electrolyte) than in solutions of the same molality of nonelectrolyte, thus confirming the hypothesis that an ion has the same primary effect as a molecule on colligative properties. It bears repeating, however, that two other effects are observed as the concentration of electrolyte is increased.

The first effect results in less than 2-, 3-, or 4-fold intensification of a colligative property. This reduction is ascribed to interionic attraction between the positive and negatively charged ions, in consequence of which the ions are not dissociated completely from each other and do not exert their full effect on vapor pressure and other colligative properties. This deviation generally increases with increasing concentration of electrolyte.

The second effect intensifies the colligative properties and is attributed to the attraction of ions for solvent molecules (called *solvation*, or, if water is the solvent, *hydration*), which holds the solvent in solution and reduces its escaping tendency, with a consequent enhancement of the vapor-pressure lowering. Solvation also reduces interionic attraction and, thereby, further lowers the vapor pressure.

CONDUCTIVITY

The ability of metals to conduct an electric current results from the mobility of electrons in the metals. This type of conductivity is called *metallic conductance*. On the other hand, various chemical compounds—notably acids, bases, and salts—conduct electricity by virtue of ions present or formed, rather than by

electrons. This is called *electrolytic conductance*, and the conducting compounds are electrolytes. Although the fact that certain electrolytes conduct electricity in the molten state is important, their behavior when dissolved in a solvent, particularly in water, is of greater concern in pharmaceutical science.

The electrical conductivity (or conductance) of a solution of an electrolyte is merely the reciprocal of the resistance of the solution. Therefore, to measure conductivity is actually to measure electrical resistance, commonly with a *Wheatstone bridge apparatus*, and then to *calculate* the conductivity. Figure 17-1 is a representation of the component parts of the apparatus.

The solution to be measured is placed in a glass or quartz cell having two inert electrodes, commonly made of platinum or gold and coated with spongy platinum to absorb gases, across which passes an alternating current generated by an oscillator at a frequency of about 1000 Hz. The reason for using alternating current is to reverse the electrolysis that occurs during flow of current that would cause polarization of the electrodes and lead to abnormal results. The size of the electrodes and their distance apart may be varied to reduce very high resistance or increase very low resistance to increase the accuracy and precision of measurement. Thus, solutions of high conductance (low resistance) are measured in cells having small electrodes relatively far apart, whereas solutions of low conductance (high resistance) are measured in cells with large electrodes placed close to each other.

Electrolytic resistance, like metallic resistance, varies directly with the length of the conducting medium and inversely with its cross-sectional area. The known resistance required for the circuit is provided by a resistance box containing calibrated coils. Balancing of the bridge may be achieved by sliding a contact over a wire of uniform resistance until no (or minimum) current flows through the circuit, as detected either visually with a cathode-ray oscilloscope or audibly with earphones.

The resistance, in ohms, is calculated by the simple procedure used in the Wheatstone bridge method. The reciprocal of the resistance is the conductivity, the units of which are *reciprocal ohms* (also called *mho*). As the numerical value of the conductivity will vary with the dimensions of the conductance cell, the value must be calculated as *specific conductance*, L , which is the conductance in a cell having electrodes of 1-cm² cross-sectional area and 1 cm apart. If the dimensions of the cell used in the experiment were known, calculating the specific conductance would be possible. Nevertheless, this information actually is not required, because calibrating a cell by measuring in it the conductivity of a standard solution of known specific conductance is possible—and much more convenient—and then calculating a *cell constant*. Because this constant is a function

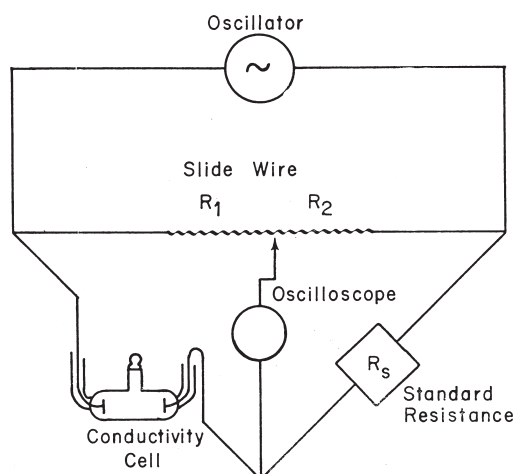


Figure 17-1. Alternating current Wheatstone bridge for measuring conductivity.

Table 17-1. Equivalent Conductances^a at 25°

G-EQ/L	HCL	HOAC	NACL	KCL	NAI	KI	NAOAC
Inf dil	426.1	390.6 ^a	126.5	149.9	126.9	150.3	91.0
0.0005	422.7	67.7	124.5	147.8	125.4	—	89.2
0.0010	421.4	49.2	123.7	146.9	124.3	—	88.5
0.0050	415.8	22.9	120.6	143.5	121.3	144.4	85.7
0.0100	412.0	16.3	118.5	141.3	119.2	142.2	83.8
0.0200	407.2	11.6	115.8	138.3	116.7	139.5	81.2
0.0500	399.1	7.4	111.1	133.4	112.8	135.0	76.9
0.1000	391.3	5.2	106.7	129.0	108.8	131.1	72.8

^a The equivalent conductance at infinite dilution for acetic acid, a weak electrolyte, is obtained by adding the equivalent conductances of hydrochloric acid and sodium acetate and subtracting that of sodium chloride.

only of the dimensions of the cell, it can be used to convert all measurements in that cell to specific conductivity. Solutions of known concentration of pure potassium chloride are used as standard solutions for this purpose.

EQUIVALENT CONDUCTANCE—In studying the variation of conductance of electrolytes with dilution it is essential to make allowance for dilution so that the comparison of conductances may be made for identical amounts of solute. This may be achieved by expressing conductance measurements in terms of *equivalent conductance*, Λ , which is obtained by multiplying the specific conductance, L , by the volume in milliliters, V_e , of a solution containing 1 g-eq of solute. Thus,

$$\Lambda = LV_e = \frac{1000L}{C} \quad (2)$$

where C is the concentration of electrolyte in the solution in g-eq/L, that is, the normality of the solution. For example, the equivalent conductance of 0.01 N potassium chloride solution, which has a specific conductance of 0.001413 mho/cm, may be calculated in either of the following ways:

$$\Lambda = 0.001413 \times 100,000 = 141.3 \text{ mho cm}^2/\text{eq}$$

or

$$\Lambda = \frac{1000 \times 0.001413}{0.01} = 141.3$$

STRONG AND WEAK ELECTROLYTES—Electrolytes are classified broadly as *strong electrolytes* and *weak electrolytes*. The former category includes solutions of strong acids, strong bases, and most salts; the latter includes weak acids and bases, primarily organic acids, amines, and a few salts. The usual criterion for distinguishing between strong and weak electrolytes is the extent of *ionization*. An electrolyte existing entirely or very largely as ions is considered a strong electrolyte, while one that is a mixture of some molecular species along with ions derived from it is a weak electrolyte. For the purposes of this discussion, classification of electrolytes as strong or weak will be based on certain conductance characteristics exhibited in aqueous solution.

The equivalent conductances of some electrolytes, at different concentrations, are given in Table 17-1 and for certain of these electrolytes again in Figure 17-2, where the equivalent conductance is plotted against the square root of concentration. By plotting the data in this manner a linear relationship is observed for strong electrolytes, while a steeply rising curve is noted for weak electrolytes; this difference is a characteristic that distinguishes strong and weak electrolytes. The interpretation of the steep rise in the equivalent conductance of weak electrolytes is that the degree of ionization increases with dilution, becoming complete at infinite dilution.

Interionic interference effects generally have a minor role in the conductivity of weak electrolytes. With strong electrolytes, which are usually completely ionized, the increase in equivalent conductance results not from increased ionization but from

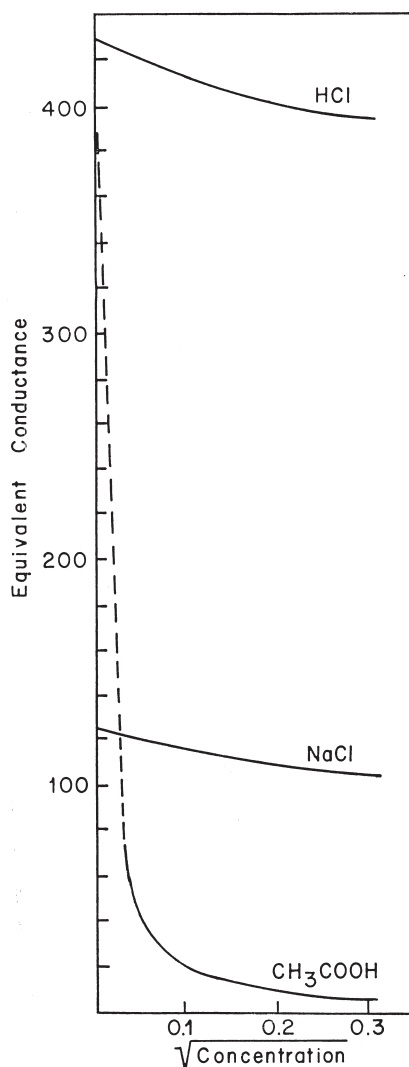


Figure 17-2. Variation of equivalent conductance with square root of concentration.

diminished ionic interference as the solution is diluted, in consequence of which ions have greater freedom of mobility (ie, increased conductance).

The value of the equivalent conductance extrapolated to infinite dilution (zero concentration), designated by the symbol Λ_0 , has special significance. It represents the equivalent conductance of the completely ionized electrolyte when the ions are so far apart that there is no interference with their migration due to interionic interactions. It has been shown, by Kohlrausch, that the equivalent conductance of an electrolyte at infinite dilution is the sum of the equivalent conductances of its component ions at infinite dilution, expressed symbolically as

$$\Lambda_0 = l_0(\text{cation}) + l_0(\text{anion}) \quad (3)$$

The significance of Kohlrausch's law is that each ion, at infinite dilution, has a characteristic value of conductance that is independent of the conductance of the oppositely charged ion with which it is associated. Thus, if the equivalent conductances of various ions are known, the conductance of any electrolyte may be calculated simply by adding the appropriate ionic conductances.

As the fraction of current carried by cations (*transference number* of the cations) and by anions (*transference number* of

anions) in an electrolyte may be determined readily by experiment, ionic conductances are known. Table 17-2 gives the equivalent ionic conductances at infinite dilution of some cations and anions. It is not necessary to have this information to calculate the equivalent conductance of an electrolyte, for Kohlrausch's law permits the latter to be calculated by adding and subtracting values of Λ_0 for appropriate electrolytes. For example, the value of Λ_0 for acetic acid may be calculated as

$$\Lambda_0(\text{CH}_3\text{COOH}) = \Lambda_0(\text{HCl}) + \Lambda_0(\text{CH}_3\text{COONa}) - \Lambda_0(\text{NaCl})$$

which is equivalent to

$$l_0(\text{H}^+) + l_0(\text{CH}_3\text{COO}^-) = l_0(\text{H}^+) + l_0(\text{Cl}^-) + (l_0(\text{Na}^+) + l_0(\text{CH}_3\text{COO}^-) - l_0(\text{Na}^+) - l_0(\text{Cl}^-))$$

This method is especially useful for calculating for weak electrolytes such as acetic acid. As evident from Figure 17-2, the Λ_0 value for acetic acid cannot be determined accurately by extrapolation because of the steep rise of conductance in dilute solutions. For strong electrolytes, on the other hand, the extrapolation can be made very accurately. Thus, in the example above, the values of for HCl, CH_3COONa , and NaCl are determined easily by extrapolation as the substances are strong electrolytes. Substitution of these extrapolated values, as given in Table 17-2, yields a value of 390.6 for the value of Λ_0 for CH_3COOH .

IONIZATION OF WEAK ELECTROLYTES—When Arrhenius introduced his theory of ionization he proposed that the degree of ionization, α , of an electrolyte is measured by the ratio

$$\alpha = \Lambda/\Lambda_0 \quad (4)$$

where Λ is the equivalent conductance of the electrolyte at any specified concentration of solution and Λ_0 is the equivalent conductance at infinite dilution. As strong electrolytes were then not recognized as being 100% ionized, and interionic interference effects had not been evaluated, he believed the equation to be applicable to both strong and weak electrolytes. It now is known that the apparent variation of ionization of strong electrolytes arises from a change in the mobility of ions at different concentrations, rather than from varying ionization, so the equation is not applicable to strong electrolytes. It does provide, however, a generally acceptable approximation of the degree of ionization of weak electrolytes, for which deviations resulting from neglect of activity coefficients and of some change of ionic mobilities with concentration are, for most purposes, negligible. The following example illustrates the use of the equation to calculate the degree of ionization of a typical weak electrolyte.

Example—Calculate the degree of ionization of $1 \times 10^{-3} N$ acetic acid, the equivalent conductance of which is 48.15 mho cm^2/eq . The equivalent conductance at infinite dilution is 390.6 mho cm^2/eq .

$$\alpha = \frac{48.15}{390.6} = 0.12$$

$$\% \text{ ionization} = 100\alpha = 12\%$$

Table 17-2. Equivalent Ionic Conductivities at Infinite Dilution, at 25°

CATIONS	l_0	ANIONS	l_0
H^+	349.8	OH^-	198.0
Li^+	38.7	Cl^-	76.3
Na^+	50.1	Br^-	78.4
K^+	73.5	I^-	76.8
NH_4^+	61.9	AcO^-	40.9
$\frac{1}{2}\text{Ca}^{2+}$	59.5	$\frac{1}{2}\text{SO}_4^{2-}$	79.8
$\frac{1}{2}\text{Mg}^{2+}$	53.0		

The degree of dissociation also can be calculated using the van't Hoff factor, i , and

$$\alpha = \frac{i - 1}{v - 1} \quad (5)$$

where v is the number of ions into which the electrolyte dissociates.

Example—A $1.0 \times 10^{-3} M$ solution of acetic acid has a van't Hoff factor equal to 1.12. Calculate the degree of dissociation of the acid at this concentration.

$$\alpha = \frac{i - 1}{v - 1} = \frac{1.12 - 1}{2 - 1} = 0.12$$

This result agrees with that obtained using equivalent conductance and Equation 4.

MODERN THEORIES

The Arrhenius theory explains why solutions of electrolytes conduct electricity, and why they exhibit enhanced colligative properties. The theory is satisfactory for solutions of weak electrolytes. Several deficiencies, however, do exist when it is applied to solutions of strong electrolytes. It does not explain the failure of strong electrolytes to follow the law of mass action as applied to ionization; discrepancies exist between the degree of ionization calculated from the van't Hoff factor and the conductivity ratio for strong electrolyte solutions having concentrations greater than about $0.5 M$.

These deficiencies can be explained by the following observations

1. In the molten state, strong electrolytes are excellent conductors of electricity. This suggests that these materials are already ionized in the crystalline state. Further support for this is given by x-ray studies of crystals, which indicate that the units comprising the basic lattice structure of strong electrolytes are ions.
2. Arrhenius neglected the fact that ions in solution, being oppositely charged, tend to associate through electrostatic attraction. In solutions of weak electrolytes, the number of ions is not large and it is not surprising that electrostatic attractions do not cause appreciable deviations from theory. In dilute solutions, in which strong electrolytes are assumed to be 100% ionized, the number of ions is large, and interionic attractions become major factors in determining the chemical properties of these solutions. These effects should, and do, become more pronounced as the concentration of electrolyte or the valence of the ions is increased.

It is not surprising, therefore, that the Arrhenius theory of partial ionization involving the law of mass action and neglecting ionic charge does not hold for solutions of strong electrolytes. Neutral molecules of strong electrolytes, if they do exist in solution, must arise from interionic attraction rather than from incomplete ionization.

ACTIVITY AND ACTIVITY COEFFICIENTS—Due to increased electrostatic attractions as a solution becomes more concentrated, the concentration of an ion becomes less efficient as a measure of its net effectiveness. A more efficient measure of the physical or chemical effectiveness of an ion is known as its *activity*, which is a measure of the concentration of an ion related to its concentration at a universally adopted reference-standard state. The relationship between the activity and the concentration of an ion can be expressed as

$$a = m\gamma \quad (6)$$

where m is the molal concentration, γ is the activity coefficient, and a is the activity. The activity also can be expressed in terms of molar concentration, c , as

$$a = fc \quad (7)$$

where f is the activity coefficient on a molar scale. In dilute solutions (below $0.01 M$) the two activity coefficients are identical, for all practical purposes.

The activity coefficient may be determined in various ways, such as measuring colligative properties, electromotive force, solubility, or distribution coefficients. For a strong electrolyte, the mean ionic activity coefficient, γ_{\pm} or f_{\pm} , provides a measure of the deviation of the electrolyte from ideal behavior. The mean ionic activity coefficients on a molal basis for several strong electrolytes are given in Table 17-2. It is characteristic of the electrolytes that the coefficients at first decrease with increasing concentration, pass through a minimum and finally increase with increasing concentration of electrolyte.

IONIC STRENGTH—Ionic strength is a measure of the intensity of the electrical field in a solution and may be expressed as

$$\mu = \frac{1}{2} \sum c_i z_i^2 \quad (8)$$

where z_i is the valence of ion i . The mean ionic activity coefficient is a function of ionic strength as are such diverse phenomena as solubilities of sparingly soluble substances, rates of ionic reactions, effects of salts on pH of buffers, electrophoresis of proteins, and so on.

The greater effectiveness of ions of higher charge on a specific property, compared with the effectiveness of the same number of singly charged ions, generally coincides with the ionic strength calculated by Equation 8. The variation of ionic strength with the valence (charge) of the ions comprising a strong electrolyte should be noted.

For univalent cations and univalent anions (called *uniunivalent* or 1-1) electrolytes, the ionic strength is identical with molarity. For bivalent cation and univalent anion (*biunivalent* or 2-1) electrolytes, or univalent cation and bivalent anion (*unibivalent* or 1-2) electrolytes, the ionic strength is three times the molarity. For bivalent cation and bivalent anion (*bibivalent* or 2-2) electrolytes, the ionic strength is four times the molarity. These relationships are evident from the following example.

Example—Calculate the ionic strength of $0.1 M$ solutions of NaCl, Na_2SO_4 , $MgCl_2$, and $MgSO_4$, respectively, for

$$NaCl \quad \mu = \frac{1}{2} (0.1 \times 1^2 + 0.1 \times 1^2) = 0.1$$

$$Na_2SO_4 \quad \mu = \frac{1}{2} (0.2 \times 1^2 + 0.1 \times 2^2) = 0.3$$

$$MgCl_2 \quad \mu = \frac{1}{2} (0.1 \times 2^2 + 0.2 \times 1^2) = 0.3$$

$$MgSO_4 \quad \mu = \frac{1}{2} (0.1 \times 2^2 + 0.1 \times 2^2) = 0.4$$

The ionic strength of a solution containing more than one electrolyte is the sum of the ionic strengths of the individual salts comprising the solution. For example, the ionic strength of a solution containing NaCl, Na_2SO_4 , $MgCl_2$, and $MgSO_4$, each at a concentration of $0.1 M$, is 1.1.

DEBYE-HUCKEL THEORY—The *Debye-Huckel equations*, which are applicable only to very dilute solutions (about 0.02μ), may be extended to somewhat more concentrated solutions (about 0.1μ) in the simplified form

$$\log f_i = \frac{-0.51 z_i^2 \sqrt{\mu}}{1 + \sqrt{\mu}} \quad (9)$$

The mean ionic activity coefficient for aqueous solutions of electrolytes at 25° can be expressed as

$$\log f_{\pm} = \frac{-0.51 z_+ z_- \sqrt{\mu}}{1 + \sqrt{\mu}} \quad (10)$$

in which z_+ is the valence of the cation and z_- is the valence of the anion. When the ionic strength of the solution becomes high (approximately 0.3 to 0.5), these equations become inadequate and a linear term in μ is added. This is illustrated for the mean ionic activity coefficient,

$$\log f_{\pm} = \frac{-0.51 z_+ z_- \sqrt{\mu}}{1 + \sqrt{\mu}} + K_s \mu \quad (11)$$

Table 17-3. Values of Some Salting-Out Constants for Various Barbiturates at 25°

BARBITURATE	KCL	KBR	NACL	NABR
Amobarbital	0.168	0.095	0.212	0.143
Aprobarbital	0.136	0.062	0.184	0.120
Barbital	0.092	0.042	0.136	0.088
Phenobarbital	0.092	0.034	0.132	0.078
Vinbarbital	0.125	0.036	0.143	0.096

in which K_s is a *salting-out* constant chosen empirically for each salt. This equation is valid for solutions with ionic strength up to approximately 1.

SALTING-OUT EFFECT—The aqueous solubility of a slightly soluble organic substance generally is affected markedly by the addition of an electrolyte. This effect is particularly noticeable when the electrolyte concentration reaches 0.5 *M* or higher. If the aqueous solution of the organic substance has a dielectric constant lower than that of pure water, its solubility is decreased and the substance is *salted-out*. The use of high concentrations of electrolytes, such as ammonium sulfate or sodium sulfate, for the separation of proteins by differential precipitation is perhaps the most striking example of this effect. The aqueous solutions of a few substances such as hydrocyanic acid, glycine, and cystine have a higher dielectric constant than that of pure water, and these substances are *salted-in*. These phenomena can be expressed empirically as

$$\log S = \log S_0 \pm K_s m \quad (12)$$

in which S_0 represents the solubility of the organic substance in pure water and S is the solubility in the electrolyte solution. The slope of the straight line obtained by plotting $\log S$ versus m is positive for salting-in and negative for salting-out. In terms of ionic strength this equation becomes

$$\log S = \log S_0 \pm K'_s \mu \quad (13)$$

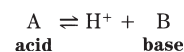
where $K'_s = K_s$ for univalent salts, $K'_s = K_s/3$ for univalent salts, and $K'_s = K_s/4$ for bivalent salts. The salting-out constant depends on the temperature as well as the nature of both the organic substance and the electrolyte. The effect of the electrolyte and the organic substance can be seen in Table 17-3. In all instances, if the anion is constant, the sodium cation has a greater salting-out effect than the potassium cation, probably due to the higher charge density of the former. Although the reasoning is less clear, it appears that, for a constant cation, chloride anion has a greater effect than bromide anion upon the salting-out phenomenon.

ACIDS AND BASES

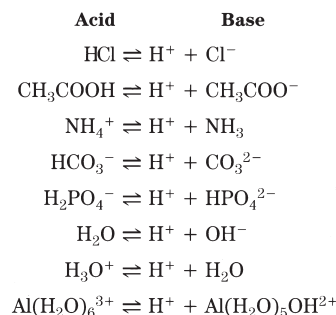
Arrhenius defined an acid as a substance that yields hydrogen ions in aqueous solution and a base as a substance that yields hydroxyl ions in aqueous solution. Except for the fact that hydrogen ions neutralize hydroxyl ions to form water, no complementary relationship between acids and bases (eg, that between oxidants and reductants) is evident in Arrhenius' definitions for these substances; rather, their oppositeness of character is emphasized. Moreover, no account is taken of the behavior of acids and bases in nonaqueous solvents. Also, although acidity is associated with so elementary a particle as the proton (hydrogen ion), basicity is attributed to so relatively complex an association of atoms as the hydroxyl ion. It would seem that a simpler concept of a base could be devised.

PROTON CONCEPT—In pondering the objections to Arrhenius' definitions, Brønsted and Bjerrum in Denmark and Lowry in England developed, and in 1923 announced, a more satisfactory, and more general, theory of acids and bases. According to this theory, an acid is a substance capable of yielding a proton (hydrogen ion), whereas a base is a substance capable

of accepting a proton. This complementary relationship may be expressed by

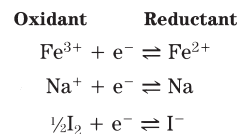


The pair of substances thus related through mutual ability to gain or lose a proton is called a conjugate acid–base pair. Specific examples of such pairs are

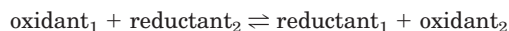


It is apparent that not only molecules, but also cations and anions, may function as acids or bases.

The complementary nature of the acid–base pairs listed is reminiscent of the complementary relationship of pairs of oxidants and reductants where, however, the ability to gain or lose one or more electrons—rather than protons—is the distinguishing characteristic.

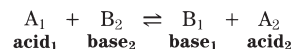


However, these examples of acid–base pairs and oxidant–reductant pairs represent reactions that are possible in principle only. Ordinarily acids will not release free protons any more than reductants will release free electrons. That is, protons and electrons, respectively, can be transferred only from one substance (an ion, atom, or molecule) to another. Thus, it is a fundamental fact of chemistry that oxidation of one substance will occur only if reduction of another substance occurs simultaneously. Stated in another way, electrons will be released from the reductant (oxidation) only if an oxidant capable of accepting electrons (reduction) is present. For this reason oxidation–reduction reactions must involve two conjugate oxidant–reductant pairs of substances:



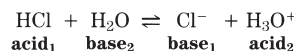
where Subscript 1 represents one conjugate oxidant–reductant pair and Subscript 2 represents the other.

Similarly, an acid will not release a proton unless a base capable of accepting it is present simultaneously. This means that any actual manifestation of acid–base behavior must involve interaction between two sets of conjugate acid–base pairs, represented as



In such a reaction, which is called *protolysis* or a *protolytic reaction*, A₁ and B₁ constitute one conjugate acid–base pair, and A₂ and B₂ the other; the proton given up by A₁ (which thereby becomes B₁) is transferred to B₂ (which becomes A₂).

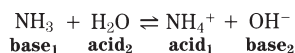
When an acid, such as hydrochloric, is dissolved in water, a *protolytic reaction* occurs.



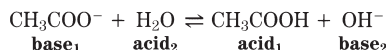
The ionic species H₃O⁺, called *hydronium* or *oxonium* ion, always is formed when an acid is dissolved in water. Often, for purposes of convenience, this is written simply as H⁺ and is

called hydrogen ion, although the "bare" ion practically is nonexistent in solution.

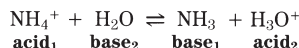
When a base (eg, ammonia) is dissolved in water, the reaction of protolysis is



The proton theory of acid–base function makes the concept of hydrolysis superfluous. When, for example, sodium acetate is dissolved in water, this acid–base interaction occurs



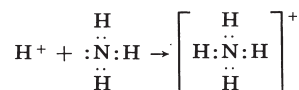
In an aqueous solution of ammonium chloride the reaction is



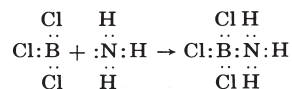
Transfer of protons (protolysis) is not limited to dissimilar conjugate acid–base pairs. In the preceding examples H₂O sometimes behaves as an acid and at other times as a base. Such an amphoteric substance is called, in Brønsted's terminology, an *amphiprotic substance*.

ELECTRON-PAIR CONCEPT—The proton concept of acids and bases provides a more general definition for these substances, but it does not indicate the basic reason for proton transfer, nor does it explain how such substances as sulfur trioxide, boron trichloride, stannic chloride, or carbon dioxide—none of which is capable of donating a proton—can behave as acids. Both deficiencies of the proton theory are avoided in the more inclusive definition of acids and bases proposed by Lewis in 1923. In 1916 he proposed that sharing of a pair of electrons by two atoms established a bond (covalent) between the atoms; therefore, an acid is a substance capable of sharing a pair of electrons made available by another substance called a base, thereby forming a *coordinate covalent bond*. The base is the substance that donates a share in its electron pair to the acid.

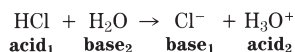
The following equation illustrates how Lewis' definitions explain the transfer of a proton (hydrogen ion) to ammonia to form ammonium ion.



The reaction of boron trichloride, which according to the Lewis theory is an acid, with ammonia is similar, for the boron lacks an electron pair if it is to attain a stable octet configuration, while ammonia has a pair of electrons that may be shared, thus,



LEVELING EFFECT OF A SOLVENT—When the strong acids such as HClO₄, H₂SO₄, HCl, or HNO₃ are dissolved in water, the solutions—if they are of identical normality and are not too concentrated—all have about the same hydrogen-ion concentration, indicating the acids to be of about the same strength. The reason for this is that each one of the acids undergoes practically complete protolysis in water.



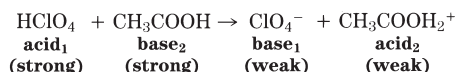
This phenomenon, called the *leveling effect of water*, occurs whenever the added acid is stronger than the hydronium ion. Such a reaction manifests the tendency of proton-transfer reactions to proceed spontaneously in the direction of forming a weaker acid or weaker base.

Since the strongest acid that can exist in an amphiprotic solvent is the conjugate acid form of the solvent, any stronger acid will undergo protolysis to the weaker solvent acid. HClO₄, H₂SO₄, HCl, or HNO₃ are all stronger acids than the

hydronium ion, so they are converted in water to the hydronium ion.

When the strong bases sodium hydride, sodium amide, or sodium ethoxide are dissolved in water, each reacts with water to form sodium hydroxide. These reactions illustrate the leveling effect of water on bases. Because the hydroxide ion is the strongest base that can exist in water, any base stronger than the hydroxide ion undergoes protolysis to hydroxide.

Intrinsic differences in the acidity of acids become evident if they are dissolved in a relatively poor proton acceptor such as anhydrous acetic acid. Perchloric acid (HClO₄), a strong acid, undergoes practically complete reaction with acetic acid to produce the *acetonium ion* (acid₂):



but sulfuric acid and hydrochloric acid behave as weak acids. It is because perchloric acid is a very strong acid when dissolved in glacial acetic acid that it has found many important applications in analytical chemistry as a titrant for a variety of substances that behave as bases in acetic acid. Because of its ability to differentiate the acidity of various acids, it is called a *differentiating solvent for acids*; this property results from its relatively weak proton-acceptor tendency. A solvent that differentiates basicity of different bases must have a weak proton-donor tendency; it is called a *differentiating solvent for bases*. Liquid ammonia is typical of solvents in this category.

Solvents that have both weak proton-donor and proton-acceptor tendencies are called *aprotic solvents* and may serve as differentiating solvents for both acids and bases; they have little if any action on solutes and serve mainly as inert dispersion media for the solutes. Useful aprotic solvents are benzene, toluene, or hexane.

IONIZATION OF ACIDS AND BASES—Acids and bases commonly are classified as strong or weak acids and strong or weak bases depending on whether they are ionized extensively or slightly in aqueous solutions. If, for example, 1 N aqueous solutions of hydrochloric acid and acetic acid are compared, it is found that the former is a better conductor of electricity, reacts much more readily with metals, catalyzes certain reactions more efficiently, and possesses a more acid taste than the latter. Both solutions, however, will neutralize identical amounts of alkali. A similar comparison of 1 N solutions of sodium hydroxide and ammonia reveals the former to be more *active* than the latter, although both solutions will neutralize identical quantities of acid.

The differences in the properties of the two acids is attributed to differences in the concentration of hydrogen (more accurately hydronium) ion, the hydrochloric acid being ionized to a greater extent and thus containing a higher concentration of hydrogen ion than acetic acid. Similarly, most of the differences between the sodium hydroxide and ammonia solutions are attributed to the higher hydroxyl-ion concentration in the former.

The ionization of incompletely ionized acids may be considered a reversible reaction of the type

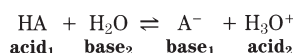


where HA is the molecular acid and A[−] is its anion. An equilibrium expression based on the law of mass action may be applied to the reaction

$$K_a = \frac{[\text{H}^+][\text{A}^-]}{[\text{HA}]} \quad (14)$$

where K_a is the ionization or dissociation constant, and the brackets signify concentration. For any given acid in any specified solvent and at any constant temperature, K_a remains relatively constant as the concentration of acid is varied, provided the acid is weakly ionized. With increasingly stronger acids, however, progressively larger deviations occur.

Although the strength of an acid commonly is measured in terms of the ionization or dissociation constant defined in Equation 14, the process of ionization probably is never as simple as shown above. A proton simply will not detach itself from one molecule unless it is accepted simultaneously by another molecule. When an acid is dissolved in water, the latter acts as a base, accepting a proton (Brønsted's definition of a base) by donating a share in a pair of electrons (Lewis' definition of a base). This reaction may be written as



Application of the law of mass action to this reaction gives

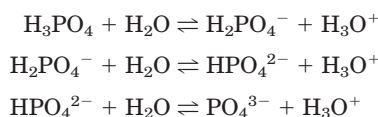
$$K = \frac{[\text{H}_3\text{O}^+][\text{A}^-]}{[\text{HA}][\text{H}_2\text{O}]} \quad (15)$$

Because $[\text{H}_2\text{O}]$ is a constant, this equation may be written

$$K_a = \frac{[\text{H}_3\text{O}^+][\text{A}^-]}{[\text{HA}]} \quad (16)$$

This equation is identical with Equation 14 because $[\text{H}_3\text{O}^+]$ is numerically equal to $[\text{H}^+]$.

Acids that are capable of donating more than one proton are termed *polyprotic*. The ionization of a polyprotic acid occurs in stages and can be illustrated by considering the equilibria involved in the ionization of phosphoric acid:



Application of the law of mass action to this series of reactions gives

$$K_1 = \frac{[\text{H}_2\text{PO}_4^-][\text{H}_3\text{O}^+]}{[\text{H}_3\text{PO}_4]} \quad (17)$$

$$K_2 = \frac{[\text{HPO}_4^{2-}][\text{H}_3\text{O}^+]}{[\text{H}_2\text{PO}_4^-]} \quad (18)$$

$$K_3 = \frac{[\text{PO}_4^{3-}][\text{H}_3\text{O}^+]}{[\text{HPO}_4^{2-}]} \quad (19)$$

If the three expressions for the ionization constants are multiplied together, an overall ionization, K , can be obtained

$$K = K_1 K_2 K_3 = \frac{[\text{PO}_4^{3-}][\text{H}_3\text{O}^+]^3}{[\text{H}_3\text{PO}_4]} \quad (20)$$

Each of the successive ionizations is suppressed by the hydronium ion formed from preceding stages according to Le Chatelier's principle. The successive dissociation constants always decrease in value, as successive protons must be removed from species that always are charged more negatively. This can be seen from the data in Table 17-4, in which K_1 for phosphoric acid is approximately 100,000 times greater than K_2 , which is in turn approximately 100,000 times greater than K_3 . Although successive dissociation constants are always smaller, the difference is not always as great as it is for phosphoric acid. Tartaric acid, for example, has $K_1 = 9.12 \times 10^{-4}$ and $K_2 = 4.27 \times 10^{-5}$.

Ionization of a base can be illustrated by using the specific substance NH_3 for an example. According to Brønsted and Lewis, when the base NH_3 is dissolved in water, the latter acts as an acid, donating a proton to NH_3 , which accepts it by offering a share in a pair of electrons on the nitrogen atom. This reaction is written

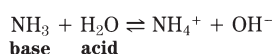


Table 17-4. Dissociation Constants in Water at 25°

SUBSTANCE		K
Weak acids		
Acetic		1.75×10^{-3}
Acetylsalicylic		3.27×10^{-4}
Barbital		1.23×10^{-8}
Barbituric		1.05×10^{-4}
Benzoic		6.30×10^{-5}
Benzyl penicillin		1.74×10^{-3}
Boric	K_1	5.8×10^{-10}
Caffeine		1×10^{-14}
Carbonic	K_1	4.31×10^{-7}
	K_2	4.7×10^{-11}
Citric (1H ₂ O)	K_1	7.0×10^{-4}
	K_2	1.8×10^{-5}
	K_3	4.0×10^{-7}
Dichloroacetic		5×10^{-2}
Ethylenediaminetetraacetic acid (EDTA)	K_1	1×10^{-2}
	K_2	2.14×10^{-3}
	K_3	6.92×10^{-7}
	K_4	5.5×10^{-11}
Formic		1.77×10^{-4}
Glycerophosphoric	K_1	3.4×10^{-2}
	K_2	6.4×10^{-7}
Glycine	K_1	4.5×10^{-3}
	K_2	1.7×10^{-10}
Lactic		1.39×10^{-4}
Mandelic		4.29×10^{-4}
Monochloroacetic		1.4×10^{-3}
Oxalic (2H ₂ O)	K_1	5.5×10^{-2}
	K_2	5.3×10^{-5}
Phenobarbital		3.9×10^{-8}
Phenol		1×10^{-10}
Phosphoric	K_1	7.5×10^{-3}
	K_2	6.2×10^{-8}
	K_3	2.1×10^{-13}
Picric		4.2×10^{-1}
Propionic		1.34×10^{-5}
Saccharin		2.5×10^{-2}
Salicylic		1.06×10^{-3}
Succinic	K_1	6.4×10^{-5}
	K_2	2.3×10^{-6}
Sulfadiazine		3.3×10^{-7}
Sulfamerazine		8.7×10^{-8}
Sulfapyridine		3.6×10^{-9}
Sulfathiazole		7.6×10^{-8}
Tartaric	K_1	9.6×10^{-4}
	K_2	4.4×10^{-5}
Trichloroacetic		1.3×10^{-1}
Weak bases		
Acetanilide		4.1×10^{-14} (40°)
Ammonia		1.74×10^{-5}
Apomorphine		1.0×10^{-7}
Atropine		4.5×10^{-5}
Benzocaine		6.0×10^{-12}
Caffeine		4.1×10^{-14} (40°)
Cocaine		2.6×10^{-6}
Codeine		9×10^{-7}
Ephedrine		2.3×10^{-5}
Morphine		7.4×10^{-7}
Papaverine		8×10^{-9}
Physostigmine	K_1	7.6×10^{-7}
	K_2	5.7×10^{-13}
Pilocarpine	K_1	7×10^{-8}
	K_2	2×10^{-13}
Procaine		7×10^{-6}
Pyridine		1.4×10^{-9}
Quinine	K_1	1.0×10^{-6}
	K_2	1.3×10^{-10}
Reserpine		4×10^{-8}
Strychnine	K_1	1×10^{-6}
	K_2	2×10^{-12}
Theobromine		4.8×10^{-14} (40°)
Thiourea		1.1×10^{-15}
Urea		1.5×10^{-14}

The equilibrium expression for this reaction is

$$K = \frac{[\text{NH}_4^+][\text{OH}^-]}{[\text{NH}_3][\text{H}_2\text{O}]} \quad (21)$$

With $[\text{H}_2\text{O}]$ constant, this expression may be written

$$K_b = \frac{[\text{NH}_4^+][\text{OH}^-]}{[\text{NH}_3]} \quad (22)$$

IONIZATION OF WATER—Although it is a poor conductor of electricity, pure water does ionize through a process known as *autoprotolysis*, in the following manner:



Application of the law of mass action to this reaction gives

$$K = \frac{[\text{H}_3\text{O}^+][\text{OH}^-]}{[\text{H}_2\text{O}]^2} \quad (23)$$

where K is the equilibrium constant for the reaction. Because the concentration of H_2O (molecular water) is very much greater than either the hydronium-ion or hydroxyl-ion concentrations, it can be considered to be constant and can be combined with K to give a new constant, K_w , known as the *ion product* of water, and Equation 23 becomes

$$K_w = [\text{H}_3\text{O}^+][\text{OH}^-] \quad (24)$$

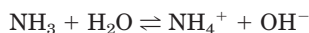
The numerical value of K_w varies with temperature; at 25° it is approximately equal to 1×10^{-14} .

Since the autoprotolysis of pure water yields one hydronium ion for each hydroxyl ion produced, $[\text{H}_3\text{O}^+]$ equal to $[\text{OH}^-]$. At 25° each has a value of 1×10^{-7} mol/L ($1 \times 10^{-7} \times 1 \times 10^{-7} = K_w = 1 \times 10^{-14}$). A solution in which $[\text{H}_3\text{O}^+]$ is equal to $[\text{OH}^-]$ is termed a *neutral* solution.

If an acid is added to water, the hydronium-ion concentration will be increased and the equilibrium between hydronium and hydroxyl ions will be disturbed *momentarily*. To restore equilibrium, some of the hydroxyl ions, originally present in the water, will combine with a *part* of the added hydronium ions to form nonionized water molecules, until the product of the concentrations of the two ions has been reduced to 10^{-14} . When equilibrium again is restored, the concentrations of the two ions no longer will be equal. If, for example, the hydronium-ion concentration is 1×10^{-3} N when equilibrium is established, the concentration of hydroxyl ion will be 1×10^{-11} (the product of the two concentrations being equal to 10^{-14}). As $[\text{H}_3\text{O}^+]$ is much greater than $[\text{OH}^-]$, the solution is said to be *acid* or *acidic*.

In a similar manner, the addition of an alkali to pure water momentarily disturbs the equilibrium between hydronium and hydroxyl ions. To restore equilibrium, some of the hydronium ions originally present in the water will combine with part of the added hydroxyl ions to form nonionized water molecules. The process continues until the product of the hydronium and hydroxyl ion concentrations again is equal to 10^{-14} . Assuming that the final hydroxyl-ion concentration is 1×10^{-4} N, the concentration of hydronium ion in the solution will be 1×10^{-10} . Because $[\text{OH}^-]$ is much greater than $[\text{H}_3\text{O}^+]$, the solution is said to be *basic* or *alkaline*.

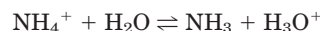
RELATIONSHIP OF K_A AND K_B —A particularly interesting and useful relationship between the strength of an acid and its conjugate base, or a base and its conjugate acid, exists. For illustration, consider the strength of the base NH_3 and its conjugate acid NH_4^+ in water. The behavior of NH_3 as a base is expressed by



for which the equilibrium, as formulated earlier, is

$$K_b = \frac{[\text{NH}_4^+][\text{OH}^-]}{[\text{NH}_3]} \quad (25)$$

The behavior of NH_4^+ as an acid is represented by



The equilibrium constant for this is

$$K_a = \frac{[\text{NH}_3][\text{H}_3\text{O}^+]}{[\text{NH}_4^+]} \quad (26)$$

Multiplying Equations 25 and 26

$$K_a K_b = \frac{[\text{NH}_3][\text{H}_3\text{O}^+][\text{NH}_4^+][\text{OH}^-]}{[\text{NH}_4^+][\text{NH}_3]} \quad (27)$$

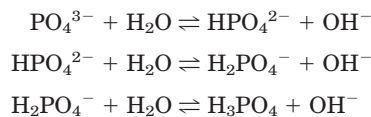
It is obvious that

$$K_w = K_a K_b \quad (28)$$

where K_w is the ion product of water as defined in Equation 24.

The utility of this relationship, which is a general one for any conjugate acid–base pair, is evident from the following deductions: (1) The strength of an acid may be expressed in terms either of the K_a or the K_b of its conjugate base, or *vice versa*; (2) the K_a of an acid may be calculated if the K_b of its conjugate base is known, or *vice versa*; and (3) the stronger an acid is, the weaker its conjugate base, or *vice versa*.

Bases that are capable of interacting with more than one proton are termed *polyacidic*, and can be illustrated by



Applying the law of mass action to this series of reactions, and using the concepts outlined in Equations 25 to 28, the relationship between the various K_a and K_b values for phosphoric acid are

$$K_w = K_{a1} \times K_{b3} = K_{a2} \times K_{b2} = K_{a3} \times K_{b1} \quad (29)$$

where K_{a1} , K_{a2} , and K_{a3} refer to the equilibria given by Equations 17, 18, and 19, respectively; K_{b1} , K_{b2} , and K_{b3} refer to the reaction of PO_4^{3-} , HPO_4^{2-} , and H_2PO_4^- , respectively, with water.

ELECTRONEGATIVITY AND DISSOCIATION CONSTANTS—Table 17-4 gives the dissociation constants of several weak acids and weak bases, in water, at 25° . Strong acids and strong bases do not obey the law of mass action, so dissociation constants cannot be formulated for these strong electrolytes.

Table 17-4 shows that great variations occur in the strength of weak acids and weak bases. The effect of various substituents on the strength of acids and bases depends on the electronegativity of the substituent atom or radical. For example, the substitution of one chlorine atom into the molecule of acetic acid increases the degree of ionization of the acid. Substitution of two chlorine atoms further increases the degree of ionization, and introduction of three chlorine atoms produces a still stronger acid. Acetic acid ionizes primarily because the oxygen atom adjacent to the hydrogen atom of the carboxyl group has a stronger affinity for electrons than the hydrogen atom. Thus, when acetic acid is dissolved in water, the polar molecules of the water have a stronger affinity for the hydrogen of acetic acid than the hydrogen atoms of water. The acetic acid ionizes as a consequence of this difference in affinities.

When an atom of chlorine is introduced into the acetic acid molecule, forming ClCH_2COOH , the electrons in the molecule are attracted very strongly to the chlorine because of its relatively high electronegativity; the bond between the hydrogen and the oxygen in the carboxyl group is thereby weakened,

and the degree of ionization increased. Introduction of two or three chlorine atoms weakens the bond further and increases the strength of the acid. On the other hand, substitution of chlorine into the molecule of ammonia reduces the strength of the base because of its decreased affinity for the hydrogen ion.

IONIC STRENGTH AND DISSOCIATION CONSTANTS—Most solutions of pharmaceutical interest are in a concentration range such that the ionic strength of the solution may have a marked effect on ionic equilibria and observed dissociation constants. One method of correcting dissociation constants for solutions with an ionic strength up to about 0.3 is to calculate an apparent dissociation constant, pK'_a , as

$$pK'_a = pK_a + \frac{0.51(2Z - 1)\sqrt{\mu}}{1 + \sqrt{\mu}} \quad (30)$$

in which pK_a is the tabulated thermodynamic dissociation constant, Z is the charge on the acid, and μ is the ionic strength.

Example—Calculate pK'_2 for succinic acid at an ionic strength of 0.1. Assume that pK_2 is 5.63. The charge on the acid species is -1 .

$$\begin{aligned} pK'_2 &= 5.63 - \frac{0.51(-2-1)\sqrt{0.1}}{1 + \sqrt{0.1}} \\ &= 5.63 - 0.37 = 5.26 \end{aligned}$$

DETERMINATION OF DISSOCIATION CONSTANTS—Although the dissociation constant of a weak acid or base can be obtained in a wide variety of ways including conductivity measurements, absorption spectrometry and partition coefficients, the most widely used method is potentiometric pH measurement (see *Potentiometry*). The simplest method involving potentiometric pH measurement is based on the measurement of the hydronium-ion concentration of a solution containing equimolar concentrations of the acid and a strong-base salt of the acid. The principle of this method is evident from an inspection of Equation 16; when equimolar concentrations of HA (the acid) and A^- (the salt) are present, the dissociation constant, K_a , numerically is equal to the hydronium-ion concentration (also, the pK_a of the acid is equal to the pH of the solution). Although this method is simple and rapid, the dissociation constant obtained is not sufficiently accurate for many purposes.

To obtain the dissociation constant of a weak acid with a high degree of accuracy and precision, a dilute solution of the acid (about 10^{-3} to 10^{-4} M) is titrated with a strong base, and the pH of the solution taken after each addition of base. The resulting data can be handled in a wide variety of ways, perhaps the best of which is the method proposed by Benet and Goyan.¹ The proton balance equation for a weak acid, HA, being titrated with a strong base such as KOH, would be

$$[K^+] + [H_3O^+] = [OH^-] + [A^-] \quad (31)$$

in which $[K^+]$ is the concentration of the base added. Equation 31 can be rearranged to give

$$Z = [A^-] = [K^+] + [H_3O^+] - [OH^-] \quad (32)$$

When a weak monoprotic acid is added to water, it can exist in the unionized form, HA, and in the ionized form, A^- . After equilibrium is established, the sum of the concentrations of both species must be equal to C_a , the stoichiometric (added) concentration of acid, or

$$C_a = [HA] + [A^-] = [HA] + Z \quad (33)$$

The term $[HA]$ can be replaced using Equation 16 to give

$$C_a = \frac{[H_3O^+]Z}{K_a} + Z \quad (34)$$

which can be rearranged to

$$Z = C_a - \frac{Z[H_3O^+]}{K_a} \quad (35)$$

According to Equation 35, if Z , which is obtained from the experimental data using Equation 32, is plotted versus the terms $Z[H_3O^+]$, a straight line results with a slope equal to $1/K_a$, and an intercept equal to C_a . In addition to obtaining an accurate estimate for the dissociation constant, the stoichiometric concentration of the substance being titrated is also obtained. This is of importance when the substance being titrated cannot be purified, or has an unknown degree of solvation. Similar equations can be developed for obtaining the dissociation constant for a weak base.¹

The dissociation constants for diprotic acids can be obtained by defining P as the average number of protons dissociated per mole of acid, or

$$P = Z/C_a \quad (36)$$

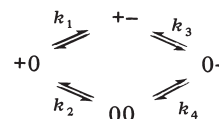
and

$$\frac{[H_3O^+]^2 P}{(2 - P)} = K_1 K_2 + \frac{K_1 [H_3O^+](1 - P)}{(2 - P)} \quad (37)$$

A plot of Equation 37 should yield a straight line with a slope equal to K_1 and an intercept of $K_1 K_2$. Dividing the intercept by the slope yields K_2 .

MICRO DISSOCIATION CONSTANTS—The dissociation constants for polyprotic acids, as determined by potentiometric titration, are known generally as *macro*, or *titration*, constants. As it is known that carboxyl groups are stronger acids than protonated amino groups, there is no difficulty in assigning K_1 and K_2 , as determined by Equation 37, to the carboxyl and amino groups, respectively, of a substance such as glycine hydrochloride.

In other chemicals or drugs such as phenylpropanolamine, in which the two acidic groups are the phenolic and the protonated amino group, the assignment of dissociation constants is more difficult. This is because, in general, both groups have dissociation constants of equal magnitude. Thus, there will be two ways of losing the first proton and two ways of losing the second, resulting in four possible species in solution. This can be illustrated using the convention of assigning a plus (+) to a positively charged group, a 0 to an uncharged group, and a minus (−) to a negatively charged group. Thus, +0 would represent the fully protonated phenylpropanolamine, +− the dipolar ion, 00 the uncharged molecule, and 0−, the anion. The total ionization scheme, therefore, can be written



The micro constants are related to the macro constants as

$$K_1 = k_1 + k_2 \quad (38)$$

$$K_1 K_2 = k_1 k_3 = k_2 k_4 \quad (39)$$

It can be seen from Equation 38 that unless k_1 or k_2 is very much smaller than the other, the observed macro constant is a composite of the two and cannot be assigned to one or the other acidic group in a nonambiguous way.

Methods for determining k_1 are given by Riegelman et al.² and Niebergall et al.³ Once k_1 , K_1 , and K_2 have been determined, all of the other micro constants can be obtained from Equations 38 and 39.

pH

The numerical values of hydronium-ion concentration may vary enormously; for a normal solution of a strong acid the value is nearly 1, while for a normal solution of a strong base it is approximately 1×10^{-14} ; there is a variation of 100,000,000,000,000 between these two limits. Because of the inconvenience of dealing with such large numbers, in 1909 Sørensen proposed that hydronium-ion concentration be expressed in terms of the logarithm (log) of its reciprocal. To this value he assigned the symbol pH. Mathematically it is written

$$\text{pH} = \log \frac{1}{[\text{H}_3\text{O}^+]} \quad (40)$$

Since the logarithm of 1 is zero, the equation also may be written

$$\text{pH} = -\log [\text{H}_3\text{O}^+] \quad (41)$$

from which it is evident that pH also may be defined as the negative logarithm of the hydronium-ion concentration. In general, this type of notation is used to indicate the negative logarithm of the term that is preceded by the p , which gives rise to the following

$$\text{pOH} = -\log [\text{OH}^-] \quad (42)$$

$$\text{p}K = -\log K \quad (43)$$

Thus, taking logarithms of Equations 28 and 24 gives

$$\text{p}K_a + \text{p}K_b = \text{p}K_w \quad (44)$$

$$\text{pH} + \text{pOH} = \text{p}K_w \quad (45)$$

The relationship of pH to hydronium-ion and hydroxyl-ion concentrations may be seen in Table 17-5.

The following examples illustrate the conversion from exponential to p notation.

1. Calculate the pH corresponding to a hydronium-ion concentration of 1×10^{-4} g-ion/L.

Solution:

$$\begin{aligned} \text{pH} &= \log \frac{1}{1 \times 10^{-4}} \\ &= \log 10,000 \text{ or } \log (1 \times 10^4) \end{aligned}$$

$$\log (1 \times 10^4) = +4$$

$$\text{pH} = 4$$

2. Calculate the pH corresponding to a hydronium ion-concentration of 0.000036 N (or g-ion/L). (*Note:* This more frequently is written as a number multiplied by a power of 10, thus, 3.6×10^{-5} for 0.000036.)

Solution:

$$\begin{aligned} \text{pH} &= \log \frac{1}{3.6 \times 10^{-5}} \\ &= \log 28,000 \text{ or } \log (2.8 \times 10^4) \end{aligned}$$

$$\log (2.8 \times 10^4) = \log 2.8 + 10^4$$

$$\log 2.8 = +0.44$$

$$\log 10^4 = +4.00$$

$$\text{pH} = 4.44$$

This problem also may be solved as follows:

$$\text{pH} = -\log (3.6 \times 10^{-5})$$

$$\log 3.6 = +0.56$$

$$\log 10^{-5} = -5.00$$

$$= -4.44 = \log (3.6 \times 10^{-5})$$

$$\text{pH} = -(-4.44) = +4.44 = 4.44$$

The following examples illustrate the conversion of p notation to exponential notation.

1. Calculate the hydronium-ion concentration corresponding to a pH of 4.44.

Solution:

$$\text{pH} = \log \frac{1}{[\text{H}_3\text{O}^+]}$$

$$4.44 = \log \frac{1}{[\text{H}_3\text{O}^+]}$$

$$\frac{1}{[\text{H}_3\text{O}^+]} = \text{antilog of } 4.44 = 28,000 \text{ (rounded off)}$$

$$[\text{H}_3\text{O}^+] = \frac{1}{28,000} = 0.000036 \text{ or } 3.6 \times 10^{-5}$$

This calculation also may be made as

$$+4.44 = -\log [\text{H}_3\text{O}^+]$$

or

$$-4.44 = +\log [\text{H}_3\text{O}^+]$$

In finding the antilog of -4.44 it should be kept in mind that the *mantissa* (the number to the right of the decimal point) of a log to the base 10 (the common or Briggsian logarithm base) is *always positive* but that the characteristic (the number to the left of the decimal point) may be *positive or negative*. As the entire log -4.44 is negative, it is obvious that one cannot look up the antilog of -0.44 . However, the number -4.44 also may be written $(-5.00 + 0.56)$, or as more often written, $\bar{5}.56$; the bar across the characteristic indicates that it alone is negative, while the rest of the number is positive. Looking up the antilog of 0.56 it is found to be 3.6; as the antilog of -5.00 is 10^{-5} , it follows that the hydronium-ion concentration must be 3.6×10^{-5} mols/L.

2. Calculate the hydronium-ion concentration corresponding to a pH of 10.17.

Solution:

$$10.17 = -\log[\text{H}_3\text{O}^+]$$

$$-10.17 = \log[\text{H}_3\text{O}^+]$$

$$-10.17 = (-11.00 + 0.83) = \bar{11}.83$$

The antilog of 0.83 = 6.8.

The antilog of $-11.00 = 10^{-11}$

The hydronium-ion concentration is therefore 6.8×10^{-11} mol/L.

Table 17-5. Hydronium-Ion and Hydroxyl-Ion Concentrations

	PH	NORMALITY IN TERMS OF HYDRONIUM ION	NORMALITY IN TERMS OF HYDROXYL ION
	0	1	10^{-14}
	1	10^{-1}	10^{-13}
	2	10^{-2}	10^{-12}
Increasing acidity	3	10^{-3}	10^{-11}
	4	10^{-4}	10^{-10}
	5	10^{-5}	10^{-9}
	6	10^{-6}	10^{-8}
Neutral point	7	10^{-7}	10^{-7}
	8	10^{-8}	10^{-6}
	9	10^{-9}	10^{-5}
	10	10^{-10}	10^{-4}
	11	10^{-11}	10^{-3}
Increasing alkalinity	12	10^{-12}	10^{-2}
	13	10^{-13}	10^{-1}
	14	10^{-14}	1

In the section *Ionization of Water*, it was shown that the hydronium-ion concentration of pure water, at 25°, is $1 \times 10^{-7} \text{ N}$, corresponding to a pH of 7.

This figure, therefore, is designated as the neutral point, and all values below a pH of 7 represent acidity—the smaller the number, the greater the acidity. Values above 7 represent alkalinity—the larger the number, the greater the alkalinity. The pH scale usually runs from 0 to 14, but mathematically there is no reason why negative numbers or numbers above 14 should not be used. In practice, however, such values are never encountered because solutions that might be expected to have such values are too concentrated to be ionized extensively or the interionic attraction is so great as to materially reduce ionic activity.

The pH of the purest water obtainable, so-called ‘conductivity water’, is 7 when the measurement is made carefully under conditions to exclude carbon dioxide and prevent errors inherent in the measuring technique (such as acidity or alkalinity of the indicator). Upon agitating this water in the presence of carbon dioxide in the atmosphere (equilibrium water), the value drops rapidly to 5.7. This is the pH of nearly all distilled water that has been exposed to the atmosphere for even a short time and often is called ‘equilibrium’ water.

It should be emphasized strongly that the generalizations stated concerning neutrality, acidity, and alkalinity hold exactly only when (1) the solvent is water, (2) the temperature is 25°, and (3) there are no other factors to cause deviation from the simply formulated equilibria underlying the definition of pH given in the preceding discussion.

SPECIES CONCENTRATION

When a weak acid, H_nA is added to water, $n + 1$ species, including the un-ionized acid, can exist. After equilibrium is established, the sum of the concentrations of all species must be equal to C_a , the stoichiometric (added) concentration of acid. Thus, for a triprotic acid H_3A ,

$$C_a = [\text{H}_3\text{A}] + [\text{H}_2\text{A}^-] + [\text{HA}^{2-}] + [\text{A}_3^-] \quad (46)$$

In addition, the concentrations of all acidic and basic species in solution vary with pH, and can be represented solely in terms of equilibrium constants and the hydronium-ion concentration. These relationships may be expressed as

$$[\text{H}_n\text{A}] = [\text{H}_3\text{O}^+]^n C_a / D \quad (47)$$

$$[\text{H}_{n-j}\text{A}^{-j}] = [\text{H}_3\text{O}^+]^{n-j} K_1, \dots, K_j C_a / D \quad (48)$$

in which n represents the total number of dissociable hydrogens in the parent acid, j is the number of protons dissociated, C_a is the stoichiometric concentration of acid, and K represents the acid dissociation constants. The term D is a power series in $[\text{H}_3\text{O}^+]$ and K , starting with $[\text{H}_3\text{O}^+]$ raised to the n th power. The last term is the product of all the dissociation constants. The intermediate terms can be generated from the last term by substituting $[\text{H}_3\text{O}^+]$ for K_n to obtain the next-to-last term, then substituting $[\text{H}_3\text{O}^+]$ for K_{n-1} to obtain the next term, and onward until the first term is reached. The following examples show the denominator, D , to be used for various types of acids:

$$\text{H}_3\text{A}: D = [\text{H}_3\text{O}^+]^3 + K_1[\text{H}_3\text{O}^+]^2 + K_1K_2[\text{H}_3\text{O}^+] + K_1K_2K_3 \quad (49)$$

$$\text{H}_2\text{A}: D = [\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1K_2 \quad (50)$$

$$\text{HA}: D = [\text{H}_3\text{O}^+] + K_a \quad (51)$$

The numerator in all instances is C_a multiplied by the term from the denominator that has $[\text{H}_3\text{O}^+]$ raised to the $n - j$ power. Thus, for diprotic acids such as carbonic, succinic, tartaric, and so on,

$$[\text{H}_2\text{A}] = \frac{[\text{H}_3\text{O}^+]^2 C_a}{[\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1K_2} \quad (52)$$

$$[\text{HA}^-] = \frac{K_1[\text{H}_3\text{O}^+]C_a}{[\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1K_2} \quad (53)$$

$$[\text{A}^{2-}] = \frac{K_1K_2C_a}{[\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1K_2} \quad (54)$$

Example—Calculate the concentrations of all succinic acid species in a $1.0 \times 10^{-3} \text{ M}$ solution of succinic acid at pH 6. Assume that $K_1 = 6.4 \times 10^{-5}$ and $K_2 = 2.3 \times 10^{-6}$.

Equations 52–54 have the same denominator, D , which can be calculated as

$$\begin{aligned} D &= [\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1K_2 \\ &= 1.0 \times 10^{-12} + 6.4 \times 10^{-5} \times 1.0 \times 10^{-6} + 6.4 \\ &\quad \times 10^{-5} \times 2.3 \times 10^{-6} \\ &= 1.0 \times 10^{-12} + 6.4 \times 10^{-11} + 14.7 \times 10^{-11} \\ &= 21.2 \times 10^{-11} \end{aligned}$$

Therefore,

$$\begin{aligned} [\text{H}_2\text{A}] &= \frac{[\text{H}_3\text{O}^+]^2 C_a}{D} \\ &= \frac{1.0 \times 10^{-12} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 4.7 \times 10^{-6} \text{ M} \end{aligned}$$

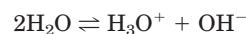
$$\begin{aligned} [\text{HA}^-] &= \frac{K_1[\text{H}_3\text{O}^+]C_a}{D} \\ &= \frac{6.4 \times 10^{-11} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 3.0 \times 10^{-4} \text{ M} \end{aligned}$$

$$\begin{aligned} [\text{A}^{2-}] &= \frac{K_1K_2C_a}{D} \\ &= \frac{14.7 \times 10^{-11} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 6.9 \times 10^{-4} \text{ M} \end{aligned}$$

PROTON-BALANCE EQUATION

In the Brønsted–Lowry system, the total number of protons released by acidic species must equal the total number of protons consumed by basic species. This results in a very useful relationship known as the *proton-balance equation* (PBE), in which the sum of the concentration terms for species that form by proton consumption is equated to the sum of the concentration terms for species that are formed by the release of protons. The PBE forms the basis of a unified approach to pH calculations, as it is an exact accounting of all proton transfers occurring in solution.

When HCl is added to water, for example, it dissociates yielding one Cl^- for each proton released. Thus, Cl^- is a species formed by the release of a proton. In the same solution, and actually in all aqueous solutions



where H_3O^+ is formed by proton consumption and OH^- is formed by proton release. Thus, the PBE is

$$[\text{H}_3\text{O}^+] = [\text{OH}^-] + [\text{Cl}^-] \quad (55)$$

In general, the PBE can be formed in the following manner:

1. Start with the species added to water.
2. Place all species that can form when protons are released on the right side of the equation.
3. Place all species that can form when protons are consumed on the left side of the equation.
4. Multiply the concentration of each species by the number of protons gained or lost to form that species.
5. Add $[\text{H}_3\text{O}^+]$ to the left side of the equation and $[\text{OH}^-]$ to the right side of the equation. These result from the interaction of two molecules of water as shown above.

Example—When H_3PO_4 is added to water, the species H_2PO_4^- forms with the release of one proton; HPO_4^{2-} forms with the release of two protons; and PO_4^{3-} forms with the release of three protons, which gives the following PBE:

$$[\text{H}_3\text{O}^+] = [\text{OH}^-] + [\text{H}_2\text{PO}_4^-] + 2[\text{HPO}_4^{2-}] + 3[\text{PO}_4^{3-}] \quad (56)$$

Example—When Na_2HPO_4 is added to water, it dissociates into two Na^+ and one HPO_4^{2-} . The sodium ion is neglected in the PBE because it is not formed from the release or consumption of protons. The species HPO_4^{2-} , however, may react with water to give H_2PO_4^- with the consumption of one proton, H_3PO_4 with the consumption of two protons, and PO_4^{3-} with the release of one proton to give the following PBE:

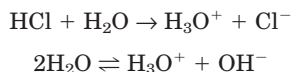
$$[\text{H}_3\text{O}^+] + [\text{H}_2\text{PO}_4^-] + 2[\text{H}_3\text{PO}_4] = [\text{OH}^-] + [\text{PO}_4^{3-}] \quad (57)$$

CALCULATIONS

The pH of solutions of acids, bases, and salts may be calculated using the concepts presented in the preceding sections.

Strong Acids or Bases

When a strong acid such as HCl is added to water, the following reactions occur:



The PBE for this system would be

$$[\text{H}_3\text{O}^+] = [\text{OH}^-] + [\text{Cl}^-] \quad (58)$$

In most instances ($C_a > 4.5 \times 10^{-7} \text{ M}$) the $[\text{OH}^-]$ would be negligible compared to the Cl^- and the equation simplifies to

$$[\text{H}_3\text{O}^+] = [\text{Cl}^-] = C_a \quad (59)$$

Thus, the hydronium-ion concentration of a solution of a strong acid would be equal to the stoichiometric concentration of the acid. This would be anticipated, because strong acids generally are assumed to be 100% ionized.

The pH of a 0.005 M solution of HCl therefore is calculated as

$$\text{pH} = -\log 0.005 = 2.30$$

In a similar manner the hydroxyl-ion concentration for a solution of a strong base such as NaOH would be

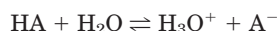
$$[\text{OH}^-] = [\text{Na}^+] = C_b \quad (60)$$

and the pH of a 0.005 M solution of NaOH would be

$$\begin{aligned} \text{pOH} &= -\log 0.005 = 2.30 \\ \text{pH} &= \text{p}K_w - \text{pOH} = 14.00 - 2.30 = 11.70 \end{aligned}$$

Weak Acids or Bases

If a weak acid, HA, is added to water, it will equilibrate with its conjugate base, A^- , as



Accounting for the ionization of water gives the following PBE for this system:

$$[\text{H}_3\text{O}^+] = [\text{OH}^-] + [\text{A}^-] \quad (61)$$

The concentration of A^- as a function of hydronium-ion concentration can be obtained as shown previously to give

$$[\text{H}_3\text{O}^+] = [\text{OH}^-] + \frac{K_a C_a}{[\text{H}_3\text{O}^+] + K_a} \quad (62)$$

Algebraic simplification yields

$$[\text{H}_3\text{O}^+] = K_a \frac{(C_a - [\text{H}_3\text{O}^+] + [\text{OH}^-])}{([\text{H}_3\text{O}^+] - [\text{OH}^-])} \quad (63)$$

In most instances for solutions of weak acids, $[\text{H}_3\text{O}^+] \gg [\text{OH}^-]$, and the equation simplifies to give

$$[\text{H}_3\text{O}^+]^2 + K_a[\text{H}_3\text{O}^+] - K_a C_a = 0 \quad (64)$$

This is a quadratic equation* that yields

$$[\text{H}_3\text{O}^+] = \frac{-K_a + \sqrt{K_a^2 + 4K_a C_a}}{2} \quad (65)$$

since $[\text{H}_3\text{O}^+]$ can never be negative. Furthermore, if $[\text{H}_3\text{O}^+]$ is less than 5% of C_a , Equation 64 is simplified further to give

$$[\text{H}_3\text{O}^+] = \sqrt{K_a C_a} \quad (66)$$

It generally is preferable to use the simplest equation to calculate $[\text{H}_3\text{O}^+]$. However, when $[\text{H}_3\text{O}^+]$ is calculated, it must be compared to C_a in order to determine whether the assumption $C_a \gg [\text{H}_3\text{O}^+]$ is valid. If the assumption is not valid, the quadratic equation should be used.

Example—Calculate the pH of a $5.00 \times 10^{-5} \text{ M}$ solution of a weak acid having a $K_a = 1.90 \times 10^{-5}$.

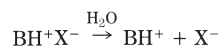
$$\begin{aligned} [\text{H}_3\text{O}^+] &= \sqrt{K_a C_a} \\ &= \sqrt{1.90 \times 10^{-5} \times 5.00 \times 10^{-5}} \\ &= 3.08 \times 10^{-5} \text{ M} \end{aligned}$$

As C_a ($5.00 \times 10^{-5} \text{ M}$) is not much greater than $[\text{H}_3\text{O}^+]$, the quadratic equation (Equation 65) should be used.

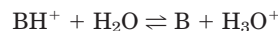
$$\begin{aligned} [\text{H}_3\text{O}^+] &= \frac{-1.90 \times 10^{-5} + \sqrt{(1.90 \times 10^{-5})^2 + 4(5.00 \times 10^{-5})}}{2} \\ &= 7.06 \times 10^{-3} \\ \text{pH} &= -\log (7.06 \times 10^{-3}) = 2.15 \end{aligned}$$

Note that the assumption $[\text{H}_3\text{O}^+] \gg [\text{OH}^-]$ is valid. The hydronium-ion concentration calculated from Equation 66 has a relative error of about 100% when compared to the correct value obtained from Equation 65.

When a salt obtained from a strong acid and a weak base—such as ammonium chloride, morphine sulfate, or pilocarpine hydrochloride—is dissolved in water, it dissociates as



in which BH^+ is the protonated form of the base B, and X^- is the anion of a strong acid. Because X^- is the anion of a strong acid, it is too weak a base to undergo any further reaction with water. The protonated base, however, can act as a weak acid to give



Thus, Equations 65 and 66 are valid, with C_a being equal to the concentration of the salt in solution. If K_a for the protonated base is not available, it can be obtained by dividing K_b for the base B, into K_w .

Example—Calculate the pH of a 0.026 M solution of ammonium chloride. Assume that K_b for ammonia is 1.74×10^{-5} and K_w is 1.00×10^{-14} .

* The general solution to a quadratic equation of the form

$$aX^2 + bX + c = 0 \quad \text{is} \quad X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$K_a = \frac{K_w}{K_b} = \frac{1.00 \times 10^{-14}}{1.74 \times 10^{-5}} = 5.75 \times 10^{-10}$$

$$\begin{aligned} [\text{H}_3\text{O}^+] &= \sqrt{K_a C_a} \\ &= \sqrt{5.75 \times 10^{-10} \times 2.6 \times 10^{-2}} \\ &= 3.87 \times 10^{-6} M \end{aligned}$$

$$\text{pH} = -\log(3.87 \times 10^{-6}) = 5.41$$

As C_a is much greater than $[\text{H}_3\text{O}^+]$ and $[\text{H}_3\text{O}^+]$ is much greater than $[\text{OH}^-]$, the assumptions are valid and the value calculated for pH is sufficiently accurate.

Weak Bases

When a weak base, B, is dissolved in water it ionizes to give the conjugate acid as



The PBE for this system is

$$[\text{BH}^+] + [\text{H}_3\text{O}^+] = [\text{OH}^-] \quad (67)$$

Substituting $[\text{BH}^+]$ as a function of hydronium-ion concentration and simplifying, in the same manner as shown for a weak acid, gives

$$[\text{OH}^-] = K_b \frac{(C_b - [\text{OH}^-] + [\text{H}_3\text{O}^+])}{([\text{OH}^-] - [\text{H}_3\text{O}^+])} \quad (68)$$

If $[\text{OH}^-] \gg [\text{H}_3\text{O}^+]$, as is true generally, then

$$[\text{OH}^-]^2 = K_b[\text{OH}^-] - K_b C_b = 0 \quad (69)$$

which is a quadratic with the following solution:

$$[\text{OH}^-] = \frac{-K_b + \sqrt{K_b^2 + 4K_b C_b}}{2} \quad (70)$$

If $C_b \gg [\text{OH}^-]$, the quadratic equation simplifies to

$$[\text{OH}^-] = \sqrt{K_b C_b} \quad (71)$$

Once $[\text{OH}^-]$ is calculated, it can be converted to pOH, which can be subtracted from $\text{p}K_w$ to give pH.

Example—Calculate the pH of a $4.50 \times 10^{-2} M$ solution of a weak base having $K_b = 2.00 \times 10^{-4}$. Assume that $K_w = 1.00 \times 10^{-14}$.

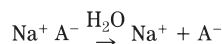
$$\begin{aligned} [\text{OH}^-] &= \sqrt{K_b C_b} \\ &= \sqrt{2.00 \times 10^{-4} \times 4.50 \times 10^{-2}} \\ &= \sqrt{9.00 \times 10^{-6}} = 3.00 \times 10^{-3} M \end{aligned}$$

Both assumptions are valid.

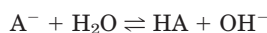
$$\text{pOH} = -\log 3.00 \times 10^{-3} = 2.52$$

$$\text{pH} = 14.00 - 2.52 = 11.48$$

When salts obtained from strong bases and weak acids (eg, sodium acetate, sodium sulfathiazole, or sodium benzoate) are dissolved in water, they dissociate as



in which A^- is the conjugate base of the weak acid, HA. The Na^+ undergoes no further reaction with water. The A^- , however, acts as a weak base to give



Thus, Equations 70 and 71 are valid, with C_b being equal to the concentration of the salt in solution. The value for K_b can be obtained by dividing K_a for the conjugate acid, HA, into K_w .

Example—Calculate the pH of a $0.05 M$ solution of sodium acetate. Assume that K_a for acetic acid = 1.75×10^{-5} and $K_w = 1.00 \times 10^{-14}$.

$$K_b = \frac{K_w}{K_a} = \frac{1.00 \times 10^{-14}}{1.75 \times 10^{-5}}$$

$$= 5.71 \times 10^{-10}$$

$$\begin{aligned} \text{OH}^- &= \sqrt{K_b C_b} = \sqrt{5.71 \times 10^{-10} \times 5.0 \times 10^{-2}} \\ &= 5.34 \times 10^{-6} M \end{aligned}$$

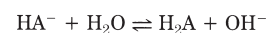
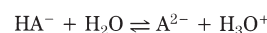
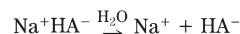
Both assumptions are valid:

$$\text{pOH} = -\log(5.34 \times 10^{-6}) = 5.27$$

$$\text{pH} = 14.00 - 5.27 = 8.73$$

Ampholytes

Substances such as NaHCO_3 and NaH_2PO_4 are termed *ampholytes* and are capable of functioning both as acids and bases. When an ampholyte of the type NaHA is dissolved in water, the following series of reactions can occur:



The total PBE for the system is

$$[\text{H}_3\text{O}^+] + [\text{H}_2\text{A}] = [\text{OH}^-] + [\text{A}^{2-}] \quad (72)$$

Substituting both $[\text{H}_2\text{A}]$ and $[\text{A}^{2-}]$ as a function of $[\text{H}_3\text{O}^+]$ (see Equations 52 and 54), yields

$$\begin{aligned} [\text{H}_3\text{O}^+] + \frac{[\text{H}_3\text{O}^+]^2 C_s}{[\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1 K_2} \\ = \frac{K_w}{[\text{H}_3\text{O}^+]} + \frac{K_1 K_2 C_s}{[\text{H}_3\text{O}^+]^2 + K_1[\text{H}_3\text{O}^+] + K_1 K_2} \end{aligned} \quad (73)$$

This gives a fourth-order equation in $[\text{H}_3\text{O}^+]$, which can be simplified using certain judicious assumptions to

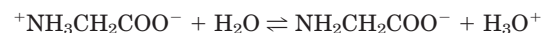
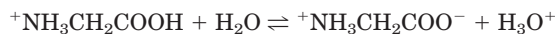
$$[\text{H}_3\text{O}^+] = \sqrt{\frac{K_1 K_2 C_s}{K_1 + C_s}} \quad (74)$$

In most instances, $C_s \gg K_1$, and the equation further simplifies to

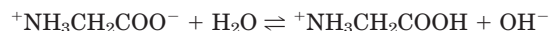
$$[\text{H}_3\text{O}^+] = \sqrt{K_1 K_2} \quad (75)$$

and $[\text{H}_3\text{O}^+]$ becomes independent of the concentration of the salt. A special property of ampholytes is that the concentration of the species HA^- is maximum at the pH corresponding to Equation 75.

When the simplest amino acid salt, glycine hydrochloride, is dissolved in water, it acts as a diprotic acid and ionizes as



The form, $^+\text{NH}_3\text{CH}_2\text{COO}^-$, is an ampholyte because it also can act as a weak base:



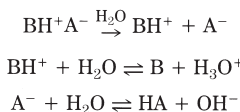
This type of substance, which carries both a charged acidic and a charged basic moiety on the same molecule is termed a *zwitterion*. Because the two charges balance each other, the molecule acts essentially as a neutral molecule. The pH at which the zwitterion concentration is maximum is known as the *isoelectric point*, which can be calculated from Equation 75.

On the acid side of the isoelectric point, amino acids and proteins are cationic and incompatible with anionic materials such as the naturally occurring gums used as suspending and/or

emulsifying agents. On the alkaline side of the isoelectric point, amino acids and proteins are anionic and incompatible with cationic materials such as benzalkonium chloride.

Salts of Weak Acids and Weak Bases

When a salt such as ammonium acetate (which is derived from a weak acid and a weak base) is dissolved in water, it undergoes the following reactions:



The total PBE for this system is

$$[\text{H}_3\text{O}^+] + [\text{HA}] = [\text{OH}^-] + [\text{B}] \quad (76)$$

Replacing [HA] and [B] as a function of $[\text{H}_3\text{O}^+]$, gives

$$[\text{H}_3\text{O}^+] + \frac{[\text{H}_3\text{O}^+]\text{C}_s}{[\text{H}_3\text{O}^+] + K_a} = [\text{OH}^-] + \frac{K'_a\text{C}_s}{[\text{H}_3\text{O}^+] + K'_a} \quad (77)$$

in which C_s is the concentration of salt, K_a is the ionization constant of the conjugate acid formed from the reaction between A^- and water, and K'_a is the ionization constant for the protonated base, BH^+ . In general, $[\text{H}_3\text{O}^+]$, $[\text{OH}^-]$, K_a , and K'_a usually are smaller than C_s and the equation simplifies to

$$[\text{H}_3\text{O}^+] = \sqrt{K_a K'_a} \quad (78)$$

Example—Calculate the pH of a 0.01 M solution of ammonium acetate. The ammonium ion has a K'_a equal to 5.75×10^{-10} , which represents K'_a in Equation 78. Acetic acid has a K_a of 1.75×10^{-5} , which represents K_a in Equation 78:

$$\begin{aligned} [\text{H}_3\text{O}^+] &= \sqrt{1.75 \times 10^{-5} \times 5.75 \times 10^{-10}} \\ &= 1.00 \times 10^{-7} \\ \text{pH} &= -\log(1.00 \times 10^{-7}) = 7.00 \end{aligned}$$

All of the assumptions are valid.

Buffers

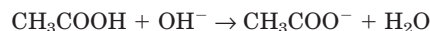
The terms *buffer*, *buffer solution*, and *buffered solution*, when used with reference to hydrogen-ion concentration or pH, refer to the ability of a system, particularly an aqueous solution, to resist a change of pH on adding acid or alkali, or on dilution with a solvent.

If an acid or base is added to water, the pH of the latter is changed markedly, for water has no ability to resist change of pH; it is completely devoid of buffer action. Even a very weak acid such as carbon dioxide changes the pH of water, decreasing it from 7 to 5.7 when the small concentration of carbon dioxide present in air is equilibrated with pure water. This extreme susceptibility of distilled water to a change of pH upon adding very small amounts of acid or base is often of great concern in pharmaceutical operations. Solutions of neutral salts, such as sodium chloride, similarly lack ability to resist change of pH on adding acid or base; such solutions are called *unbuffered*.

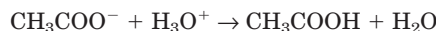
Characteristic of *buffered solutions*, which undergo small changes of pH on addition of acid or base, is the presence either of a weak acid and a salt of the weak acid, or a weak base and a salt of the weak base. An example of the former system is acetic acid and sodium acetate; and of the latter, ammonium hydroxide and ammonium chloride. From the proton concept of acids and bases discussed earlier, it is apparent that such buffer action involves a conjugate acid–base pair in the solution. It will be recalled that acetate ion is the conjugate base of acetic acid, and that ammonium ion is the conjugate acid of am-

monia (the principal constituent of what commonly is called ammonium hydroxide).

The mechanism of action of the acetic acid–sodium acetate buffer pair is that the acid, which exists largely in molecular (nonionized) form, combines with hydroxyl ion that may be added to form acetate ion and water; thus,

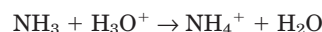


The acetate ion, which is a base, combines with the hydrogen (more exactly hydronium) ion that may be added to form essentially nonionized acetic acid and water, represented as

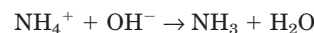


As will be illustrated later by an example, the change of pH is slight as long as the amount of hydronium or hydroxyl ion added does not exceed the capacity of the buffer system to neutralize it.

The ammonia–ammonium chloride pair functions as a buffer because the ammonia combines with hydronium ion that may be added to form ammonium ion and water; thus,

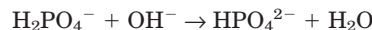


Ammonium ion, which is an acid, combines with added hydroxyl ion to form ammonia and water, as

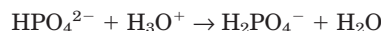


Again, the change of pH is slight if the amount of added hydronium or hydroxyl ion is not in excess of the capacity of the system to neutralize it.

Besides these two general types of buffers, a third appears to exist. This is the buffer system composed of two salts, as monobasic potassium phosphate, KH_2PO_4 , and dibasic potassium phosphate, K_2HPO_4 . This is not, however, a new type of buffer; it is actually a weak-acid/conjugate-base buffer in which an ion, H_2PO_4^- , serves as the weak acid, and HPO_4^{2-} is its conjugate base. When hydroxyl ion is added to this buffer the following reaction takes place:



and when hydronium ion is added,



It is apparent that the mechanism of action of this type of buffer is essentially the same as that of the weak-acid/conjugate-base buffer composed of acetic acid and sodium acetate.

CALCULATIONS—A buffer system composed of a conjugate acid–base pair, NaA-HA (such as sodium acetate and acetic acid), would have a PBE of

$$[\text{H}_3\text{O}^+] + [\text{HA}] = [\text{OH}^-] + [\text{A}^-] \quad (79)$$

Replacing [HA] and $[\text{A}^-]$ as a function of hydronium-ion concentration gives

$$[\text{H}_3\text{O}^+] + \frac{[\text{H}_3\text{O}^+]\text{C}_b}{[\text{H}_3\text{O}^+] + K_a} = [\text{OH}^-] + \frac{K_a\text{C}_a}{[\text{H}_3\text{O}^+] + K_a} \quad (80)$$

where C_b is the concentration of the salt, NaA , and C_a is the concentration of the weak acid, HA . This equation can be rearranged to give

$$[\text{H}_3\text{O}^+] = K_a \frac{(C_a - [\text{H}_3\text{O}^+] + [\text{OH}^-])}{(C_b + [\text{H}_3\text{O}^+] - [\text{OH}^-])} \quad (81)$$

In general, both C_a and C_b are much greater than $[\text{H}_3\text{O}^+]$, which is in turn much greater than $[\text{OH}^-]$ and the equation simplifies to

$$[\text{H}_3\text{O}^+] = \frac{K_a C_a}{C_b} \quad (82)$$

or, expressed in terms of pH, as

$$\text{pH} = \text{p}K_a + \log \frac{C_b}{C_a} \quad (83)$$

This equation generally is called the *Henderson-Hasselbalch equation*. It applies to all buffer systems formed from a single conjugate acid–base pair, regardless of the nature of the salts. For example, it applies equally well to the following buffer systems: ammonia–ammonium chloride, monosodium phosphate–disodium phosphate, and phenobarbital–sodium phenobarbital. In the ammonia–ammonium chloride system, ammonia is obviously the base and the ammonium ion is the acid (C_a equal to the concentration of the salt). In the phosphate system, monosodium phosphate is the acid and disodium phosphate is the base. For the phenobarbital buffer system, phenobarbital is the acid and the phenobarbital anion is the base (C_b equal to the concentration of sodium phenobarbital).

As an example of the application of this equation, the pH of a buffer solution containing acetic acid and sodium acetate, each in 0.1 *M* concentration, may be calculated. The K_a of acetic acid, as defined above, is 1.8×10^{-5} , at 25°.

Solution:

First, the $\text{p}K_a$ of acetic acid is calculated:

$$\begin{aligned} \text{p}K_a &= -\log K_a = -\log 1.8 \times 10^{-5} \\ &= -\log 1.8 - \log 10^{-5} \\ &= -0.26 - (-5) = +4.74 \end{aligned}$$

Substituting this value into Equation 83:

$$\text{pH} = \log \frac{0.1}{0.1} + 4.74 = +4.74$$

The Henderson-Hasselbalch equation predicts that any solutions containing the same molar concentration of acetic acid as of sodium acetate will have the same pH. Thus, a solution of 0.01 *M* concentration of each will have the same pH, 4.74, as one of 0.1 *M* concentration of each component. Actually, there will be some difference in the pH of the solutions, for the *activity coefficient* of the components varies with concentration. For most practical purposes, however, the approximate values of pH calculated by the equation are satisfactory. It should be pointed out that the buffer of higher concentration of each component will have a much greater capacity for neutralizing added acid or base and this point will be discussed further in the discussion of buffer capacity.

The Henderson-Hasselbalch equation is useful also for calculating the ratio of molar concentrations of a buffer system required to produce a solution of specific pH. As an example, suppose that an acetic acid–sodium acetate buffer of pH 4.5 must be prepared. What ratio of the buffer components should be used?

Solution:

Rearranging Equation 83, which is used to calculate the pH of weak acid–salt type buffers, gives

$$\begin{aligned} \log \frac{[\text{base}]}{[\text{acid}]} &= \text{pH} - \text{p}K_a \\ &= 4.5 - 4.76 = -0.24 = (9.76 - 10) \\ \frac{[\text{base}]}{[\text{acid}]} &= \text{antilog of } (9.76 - 10) = 0.575 \end{aligned}$$

The interpretation of this result is that the *proportion* of sodium acetate to acetic acid should be 0.575 mol of the former to 1 mol of the latter to produce a pH of 4.5. A solution containing 0.0575 mol of sodium acetate and 0.1 mol of acetic acid per liter would meet this requirement, as would also one containing 0.00575 mol of sodium acetate and 0.01 mol of acetic acid per liter. The actual concentration selected would depend chiefly on the desired buffer capacity.

BUFFER CAPACITY—The ability of a buffer solution to resist changes in pH upon addition of acid or alkali may be measured in terms of *buffer capacity*. In the preceding discussion of buffers, it has been seen that, in a general way, the concentra-

tion of acid in a weak-acid/conjugate-base buffer determines the capacity to “neutralize” added base, while the concentration of salt of the weak acid determines the capacity to neutralize added acid. Similarly, in a weak-base/conjugate-acid buffer the concentration of the weak base establishes the buffer capacity toward added acid, while the concentration of the conjugate acid of the weak base determines the capacity toward added base. When the buffer is equimolar in the concentrations of weak acid and conjugate base, or of weak base and conjugate acid, it has equal buffer capacity toward added strong acid or strong base.

Van Slyke, the biochemist, introduced a quantitative expression for evaluating buffer capacity. This may be defined as the amount, in gram-equivalents (g-eq) per liter, of strong acid or strong base required to be added to a solution to change its pH by 1 unit; a solution has a buffer capacity of 1 when 1 L requires 1 g-eq of strong base or acid to change the pH 1 unit. (In practice, considerably smaller increments are measured, expressed as the ratio of acid or base added to the change of pH produced.) From this definition it is apparent that the smaller the pH change in a solution caused by the addition of a specified quantity of acid or alkali, the greater the buffer capacity of the solution.

The following examples illustrate certain basic principles and calculations concerning buffer action and buffer capacity.

Example 1—What is the change of pH on adding 0.01 mol of NaOH to 1 L of 0.10 *M* acetic acid?

(a) Calculate the pH of a 0.10 molar solution of acetic acid:

$$\begin{aligned} [\text{H}_3\text{O}^+] &= \sqrt{K_a C_a} = \sqrt{1.75 \times 10^{-4} \times 1.0 \times 10^{-1}} = 4.18 \times 10^{-3} \\ \text{pH} &= -\log 4.18 \times 10^{-3} = 2.38 \end{aligned}$$

(b) On adding 0.01 mol of NaOH to a liter of this solution, 0.01 mol of acetic acid is converted to 0.01 mol of sodium acetate, thereby decreasing C_a to 0.09 *M*, and $C_b = 1.0 \times 10^{-2} *M*. Using the Henderson-Hasselbalch equation gives$

$$\text{pH} = 4.76 + \log \frac{0.01}{0.09} = 4.76 - 0.95 = 3.81$$

The pH change is, therefore, 1.43 unit. The buffer capacity as defined above is calculated to be

$$\frac{\text{mols of NaOH added}}{\text{change in pH}} = 0.011$$

Example 2—What is the change of pH on adding 0.1 mol of NaOH to 1 L of buffer solution 0.1 *M* in acetic acid and 0.1 *M* in sodium acetate?

(a) The pH of the buffer solution before adding NaOH is

$$\begin{aligned} \text{pH} &= \log \frac{[\text{base}]}{[\text{acid}]} + \text{p}K_a \\ &= \log \frac{0.1}{0.1} + 4.76 = 4.76 \end{aligned}$$

(b) On adding 0.01 mol of NaOH per liter to this buffer solution, 0.01 mol of acetic acid is converted to 0.01 mol of sodium acetate, thereby decreasing the concentration of acid to 0.09 *M* and increasing the concentration of base to 0.11 *M*. The pH is calculated as

$$\begin{aligned} \text{pH} &= \log \frac{0.11}{0.09} + 4.76 \\ &= 0.087 + 4.76 = 4.85 \end{aligned}$$

The change of pH in this case is only 0.09 unit, about 1/10 the change in the preceding example. The buffer capacity is calculated as

$$\frac{\text{mols of NaOH added}}{\text{change of pH}} = \frac{0.01}{0.09} = 0.11$$

Thus, the buffer capacity of the acetic acid–sodium acetate buffer solution is approximately 10 times that of the acetic acid solution.

As is in part evident from these examples, and may be further evidenced by calculations of pH changes in other systems, the degree of buffer action and, therefore, the buffer capacity, depend on the kind and concentration of the buffer components, the pH region involved and the kind of acid or alkali added.

STRONG ACIDS AND BASES AS “BUFFERS”—In the foregoing discussion, buffer action was attributed to systems of (1) weak acids and their conjugate bases, (2) weak bases and their conjugate acids, and (3) certain acid–base pairs that can function in the manner either of system 1 or 2.

The ability to resist change in pH on adding acid or alkali is possessed also by relatively concentrated solutions of strong acids and strong bases. If to 1 L of pure water having a pH of 7 is added 1 mL of 0.01 M hydrochloric acid, the pH is reduced to about 5. If the same volume of the acid is added to 1 L of 0.001 M hydrochloric acid, which has a pH of about 3, the hydronium-ion concentration is increased only about 1% and the pH is reduced hardly at all. The nature of this buffer action is quite different from that of the true buffer solutions. The very simple explanation is that when 1 mL of 0.01 M HCl, which represents 0.00001 g-eq of hydronium ions, is added to the 0.0000001 g-eq of hydronium ions in 1 L of pure water, the hydronium-ion concentration is increased 100-fold (equivalent to two pH units), but when the same amount is added to the 0.001 g-eq of hydronium ions in 1 L of 0.001 M HCl, the increase is only 1/100 the concentration already present. Similarly, if 1 mL of 0.01 M NaOH is added to 1 L of pure water, the pH is increased to 9, while if the same volume is added to 1 L of 0.001 molar NaOH, the pH is increased almost immeasurably.

In general, solutions of strong acids of pH 3 or less, and solutions of strong bases of pH 11 or more, exhibit this kind of buffer action by virtue of the relatively high concentration of hydronium or hydroxyl ions present. The USP includes among its Standard Buffer Solutions a series of hydrochloric acid buffers, covering the pH range 1.2 to 2.2, which also contain potassium chloride. The salt does not participate in the buffering mechanism, as is the case with salts of weak acids; instead, it serves as a nonreactive constituent required to maintain the proper electrolyte environment of the solutions.

DETERMINATION OF pH

Colorimetry

A relatively simple and inexpensive method for determining the approximate pH of a solution depends on the fact that some conjugate acid–base pairs (indicators) possess one color in the acid form and another color in the base form. Assume that the acid form of a particular indicator is red, and the base form is yellow. The color of a solution of this indicator will range from red when it is sufficiently acid, to yellow when it is sufficiently alkaline.

In the intermediate pH range (the transition interval) the color will be a blend of red and yellow depending upon the ratio of the base to the acid form. In general, although there are slight differences between indicators, color changes apparent to the eye cannot be discerned when the ratio of base to acid form, or acid to base form exceeds 10:1. The use of Equation 83 indicates that the transition range of most indicators is equal to the pK_a of the indicator ± 1 pH unit, or a useful range of approximately two pH units. Standard indicator solutions can be made at known pH values within the transition range of the indicator, and the pH of an unknown solution can be determined by adding the indicator to it and comparing the resulting color with the standard solutions.

Another method for using these indicators is to apply them to thin strips of filter paper. A drop of the unknown solution is placed on a piece of the indicator paper and the resulting color is compared to a color chart supplied with the indicator paper. These papers are available in a wide variety of pH ranges.

Potentiometry

Electrometric methods for the determination of pH are based on the fact that the difference of electrical potential between

two suitable electrodes dipping into a solution containing hydronium ions depends on the concentration (or activity) of the latter. The development of a potential difference is not a specific property of hydronium ions. A solution of any ion will develop a potential proportional to the concentration of that ion if a suitable pair of electrodes is placed in the solution.

The relationship between the potential difference and concentration of an ion in equilibrium with the electrodes may be derived as follows. When a metal is immersed into a solution of one of its salts, there is a tendency for the metal to go into solution in the form of ions. This tendency is known as the *solution pressure* of the metal and is comparable to the tendency of sugar molecules (eg, to dissolve in water). The metallic ions in solution tend, on the other hand, to become discharged by forming atoms, this effect being proportional to the *osmotic pressure* of the ions.

For an atom of a metal to go into solution as a positive ion, electrons, equal in number to the charge on the ion, must be left behind on the metal electrode with the result that the latter becomes negatively charged. The positively charged ions in solution, however, may become discharged as atoms by taking up electrons from the metal electrode. Depending on which effect predominates, the electrical charge on the electrode will be either positive or negative and may be expressed quantitatively by the following equation proposed by Nernst in 1889:

$$E = \frac{RT}{nF} \ln \frac{p}{P} \quad (84)$$

where E is the potential difference or electromotive force, R is the gas constant (8.316 joules), T is the absolute temperature, n is the valence of the ion, F is the Faraday of electricity (96,500 coulombs), p is the osmotic pressure of the ions, and P is the solution pressure of the metal.

Inasmuch as it is impossible to measure the potential difference between one electrode and a solution with any degree of certainty, it is customary to use two electrodes and to measure the potential difference between them. If two electrodes, both of the same metal, are immersed in separate solutions containing ions of that metal—at osmotic pressure p_1 and p_2 , respectively—and are connected by means of a tube containing a nonreacting salt solution (a so-called *salt bridge*), the potential developed across the two electrodes will be equal to the difference between the potential differences of the individual electrodes; thus,

$$E = E_1 - E_2 = \frac{RT}{nF} \ln \frac{p_1}{P_1} - \frac{RT}{nF} \ln \frac{p_2}{P_2} \quad (85)$$

As both electrodes are of the same metal, $P_1 = P_2$ and the equation may be simplified to

$$E = \frac{RT}{nF} \ln p_1 - \frac{RT}{nF} \ln p_2 = \frac{RT}{nF} \ln \frac{p_1}{p_2} \quad (86)$$

In place of osmotic pressures it is permissible, for dilute solutions, to substitute the concentrations c_1 and c_2 that were found (see Chapter 16), to be proportional to p_1 and p_2 . The equation then becomes

$$E = \frac{RT}{nF} \ln \frac{c_1}{c_2} \quad (87)$$

If either c_1 or c_2 is known, it is obvious that the value of the other may be found if the potential difference, E , of this cell can be measured.

For the determination of hydronium-ion concentration or pH, an electrode at which an equilibrium between hydrogen gas and hydronium ion can be established must be used in place of metallic electrodes. Such an electrode may be made by electrolytically coating a strip of platinum, or other noble metal, with platinum black and saturating the latter with pure hydrogen gas. This device functions as a *hydrogen electrode*. Two such electrodes may be assembled as shown in Figure 17-3.

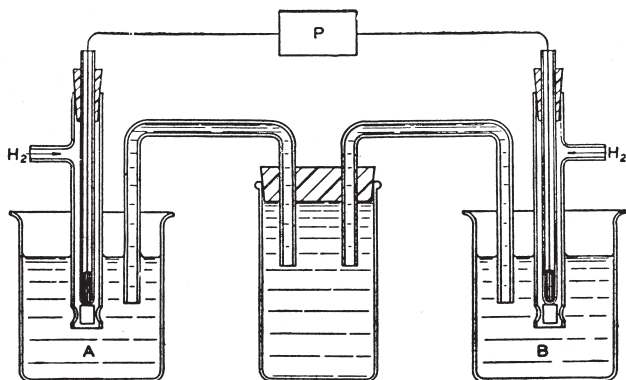


Figure 17-3. Hydrogen-ion concentration chain.

In this diagram one electrode dips into Solution A, containing a known hydronium-ion concentration, and the other electrode dips into Solution B, containing an unknown hydronium-ion concentration. The two electrodes and solutions, sometimes called *half-cells*, then are connected by a bridge of neutral salt solution, which has no significant effect on the solutions it connects. The potential difference across the two electrodes is measured by means of a potentiometer, *P*. If the concentration, c_1 , of hydronium ion in Solution A is 1 *N*, Equation 87 simplifies to

$$E = \frac{RT}{nF} \ln \frac{1}{c_2} \quad (88)$$

or in terms of Briggian logarithms

$$E = 2.303 \frac{RT}{nF} \log_{10} \frac{1}{c_2} \quad (89)$$

If for $\log_{10} 1/c_2$ there is substituted its equivalent pH, the equation becomes

$$E = 2.303 \frac{RT}{nF} \text{pH} \quad (90)$$

and finally by substituting numerical values for R , n , T , and F , and assuming the temperature to be 20°, the following simple relationship is derived:

$$E = 0.0581 \text{ pH or } \text{pH} = \frac{E}{0.0581} \quad (91)$$

The hydrogen electrode dipping into a solution of known hydronium-ion concentration, called the *reference electrode*, may be replaced by a calomel electrode, one type of which is shown in Figure 17-4. The elements of a calomel electrode are mercury and calomel in an aqueous solution of potassium chloride. The potential of this electrode is constant, regardless of the hydronium-ion concentration of the solution into which it dips. The potential depends on the equilibrium that is set up between mercury and mercurous ions from the calomel, but the concentration of the latter is governed, according to the solubility-product principle, by the concentration of chloride ions, which are derived mainly from the potassium chloride in the solution. Therefore, the potential of this electrode varies with the concentration of potassium chloride in the electrolyte.

Because the calomel electrode always indicates voltages that are higher, by a constant value, than those obtained when the normal hydrogen electrode chain shown in Figure 17-3 is used, it is necessary to subtract the potential due to the calomel electrode itself from the observed voltage. As the magnitude of this voltage depends on the concentration of potassium chloride in the calomel-electrode electrolyte, it is necessary to know the concentration of the former. For most purposes a saturated potassium chloride solution is used that

produces potential difference of 0.2488 V. Accordingly, before using Equation 86 for the calculation of pH from the voltage of a cell made up of a calomel and a hydrogen electrode dipping into the solution to be tested, 0.2488 V must be subtracted from the observed potential difference. Expressed mathematically, Equation 92 is used for calculating pH from the potential difference of such a cell.

$$\text{pH} = \frac{E - 0.2488}{0.0581} \quad (92)$$

In measuring the potential difference between the electrodes, it is imperative that very little current be drawn from the cell, for with current flowing the voltage changes, owing to polarization effects at the electrode. Because of this it is not possible to make accurate measurements with a voltmeter that requires appreciable current to operate it. In its place a potentiometer is used that does not draw a current from the cell being measured.

There are many limitations to the use of the hydrogen electrode:

- It cannot be used in solutions containing strong oxidants such as ferric iron, dichromates, nitric acid, peroxide, or chlorine or reductants such as sulfurous acid and hydrogen sulfide.
- It is affected by the presence of organic compounds that are reduced fairly easily.
- It cannot be used successfully in solutions containing cations that fall below hydrogen in the electrochemical series.
- Erratic results are obtained in the measurement of unbuffered solutions unless special precautions are taken.
- It is troublesome to prepare and maintain.

As other electrodes more convenient to use now are available, the hydrogen electrode today is used rarely. Nevertheless, it is the ultimate standard for pH measurements.

To avoid some of the difficulties with the hydrogen electrode, the *quinhydrone* electrode was introduced and was popular for a long time, particularly for measurements of acid solutions. The unusual feature of this electrode is that it consists of a piece of gold or platinum wire or foil dipping into the solution to be tested, in which has been dissolved a small quantity of quinhydrone. A calomel electrode may be used for reference, just as in determinations with the hydrogen electrode.

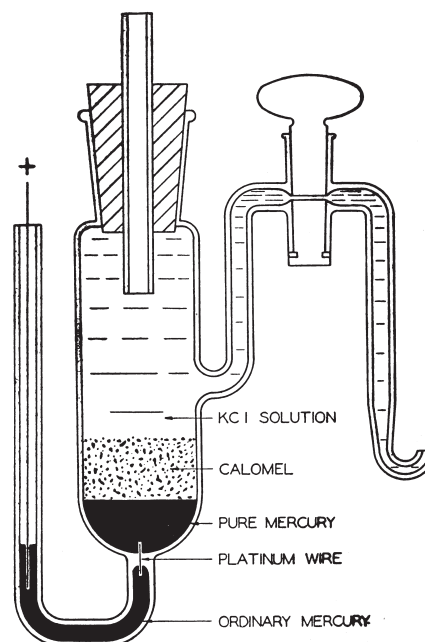
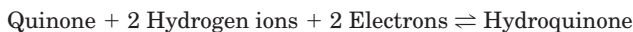


Figure 17-4. Calomel electrode.

Quinhydrone consists of an equimolecular mixture of quinone and hydroquinone; the relationship between these substances and hydrogen-ion concentration is



In a solution containing hydrogen ions the potential of the quinhydrone electrode is related logarithmically to hydronium-ion concentration if the ratio of the hydroquinone concentration to that of quinone is constant and practically equal to 1. This ratio is maintained in an acid solution containing an excess of quinhydrone, and measurements may be made quickly and accurately; however, quinhydrone cannot be used in solutions more alkaline than pH 8.

An electrode that, because of its simplicity of operation and freedom from contamination or change of the solution being tested, has replaced both the hydrogen and quinhydrone electrodes is the *glass electrode*. It functions because when a thin membrane of a special composition of glass separates two solutions of different pH, a potential difference develops across the membrane that depends on the pH of both solutions. If the pH of one of the solutions is known, the other may be calculated from the potential difference.

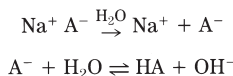
In practice, the glass electrode usually consists of a bulb of the special glass fused to the end of a tube of ordinary glass. Inside the bulb is placed a solution of known pH, in contact with an internal silver-silver chloride or other electrode. This glass electrode and another reference electrode are immersed in the solution to be tested and the potential difference is measured. A potentiometer providing electronic amplification of the small current produced is employed. The modern instruments available permit reading the pH directly and provide also for compensation of variations due to temperature in the range of 0° to 50° and to the small but variable asymmetry potential inherent in the glass electrode.

PHARMACEUTICAL SIGNIFICANCE

In the broad realm of knowledge concerning the preparation and action of drugs few, if any, variables are so important as pH. For the purpose of this presentation, four principal types of pH-dependence of drug systems will be discussed: solubility, stability, activity, and absorption.

Drug Solubility

If a salt, NaA, is added to water to give a concentration C_s , the following reactions occur:



If the pH of the solution is lowered, more of the A^- would be converted to the unionized acid, HA, in accordance with Le Chatelier's principle. Eventually, a pH will be obtained, below which the amount of HA formed exceeds its aqueous solubility, S_0 , and the acid will precipitate from solution; this pH can be designated as pH_p . At this point, at which the amount of HA formed just equals S_0 , a mass balance on the total amount of drug in solution yields

$$C_s = [\text{HA}] + [\text{A}^-] = S_0 + [\text{A}^-] \quad (93)$$

Replacing $[\text{A}^-]$ as a function of hydronium-ion concentration gives

$$C_s = S_0 + \frac{K_a C_s}{[\text{H}_3\text{O}^+]_p + K_a} \quad (94)$$

where K_a is the ionization constant for the conjugate acid, HA, and $[\text{H}_3\text{O}^+]_p$ refers to the hydronium-ion concentration above

which precipitation will occur. This equation can be rearranged to give

$$[\text{H}_3\text{O}^+]_p = K_a \frac{S_0}{C_s - S_0} \quad (95)$$

Taking logarithms gives

$$\text{pH}_p = \text{p}K_a + \log \frac{C_s - S_0}{S_0} \quad (96)$$

Thus, the pH below which precipitation occurs is a function of the amount of salt added initially, the $\text{p}K_a$ and the solubility of the free acid formed from the salt.

The analogous equation for salts of weak bases and strong acids (such as pilocarpine hydrochloride, cocaine hydrochloride, or codeine phosphate) would be

$$\text{pH}_p = \text{p}K_a + \log \frac{S_0}{C_s - S_0} \quad (97)$$

in which $\text{p}K_a$ refers to the protonated form of the weak base.

Example—Below what pH will free phenobarbital begin to precipitate from a solution initially containing 1.3 g of sodium phenobarbital/100 mL at 25°? The molar solubility of phenobarbital is 0.0050 and its $\text{p}K_a$ is 7.41. The molecular weight of sodium phenobarbital is 254.

The molar concentration of salt initially added is

$$\begin{aligned} C_s &= \frac{\text{g/L}}{\text{mol wt}} = \frac{13}{254} = 0.051 \text{ M} \\ \text{pH}_p &= 7.41 + \log \frac{0.051 - 0.005}{0.005} \\ &= 7.41 + 0.96 = 8.37 \end{aligned}$$

Example—Above what pH will free cocaine begin to precipitate from a solution initially containing 0.0294 mol of cocaine hydrochloride per liter? The $\text{p}K_b$ of cocaine is 5.59, and its molar solubility is 5.60×10^{-3} .

$$\begin{aligned} \text{p}K_a &= \text{p}K_w - \text{p}K_b = 14.00 - 5.59 = 8.41 \\ \text{pH}_p &= 8.41 + \log \frac{0.0056}{0.0294 - 0.0056} \\ &= 8.41 + (-0.63) = 7.78 \end{aligned}$$

Drug Stability

One of the most diversified and fruitful areas of study is the investigation of the effect of hydrogen-ion concentration on the stability or, in more general terms, the reactivity of pharmaceutical systems. The evidence for enhanced stability of systems when these are maintained within a narrow range of pH, as well as of progressively decreasing stability as the pH departs from the optimum range, is abundant. Stability (or instability) of a system may result from gain or loss of a proton (hydrogen ion) by a substrate molecule—often accompanied by an electronic rearrangement—that reduces (or increases) the reactivity of the molecule. *Instability* results when the substance desired to remain unchanged is converted to one or more other, unwanted, substances. In aqueous solution, instability may arise through the catalytic effect of acids or bases—the former by transferring a proton to the substrate molecule, the latter by accepting a proton.

Specific illustrations of the effect of hydrogen-ion concentration on the stability of medicinals are myriad; only a few will be given here, these being chosen to show the importance of pH adjustment of solutions that require sterilization.

Morphine solutions are not decomposed during a 60-min exposure at a temperature of 100° if the pH is less than 5.5; neutral and alkaline solutions, however, are highly unstable. Minimum hydrolytic decomposition of solutions of cocaine

occurs in the range of pH of 2 to 5; in one study a solution of cocaine hydrochloride, initially at a pH of 5.7, remained stable during 2 months (although the pH dropped to 4.2 in this time), while another solution buffered to about pH 6 underwent approximately 30% hydrolysis in the same time. Similarly, solutions of procaine hydrochloride containing some hydrochloric acid showed no appreciable decomposition; when dissolved in water alone, 5% of the procaine hydrochloride hydrolyzed, whereas when buffered to pH 6.5, from 19 to 35% underwent decomposition by hydrolysis. Solutions of thiamine hydrochloride may be sterilized by autoclaving without appreciable decomposition if the pH is below 5; above this, thiamine hydrochloride is unstable.

The stability of many disperse systems, and especially of certain emulsions, is often pH dependent. Information concerning specific emulsion systems, and the effect of pH upon them, may be found in Chapter 21.

Drug Activity

Drugs that are weak acids or weak bases—and hence may exist in ionized or nonionized form (or a mixture of both)—may be *active* in one form but not in the other; often such drugs have an optimum pH range for maximum activity. Thus, mandelic acid, benzoic acid, or salicylic acid have pronounced antibacterial activity in nonionized form but have practically no such activity in ionized form. Accordingly, these substances require an acid environment to function effectively as antibacterial agents. For example, sodium benzoate is effective as a preservative in 4% concentration at pH 7, in 0.06 to 0.1% concentration at pH 3.5 to 4, and in 0.02 to 0.03% concentration at pH 2.3 to 2.4. Other antibacterial agents are active principally, if not entirely, in cationic form. Included in this category are the acridines and quaternary ammonium compounds.

Drug Absorption

The degree of ionization and lipid solubility of a drug are two important factors that determine the rate of *absorption* of drugs from the gastrointestinal tract, and indeed their passage through cellular membranes generally. Drugs that are weak organic acids or bases, and that in nonionized form are soluble in lipids, apparently are absorbed through cellular membranes by virtue of the lipoidal nature of the membranes. Completely ionized drugs, on the other hand, are absorbed poorly, if at all. Rates of absorption of a variety of drugs are related to their ionization constants and in many cases may be predicted quantitatively on the basis of this relationship. Thus, not only the degree of the acidic or basic character of a drug, but also consequently the pH of the physiological medium (eg, gastric or intestinal fluid, plasma, cerebrospinal fluid) in which a drug is dissolved or dispersed—because this pH determines the extent to which the drug will be converted to ionic or nonionic form—become important parameters of drug absorption. Further information on drug absorption is given in Chapter 58.

REFERENCES

1. Benet LZ, Goyan JE. *J Pharm Sci* 1965; 54:1179.
2. Riegelman S et al. *J Pharm Sci* 1962; 51:129.
3. Niebergall PJ et al. *J Pharm Sci* 1972; 61:232.

BIBLIOGRAPHY

- Conway BE. *Ionic Hydration in Chemistry and Biophysics*. Amsterdam: Elsevier, 1980.
- Denbigh K. *The Principles of Chemical Equilibrium*, 4th ed. London: Cambridge University Press, 1981.
- Freiser H, Fernando Q. *Ionic Equilibria in Analytical Chemistry*. New York: Wiley, 1966.
- Harned HS, Owen BB. *The Physical Chemistry of Electrolytic Solutions*. New York, Reinhold, 1958.

ACKNOWLEDGMENTS—Paul J Niebergall, PhD is acknowledged for his efforts in previous editions of this work.

Tonicity, Osmoticity, Osmolality, and Osmolarity

Cathy Y Poon, PharmD



BASIC DEFINITIONS

If a solution is placed in contact with a membrane that is permeable to molecules of the solvent, but not to molecules of the solute, the movement of solvent through the membrane is called *osmosis*. Such a membrane often is called *semi-permeable*. As the several types of membranes of the body vary in their permeability, it is well to note that they are *selectively* permeable. Most normal living-cell membranes maintain various solute concentration gradients. A selectively permeable membrane may be defined either as one that does not permit free, unhampered diffusion of all the solutes present, or as one that maintains at least one solute concentration gradient across itself. Osmosis, then, is the diffusion of water through a membrane that maintains at least one solute concentration gradient across itself.

Assume that Solution A is on one side of the membrane, and Solution B of the same solute but of a higher concentration is on the other side; the solvent will tend to pass into the more concentrated solution until equilibrium has been established. The pressure required to prevent this movement is the osmotic pressure. It is defined as the excess pressure, or pressure greater than that above the pure solvent, that must be applied to Solution B to prevent passage of solvent through a perfect semipermeable membrane from A to B. The concentration of a solution with respect to effect on osmotic pressure is related to the number of particles (unionized molecules, ions, macromolecules, aggregates) of solute(s) in solution and thus is affected by the degree of ionization or aggregation of the solute. See Chapter 16 for review of colligative properties of solutions.

Body fluids, including blood and lacrimal fluid, normally have an osmotic pressure that often is described as corresponding to that of a 0.9% solution of sodium chloride. The body also attempts to keep the osmotic pressure of the contents of the gastrointestinal (GI) tract at about this level, but there the normal range is much wider than that of most body fluids. The 0.9% sodium chloride solution is said to be *iso-osmotic* with physiological fluids. In medicine, the term *isotonic*, meaning equal tone, is commonly used interchangeably with *iso-osmotic*. However, terms such as isotonic and tonicity should be used *only* with reference to a physiological fluid. *Iso-osmotic* actually is a physical term that compares the osmotic pressure (or another colligative property, such as freezing-point depression) of two liquids, neither of which may be a physiological fluid, or which may be a physiological fluid only under certain circumstances. For example, a solution of boric acid that is *iso-osmotic* with both blood and lacrimal fluid is *isotonic* only with the lacrimal fluid. This solution causes hemolysis of red blood cells because molecules of boric acid pass freely through the erythrocyte membrane regardless of concentration. Thus, *isotonicity* infers a sense of physiological compatibility where *iso-osmoticity* need not. As another example, a *chemically de-*

finied elemental diet or enteral nutritional fluid can be *iso-osmotic* with the contents of the GI tract, but would not be considered a physiological fluid, or suitable for parenteral use.

A solution is *isotonic* with a living cell if there is no net gain or loss of water by the cell, or other change in the cell, when it is in contact with that solution. Physiological solutions with an osmotic pressure lower than that of body fluids, or of 0.9% sodium chloride solution, are referred to commonly as being *hypotonic*. Physiological solutions having a greater osmotic pressure are termed *hypertonic*.

Such qualitative terms are of limited value, and it has become necessary to state osmotic properties in quantitative terms. To do so, a term must be used that will represent all the particles that may be present in a given system. The term used is *osmol*: the weight, in grams, of a solute, existing in a solution as molecules (and/or ions, macromolecules, aggregates, etc), which is osmotically equivalent to a mole of an ideally behaving nonelectrolyte. Thus, the *osmol* weight of a nonelectrolyte, in a dilute solution, generally is equal to its gram molecular weight. A *milliosmol*, abbreviated *mOsm*, is the weight stated in milligrams.

If one extrapolates this concept of relating an *osmol* and a mole of a nonelectrolyte as being equivalent, then one also may define an *osmol* in the following ways. It is the amount of solute that will provide 1 Avogadro's number (6.02×10^{23}) of particles in solution and it is the amount of solute that, on dissolution in 1 kg of water, will result in an osmotic pressure increase of 17,000 torr at 0° or 19,300 torr at 37°. One *mOsmol* is 1/1000 of an *osmol*. For example, 1 mol of anhydrous dextrose is equal to 180 g. One *osmol* of this nonelectrolyte is also 180 grams. One *mOsmol* would be 180 mg. Thus, 180 mg of this solute dissolved in 1 kg of water will produce an increase in osmotic pressure of 19.3 torr at body temperature.

For a solution of an electrolyte such as sodium chloride, one molecule of sodium chloride represents one sodium and one chloride ion. Hence, 1 mol will represent 2 *osmol* of sodium chloride theoretically. Accordingly, 1 *osmol* NaCl = 58.5 g/2 or 29.25 g. This quantity represents the sum total of 6.02×10^{23} ions as the total number of particles. Ideal solutions infer very dilute solutions or infinite dilution.

However, as the concentration is increased, other factors enter. With strong electrolytes, interionic attraction causes a decrease in their effect on colligative properties. In addition, and in opposition, for all solutes, including nonelectrolytes, solvation and possibly other factors operate to intensify their colligative effect. Therefore, it is very difficult and often impossible to predict accurately the osmoticity of a solution. It may be possible to do so for a dilute solution of a single pure and well-characterized solute, but not for most parenteral and enteral medicinal and/or nutritional fluids; experimental determination likely is required.

THERAPEUTIC CONSIDERATIONS

It generally is accepted that osmotic effects have a major place in the maintenance of homeostasis (the state of equilibrium in the living body with respect to various functions and to the chemical composition of the fluids and tissues, eg, temperature, heart rate, blood pressure, water content, or blood sugar). To a great extent these effects occur within or between cells and tissues where they cannot be measured. One of the most troublesome problems in clinical medicine is the maintenance of adequate body fluids and proper balance between extracellular and intracellular fluid volumes in seriously ill patients. It should be kept in mind, however, that fluid and electrolyte abnormalities are not diseases, but are the manifestations of disease.

The physiological mechanisms that control water intake and output appear to respond primarily to serum osmoticity. Renal regulation of output is influenced by variation in rate of release of pituitary antidiuretic hormone (ADH) and other factors in response to changes in serum osmoticity. Osmotic changes also serve as a stimulus to moderate thirst. This mechanism is sufficiently sensitive to limit variations in osmoticity in the normal individual to less than about 1%. Body fluid continually oscillates within this narrow range. An increase of plasma osmoticity of 1% will stimulate ADH release, result in reduction of urine flow, and, at the same time, stimulate thirst that results in increased water intake. Both the increased renal reabsorption of water (without solute) stimulated by circulating ADH and the increased water intake tend to lower serum osmoticity.

The transfer of water through the cell membrane occurs so rapidly that any lack of osmotic equilibrium between the two fluid compartments in any given tissue usually is corrected within a few seconds and, at most, within a minute or so. However, this rapid transfer of water does not mean that complete equilibration occurs between the extracellular and intracellular compartments throughout the entire body within this same short period of time. The reason is that fluid usually enters the body through the gut and then must be transported by the circulatory system to all tissues before complete equilibration can occur. In the normal person it may require 30 to 60 min to achieve reasonably good equilibration throughout the body after drinking water. Osmoticity is the property that largely determines the physiological acceptability of a variety of solutions used for therapeutic and nutritional purposes.

Pharmaceutical and therapeutic consideration of osmotic effects has been, to a great extent, directed toward the side effects of ophthalmic and parenteral medicinals due to abnormal osmoticity, and either to formulating to avoid the side effects or to finding methods of administration to minimize them. More recently this consideration has been extended to total (central) parenteral nutrition, to enteral hyperalimentation ("tube" feeding), and to concentrated-fluid infant formulas.¹ Also, in recent years, the importance of osmometry of serum and urine in the diagnosis of many pathological conditions has been recognized.

There are a number of examples of the direct therapeutic effect of osmotic action, such as the *intravenous* (IV) use of mannitol as a diuretic that is filtered at the glomeruli and thus increases the osmotic pressure of tubular urine. Water must then be reabsorbed against a higher osmotic gradient than otherwise, so reabsorption is slower and diuresis is observed. The same fundamental principle applies to the IV administration of 30% urea used to affect intracranial pressure in the control of cerebral edema. Peritoneal dialysis fluids tend to be somewhat hyperosmotic to withdraw water and nitrogenous metabolites. Two to 5% sodium chloride solutions or dispersions in an oleaginous base (Muro, *Bausch & Lomb*) and a 40% glucose ointment are used topically for corneal edema. Ophthalgan (*Wyeth-Ayerst*) is ophthalmic glycerin employed for its osmotic effect to clear edematous cornea to facilitate an ophthalmoscopic or gonioscopic examination. Glycerin solutions in 50% concentra-

tion Osmoglyn (*Alcon*) and isosorbide solution Ismotiv (*Alcon*) are oral osmotic agents for reducing intraocular pressure.

The osmotic principle also applies to plasma extenders such as polyvinylpyrrolidone and to saline laxatives such as magnesium sulfate, magnesium citrate solution, magnesium hydroxide (via gastric neutralization), sodium sulfate, sodium phosphate, and sodium biphosphate oral solution, and enema (*Fleet*).

An interesting osmotic laxative that is a nonelectrolyte is a lactulose solution. Lactulose is a nonabsorbable disaccharide that is colon-specific, wherein colonic bacteria degrade some of the disaccharide to lactic and other simple organic acids. These, in toto, lead to an osmotic effect and laxation. An extension of this therapy is illustrated by Cephulac (*Marion Merrell Dow*) solution, which uses the acidification of the colon via lactulose degradation to serve as a trap for ammonia migrating from the blood to the colon. The conversion of ammonia of blood to the ammonium ion in the colon ultimately is coupled with the osmotic effect and laxation, thus expelling undesirable levels of blood ammonia. This product is employed to prevent and treat frontal systemic encephalopathy.

Osmotic laxation is observed with the oral or rectal use of glycerin and sorbitol. Epsom salt has been used in baths and compresses to reduce edema associated with sprains. Another approach is the indirect application of the osmotic effect in therapy via osmotic pump drug delivery systems.²

OSMOLALITY AND OSMOLARITY

It is necessary to use several additional terms to define expressions of concentration in reflecting the osmoticity of solutions. The terms include *osmolality*, the expression of osmolal concentration, and *osmolality*, the expression of osmolar concentration.

OSMOLALITY—A solution has an osmolal concentration of one when it contains 1 osmol of solute/kg of water. A solution has an osmolality of n when it contains n osmol/kg of water. Osmolal solutions, like their counterpart molal solutions, reflect a weight-to-weight relationship between the solute and the solvent. Because an osmol of any nonelectrolyte is equivalent to 1 mol of that compound, then a 1 osmolal solution is synonymous to a 1 molal solution for a typical nonelectrolyte.

With a typical electrolyte like sodium chloride, 1 osmol is approximately 0.5 mol of sodium chloride. Thus, it follows that a 1 osmolal solution of sodium chloride essentially is equivalent to a 0.5 molal solution. Recall that a 1 osmolal solution of dextrose or sodium chloride each will contain the same particle concentration. In the dextrose solution there will be 6.02×10^{23} molecules/kg of water and in the sodium chloride solution one will have 6.02×10^{23} total ions/kg of water, one-half of which are Na^+ ions and the other half Cl^- ions.

As in molal solutions, osmolal solutions usually are employed where quantitative precision is required, as in the measurement of physical and chemical properties of solutions (ie, colligative properties). The advantage of the w/w relationship is that the concentration of the system is not influenced by temperature.

OSMOLARITY—The relationship observed between molality and osmolality is shared similarly between molarity and osmolality. A solution has an osmolar concentration of 1 when it contains 1 osmol of solute per liter of solution. Likewise, a solution has an osmolality of n when it contains n osmols/L of solution. Osmolar solutions, unlike osmolal solution, reflect a weight in volume relationship between the solute and final solution. A 1 molar and 1 osmolar solution would be identical for nonelectrolytes. For sodium chloride a 1 osmolar solution would contain 1 osmol of sodium chloride per liter which approximates a 0.5 molar solution. The advantage of employing osmolar concentrations over osmolal concentrations is the ability to relate a specific number of osmols or milliosmols to a volume, such as a liter or milliliter. Thus, the osmolar concept is simpler and more practical. Volumes of solution, rather than weights of solution, are more practical in the delivery of liquid dosage forms.

Many health professionals do not have a clear understanding of the difference between osmolality and osmolarity. In fact, the terms have been used interchangeably. A 1 osmolar solution of a solute always will be more concentrated than a 1 osmolal solution. With dilute solutions the difference may be acceptably small. For example, a 0.9% *w/v* solution of sodium chloride in water contains 9 g of sodium chloride/L of solution, equivalent to 0.308 osmolar; or 9 g of sodium chloride/996.5 g of water, equivalent to 0.309 osmolal, less than a 1% error. For concentrated solutions the percent difference between osmolarity and osmolality is much greater and may be highly significant; 3.5% for 5% *w/v* dextrose solution and 25% for 25% *w/v* dextrose solution. One should be alerted to the sizable errors that may occur with concentrated solutions or fluids, such as those employed in total parenteral nutrition, enteral hyperalimentation, and oral nutritional fluids for infants.

Reference has been made to the terms hypertonic and hypotonic. Analogous terms are hyperosmotic and hypo-osmotic. Assuming normal serum osmolality to be 285 mOsmol/kg, as serum osmolality increases due to water deficit, the following signs and symptoms usually are found to accumulate progressively at approximately these values: 294 to 298—thirst (if the patient is alert and communicative); 299 to 313—dry mucous membranes; 314 to 329—weakness, doughy skin; above 330—disorientation, postural hypotension, severe weakness, fainting, CNS changes, stupor, and coma. As serum osmolality decreases due to water excess the following may occur: 275 to 261—headache; 262 to 251—drowsiness, weakness; 250 to 233—disorientation, cramps; below 233—seizures, stupor, and coma.

As indicated previously, the mechanisms of the body actively combat such major changes by limiting the variation in osmolality for normal individuals to less than about 1% (approximately in the range 282–288 mOsmol/kg, based on the above assumption).

The value given for normal serum osmolality above was described as an assumption because of the variety of values found in the literature. Serum osmolality often is stated loosely to be about 300 mOsmol/L. Various references report 280 to 295 mOsmol/L, 275 to 300 mOsmol/L, 290 mOsmol/L, 306 mOsmol/L, and 275 to 295 mOsmol/kg.

In recent years, much attention has been directed at determining osmoticity of total parenteral nutrition solutions, enteral formulas, and parenteral and enteral medications.^{3–5} Hyperosmoticity of parenteral and enteral formulas and medications serves as an indicator for potential risks, including thrombophlebitis, pain at injection site, diarrhea, and abdominal cramping. However, the terms osmolality and osmolarity often have been used interchangeably and caused much confusion for practitioners. Often, when the term osmolarity is used, one cannot discern whether this simply is incorrect terminology, or if osmolarity actually has been calculated from osmolality.

Another current practice that can cause confusion is the use of the terms *normal* or *physiological* for isotonic sodium chloride solution (0.9%). The solution surely is iso-osmotic. However, as to being physiological, the concentration of ions are each of 154 mEq/L whereas serum contains about 140 mEq of sodium and about 103 mEq of chloride.

The range of mOsmol values found for serum raises the question as to what really is meant by the terms hypotonic and hypertonic for medicinal and nutritional fluids. One can find the statement that fluids with an osmolality of 50 mOsmol or more above normal are hypertonic; and, if they are 50 mOsmol or more below normal, they are hypotonic. One also can find the statement that peripheral infusions should not have an osmolarity exceeding 700 to 800 mOsmol/L.⁶ Examples of osmol concentrations of solutions used in peripheral infusions are (D5W) 5% dextrose solution, 252 mOsmol/L; (D10W) 10% dextrose solution, 505 mOsmol/L; and Lactated Ringer's 5% Dextrose, 525 mOsmol/L. When a fluid is hypertonic, undesirable effects often can be decreased by using relatively slow rates of infusion, and/or relatively short periods of infusion. For example, 25%

dextrose solution (D25W)—4.25% Amino Acids is a representative of a highly osmotic hyperalimentation solution. It has been stated that when osmolal loading is needed, a maximum safe tolerance for a normally hydrated subject would be an approximate increase of 25 mOsmol/kg of water over 4 hours.⁷

COMPUTATION OF OSMOLARITY

Several methods are used to obtain numerical values of osmolarity. The osmolar concentration, sometimes referred to as the *theoretical osmolarity*, is calculated from the *w/v* concentration using the following equation:

$$\frac{\text{g}}{\text{L}} \times \frac{\text{mols}}{\text{g}} \times \frac{\text{osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{osmol}} = \frac{\text{mOsmol}}{\text{L}} \quad (1)$$

The number of osmol/mol is equal to 1 for nonelectrolytes and is equal to the number of ions per molecule for strong electrolytes.

This calculation omits consideration of factors such as solvation and interionic forces. By this method of calculation, 0.9% sodium chloride has an osmolar concentration of 308 mOsmol/L and a concentration of 154 mOsmol/L in either sodium or chloride ion.

Two other methods compute osmolarity from values of osmolality. The determination of osmolality will be discussed later. One method has a strong theoretical basis of physical-chemical principles⁸ using values of the partial molal volume(s) of the solute(s). A 0.9% sodium chloride solution, found experimentally to have an osmolality of 286 mOsmol/kg, was calculated to have an osmolarity of 280 mOsmol/L, rather different from the value of 308 mOsmol/L calculated as above. The method, using partial molal volumes, is relatively rigorous, but many systems appear to be too complex and/or too poorly defined to be dealt with by this method.

The other method is based on calculating the weight of water from the solution density and concentration

$$\frac{\text{g water}}{\text{mL solution}} = \frac{\text{g solution}}{\text{mL solution}} - \frac{\text{g solute}}{\text{mL solution}}$$

then

$$\begin{aligned} \text{osmolarity} \left(\frac{\text{mOsmol}}{\text{L solution}} \right) \\ = \text{osmolality} \left(\frac{\text{mOsmol}}{1000 \text{ g water}} \right) \times \frac{\text{g water}}{\text{mL solution}} \end{aligned}$$

The experimental value for the osmolality of 0.9% sodium chloride solution was 292.7 mOsmol/kg; the value computed for osmolarity was 291.4 mOsmol/L. This method uses easily obtained values of density of the solution and of its solute content and can be used with all systems. For example, the osmolality of a nutritional product was determined by the freezing-point depression method to be 625 mOsmol/kg¹⁰; its osmolarity was calculated as 625 × 0.839 = 524 mOsmol/L.

Monographs in the USP for solutions provide IV replenishment of fluid, nutrients, or electrolytes, and for osmotic diuretics such as Mannitol Injection, require the osmolar concentration be stated on the label in osmol/L; however, when the contents are less than 100 mL, or when the label states the article is not for direct injection but is to be diluted before use, the label alternatively may state the total osmolar concentration in mOsmol/mL.

An example of the use of the first method described above is the computation of the approximate osmolar concentration (*theoretical osmolarity*) of a Lactated Ringer's 5% Dextrose Solution (Abbott), which is labeled to contain, per liter, dextrose (hydrous) 50 g, sodium chloride 6 g, potassium chloride 300 mg, calcium chloride 200 mg, and sodium lactate 3.1 g. Also stated is that the total osmolar concentration of the solution is approximately 524 mOsmol/L in part contributed by 130 mEq of Na⁺, 109 mEq of Cl⁻, 4 mEq of K⁺, 3 mEq of Ca²⁺, and 28 mEq of lactate ion.

The derivation of the osmolar concentrations from the stated composition of the solution may be verified by calculations using Equation 1.

Dextrose

$$\frac{50 \text{ g}}{\text{L}} \times \frac{1 \text{ mol}}{198 \text{ g}} \times \frac{1 \text{ osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{Osmol}} = 252 \text{ mOsmol/L}$$

Sodium Chloride

$$\begin{aligned} \frac{6 \text{ g}}{\text{L}} \times \frac{1 \text{ mol}}{58.4 \text{ g}} \times \frac{2 \text{ osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{osmol}} \\ = 205 \frac{\text{mOsmol}}{\text{L}} \left\{ \begin{array}{l} (102.7 \text{ mOsmol Na}^+) \\ (102.7 \text{ mOsmol Cl}^-) \end{array} \right. \end{aligned}$$

Potassium Chloride

$$\begin{aligned} \frac{0.3 \text{ g}}{\text{L}} \times \frac{1 \text{ mol}}{74.6 \text{ g}} \times \frac{2 \text{ osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{osmol}} \\ = \frac{8.04 \text{ mOsmol}}{\text{L}} \left\{ \begin{array}{l} (4.02 \text{ mOsmol K}^+) \\ (4.02 \text{ mOsmol Cl}^-) \end{array} \right. \end{aligned}$$

Calcium Chloride

$$\begin{aligned} \frac{0.2 \text{ g}}{\text{L}} \times \frac{1 \text{ mol}}{111 \text{ g}} \times \frac{3 \text{ osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{osmol}} \\ = \frac{5.41 \text{ mOsmol}}{\text{L}} \left\{ \begin{array}{l} (1.80 \text{ mOsmol Ca}^{2+}) \\ (3.61 \text{ mOsmol Cl}^-) \end{array} \right. \end{aligned}$$

Sodium Lactate

$$\begin{aligned} \frac{3.1 \text{ g}}{\text{L}} \times \frac{1 \text{ mol}}{112 \text{ g}} \times \frac{2 \text{ osmol}}{\text{mol}} \times \frac{1000 \text{ mOsmol}}{\text{osmol}} \\ = \frac{55.4 \text{ mOsmol}}{\text{L}} \left\{ \begin{array}{l} (27.7 \text{ mOsmol Na}^+) \\ (27.7 \text{ mOsmol lactate}) \end{array} \right. \end{aligned}$$

The total osmolar concentration of the five solutes in the solution is 526, in good agreement with the labeled total osmolar concentration of approximately 524 mOsmol/L.

The mOsmol of sodium in 1 L of the solution is the sum of the mOsmol of the ion from sodium chloride and sodium lactate: $102 + 27.6 = 129.6$ mOsmol. Chloride ions come from the sodium chloride, potassium chloride, and calcium chloride, the total osmolar concentration being $102 + 4.02 + 3.61 = 109.6$ mOsmol. The mOsmol values of potassium, calcium, and lactate are calculated to be 4.02, 1.80, and 27.6, respectively.

The osmolality of a mixture of complex composition, such as an enteral hyperalimentation fluid, cannot be calculated with any acceptable degree of certainty; therefore, the *osmolality* of such preparations should be determined experimentally.

OSMOMETRY AND THE CLINICAL LABORATORY

Serum and urine osmometry may assist in the diagnosis of certain fluid and electrolyte problems. However, osmometry values have little meaning unless the clinical situation is known. Osmometry is used in renal dialysis as a check on the electrolyte composition of the fluid. In the clinical laboratory, as stated above, the term *osmolality* is used generally, but usually is reported as mOsmol/L. It may seem unnecessary to mention that osmolality depends not only on the number of solute particles, but also on the quantity of water in which they are dissolved. However, it may help one to understand the statement that the normal range of urine osmolality is 50 to 1400 mOsmol/L, and for a random specimen is 500 to 800 mOsmol/L.

Serum Osmoticity

Sodium is by far the principal solute involved in serum osmoticity. Therefore, abnormal serum osmoticity is most likely

to be associated with conditions that cause abnormal sodium concentration and/or abnormal water volume.

Thus, hyperosmotic serum is likely to be caused by an increase in serum sodium and/or loss of water. It may be associated with diabetes insipidus, hypercalcemia, diuresis during severe hyperglycemia, or with early recovery from renal shutdown. Alcohol ingestion is said to be the most common cause of the hyperosmotic state and of coexisting coma and the hyperosmotic state. An example of hyperosmoticity is a comatose diabetic with a serum osmoticity of 365 mOsmol/L.

In a somewhat analogous fashion, hypo-osmotic serum is likely to be due to decrease in serum sodium and/or excess of water. It may be associated with the postoperative state (especially with excessive water replacement therapy), treatment with diuretic drugs and low-salt diet (as with patients with heart failure, cirrhosis, etc), adrenal disease (eg, Addison's disease, adrenogenital syndrome), or SIADH (syndrome of inappropriate ADH secretion). There are many diseases that cause ADH to be released inappropriately (ie, in spite of serum osmoticity and volume having been normal initially). These include oat-cell carcinoma of the lung, bronchogenic carcinoma, congestive heart failure, inflammatory pulmonary lesions, porphyria, severe hypothyroidism, or cerebral disease (such as tumor, trauma, infection, and vascular abnormalities). It also may be found with some patients with excessive diuretic use. Serum and urine osmoticity are measured when SIADH is suspected. In SIADH there is hypo-osmoticity of the blood in association with a relative hyperosmoticity of urine. The usual cause is a malfunction of the normal osmotic response of osmoreceptors, an excess of exogenous vasopressin, or a production of a vasopressin-like hormone that is not under the regular control of serum osmoticity. The diagnosis is made by simultaneous measurement of urine and serum osmolality. The serum osmolality will be lower than normal and much lower than the urine osmolality, indicating inappropriate secretion of a concentrated urine in the presence of a dilute serum.

Cardiac, renal, and hepatic disease characteristically reduce the sodium/osmolality ratio, this being partially attributed to the effects of increased blood sugar, urea, or unknown metabolic products. Patients in shock may develop disproportionately elevated measured osmolality compared to calculated osmolality, which points toward the presence of circulating metabolic products.

There are several approximate methods for estimating serum osmolality from clinical laboratory values for sodium ion. They may be of considerable value in an emergency situation.

1. Serum osmolality may be estimated from

$$\text{mOsmol} = (1.86 \times \text{sodium}) + \frac{\text{blood sugar}}{18} + \frac{\text{BUN}}{2.8} + 5$$

(Na in mEq/L, blood sugar and BUN in mg/100 mL).

2. A quick approximation is

$$\text{mOsmol} = 2 \text{ Na} + \frac{\text{BS}}{20} + \frac{\text{BUN}}{3}$$

3. The osmolality is usually, *but not always*, very close to two times the sodium reading plus 10.

Urine Osmoticity

The two main functions of the kidney are glomerular filtration and tubular reabsorption. Clinically, tubular function is measured best by tests that determine the ability of the tubules to concentrate and dilute the urine. Tests of urinary dilution are not as sensitive in the detection of disease, as are tests of urinary concentration. As concentration of urine occurs in the renal medulla (interstitial fluids, loops of Henle, capillaries of the medulla, and collecting tubules), the disease processes that disturb the function or structure of the medulla produce early impairment of the concentrating power of the kidney.

Such diseases include acute tubular necrosis, obstructive uropathy, pyelonephritis, papillary necrosis, medullary cysts, hypokalemic and hypercalcemic nephropathy, and sickle cell disease.

Measurement of urine osmolality is an accurate test for the diluting and concentrating ability of the kidneys. In the absence of ADH, the daily urinary output is likely to be 6 to 8 liters or more. The normal urine osmolality depends on the clinical setting; normally, with maximum ADH stimulation, it can be as much as 1200 mOsmol/kg, and with maximum ADH suppression as little as 50 mOsmol/kg. Simultaneous determination of serum and urine osmolality often is valuable in assessing the distal tubular response to circulating ADH. For example, if the patient's serum is hyperosmolar, or in the upper limits of normal ranges, and the patient's urine osmolality measured at the same time is much lower, a decreased responsiveness of the distal tubules to circulating ADH is suggested.

Measurement of urine osmolality during water restriction is an accurate, sensitive test of decreased renal function. For example, under the conditions of one test, normal osmolality would be greater than 800 mOsmol/kg. With severe impairment the value would be less than 400 mOsmol/kg. Knowledge of urine osmolality may point to a problem even though other tests are normal (eg, the Fishberg concentration test, blood urea nitrogen, PSP excretion, creatinine clearance, or IV pyelogram). Knowledge of its value may be useful especially in diabetes mellitus, essential hypertension, and silent pyelonephritis. The urine/serum osmolality ratio should be calculated and should be equal to or greater than 3.

UNDESIRABLE EFFECTS OF ABNORMAL OSMOTICITY

OPHTHALMIC MEDICATION—It is generally accepted that ophthalmic preparations intended for instillation into the cul-de-sac of the eye should, if possible, be approximately isotonic to avoid irritation (see Chapter 43). It also has been stated that the abnormal tonicity of contact lens solutions can cause the lens to adhere to the eye and/or cause burning or dryness and photophobia.

PARENTERAL MEDICATION—Osmoticity is of great importance in parenteral injections, its effects depending on such factors as the degree of deviation from tonicity, the concentration, the location of the injection, the volume injected, the speed of the injection, and the rapidity of dilution and diffusion, etc. When formulating parenterals, solutions otherwise hypotonic usually have their tonicity adjusted by the addition of dextrose or sodium chloride. Hypertonic parenteral drug solutions cannot be adjusted. Hypotonic and hypertonic solutions usually are administered slowly in small volumes, or into a large vein such as the subclavian, where dilution and distribution occur rapidly. Solutions that differ from the serum in tonicity generally cause tissue irritation, pain on injection, and electrolyte shifts, the effect depending on the degree of deviation from tonicity:

Excessive infusion of *hypotonic* fluids may cause swelling of red blood cells, hemolysis, and water invasion of the body's cells in general. When this is beyond the body's tolerance for water, water intoxication results, with convulsions and edema, such as pulmonary edema.

Excessive infusion of *isotonic* fluids can cause an increase in extracellular fluid volume, which can result in circulatory overload.

Excessive infusion of *hypertonic* fluids leads to a wide variety of complications. For example, the sequence of events when the body is presented with a large IV load of hypertonic fluid, rich in dextrose, is as follows: hyperglycemia, glycosuria and intracellular dehydration, osmotic diuresis, loss of water and electrolytes, dehydration, and coma.

One cause of osmotic diuresis is the infusion of dextrose at a rate faster than the ability of the patient to metabolize it (as

greater than perhaps 400–500 mg/kg per hour for an adult on total parenteral nutrition). A heavy load of unmetabolizable dextrose increases the osmoticity of blood and acts as a diuretic; the increased solute load requires more fluid for excretion, 10 to 20 mL of water being required to excrete each gram of dextrose. Solutions such as those for total parenteral nutrition should be administered by means of a metered constant-infusion apparatus over a lengthy period (usually more than 24 hr) to avoid sudden hyperosmotic dextrose loads. Such solutions may cause osmotic diuresis; if this occurs, water balance is likely to become negative because of the increased urinary volume, and electrolyte depletion may occur because of excretion of sodium and potassium secondary to the osmotic diuresis. If such diuresis is marked, body weight falls abruptly and signs of dehydration appear. Urine should be monitored for signs of osmotic diuresis, such as glycosuria and increased urine volume.

If the IV injection rate of hypertonic solution is too rapid, there may be catastrophic effects on the circulatory and respiratory systems. Blood pressure may fall to dangerous levels, cardiac irregularities or arrest may ensue, respiration may become shallow and irregular, and there may be heart failure and pulmonary edema. Probably the precipitating factor is a bolus of concentrated solute suddenly reaching the myocardium and the chemoreceptors in the aortic arch and carotid sinus.⁷

Abrupt changes in serum osmoticity can lead to cerebral hemorrhage. It has been shown experimentally that rapid infusions of therapeutic doses of hypertonic saline with osmotic loads produce a sudden rise in cerebrospinal fluid (CSF) pressure and venous pressure (VP) followed by a precipitous fall in CSF pressure. This particularly may be conducive to intracranial hemorrhage, as the rapid infusion produces an increase in plasma volume and venous pressure at the same time the CSF pressure is falling. During the CSF pressure rise, there is a drop in hemoglobin and hematocrit, reflecting a marked increase in blood volume.

Hyperosmotic medications, such as sodium bicarbonate (osmolarity of 1560 at 1 mEq/mL), which are administered intravenously, should be diluted prior to use and should be injected slowly to allow dilution by the circulating blood. Rapid *push* injections may cause a significant increase in blood osmoticity.⁸

As to other possibilities, there may be crenation of red blood cells and general cellular dehydration. Hypertonic dextrose or saline infused through a peripheral vein with small blood volume may traumatize the vein and cause thrombophlebitis. Infiltration can cause trauma and necrosis of tissues. Safety, therefore, demands that all IV injections, especially highly osmotic solutions, be performed slowly, usually being given preferably over a period not less than that required for a complete circulation of the blood, for example, 1 min. The exact danger point varies with the state of the patient, the concentration of the solution, the nature of the solute, and the rate of administration.

Hyperosmotic solutions also should not be discontinued suddenly. In dogs, marked increase in levels of intracranial pressure occur when hyperglycemia produced by dextrose infusions is reversed suddenly by stopping the infusion and administering saline. It also has been shown that the CSF pressure in humans rises during treatment of diabetic ketoacidosis in association with a fall in the plasma concentration of dextrose and a fall in plasma osmolality. These observations may be explained by the different rates of decline in dextrose content of the brain and of plasma. The concentration of dextrose in the brain may fall more slowly than in the plasma, causing a shift of fluid from the extracellular fluid space to the intracellular compartment of the CNS, resulting in increased intracranial pressure.

Clinical Applications

Although there are many issues with abnormal osmoticity, most pharmacists are concerned with preventable adverse ef-

fects such as thrombophlebitis and pain at the injection site. The understanding of these potential risks from hyperosmotic parenteral medications has fine-tuned IV administration techniques. The site of administration—peripheral versus central venous catheter—plays a significant role in determining the final concentration of parenteral medications infused IV. Attention should be directed toward establishing the optimal osmolality of IV administered parenteral medications via the peripheral venous route that will result in the least adverse effects.

Since the introduction of parenteral nutrition support, hyperosmoticity of these nutrition solutions remains a concern. The commonly accepted osmolality of less than 900 mOsmol/L has been quoted for safe peripheral administration of parenteral nutrition solutions.^{11,12} All attempts should be made to prepare solutions with osmoticity close to that of serum osmoticity or no greater than 900 mOsmol/L. This can be achieved by carefully selecting the diluent for dilution and determining the final concentration of the parenteral medication. Dextrose 5% in Water for Injection and Sodium Chloride 0.9% have been used routinely as diluents. When comparing the two diluents, parenteral medications diluted with Dextrose 5% in Water for Injection have a lower osmolality than do solutions diluted with Sodium Chloride 0.9% at the same final concentration.

Several studies have been conducted to determine optimal final concentration of commonly used parenteral medications.^{3–5} The published final concentrations for most parenteral medications are recommended for peripheral as well as central venous catheter IV administration for patients with no special needs, such as fluid restriction. In the event that fluid restriction is required or the recommended final concentration is not achievable, the parenteral medication should be administered via a central venous catheter, where immediate dilution and distribution is achieved rapidly. This will minimize potential for the phlebitis and pain at the injection site.

Osmoticity issues associated with parenteral medications are also applicable to *total parenteral nutrition* (TPN) solutions, especially via peripheral venous administration. Peripheral parenteral nutrition support remains an integral part of therapeutic options for hospitalized patients. The peripheral route of administration often is preferred for patients who require short-term therapy or supplemental nutrition support.

In clinical practice, however, many institutions use the macronutrient dextrose as the sole determinant for the safety of peripheral parenteral nutrition administration. For example, the approximate osmolality of dextrose is 50 mOsmol/% of dextrose. Thus, a 10% dextrose solution equals 500 mOsmol/L. It is assumed that with *normal* protein and micronutrient requirements, the final osmolality is estimated to be approximately 900 mOsmol/L. Therefore, guidelines for most institutions recommend any parenteral nutrition solution with a dextrose concentration less than or equal to 10% is safe for peripheral administration, irrespective of other components. Conversely, a parenteral nutrition solution with a final dextrose concentration greater than 10% should not be administered peripherally and should be considered for central venous catheter administration. Although this method appears to be practical and provides quick decision-making ability, it ignores the contributions of the other components, restricts its validity to adult parenteral nutrition solutions with *normal* protein and micronutrient requirements, and does not address neonatal and pediatric parenteral nutrition solutions. Because of the different fluid and nutrient requirements of neonates and pediatric patients, the final concentration of dextrose and amino acids is generally greater to provide the calories and protein requirements in a smaller volume of liquid. For example, protein requirements of neonates are much higher compared with adult requirements, 3 g/kg/day versus 1 g/k/day. Thus, the final percentage of amino acid in neonatal parenteral nutrition solution is generally higher. Coupled with an approximate osmolality of amino acid equal to 100 mOsmol/%, amino acids may

contribute equally to the final osmolality of a parenteral nutrition solution. Therefore, components other than dextrose cannot be ignored.

Currently, most institutions use automated compounding systems to prepare parenteral nutrition solutions. These systems often are computerized and include programs that will calculate the osmolality of the final parenteral nutrition solution. This has helped clinicians determine the safety of parenteral nutrition solutions with various macro- and micronutrient combinations, thereby accounting for all components of parenteral nutrition solutions.

OSMOTICITY AND ENTERAL HYPERALIMENTATION

Some aspects of nutrition are discussed briefly here because of the potential major side effects due to abnormal osmoticity of nutritional fluids, and because there exists increasing dialogue on nutrition among pharmacists, dietitians, nurses, and physicians. The professional organization ASPEN (The American Society for Parenteral and Enteral Nutrition), for example, has a membership open to all of the above health practitioners. Pharmacists should be able to discuss these matters with other health professionals in terms of nutrition as well as medicine.

Osmoticity has been of special importance in the IV infusion of large volumes of highly concentrated nutritional solutions. Their hyperosmoticity has been a major factor in the requirement that they be injected centrally into a large volume of rapidly moving blood, instead of using peripheral infusion. Use of such solutions and knowledge of their value have led, more recently, to the use of similar formulations administered, not parenterally, but by instillation into some part of the GI tract, orally, by nasogastric tube, via feeding gastrostomy, or by needle-catheter jejunostomy. This method has given excellent total nutrition, for a period of time, to many patients and obviously avoids some of the problems associated with injections.

Enteral nutritional formulas can be modular, allowing individual supplementation of protein, carbohydrate, or fat. Other formulas are called *defined formula diets* and contain protein, carbohydrate, fat, minerals, and vitamins. These nutritionally complete formulations can be monomeric (or oligomeric), based on amino acids, short peptides, and simple carbohydrates, or can be polymeric, based on complex protein and carbohydrates.

These diets are necessarily relatively high in osmoticity because their smaller molecules result in more particles per gram than in normal foods. An example is a fluid consisting of L-amino acids, dextrose oligosaccharides, vitamins (including fat-soluble vitamins), fat as a highly purified safflower oil or soybean oil, electrolytes, trace minerals, and water. As it contains fat, that component is not in solution and therefore should have no direct effect on osmoticity. However, the potential for interactions can cause some significant changes in total particle concentration and indirectly affect the osmoticity.¹³

Although it is easily digested, dextrose contributes more particles than most other carbohydrate sources such as starch, and is more likely to cause osmotic diarrhea, especially with bolus feeding. Osmoticity is improved (decreased) by replacing dextrose with dextrose oligosaccharides (carbohydrates that yield on hydrolysis 2 to 10 monosaccharides). Flavoring also increases the osmoticity of a product, different flavors causing varying increases.

Commercial diets are packaged as fluids or as powders for reconstitution. Reconstitution is usually with water. These products are categorized on caloric density, (calories/mL), protein content, or osmolality (mOsm/kg of H₂O). Parenteral nutritional products, on the other hand, are labeled in terms of osmolality (mOsm/L).

The enteric route for hyperalimentation frequently is overlooked in many diseases or post-trauma states, if the patient

is not readily responsive to traditional oral feedings. Poor appetite, chronic nausea, general apathy, and a degree of somnolence or sedation are common concomitants of serious disease. This frequently prevents adequate oral alimentation and results in progressive energy and nutrient deficits. Often, supplementary feedings of a highly nutritious formula are taken poorly or refused entirely. However, the digestive and absorptive capabilities of the GI tract are frequently intact and, when challenged with appropriate nutrient fluids, can be used effectively. By using an intact GI tract for proper alimentation, the major problems of sepsis and metabolic derangement that relate to IV hyperalimentation largely are obviated, and adequate nutritional support is simplified greatly. Because of this increased safety and ease of administration, the enteric route for hyperalimentation should be used whenever possible.¹⁴

When certain foods are ingested in large amounts or as concentrated fluids, their osmotic characteristics can cause an upset in the normal water balance within the body. For a given weight of solute the osmolality of the solution is inversely proportional to the size of the particles. Nutritional components can be listed in an approximate order of decreasing osmotic effect per gram, as¹⁵

1. Electrolytes such as sodium chloride
2. Relatively small organic molecules such as dextrose (glucose) and amino acids
3. Dextrose oligosaccharides
4. Starches
5. Proteins
6. Fats (as fats are not water soluble, they have no osmotic effect)

Thus, in foods, high proportions of electrolytes, amino acids, and simple sugars have the greatest effect on osmolality and, as a result, on tolerance. The approximate osmolality of a few common foods and beverages is

	mOsmol/kg
Whole milk	295
Tomato juice	595
Orange juice	935
Ice cream	1150

When nutrition of high osmoticity is ingested, large amounts of water will transfer to the stomach and intestines from the fluid surrounding those organs in an attempt to lower the osmoticity. The higher the osmoticity, the larger the amount of water required; a large amount of water in the GI tract can cause distention, cramps, nausea, vomiting, hypermotility, and shock. The food may move through the tract too rapidly for the water to be reabsorbed, and result in diarrhea; severe diarrhea can cause dehydration. The hyperosmotic enteral effects have been observed by the administration of undiluted hypertonic oral medication.¹⁶⁻¹⁷ Table I from this work lists average osmolality values of some commercially available drug solutions and suspensions. Thus, there is some analogy to the effect of hyperosmotic IV infusions.

Hyperosmotic feedings may result in mucosal damage in the GI tract. Rats given hyperosmotic feeding showed transient decrease in disaccharidase activity, and an increase in alkaline phosphatase activity. They also showed morphological alterations in the microvilli of the small intestines. After a period of severe gastroenteritis, the bowel may be unusually susceptible to highly osmotic formulas, and their use may increase the frequency of diarrhea. Infant formulas that are hyperosmotic may affect preterm infants adversely during the early neonatal period, and they may produce or predispose neonates to necrotizing enterocolitis when the formulas delivered to the jejunum through a nasogastric tube. The body attempts to keep the osmoticity of the contents of the stomach and intestines at approximately the same level as that of the fluid surrounding them. As a fluid of lower osmoticity requires the transfer of less water to dilute it, it should be tolerated better than one of higher osmoticity.

As to tolerance, there is a great variation from one individual to another in sensitivity to the osmoticity of foods. The majority of patients receiving nutritional formulas, either orally or by tube, are able to tolerate feedings with a wide range of osmoticities when the formulas are administered slowly and when adequate additional fluids are given. However, certain patients are more likely to develop symptoms of intolerance when receiving fluids of high osmoticity. These include debilitated patients, patients with GI disorders, pre- and postoperative patients, gastrostomy- and jejunostomy-fed patients, and patients whose GI tracts have not been challenged for an extended period of time. Thus, osmoticity always should be considered in the selection of the formula for each individual patient.

With all products, additional fluid intake may be indicated for individuals with certain clinical conditions. Frequent feedings of small volume or a continual instillation (pumped) may be of benefit initially in establishing tolerance to a formula. For other than iso-osmotic formulas, feedings of reduced concentration (osmolality less than 400 mOsmol/kg) also may be helpful initially if tolerance problems arise in sensitive individuals. Concentration and size of feeding then can be increased gradually to normal as tolerance is established.

A common disturbance of intake encountered in elderly individuals relates to excess solid intake rather than to reduced water intake. For example, an elderly victim of a cerebral vascular accident who is being fed by nasogastric tube may be given a formula whose solute load requires a greatly increased water intake. Thus, tube feeding containing 120 g of protein and 10 g of salt will result in the excretion of more than 1000 mOsmol of solute. This requires the obligatory excretion of a volume of urine between 1200 and 1500 mL when the kidneys are capable of normal concentration ability. As elderly individuals often have significant impairment in renal function, water loss as urine may exceed 2000 to 2500 mL per day. Such an individual would require 3 to 4 liters of water per day simply to meet the increased demand created by this high solute intake. Failure of the physician to provide such a patient with the increased water intake needed will result in a progressive water deficit that rapidly may become critical. The importance of knowing the complete composition of the tube feeding formulas used for incapacitated patients cannot be overemphasized.

OSMOLALITY DETERMINATION

The need for experimental determination of osmolality has been established. In regard to this there are four properties of solutions that depend only on the number of *particles* in the solution. They are *osmotic-pressure elevation*, *boiling-point elevation*, *vapor-pressure depression*, and *freezing-point depression*. These are called *colligative properties*, and if one of them is known, the others can be calculated from its value. Osmotic-pressure elevation is the most difficult to measure satisfactorily. The boiling-point elevation may be determined, but the values are rather sensitive to changes in barometric pressure. Also, for an aqueous solution the molal boiling point elevation is considerably less than the freezing-point depression. Thus, it is less accurate than the freezing-point method. Determinations of vapor-pressure lowering are quite easy, rapid, and convenient. A vapor pressure osmometer with a precision of <2 mOsmol/kg is reported by Dickerson et al.¹⁶ Another commonly used method is that of freezing-point depression, which can be determined quite readily with a fair degree of accuracy (see *Freezing-Point Depression* in Chapter 16). It should be noted that the data in Appendix A can be converted readily to vapor-pressure lowering if desired.

The results of investigations by Lund et al¹⁸ indicate that the freezing point of normal, healthy human blood is -0.52° . Inasmuch as water is the medium in which the various constituents

of blood are either suspended or dissolved in this method, it is assumed that *any aqueous solution* freezing at -0.52° is *isotonic with blood*. Now it is rare that a simple aqueous solution of the therapeutic agent to be injected parenterally has a freezing point of -0.52° , and to obtain this freezing point it is necessary either to add some other therapeutically inactive solute if the solution is hypotonic (freezing point above -0.52°) or to dilute the solution if it is hypertonic (freezing point below -0.52°). The usual practice is to add either sodium chloride or dextrose to adjust hypotonic parenteral solutions to isotonicity. Certain solutes, including ammonium chloride, boric acid, urea, glycerin, and propylene glycol, cause hemolysis even when they are present in a concentration that is iso-osmotic, and such solutions obviously are not isotonic. See Appendix A.

In a similar manner solutions intended for ophthalmic use may be adjusted to have a freezing point identical to that of lacrimal fluid, namely -0.52° . Ophthalmic solutions with higher freezing points usually are made isotonic by the addition of boric acid or sodium chloride.

In laboratories where the necessary equipment is available, the method usually followed for adjusting hypotonic solutions is to determine the freezing-point depression produced by the ingredients of a given prescription or formula, and then to add a quantity of a suitable inert solute calculated to lower the freezing point to -0.52° , whether the solution is for parenteral injection or ophthalmic application. A final determination of the freezing-point depression may be made to verify the accuracy of the calculation. If the solution is hypertonic, it must be diluted if an isotonic solution is to be prepared, but it must be remembered that some solutions cannot be diluted without impairing their therapeutic activity. For example, solutions to be used for treating varicose veins require a high concentration of the active ingredient (solute) to make the solution effective. Dilution to isotonic concentration is not indicated in such cases.

FREEZING-POINT CALCULATIONS

As explained in the preceding section, freezing-point data often may be employed in solving problems of isotonicity adjustment. Obviously, the utility of such data is limited to those solutions where the solute does not penetrate the membrane of the tissue (eg, red blood cells) with which it is in contact. In such cases, Appendix A, which gives the freezing-point depression of solutions of different concentrations of various substances, provides information essential for solving the problem.

For most substances listed in the table, the concentration of an isotonic solution (one that has a freezing point of -0.52°) is given. If this is not listed in the table, it may be determined with sufficient accuracy by simple proportion using, as the basis for calculation, the figure that most nearly produces an isotonic solution. Actually the depression of the freezing point of a solution of an electrolyte is not absolutely proportional to the concentration but varies according to dilution; for example, a solution containing 1 g of procaine hydrochloride in 100 mL has a freezing-point depression of 0.12° , whereas a solution containing 3 g of the same salt in 100 mL has a freezing-point depression of 0.33° , not 0.36° ($3 \times 0.12^\circ$). Because the adjustment to isotonicity need not be absolutely exact, approximations may be made. Nevertheless, adjustments to isotonicity should be as exact as practicable.

EFFECT OF SOLVENTS—Besides water, certain other solvents frequently are employed in nose drops, eardrops, and other preparations to be used in various parts of the body. Liquids such as glycerin, propylene glycol, or alcohol may compose part of the solvent. In solving isotonicity adjustment problems for such solutions, it should be kept in mind that these solvent components contribute to the freezing-point depression but they may or may not have an effect on the *tone* of the tissue to which they are applied; thus, an *iso-osmotic* solution may not be *isotonic*. In such cases, it is apparent that the utility of the

methods described above—or for that matter, of any other method of evaluating *tonicity*—is questionable.

TONICITY TESTING BY OBSERVING ERYTHROCYTE CHANGES

Observation of the behavior of human erythrocytes when suspended in a solution is the ultimate and direct procedure for determining whether the solution is isotonic, hypotonic, or hypertonic. If hemolysis or marked change in the appearance of the erythrocytes occurs, the solution is not isotonic with the cells. If the cells retain their normal characteristics, the solution is isotonic.

Hemolysis may occur when the osmotic pressure of the fluid in the erythrocytes is greater than that of the solution in which the cells are suspended, but the specific chemical reactivity of the solute in the solution often is far more important in producing hemolysis than is the osmotic effect. There is no certain evidence that any single mechanism of action causes hemolysis. The process appears to involve such factors as pH, lipid solubility, molecular and ionic sizes of solute particles, and possibly inhibition of cholinesterase in cell membranes and denaturing action on plasma membrane protein.

Some investigators test the tonicity of injectable solutions by observing variations of red blood cell volume produced by these solutions. This method appears to be more sensitive to small differences in tonicity than those based on observation of a hemolytic effect. Much useful information concerning the effect of various solutes on erythrocytes has been obtained by this procedure.

METHODS OF ADJUSTING TONICITY

There are several methods for adjusting the tonicity of an aqueous solution, provided, of course, that the solution is hypotonic when the drug and additives are dissolved. The most prominent of these methods are the freezing-point depression method, the sodium chloride equivalent method, and the isotonic solution *V*-value method. The first two of these methods can be used with a three-step problem-solving process based on sodium chloride.

1. Identify a reference solution and the associated tonicity parameter.
2. Determine the contribution of the drug(s) and additive(s) to the total tonicity.
3. Determine the amount of sodium chloride needed by subtracting the contribution of the actual solution from the reference solution.

The result of the third step also indicates whether the actual solution is hypotonic, isotonic, or hypertonic. If the actual solution contributes less to the total tonicity than the reference solution, then the actual solution is hypotonic. If, however, the actual solution contributes a greater amount to tonicity than the reference solution, the actual solution is hypertonic and can be adjusted to isotonicity only by dilution. This may not be possible on therapeutic grounds.

The amount of sodium chloride resulting in the third step also can be converted into an amount of other materials, such as dextrose, to render the actual solution isotonic.

FREEZING-POINT-DEPRESSION METHOD—The freezing-point method makes use of a *D* value (found in Appendix A) which has the units of degree centigrade/(*x*% drug). For example, in Appendix A, dexamethasone sodium phosphate has *D* values of $0.050^\circ/(0.5\% \text{ drug})$, $0.180^\circ/(2.0\% \text{ drug})$, $0.52^\circ/(6.75\% \text{ drug})$, etc. It is apparent that the *D* value is nearly proportional to concentration. If a *D* value is needed for a concentration of drug not listed in Appendix A, a *D* value can be calculated from the appendix by direct proportion, using a *D* value closest to the concentration of drug in the actual solution.

The reference solution for the freezing-point-depression method is 0.9% sodium chloride, which has a freezing-point de-

pression of $\Delta T_f = 0.52^\circ$. Using the three steps described above, the dexamethasone sodium phosphate solution in Example 1 can be rendered isotonic as follows:

EXAMPLE 1

Dexamethasone Sodium Phosphate	0.1%
Purified Water qs	30 mL

Mft Isotonic Solution

Step 1—Reference solution: 0.9% sodium chloride.

$$\Delta T_f = 0.52^\circ$$

$$D = 0.050/0.5\% \text{ (dexamethasone sodium phosphate)}$$

Step 2—Contribution of drug.

$$\frac{0.050^\circ}{0.5\% \text{ drug}} \times 0.1\% \text{ drug} = 0.010^\circ$$

Step 3—Reference solution – Actual solution.

$$0.52^\circ - 0.01^\circ = 0.51^\circ$$

Sodium chloride needed.

$$\frac{0.9\% \text{ NaCl}}{0.52^\circ} \times 0.51^\circ = 0.883\% \text{ NaCl}$$

$$\frac{0.883 \text{ g NaCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.265 \text{ g NaCl}$$

The above solution could be made isotonic with any appropriate material other than sodium chloride by using the D value for that material. For example, to make the solution isotonic with dextrose with a D value, $D = 0.091^\circ/1\%$;

$$\frac{1\% \text{ Dextrose}}{0.091^\circ} \times 0.51^\circ = 5.60\% \text{ Dextrose}$$

$$\frac{5.60 \text{ g Dextrose}}{100 \text{ mL}} \times 30 \text{ mL} = 1.68 \text{ g Dextrose}$$

EXAMPLE 2

Naphazoline HCl (N.HCl)	0.02%
Zinc Sulfate	0.25%
Purified Water qs	30 mL

Mft Isotonic solution

Step 1—Reference solution: 0.9% sodium chloride.

$$\Delta T_f = 0.52^\circ$$

$$D = 0.14^\circ/1\% \text{ (naphazoline HCl)}$$

$$D = 0.086^\circ/1\% \text{ (zinc sulfate)}$$

Step 2—Contribution of drugs.

$$\frac{0.14^\circ}{1\% \text{ N.HCl}} \times 0.02\% \text{ N.HCl} = 0.003^\circ$$

$$\frac{0.086^\circ}{1\% \text{ ZnSO}_4} \times 0.25\% \text{ ZnSO}_4 = 0.022^\circ$$

$$0.003^\circ + 0.022^\circ = 0.025^\circ$$

Step 3—Reference solution – actual solution.

$$0.52^\circ - 0.025^\circ = 0.495^\circ$$

Sodium chloride needed.

$$\frac{0.9\% \text{ NaCl}}{0.52^\circ} \times 0.495^\circ = 0.857\% \text{ NaCl}$$

$$\frac{0.857 \text{ g NaCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.257 \text{ g NaCl}$$

The above solution could be made isotonic with any appropriate material other than sodium chloride by using the D value for that material.

For example, to make the solution isotonic with dextrose with a D value, $D = 0.091^\circ/1\%$;

$$\frac{1\% \text{ Dextrose}}{0.091^\circ} \times 0.495^\circ = 5.44\% \text{ Dextrose}$$

$$\frac{5.44 \text{ g Dextrose}}{100 \text{ mL}} \times 30 \text{ mL} = 1.63 \text{ g Dextrose}$$

SODIUM CHLORIDE EQUIVALENT METHOD—A sodium chloride equivalent, E value, is defined as the weight of sodium chloride that will produce the same osmotic effect as 1 g of the drug. For example, in Appendix A, dexamethasone sodium phosphate has an E value of 0.18 g NaCl/g drug at 0.5% drug concentration, 0.17 g NaCl/g drug at 1% drug concentration and a value of 0.16 g NaCl/g drug at 2% drug. This slight variation in the sodium chloride equivalent with concentration is due to changes in interionic attraction at different concentration of drug; the E value is not directly proportional to concentration as was the freezing-point-depression.

The reference solution for the sodium chloride equivalent method is 0.9% sodium chloride as it was for the freezing-point-depression method.

The dexamethasone sodium phosphate solution in Example 1 can be rendered isotonic using the sodium chloride equivalent method as follows:

EXAMPLE 1

Dexamethasone Sodium Phosphate	0.1%
Purified Water qs	30 mL

Mft Isotonic Solution

Step 1—Reference solution: 0.9% sodium chloride.

$$\frac{0.9 \text{ g NaCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.270 \text{ g NaCl}$$

$$E = 0.18 \text{ g NaCl/g drug}$$

Step 2—Contribution of drug.

$$\frac{0.18 \text{ g NaCl}}{1 \text{ g drug}} \times \frac{0.1 \text{ g drug}}{100 \text{ mL}} \times 30 \text{ mL} = 0.0054 \text{ g NaCl}$$

Step 3—Reference solution – Actual solution.

$$0.270 \text{ g NaCl} - 0.0054 \text{ g NaCl} = 0.265 \text{ g NaCl}$$

The above solution can be made isotonic with a material other than sodium chloride, such as dextrose, by using the E value of that material. For example, to make the solution isotonic with dextrose, $E = 0.16 \text{ g NaCl/g dextrose}$, the amount of sodium chloride needed in Step 3, can be converted to dextrose as follows:

$$\frac{1 \text{ g Dextrose}}{0.16 \text{ g NaCl}} \times 0.265 \text{ g NaCl} = 1.66 \text{ g Dextrose}$$

EXAMPLE 2

Naphazoline HCl (N.HCl)	0.02%
Zinc Sulfate	0.25%
Purified Water qs	30 mL

Mft Isotonic Solution

Step 1—Reference solution: 0.9% sodium chloride.

$$\frac{0.9 \text{ g NaCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.270 \text{ g NaCl}$$

$$E = 0.27 \text{ g NaCl/g N.HCl}$$

$$E = 0.15 \text{ g NaCl/g ZnSO}_4$$

Step 2—Contribution of drugs.

$$\frac{0.27 \text{ g NaCl}}{1 \text{ g N.HCl}} \times \frac{0.02 \text{ g N.HCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.002 \text{ g NaCl}$$

$$\frac{0.15 \text{ g NaCl}}{1 \text{ g ZnSO}_4} \times \frac{0.25 \text{ g ZnSO}_4}{100 \text{ mL}} \times 30 \text{ mL} = 0.011 \text{ g NaCl}$$

$$0.002 \text{ g NaCl} + 0.011 \text{ g NaCl} = 0.013 \text{ g NaCl}$$

Step 3—Reference solution – actual solution.

$$0.270 \text{ g NaCl} - 0.013 \text{ g NaCl} = 0.257 \text{ g NaCl}$$

The above solution can be made isotonic with a material other than sodium chloride, such as dextrose, by using the E value of that material. For example, to make the solution isotonic with dextrose, $E = 0.16 \text{ g NaCl/g dextrose}$, the amount of sodium chloride needed in Step 3 can be converted to dextrose as follows:

$$\frac{1 \text{ g Dextrose}}{0.16 \text{ g NaCl}} \times 0.257 \text{ g NaCl} = 1.61 \text{ g Dextrose}$$

ISOTONIC SOLUTION V VALUES—The V value of a drug is the volume of water to be added to a specified weight of drug (0.3 g or 1.0 g, depending on the table used) to prepare an isotonic solution. Appendix B gives such values for some commonly used drugs. The reason for providing data for 0.3 g of drug is for convenience in preparing 30 mL (approximately 1 fluidounce) of solution, a commonly prescribed volume. The basic principle underlying the use of V values is to prepare an isotonic solution of the prescribed drug and then dilute this solution to final volume with a suitable isotonic vehicle.

The two solutions in the previous examples can be prepared as follows using the V -value method:

EXAMPLE 1

Dexamethasone Sodium Phosphate	0.1%
Purified Water qs	30 mL

Mft Isotonic Solution

Step 1—The V value for dexamethasone sodium phosphate can be calculated from the sodium chloride equivalent, E , as outlined in the footnote in Appendix B.

$$\frac{100 \text{ mL Soln}}{0.9 \text{ g NaCl}} \times \frac{0.17 \text{ g NaCl}}{1 \text{ g drug}} \times 0.3 \text{ g drug} = 5.67 \text{ mL Soln}$$

for a dilute solution:

$$5.67 \text{ mL Soln} \cong 5.67 \text{ mL H}_2\text{O} \text{ there } V \\ = (5.67 \text{ mL H}_2\text{O}) / (0.3 \text{ g drug})$$

Step 2—Amount of drug needed.

$$\frac{0.1 \text{ g drug}}{100 \text{ mL}} \times 30 \text{ mL} = 0.030 \text{ g drug}$$

Volume of water needed to prepare an isotonic solution.

$$\frac{5.67 \text{ mL H}_2\text{O}}{0.3 \text{ g drug}} \times 0.030 \text{ g drug} = 0.57 \text{ mL H}_2\text{O}$$

Step 3—To prepare the solution, dissolve 0.030 g of drug in 0.57 mL water, and qs to volume with a suitable isotonic vehicle such as 0.9% sodium chloride solution, 5.51% dextrose, or an isotonic phosphate buffer.

EXAMPLE 2

Naphazoline HCl (N.HCl)	0.02%
Zinc Sulfate	0.25%
Purified Water qs	30 mL

Mft Isotonic Solution

Step 1—The V value for naphazoline HCl can be calculated from the sodium chloride equivalent, E , as outlined in the footnote in Appendix B; the V value for zinc sulfate is taken directly from Appendix B.

$$\frac{100 \text{ mL Soln}}{0.9 \text{ g NaCl}} \times \frac{0.27 \text{ g NaCl}}{1 \text{ g N.HCl}} \times 0.3 \text{ g N.HCl} = 9.00 \text{ mL Soln}$$

for a dilute solution:

$$9.00 \text{ mL Soln} \cong 9.00 \text{ mL H}_2\text{O} \text{ there } V \\ = (9.00 \text{ mL H}_2\text{O}) / (0.3 \text{ g N.HCl}) \\ V = 5.00 \text{ mL H}_2\text{O} / 0.3 \text{ g ZnSO}_4$$

Step 2—Amount of drugs needed.

$$\frac{0.02 \text{ g N.HCl}}{100 \text{ mL}} \times 30 \text{ mL} = 0.006 \text{ g N.HCl}$$

$$\frac{0.25 \text{ g ZnSO}_4}{100 \text{ mL}} \times 30 \text{ mL} = 0.075 \text{ g ZnSO}_4$$

Volume of water needed to prepare an isotonic solution.

$$\frac{9.00 \text{ mL H}_2\text{O}}{0.3 \text{ g N.HCl}} \times 0.006 \text{ g drug} = 0.18 \text{ mL H}_2\text{O}$$

$$\frac{5.00 \text{ mL H}_2\text{O}}{0.3 \text{ g ZnSO}_4} \times 0.075 \text{ g ZnSO}_4 = 1.25 \text{ mL H}_2\text{O}$$

Step 3—To prepare the solution, dissolve 0.006 g of naphazoline HCl and 0.075 g zinc sulfate in 1.43 mL water, and qs to volume with a suitable isotonic vehicle such as 0.9% sodium chloride solution, 5.51% dextrose, or an isotonic phosphate buffer.

REFERENCES

1. Kaminski MV. *Surg Gynecol Obstet* 1976; 143:12.
2. Theeuwes F. *J Pharm Sci* 1975; 64:1987.
3. Wermeling DP et al. *Am J Hosp Pharm* 1985; 1739:42.
4. Crane VS. *Drug Intell Clin Pharm* 1987; 21: 830.
5. Santeiro ML et al. *Am J Hosp Pharm* 1990; 47:1359.
6. McDuffee L. *IL Council Hosp Pharm Drug Inf Newsl* 1978; 8 (Oct–Nov).
7. Zenk K, Huxtable RF. *Hosp Pharm* 1978; 13:577.
8. Streng WH et al. *J Pharm Sci* 1978; 67:384.
9. Murty BSR et al. *Am J Hosp Pharm* 1976; 33:546.
10. Bray AJ. Personal communication. Evansville, IN: Mead Johnson Nutritional Division, 1978.
11. Payne-James JJ et al. *J Parenter Enter Nutr* 1993; 17:468.
12. Miller SJ. *Hosp Pharm* 1991; 26:796.
13. Andrassy RJ et al. *Surgery* 1977; 82:205.
14. Dobbie RP, Hoffmeister JA. *Surg Gynecol Obstet* 1976; 143:273.
15. *Osmolality*. Minneapolis: Doyle Pharmaceutical, 1978.
16. Dickerson RN, Melnik G. *Am J Hosp Pharm* 1988; 45:832.
17. Holtz L, Milton J, Sturek JK. *J Parenter Enter Nutr* 1987; 11:183.
18. Lund CG et al. *The Preparation of Solutions Iso-osmotic with Blood, Tears, and Tissue*. Copenhagen: Danish Pharmacopoeial Commission, Einar Munksgaard, 1947.
19. Hammarlund ER et al. *J Pharm Sci* 1965; 54:160.
20. Hammarlund ER, Pedersen-Bjergaard K. *J APhA Sci Ed* 1958; 47:107.
21. Hammarlund ER, Pedersen-Bjergaard K. *J Pharm Sci* 1961; 50:24.
22. Hammarlund ER, Van Pevenage GL. *J Pharm Sci* 1966; 55:1448.
23. Sapp C et al. *J Pharm Sci* 1975; 64:1884.
24. *British Pharmaceutical Codex*. London: Pharmaceutical Press, 1973.
25. Fassett WE et al. *J Pharm Sci* 1969; 58:1540.
26. Kagan DG, Kinsey VE. *Arch Ophthalmol* 1942; 27:696.

BIBLIOGRAPHY

- Alberty RA, Daniels F. *Physical Chemistry*, 7th ed. New York: Wiley, 1987.
- Cowan G, Scheetz W, eds. *Intravenous Hyperalimentation*. Philadelphia: Lea & Febiger, 1972.
- Garb S. *Laboratory Tests in Common Use*, 6th ed. New York: Springer, 1976.
- Hall WE. *Am J Pharm Ed* 1970; 34:204.
- Harvey AM, Johns RJ, Owens AH, et al. *The Principles and Practice of Medicine*, 18th ed. New York: Appleton Century Crofts, 1972.
- Martin AN, Swarbrick J, Cammarata A. *Physical Pharmacy*, 4th ed. Philadelphia: Lea & Febiger, 1993.
- Plumer AL. *Principles and Practice of Intravenous Therapy*, 4th ed. Boston: Little, Brown, 1987.
- Ravel R. *Clinical Laboratory Medicine*, 5th ed. St Louis: Mosby, 1988.
- Shizgal HM. *Ann Rev Med* 1991; 42:549.
- Tilkian SM, Conover MH. *Clinical Implications of Laboratory Tests*, 4th ed. St Louis: Mosby, 1987.
- Turco S, King RE. *Sterile Dosage Forms*, 3rd ed. Philadelphia: Lea & Febiger, 1987.
- Wallach J. *Interpretation of Diagnostic Tests*, 4th ed. Boston: Little, Brown, 1986.

Appendix A Sodium Chloride Equivalents, Freezing-Point Depressions, and Hemolytic Effects of Certain Medicinals in Aqueous Solution

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				
	E	D	E	D	E	D	E	D	E	D	%	E	D	H	pH
Acetrizoate methylglucamine	0.09		0.08		0.08		0.08		0.08	12.12	0.07		0		7.1
Acetrizoate sodium	0.10	0.027	0.10	0.055	0.10	0.109	0.10	0.163	0.10	0.273	9.64	0.09	0.52	0	6.9 [†]
Acetylcysteine	0.20	0.055	0.20	0.113	0.20	0.227	0.20	0.341			4.58	0.20	0.52	100*	2.0
Adrenaline HCl											4.24			68	4.5
Alphaprodine HCl	0.19	0.053	0.19	0.105	0.18	0.212	0.18	0.315			4.98	0.18	0.52	100	5.3
Alum (potassium)			0.18				0.15		0.15		6.35		0.14	24*	3.4
Amantadine HCl	0.31	0.090	0.31	0.180	0.31	0.354					2.95	0.31	0.52	91	5.7
Aminoacetic acid	0.42	0.119	0.41	0.235	0.41	0.470					2.20	0.41	0.52	0*	6.2
Aminohippuric acid	0.13	0.035	0.13	0.075											
Aminophylline				0.098 ^c											
Ammonium carbonate	0.70	0.202	0.70	0.405							1.29	0.70	0.52	97	7.7
Ammonium chloride			1.12								0.8	1.12	0.52	93	5.0
Ammonium lactate	0.33	0.093	0.33	0.185	0.33	0.370					2.76	0.33	0.52	98	5.9
Ammonium nitrate	0.69	0.200	0.69	0.400							1.30	0.69	0.52	91	5.3
Ammonium phosphate, dibasic	0.58	0.165	0.55	0.315							1.76	0.51	0.52	0	7.9
Ammonium sulfate	0.55	0.158	0.55	0.315							1.68	0.54	0.52	0	5.3
Amobarbital sodium			0.25	0.143 ^c			0.25				3.6	0.25	0.52	0	9.3
d-Amphetamine HCl											2.64			98	5.7
Amphetamine phosphate			0.34	0.20			0.27	0.47			3.47	0.26	0.52	0	4.5
Amphetamine sulfate			0.22	0.129 ^c			0.21	0.36			4.23	0.21	0.52	0	5.9
Amprotropine phosphate											5.90			0	4.2
Amylcaïne HCl			0.22				0.19				4.98	0.18		100	5.6
Anileridine HCl	0.19	0.052	0.19	0.104	0.19	0.212	0.18	0.316	0.18	0.509	5.13	0.18	0.52	12	2.6
Antazoline phosphate											6.05			90	4.0
Antimony potassium tartrate			0.18				0.13	0.10							
Antipyrine			0.17	0.10			0.14	0.24	0.14	0.40	6.81	0.13	0.52	100	6.1
Apomorphine HCl			0.14	0.080 ^c											
Arginine glutamate	0.17	0.048	0.17	0.097	0.17	0.195	0.17	0.292	0.17	0.487	5.37	0.17	0.52	0	6.9
Ascorbic acid				0.105 ^c							5.05	0.52 ^b	100*	2.2	
Atropine methylbromide			0.14				0.13		0.13		7.03	0.13			
Atropine methylnitrate											6.52			0	5.2
Atropine sulfate			0.13	0.075			0.11	0.19	0.11	0.32	8.85	0.10	0.52	0	5.0
Bacitracin			0.05	0.03			0.04	0.07	0.04	0.12					
Barbital sodium			0.30	0.171 ^c			0.29	0.50			3.12	0.29	0.52	0	9.8
Benzalkonium chloride			0.16				0.14		0.13						
Benztropine mesylate	0.26	0.073	0.21	0.115	0.15	0.170	0.12	0.203	0.09	0.242					
Benzyl alcohol			0.17	0.09 ^c			0.15								
Bethanechol chloride	0.50	0.140	0.39	0.225	0.32	0.368	0.30	0.512			3.05	0.30		0	6.0
Bismuth potassium tartrate			0.09				0.06		0.05						
Bismuth sodium tartrate			0.13				0.12		0.11		8.91	0.10		0	6.1
Boric acid	0.50	0.288 ^c									1.9	0.47	0.52	100	4.6
Brompheniramine maleate	0.10	0.026	0.09	0.050	0.08	0.084									
Bupivacaine HCl	0.17	0.048	0.17	0.096	0.17	0.193	0.17	0.290	0.17	0.484	5.38	0.17	0.52	83	6.8
Butabarbital sodium	0.27	0.078	0.27	0.155	0.27	0.313	0.27	0.470			3.33	0.27	0.52	0	6.8
Butacaine sulfate			0.20	0.12			0.13	0.23	0.10	0.29					
Caffeine and sodium benzoate			0.26	0.15			0.23	0.40			3.92	0.23	0.52	0	7.0
Caffeine and sodium salicylate			0.12	0.12			0.17	0.295	0.16	0.46	5.77	0.16	0.52	0	6.8
Calcium aminosalicylate											4.80			0	6.0
Calcium chloride			0.51	0.298 ^c							1.70	0.53	0.52	0	5.6
Calcium chloride (6 H ₂ O)			0.35	0.20							2.5	0.36	0.52	0	5.7
Calcium chloride, anhydrous			0.68	0.39							1.3	0.69	0.52	0	5.6
Calcium disodium edetate	0.21	0.061	0.21	0.120	0.21	0.240	0.20	0.357			4.50	0.20	0.52	0	6.1
Calcium gluconate			0.16	0.091 ^c			0.14	0.24							
Calcium lactate			0.23	0.13			0.12	0.36			4.5	0.20	0.52	0	6.7
Calcium lactobionate	0.08	0.022	0.08	0.043	0.08	0.085	0.07	0.126	0.07	0.197					
Calcium levulinate			0.27	0.16			0.25	0.43			3.58			0	7.2
Calcium pantothenate											5.50			0	7.4
Camphor			0.12 ^d												
Capreomycin sulfate	0.04	0.011	0.04	0.020	0.04	0.042	0.04	0.063	0.04	0.106					
Carbachol				0.205 ^c							2.82			0	5.9
Carbenicillin sodium	0.20	0.059	0.20	0.118	0.20	0.236	0.20	0.355			4.40	0.20	0.52	0	6.6
Carboxymethylcellulose sodium	0.03	0.007	0.03	0.017	0.145										
Cephaloridine	0.09	0.023	0.07	0.041	0.06	0.074	0.06	0.106	0.05						
Chloramine-T											4.10			100*	9.1
Chloramphenicol				0.06 ^d											
Chloramphenicol sodium succinate	0.14	0.038	0.14	0.078	0.14	0.154	0.13	0.230	0.13	0.382	6.83	0.13	0.52	partial	6.1
Chlordiazepoxide HCl	0.24	0.068	0.22	0.125	0.19	0.220	0.18	0.315	0.17	0.487	5.50	0.16	0.52	66	2.7
Chlorobutanol (hydrated)			0.24	0.14											
Chloroprocaine HCl	0.20	0.054	0.20	0.108	0.18	0.210									
Chloroquine phosphate	0.14	0.039	0.14	0.082	0.14	0.162	0.14	0.242	0.13	0.379	7.15	0.13	0.52	0	4.3
Chloroquine sulfate	0.10	0.028	0.09	0.050	0.08	0.090	0.07	0.127	0.07	0.195					
Chlorpheniramine maleate	0.17	0.048	0.15	0.085	0.14	0.165	0.13	0.220	0.09	0.265					
Chlortetracycline HCl	0.10	0.030	0.10	0.061	0.10	0.121									
Chlortetracycline sulfate			0.13	0.08			0.10	0.17							

Appendix A Continued

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				
	E	D	E	D	E	D	E	D	E	D	%	E	D	H	pH
Citric acid			0.18	0.10			0.17	0.295	0.16	0.46	5.52	0.16	0.52	100*	1.8
Clindamycin phosphate	0.08	0.022	0.08	0.046	0.08	0.095	0.08	0.144	0.08	0.242	10.73	0.08	0.52	58*	6.8
Cocaine HCl			0.16	0.090 ^c			0.15	0.26	0.14	0.40	6.33	0.14	0.52	47	4.4
Codeine phosphate			0.14	0.080 ^c			0.13	0.23	0.13	0.38	7.29	0.12	0.52	0	4.4
Colistimethate sodium	0.15	0.045	0.15	0.085	0.15	0.170	0.15	0.253	0.14	0.411	6.73	0.13	0.52	0	7.6
Cupric sulfate			0.18	0.100 ^c			0.15		0.14		6.85	0.13		trace*	3.9
Cyclizine HCl	0.20	0.060													
Cyclophosphamide	0.10	0.031	0.10	0.061	0.10	0.125									
Cytarabine	0.11	0.034	0.11	0.066	0.11	0.134	0.11	0.198	0.11	0.317	8.92	0.10	0.52	0	8.0
Deferoxamine mesylate	0.09	0.023	0.09	0.047	0.09	-0.093	0.09	0.142	0.09	0.241					
Demecarium bromide	0.14	0.038	0.12	0.069	0.10	0.108	0.08	0.139	0.07	0.192					
Dexamethasone sodium phosphate	0.18	0.050	0.17	0.095	0.16	0.180	0.15	0.260	0.14	0.410	6.75	0.13	0.52	0	8.9
Dextroamphetamine HCl	0.34	0.097	0.34	0.196	0.34	0.392					2.64	0.34	0.52		
Dextroamphetamine phosphate			0.25	0.14			0.25	0.44			3.62	0.25	0.52	0	4.7
Dextroamphetamine sulfate	0.24	0.069	0.23	0.134	0.22	0.259	0.22	0.380			4.16	0.22	0.52	0	5.9
Dextrose			0.16	0.091 ^c			0.16	0.28	0.16	0.46	5.51	0.16	0.52	0	5.9
Dextrose (anhydrous)			0.18	0.101 ^c			0.18	0.31			5.05	0.18	0.52	0	6.0
Diatrizoate sodium	0.10	0.025	0.09	0.049	0.09	0.098	0.09	0.149	0.09	0.248	10.55	0.09	0.52	0	7.9
Dibucaine HCl				0.074 ^c											
Dicloxacillin sodium (1 H ₂ O)	0.10	0.030	0.10	0.061	0.10	0.122	0.10	0.182							
Diethanolamine	0.31	0.089	0.31	0.177	0.31	0.358					2.90	0.31	0.52	100	11.3
Dihydrostreptomycin sulfate			0.06	0.03			0.05	0.09	0.05	0.14	19.4	0.05	0.52	0	6.1
Dimethylpyrindene maleate	0.13	0.039	0.12	0.070	0.11	0.120									
Dimethyl sulfoxide	0.42	0.122	0.42	0.245	0.42	0.480					2.16	0.42	0.52	100	7.6
Diperodon HCl	0.15	0.045	0.14	0.079	0.13	0.141									
Diphenhydramine HCl				0.161 ^c							5.70			88*	5.5
Diphenidol HCl	0.16	0.045	0.16	0.09	0.16	0.180									
Doxapram HCl	0.12	0.035	0.12	0.070	0.12	0.140	0.12	0.210							
Doxycycline hyclate	0.12	0.035	0.12	0.072	0.12	0.134	0.11	0.186	0.09	0.264					
Dyphylline	0.10	0.025	0.10	0.052	0.09	0.104	0.09	0.155	0.08	0.245					
Echothiophate iodide	0.16	0.045	0.16	0.090	0.16	0.179									
Edetate disodium	0.24	0.070	0.23	0.132	0.22	0.248	0.21	0.360			4.44	0.20	0.52	0	4.7
Edetate trisodium monohydrate	0.29	0.079	0.29	0.158	0.28	0.316	0.27	0.472			3.31	0.27	0.52	0	8.0
Emetine HCl				0.058 ^c				0.17		0.29					
Ephedrine HCl			0.30	0.165 ^c			0.28				3.2	0.28		96	5.9
Ephedrine sulfate			0.23	0.13			0.20	0.35			4.54	0.20	0.52	0	5.7
Epinephrine bitartrate			0.18	0.104			0.16	0.28	0.16	0.462	5.7	0.16	0.52	100*	3.4
Epinephrine hydrochloride			0.29	0.16 ^b			0.26				3.47	0.26			
Ergonovine maleate				0.089 ^c											
Erythromycin lactobionate	0.08	0.020	0.07	0.040	0.07	0.078	0.07	0.115	0.06	0.187				100	6.0
Ethyl alcohol											1.39				
Ethylenediamine				0.253 ^c							2.08			100*	11.4
Ethylmorphine HCl			0.16	0.088 ^c			0.15	0.26	0.15	0.43	6.18	0.15	0.52	38	4.7
Eucatropine HCl				0.11 ^d											
Ferric ammonium citrate (green)											6.83			0	5.2
Floxuridine	0.14	0.040	0.13	0.076	0.13	0.147	0.12	0.213	0.12	0.335	8.47	0.12	0.52	3*	4.5
Fluorescein sodium			0.31	0.181 ^c			0.27	0.47			3.34	0.27	0.52	0	8.7
Fluphenazine 2-HCl	0.14	0.041	0.14	0.082	0.12	0.145	0.09	0.155						0*	5.9
<i>d</i> -Fructose											5.05				
Furtrethonium iodide	0.24	0.070	0.24	0.133	0.22	0.250	0.21	0.360			4.44	0.20	0.52	0	5.4
Galactose											4.92			0	5.9
Gentamicin sulfate	0.05	0.015	0.05	0.030	0.05	0.060	0.05	0.093	0.05	0.153					
<i>D</i> -Glucuronic acid											5.02			48*	1.6
Glycerin			0.203 ^c								2.6			100	5.9
Glycopyrrolate	0.15	0.042	0.15	0.084	0.15	0.166	0.14	0.242	0.13	0.381	7.22	0.12	0.52	92*	4.0
Gold sodium thiomalate	0.10	0.032	0.10	0.061	0.10	0.111	0.09	0.159	0.09	0.250					
Hetacillin potassium	0.17	0.048	0.17	0.095	0.17	0.190	0.17	0.284	0.17	0.474	5.50	0.17	0.52	0	6.3
Hexafluorenum bromide	0.12	0.033	0.11	0.065											
Hexamethonium tartrate	0.16	0.045	0.16	0.089	0.16	0.181	0.16	0.271	0.16	0.456	5.68	0.16	0.52		
Hexamethylene sodium acetaminosalicylate	0.18	0.049	0.18	0.099	0.17	0.199	0.17	0.297	0.16	0.485	5.48	0.16	0.52	0*	4.0
Hexobarbital sodium				0.15 ^c											
Hexylcaine HCl											4.30			100	4.8
Histamine 2HCl	0.40	0.115	0.40	0.233	0.40	0.466					2.24	0.40	0.52	79*	3.7
Histamine phosphate				0.149 ^c							4.10	0	4.6		
Histidine HCl											3.45			40	3.9
Holocaine HCl			0.20	0.12											
Homatropine hydrobromide			0.17	0.097 ^c			0.16	0.28	0.16	0.46	5.67	0.16	0.52	92	5.0
Homatropine methylbromide			0.19	0.11			0.15	0.26	0.13	0.38					
4-Homosulfanilamide HCl											3.69			0	4.9
Hyaluronidase	0.01	0.004	0.01	0.007	0.01	0.013	0.01	0.020	0.01	0.033					
Hydromorphone HCl											6.39			64	5.6

Appendix A Continued

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				
	E	D	E	D	E	D	E	D	E	D	%	E	D	H	pH
Hydroxyamphetamine HBr				0.15 ^d							3.71			92	5.0
8-Hydroxyquinoline sulfate											9.75			59*	2.5
Hydroxystilbamidine isethionate	0.20	0.060	0.16	0.090	0.12	0.137	0.10	0.170	0.07	0.216					
Hyoscyamine hydrobromide											6.53			68	5.9
Imipramine HCl	0.20	0.058	0.20	0.110	0.13	0.143									
Indigotindisulfonate sodium	0.30	0.085	0.30	0.172											
Intracaine HCl											4.97			85	5.0
Iodophthalein sodium				0.07 ^c							9.58			100	9.4
Isometheptene mucate	0.18	0.048	0.18	0.095	0.18	0.196	0.18	0.302			4.95	0.18	0.52	0	6.2
Isoproterenol sulfate	0.14	0.039	0.14	0.078	0.14	0.156	0.14	0.234	0.14	0.389	6.65	0.14	0.52	trace	4.5
Kanamycin sulfate	0.08	0.021	0.07	0.041	0.07	0.083	0.07	0.125	0.07	0.210					
Lactic acid				0.239 ^c							2.30			100*	2.1
Lactose			0.07	0.040 ^c			0.08		0.09		9.75	0.09		0*	5.8
Levallorphan tartrate	0.13	0.036	0.13	0.073	0.13	0.143	0.12	0.210	0.12	0.329	9.40	0.10	0.52	59*	6.9
Levorphanol tartrate	0.12	0.033	0.12	0.067	0.12	0.136	0.12	0.203							
Lidocaine HCl				0.13 ^c							4.42			85	4.3
Lircomycin HCl	0.16	0.045	0.16	0.090	0.15	0.170	0.14	0.247	0.14	0.400	6.60	0.14	0.52	0	4.5
Lobeline HCl				0.09 ^b											
Lyapolate sodium	0.10	0.025	0.09	0.051	0.09	0.103	0.09	0.157	0.09	0.263	9.96	0.09	0.52	0	6.5 [†]
Magnesium chloride				0.45							2.02	0.45		0	6.3
Magnesium sulfate			0.17	0.094 ^c			0.15	0.26	0.15	0.43	6.3	0.14	0.52	0	6.2
Magnesium sulfate, anhydrous	0.34	0.093	0.32	0.184	0.30	0.345	0.29	0.495			3.18	0.28	0.52	0	7.0
Mannitol				0.098 ^c						5.07				0*	6.2
Maphenide HCl		0.27	0.075	0.27	0.153	0.27	0.303	0.26	0.448		3.55	0.25	0.52		
Menadiol sodium diphosphate											4.36			0	8.2
Menadione sodium bisulfite											5.07			0	5.3
Menthol					0.12 ^d										
Meperidine HCl					0.125 ^c						4.80			98	5.0
Mepivacaine HCl	0.21	0.060	0.21	0.116	0.20	0.230	0.20	0.342			4.60	0.20	0.52	45	4.5
Merbromin				0.08 ^b											
Mercuric cyanide			0.15				0.14		0.13						
Mersalyl				0.06 ^b											
Mesoridazine besylate	0.10	0.024	0.07	0.040	0.05	0.058	0.04	0.071	0.03	0.087					
Metaraminol bitartrate	0.20	0.060	0.20	0.112	0.19	0.210	0.18	0.308	0.17	0.505	5.17	0.17	0.52	59	3.8
Methacholine chloride				0.184 ^c							3.21			0	4.5
Methadone HCl				0.101 ^c							8.59			100*	5.0
Methamphetamine HCl				0.213 ^c							2.75			97	5.9
Methdilazine HCl	0.12	0.035	0.10	0.056	0.08	0.080	0.06	0.093	0.04	0.112					
Methenamine				0.23			0.24				3.68	0.25		100	8.4
Methiodal sodium	0.24	0.068	0.24	0.136	0.24	0.274	0.24	0.410			3.81	0.24	0.52	0	5.9
Methital sodium	0.26	0.074	0.25	0.142	0.24	0.275	0.23	0.407			3.85	0.23	0.52	78	9.8
Methocarbamol	0.10	0.030	0.10	0.060											
Methotrimeprazine HCl	0.12	0.034	0.10	0.060	0.07	0.077	0.06	0.094	0.04	0.125					
Methoxyphenamine HCl	0.26	0.075	0.26	0.150	0.26	0.300	0.26	0.450			3.47	0.26	0.52	96	5.4
p-Methylaminoethanolphenol tartrate	0.18	0.048	0.17	0.095	0.16	0.190	0.16	0.282	0.16	0.453	5.83	0.16	0.52	0	6.2
Methyl Dopate HCl	0.21	0.063	0.21	0.122	0.21	0.244	0.21	0.365			4.28	0.21	0.52	partial	3.0
Methylergonovine maleate	0.10	0.028	0.10	0.056											
N-Methylglucamine	0.20	0.057	0.20	0.111	0.18	0.214	0.18	0.315	0.18	0.517	5.02	0.18	0.52	4	11.3
Methylphenidate HCl	0.22	0.065	0.22	0.127	0.22	0.258	0.22	0.388			4.07	0.22	0.52	66	4.3
Methylprednisolone Na succinate	0.10	0.025	0.09	0.051	0.09	0.102	0.08	0.143	0.07	0.200					
Minocycline HCl	0.10	0.030	0.10	0.058	0.09	0.107	0.08	0.146							
Monoethanolamine	0.53	0.154	0.53	0.306							1.70	0.53	0.52	100	11.4
Morphine HCl				0.086 ^c			0.14								
Morphine sulfate				0.079 ^c			0.11	0.19	0.09	0.26					
Nalorphine HCl	0.24	0.070	0.21	0.121	0.18	0.210	0.17	0.288	0.15	0.434	6.36	0.14	0.52	63	4.1
Naloxone HCl	0.14	0.042	0.14	0.083	0.14	0.158	0.13	0.230	0.13	0.367	8.07	0.11	0.52	35	5.2
Naphazoline HCl			0.27	0.14 ^d			0.24				3.99	0.22		100	5.3
Neoarsphenamine											2.32		17	7.8	
Neomycin sulfate			0.11	0.063 ^c			0.09	0.16	0.08	0.232					
Neostigmine bromide			0.22	0.127 ^c			0.19				4.98			0	4.6
Neostigmine methylsulfate			0.20	0.115 ^c			0.18		0.17		5.22	0.17			
Nicotinamide			0.26	0.148 ^c			0.21	0.36			4.49	0.20	0.52	100	7.0
Nicotinic acid			0.25	0.144 ^c											
Nikethamide				0.100 ^c							5.94			100	6.9
Novobiocin sodium	0.12	0.033	0.10	0.057	0.07	0.073									
Oleandomycin phosphate	0.08	0.017	0.08	0.038	0.08	0.084	0.08	0.129	0.08	0.255	10.82	0.08	0.52	0	5.0
Orphenadrine citrate	0.13	0.037	0.13	0.074	0.13	0.144	0.12	0.204	0.10	0.285					
Oxophenarsine HCl											.67			trace*	2.3
Oxymetazoline HCl	0.22	0.063	0.22	0.124	0.20	0.232	0.19	0.335			4.92	0.18	0.52	86	5.7
Oxyquinoline sulfate	0.24	0.068	0.21	0.113	0.16	0.182	0.14	0.236	0.11	0.315					
d-Pantothenyl alcohol	0.20	0.053	0.18	0.100	0.17	0.193	0.17	0.283	0.16	0.468	5.60	0.16	0.52	92	6.8
Papaverine HCl				0.061 ^c											
Paraldehyde	0.25	0.071	0.25	0.142	0.25	0.288	0.25	0.430			3.65	0.25	0.52	97	5.3
Pargyline HCl	0.30	0.083	0.29	0.165	0.29	0.327	0.28	0.491			3.18	0.28	0.52	91	3.8

Appendix A Continued

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				
	E	D	E	D	E	D	E	D	E	D	%	E	D	H	pH
Penicillin G, potassium			0.18	0.102 ^c			0.17	0.29	0.16	0.46	5.48	0.16	0.52	0	6.2
Penicillin G, procaine				0.06 ^d											
Penicillin G, sodium			0.18	0.100 ^c			0.16	0.28	0.16	0.46	5.90			18	5.2
Pentazocine lactate	0.15	0.042	0.15	0.085	0.15	0.169	0.15	0.253	0.15	0.420					
Pentobarbital sodium				0.145 ^c							4.07			0	9.9
Pentolinium tartrate											5.95			55*	3.4
Phenacaine HCl				0.09 ^d											
Pheniramine maleate				0.09 ^d											
Phenobarbital sodium			0.24	0.135 ^c			0.23	0.40			3.95	0.23	0.52	0	9.2
Phenol	0.35	0.20									2.8	0.32	0.52	0*	5.6
Phentolamine mesylate	0.18	0.052	0.17	0.096	0.16	0.173	0.14	0.244	0.13	0.364	8.23	0.11	0.52	83	3.5
Phenylephrine HCl			0.32	0.184 ^c			0.30				3.0	0.30		0	4.5
Phenylephrine tartrate												5.90		58*	5.4
Phenylethyl alcohol	0.25	0.070	0.25	0.141	0.25	0.283									
Phenylpropanolamine HCl				0.38							2.6	0.35		95	5.3
Physostigmine salicylate			0.16	0.219 ^c											
Physostigmine sulfate				0.090 ^c											
				0.074 ^c											
Pilocarpine HCl			0.24	0.138 ^c			0.22	0.38			4.08	0.22	0.52	89	4.0
Pilocarpine nitrate			0.23	0.132 ^c			0.20	0.35			4.84	0.20	0.52	88	3.9
Piperocaine HCl				0.12 ^d							5.22			65	5.7
Polyethylene glycol 300	0.12	0.034	0.12	0.069	0.12	0.141	0.12	0.216	0.13	0.378	6.73	0.13	0.52	53	3.8
Polyethylene glycol 400	0.08	0.022	0.08	0.047	0.09	0.098	0.09	0.153	0.09	0.272	8.50	0.11	0.52	0	4.4
Polyethylene glycol 1500	0.06	0.015	0.06	0.036	0.07	0.078	0.07	0.120	0.07	0.215	10.00	0.09	0.52	4	4.1
Polyethylene glycol 1540	0.02	0.005	0.02	0.012	0.02	0.028	0.03	0.047	0.03	0.094					
Polyethylene glycol 4000	0.02	0.004	0.02	0.008	0.02	0.020	0.02	0.033	0.02	0.067					
Polymyxin B sulfate			0.09	0.052			0.06	0.10	0.04	0.12					
Polysorbate 80	0.02	0.005	0.02	0.010	0.02	0.020	0.02	0.032	0.02	0.055					
Polyvinyl alcohol (99% hydrol)	0.02	0.004	0.02	0.008	0.02	0.020	0.02	0.035	0.03	0.075					
Polyvinylpyrrolidone	0.01	0.003	0.01	0.006	0.01	0.010	0.01	0.017	0.01	0.035					
Potassium acetate	0.59	0.172	0.59	0.342							1.53	0.59	0.52	0	7.6
Potassium chlorate											1.88			0	6.9
Potassium chloride			0.76	0.439 ^c							1.19	0.76	0.52	0	5.9
Potassium iodide			0.34	0.196 ^c							2.59	0.34	0.52	0	7.0
Potassium nitrate			0.56	0.324 ^c							1.62	0.56	0	5.9	
Potassium phosphate			0.46	0.27							2.08	0.43	0.52	0	8.4
Potassium phosphate, monobasic			0.44	0.25							2.18	0.41	0.52	0	4.4
Potassium sulfate			0.44								2.11	0.43		0	6.6
Pralidoxime chloride	0.32	0.092	0.32	0.183	0.32	0.364					2.87	0.32	0.52	0	4.6
Prilocaine HCl	0.22	0.062	0.22	0.125	0.22	0.250	0.22	0.375			4.18	0.22	0.52	45	4.6
Procainamide HCl			0.22	0.13			0.19	0.33	0.17	0.49					
Procaine HCl			0.21	0.122 ^c			0.19	0.33	0.18		5.05	0.18	0.52	91	5.6
Prochlorperazine edisylate	0.08	0.020	0.06	0.033	0.05	0.048	0.03	0.056	0.02	0.065					
Promazine HCl	0.18	0.050	0.13	0.077	0.09	0.102	0.07	0.112	0.05	0.137					
Proparacaine HCl	0.16	0.044	0.15	0.086	0.15	0.169	0.14	0.247	0.13	0.380	7.46	0.12	0.52		
Propiomazine HCl	0.18	0.050	0.15	0.084	0.12	0.133	0.10	0.165	0.08	0.215					
Propoxycaïne HCl											6.40			16	5.3
Propylene glycol											2.00			100	5.5
Pyrathiazine HCl	0.22	0.065	0.17	0.095	0.11	0.123	0.08	0.140	0.06	0.170					
Pyridostigmine bromide	0.22	0.062	0.22	0.125	0.22	0.250	0.22	0.377			4.13	0.22	0.52	0	7.2
Pyridoxine HCl											3.05			31*	3.2
Quinacrine methanesulfonate				0.06 ^c											
Quinine bisulfate			0.09	0.05			0.09	0.16							
Quinine dihydrochloride			0.23	0.130 ^c			0.19	0.33	0.18		5.07	0.18	0.52	trace*	2.5
Quinine hydrochloride			0.14	0.077 ^c			0.11	0.19							
Quinine and urea HCl			0.23	0.13			0.21	0.36			4.5	0.20	0.52	64	2.9
Resorcinol		0.161 ^c									3.30			96	5.0
Rolitetracline	0.11	0.032	0.11	0.064	0.10	0.113	0.09	0.158	0.07	0.204					
Rose Bengal	0.08	0.020	0.07	0.040	0.07	0.083	0.07	0.124	0.07	0.198	14.9	0.06	0.52		
Rose Bengal B	0.08	0.022	0.08	0.044	0.08	0.087	0.08	0.131	0.08	0.218					
Scopolamine HBr			0.12	0.07			0.12	0.21	0.12	0.35	7.85	0.11	0.52	8	4.8
Scopolamine methylnitrate			0.16				0.14		0.13	6.95	0.13	0	6.0		
Secobarbital sodium			0.24	0.14			0.23	0.40			3.9	0.23	0.52	trace	9.8
Silver nitrate			0.33	0.190 ^c							2.74	0.33	0.52	0*	5.0
Silver protein, mild			0.17	0.10			0.17	0.29	0.16	0.46	5.51	0.16	0.52	0	9.0
Silver protein, strong				0.06 ^d											
Sodium acetate			0.46	0.267							2.0	0.45	0.52		
Sodium acetazolamide	0.24	0.068	0.23	0.135	0.23	0.271	0.23	0.406			3.85	0.23	0.52		
Sodium aminosaliclylate				0.170 ^c							3.27			0	7.3
Sodium ampicillin	0.16	0.045	0.16	0.090	0.16	0.181	0.16	0.072	0.16	0.451	5.78	0.16	0.52	0	8.5
Sodium ascorbate											3.00			0	6.9
Sodium benzoate			0.40	0.230 ^c							2.25	0.40	0.52	0	7.5
Sodium bicarbonate			0.65	0.375							1.39	0.65	0.52	0	8.3
Sodium biphosphate (H ₂ O)			0.40	0.23							2.45	0.37	0.52	0	4.1
Sodium biphosphate(2 H ₂ O)			0.36								2.77	0.32		0	4.0
Sodium bismuth thioglycollate	0.20	0.055	0.19	0.107	0.18	0.208	0.18	0.303	0.17	0.493	5.29			0	8.3
Sodium bisulfite			0.61	0.35							1.5	0.61	0.52	0*	3.0

Appendix A Continued

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				pH	
	E	D	E	D	E	D	E	D	E	D	%	E	D	H		
Sodium borate			0.42	0.241 ^c								2.6	0.35	0.52	0	9.2
Sodium bromide											1.60				0	6.1
Sodium cacodylate			0.32				0.28				3.3	0.27			0	8.0
Sodium carbonate, monohydrated			0.60	0.346							1.56	0.58	0.52	100		11.1
Sodium cephalothin	0.18	0.050	0.17	0.095	0.16	0.179	0.15	0.259	0.14	0.400	6.80	0.13	0.52	partial		8.5
Sodium chloride			1.00	0.576 ^c				1.00	1.73	1.00	2.88	0.9	1.00	0.52	0	6.7
Sodium citrate			0.31	0.178 ^c				0.30	0.52		3.02	0.30			0	7.8
Sodium colistimethate	0.16	0.045	0.15	0.087	0.14	0.161	0.14	0.235	0.13	0.383	6.85	0.13	0.52	0		8.4
Sodium hypophosphite											1.60				0	7.3
Sodium iodide			0.39	0.222 ^c							2.37	0.38	0.52		0	6.9
Sodium iodohippurate											5.92				0	7.3
Sodium lactate											1.72				0	6.5
Sodium lauryl sulfate	0.10	0.029	0.08	0.046	0.07	0.068	0.05	0.086								
Sodium mercaptomerin											5.30				0	8.4
Sodium metabisulfite			0.67	0.386 ^c							1.38	0.65	0.52	5*		4.5
Sodium methicillin	0.18	0.050	0.18	0.099	0.17	0.192	0.16	0.281	0.15	0.445	6.00	0.15	0.52	0		5.8
Sodium nafcillin	0.14	0.039	0.14	0.078	0.14	0.158	0.13	0.219	0.10	0.285						
Sodium nitrate				0.68							1.36	0.66			0	6.0
Sodium nitrite				0.84							1.08	0.83			0*	8.5
Sodium oxacillin	0.18	0.050	0.17	0.095	0.16	0.177	0.15	0.257	0.14	0.408	6.64	0.14	0.52	0		6.0
Sodium phenylbutazone	0.19	0.054	0.18	0.104	0.17	0.202	0.17	0.298	0.17	0.488	5.34	0.17	0.52			
Sodium phosphate			0.29	0.168				0.27	0.47		3.33	0.27	0.52	0		9.2
Sodium phosphate, dibasic (2 H ₂ O)			0.42	0.24							2.23	0.40	0.52	0		9.2
Sodium phosphate, dibasic (12 H ₂ O)			0.22				0.21				4.45	0.20			0	9.2
Sodium propionate			0.61	0.35							1.47	0.61	0.52	0		7.8
Sodium salicylate			0.36	0.210 ^c							2.53	0.36	0.52	0		6.7
Sodium succinate	0.32	0.092	0.32	0.184	0.31	0.361					2.90	0.31	0.52	0		8.5
Sodium sulfate, anhydrous				0.58							1.61	0.56	0.52	0		6.2
Sodium sulfite, exsiccated				0.65							1.45				0	9.6
Sodium sulfobromophthalein	0.07	0.019	0.06	0.034	0.05	0.060	0.05	0.084	0.04	0.123						
Sodium tartrate	0.33	0.098	0.33	0.193	0.33	0.385					2.72	0.33	0.52	0		7.3
Sodium thiosulfate				0.31							2.98	0.30	0.52	0		7.4
Sodium warfarin	0.18	0.049	0.17	0.095	0.16	0.181	0.15	0.264	0.15	0.430	6.10	0.15	0.52	0		8.1
Sorbitol (½ H ₂ O)											5.48				0	5.9
Sparteine sulfate	0.10	0.030	0.10	0.056	0.10	0.111	0.10	0.167	0.10	0.277	9.46	0.10	0.52	19*		3.5
Spectinomycin HCl	0.16	0.045	0.16	0.092	0.16	0.185	0.16	0.280	0.16	0.460	5.66	0.16	0.52	3		4.4
Streptomycin HCl				0.17				0.16	0.16							
Streptomycin sulfate				0.07				0.06	0.10	0.06	0.17					
Sucrose				0.08				0.09	0.16	0.09	0.26	9.25	0.10	0.52	0	6.4
Sulfacetamide sodium				0.23				0.23	0.40		3.85	0.23	0.52	0		8.7
Sulfadiazine sodium				0.24				0.24	0.38		4.24	0.21	0.52	0		9.5
Sulfamerazine sodium				0.23				0.21	0.36		4.53	0.20	0.52	0		9.8
Sulfapyridine sodium				0.23				0.21	0.36		4.55	0.20	0.52	5		10.4
Sulfathiazole sodium				0.22				0.20	0.35		4.82	0.19	0.52	0		9.9
Tartaric acid				0.143 ^c							3.90				75*	1.7
Tetracaine HCl			0.18	0.109 ^c				0.15	0.26	0.12	0.35					
Tetracycline HCl			0.14	0.081 ^c		0.10										
Tetrahydrozoline HCl											4.10				60*	6.7
Theophylline				0.02 ^b												
Theophylline sodium glycinat											2.94				0	8.9
Thiamine HCl				0.139 ^c							4.24				87*	3.0
Thiethylperazine maleate	0.10	0.030	0.09	0.050	0.08	0.089	0.07	0.119	0.05	0.153						
Thiopental sodium				0.155 ^c							3.50				74	10.3
Thiopropazate diHCl	0.20	0.053	0.16	0.090	0.12	0.137	0.10	0.170	0.08	0.222						
Thioridazine HCl	0.06	0.015	0.05	0.025	0.04	0.042	0.03	0.055	0.03	0.075						
Thiotepa	0.16	0.045	0.16	0.090	0.16	0.182	0.16	0.278	0.16	0.460	5.67	0.16	0.52	10*		8.2
Tridihexethyl chloride	0.16	0.047	0.16	0.096	0.16	0.191	0.16	0.280	0.16	0.463	5.62	0.16	0.52	97		5.4
Triethanolamine	0.20	0.058	0.21	0.121	0.22	0.252	0.22	0.383			4.05	0.22	0.52	100		10.7
Trifluoperazine 2HCl	0.18	0.052	0.18	0.100	0.13	0.144										
Triflupromazine HCl	0.10	0.031	0.09	0.051	0.05	0.061	0.04	0.073	0.03	0.092						
Trimeprazine tartrate	0.10	0.023	0.06	0.035	0.04	0.045	0.03	0.052	0.02	0.061						
Trimethadione	0.23	0.069	0.23	0.133	0.22	0.257	0.22	0.378			4.22	0.21	0.52	100		6.0
Trimethobenzamide HCl	0.12	0.033	0.10	0.062	0.10	0.108	0.09	0.153	0.08	0.232						
Tripelennamine HCl				0.13 ^d							5.50				100	6.3
Tromethamine	0.26	0.074	0.26	0.150	0.26	0.300	0.26	0.450			3.45	0.26	0.52	0		10.2
Tropicamide	0.10	0.030	0.09	0.050												
Trypan blue	0.26	0.075	0.26	0.150												
Tryparsamide				0.11 ^c												
Tubocurarine chloride				0.076 ^c												
Urea			0.59	0.34							1.63	0.55	0.52	100		6.6
Urethan				0.18 ^b							2.93				100	6.3
Uridine	0.12	0.035	0.12	0.069	0.12	0.138	0.12	0.208	0.12	0.333	8.18	0.11	0.52	0*		6.1
Valethamate bromide	0.16	0.044	0.15	0.085	0.15	0.168	0.14	0.238	0.11	0.324						
Vancomycin sulfate	0.06	0.015	0.05	0.028	0.04	0.049	0.04	0.066	0.04	0.098						

Appendix A Continued

	0.5%		1%		2%		3%		5%		ISO-OSMOTIC CONCENTRATION ^a				
	E	D	E	D	E	D	E	D	E	D	%	E	D	H	pH
Viomycin sulfate			0.08	0.05			0.07	0.12	0.07	0.20					
Xylometazoline HCl	0.22	0.065	0.21	0.121	0.20	0.232	0.20	0.342			4.68	0.19	0.52	88	5.0
Zinc phenolsulfonate											5.40			0*	5.4
Zinc sulfate			0.15	0.086 ^c			0.13	0.23	0.12	0.35	7.65	0.12	0.52		

^a The unmarked values were taken from Hammarlund *et al.*^{19–22} and Sapp *et al.*²³

^b Adapted from Lund *et al.*¹⁷

^c Adapted from *British Pharmaceutical Codex*.²⁴

^d Obtained from several sources.

^e E, sodium chloride equivalents; D, freezing-point depression, °C; H, hemolysis, %, at the concentration that is iso-osmotic with 0.9% NaCl, based on freezing-point determination or equivalent test; pH, approximate pH of solution studied for hemolytic action; *, change in appearance of erythrocytes and/or solution^{23–25}; †, pH determined after addition of blood.

Note: See also Budavari S, ed, *Merck Index*, 11th ed, Rahway, NJ: Merck, 1988: pp MISC 79–103.

Appendix B Isotonic Solution V—Values^{26, a, b}

DRUG (0.3 g)	WATER NEEDED FOR ISOTONICITY (mL)	DRUG (0.3 g)	WATER NEEDED FOR ISOTONICITY (mL)	DRUG (0.3 g)	WATER NEEDED FOR ISOTONICITY (mL)
Alcohol	21.7	Epinephrine hydrochloride	9.7	Silver nitrate	11.0
Ammonium chloride	37.3	Ethylmorphine hydrochloride	5.3	Silver protein, mild	5.7
Amobarbital sodium	8.3	Fluorescein sodium	10.3	Sodium acetate	15.3
Amphetamine phosphate	11.3	Glycerin	11.7	Sodium bicarbonate	21.7
Amphetamine sulfate	7.3	Holocaine hydrochloride	6.7	Sodium biphosphate, anhydrous	15.3
Antipyrine	5.7	Homatropine hydrobromide	5.7	Sodium biphosphate	13.3
Apomorphine hydrochloride	4.7	Homatropine methylbromide	6.3	Sodium bisulfite	20.3
Ascorbic acid	6.0	Hyoscyamine sulfate	4.7	Sodium borate	14.0
Atropine methylbromide	4.7	Neomycin sulfate	3.7	Sodium iodide	13.0
Atropine sulfate	4.3	Oxytetracycline hydrochloride	4.3	Sodium metabisulfite	22.3
Bacitracin	1.7	Penicillin G, potassium	6.0	Sodium nitrate	22.7
Barbital sodium	10.0	Penicillin G, sodium	6.0	Sodium phosphate	9.7
Bismuth potassium tartrate	3.0	Pentobarbital sodium	8.3	Sodium propionate	20.3
Boric acid	16.7	Phenobarbital sodium	8.0	Sodium sulfite, exsiccated	21.7
Butacaine sulfate	6.7	Physostigmine salicylate	5.3	Sodium thiosulfate	10.3
Caffeine and sodium benzoate	8.7	Pilocarpine hydrochloride	8.0	Streptomycin sulfate	2.3
Calcium chloride	17.0	Pilocarpine nitrate	7.7	Sulfacetamide sodium	7.7
Calcium chloride (6 H ₂ O)	11.7	Piperocaine hydrochloride	7.0	Sulfadiazine sodium	8.0
Chlorobutanol (hydrated)	8.0	Polymyxin B sulfate	3.0	Sulfamerazine sodium	7.7
Chlortetracycline sulfate	4.3	Potassium chloride	25.3	Sulfapyridine sodium	7.7
Cocaine hydrochloride	5.3	Potassium nitrate	18.7	Sulfathiazole sodium	7.3
Cupric sulfate	6.0	Potassium phosphate, monobasic	14.7	Tetracaine hydrochloride	6.0
Dextrose, anhydrous	6.0	Procainamide hydrochloride	7.3	Tetracycline hydrochloride	4.7
Dibucaine hydrochloride	4.3	Procaine hydrochloride	7.0	Viomycin sulfate	2.7
Dihydrostreptomycin sulfate	2.0	Scopolamine hydrobromide	4.0	Zinc chloride	20.3
Ephedrine hydrochloride	10.0	Scopolamine methylnitrate	5.3	Zinc sulfate	5.0
Ephedrine sulfate	7.7	Secobarbital sodium	8.0		

^a This table of *Isotonic Solution Values* shows volumes in mL of water to be added to 300 mg of the specified drug in sterile water to produce an isotonic solution. The addition of an isotonic vehicle (commonly referred to as diluting solution) to make 30 mL yields a 1% solution. Solutions prepared as directed above are iso-osmotic with 0.9% sodium chloride solution but may not be isotonic with blood (see Appendix A for hemolysis data).

^b The V values for drugs that do not appear in Appendix B but are listed in Appendix A can be calculated from the sodium chloride equivalent for 1% drug. Example—Calculate the V value for anileridine HCl (Appendix A defines E = 0.19).

$$\frac{100 \text{ mL Soln}}{0.9 \text{ NaCl}} \times \frac{0.19 \text{ g NaCl}}{1 \text{ g drug}} \times 0.3 \text{ g drug} = 6.33 \text{ mL Soln}$$

for dilute solution

$$6.33 \text{ mL soln} \cong 6.33 \text{ mL water} \therefore V = 6.33 \text{ mL water}/0.3 \text{ g drug}.$$

Chemical Kinetics

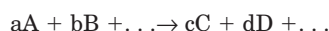
Rodney J Wigent, PhD

Thermodynamic parameters, such as ΔG , ΔE , ΔH , and ΔS , are state functions that only depend on the initial and final states of a chemical process—reactants and products—and are independent of the pathway taken to get to the final state from the initial state. *Chemical kinetics* is the discipline that is concerned with the mechanism by which a chemical process gets to its final state from its initial state and the rate in which this reaction proceeds. Therefore, chemical kinetics involves the study of rate of chemical change and the way in which this rate is influenced by the conditions of the concentration of reactants, products, and other chemical species that may be present, and by factors such as solvent, pressure, and temperature. From these studies, one or more mechanisms involving a series of elementary processes may be postulated to explain how the reactants are converted to products during a chemical process. Applied to pharmaceuticals, such information permits a rational approach to the stabilization of drug products, and prediction of shelf life and optimum storage conditions.

This chapter is intended as a general introduction to this subject. A comprehensive review of experimental approaches and interpretation of data can be found in several texts, such as the books by House, Epenson and Houston, and the compilation of information relative to kinetic studies on pharmaceuticals by Garrett.¹

REACTION RATE

The rate of a reaction is the velocity with which a reactant or reactants undergoes a chemical change. Experimentally, the rate of a reaction must be determined by directly or indirectly following the change in the concentration of the reactants or products as a function of time. When there is more than one reactant, such changes need to be normalized according to the stoichiometry of the reaction. For a reaction of the type



where the uppercase letters represent chemical species and the lowercase letters represent stoichiometric coefficients, the rate in which reactants go to products can be determined by following the rate of the disappearance of the reactants as a function of time

$$\text{Rate} = -\frac{1}{a} \frac{d[A]}{dt} = -\frac{1}{b} \frac{d[B]}{dt} \quad (1)$$

The brackets denote concentration (usually molar concentration unless otherwise indicated) and d represents the derivative function. The negative sign signifies that the concentration of the reactants is decreasing, as the rate must always be positive as long as the reaction is progressing from reactants to products.

The rate at which a reaction proceeds for the reaction type shown above also can be determined by following the appearance of the products as a function of time:

$$\text{Rate} = +\frac{1}{c} \frac{d[C]}{dt} = +\frac{1}{d} \frac{d[D]}{dt} \quad (2)$$

where the positive signs indicate that the concentrations of the products are increasing. Note that these two expressions for rate are only for the type of reaction where the reactants go irreversibly to products, without going through any intermediates.

If $[A]_0$, $[B]_0$, $[C]_0$, and $[D]_0$ represent the initial concentration (ie, $t = 0$) of each of the reactants and products, at some time t (ie, $t = t$), the concentration of A decreases by aX (ie, $[A]_t = [A]_0 - aX$) and the concentration of B decreases by bX (ie, $[B]_t = [B]_0 - bX$).

Similarly, the concentrations of the products C and D increase by cX and dX , respectively (ie, $[C]_t = [C]_0 + cX$ and $[D]_t = [D]_0 + dX$). Thus, upon normalization, the rate expressed in Equations 1 and 2 reduces to Equation 3.

$$\text{Rate} = +\frac{dX}{dt} \quad (3)$$

The *law of mass action* relates these experimentally determined rates to the concentration of all of the reacting species. This law states that, at a given temperature, the rate of the reaction is at each instant proportional to the product of the concentration of each of the reacting species raised to a power equal to the number of molecules of each species participating in the process. Accordingly, the law of mass action applied to the above reaction gives the following rate equation,

$$\text{Rate} = k[A]^n[B]^m \dots \quad (4)$$

where the proportionality constant k (referred to as the *specific rate constant* or as the *rate constant*) should be independent of the concentrations of all chemical species. The exponents n and m are known as the *order of the reaction* with respect to the components A and B, respectively; their sum represents the overall order of the reaction.

It is important to note that for a *net equation*, which is the sum of two or more elementary equations, there is no requirement that the order of the reaction with respect to a chemical species be identical to its stoichiometric coefficient in the net equation. Further, a proper rate equation should only consist of chemical species that are either reactants or products and should not contain any chemical species that are an intermediate during a chemical reaction.

It should be noted that unless the stoichiometric coefficient of the reactant or product that is being followed to determine the rate of the reaction is *unity* (one), the rate of the reaction is not equivalent to the change in the concentration of the chemical species with respect to time. For the case where there is only one chemical reactant, which has a stoichiometric coefficient that is greater than one, authors of articles and textbooks on kinetics often base the reaction rate only on the disappearance of the reactant without accounting for the stoichiometry. When this occurs, the resulting rate constant will be greater than the true rate constant by a factor equal to the stoichiometric coefficient. Thus, care must be taken to determine how the rates of reactions were determined when comparing rate constants of a reaction.

FIRST-ORDER REACTIONS

When the rate of a reaction is proportional to the first power of the concentration of a reactant, the rate equation is given by

$$\frac{dX}{dt} = k[A]_t = k([A]_0 - aX) \quad (5)$$

where a represents the stoichiometric coefficient for reactant A . For the case where $a = 1$, rearrangement of Equation 5 gives

$$\int \frac{dX}{([A]_0 - X)} = k \int dt \quad (6)$$

When Equation 6 is integrated over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$), the following first-order integrated rate equation is obtained:

$$[A]_t = [A]_0 e^{-kt} \quad (7)$$

Figure 19-1 shows a typical plot where reactant A exponentially decays to products according to Equation 7. The rate of the reaction—that is, the negative value of the tangent of this curve at any time—decreases with time as the concentration of the reactant decreases. Equation 7 can be linearized by rearrangement to give Equation 8.

$$\ln [A]_t = -kt + \ln [A]_0 \quad (8)$$

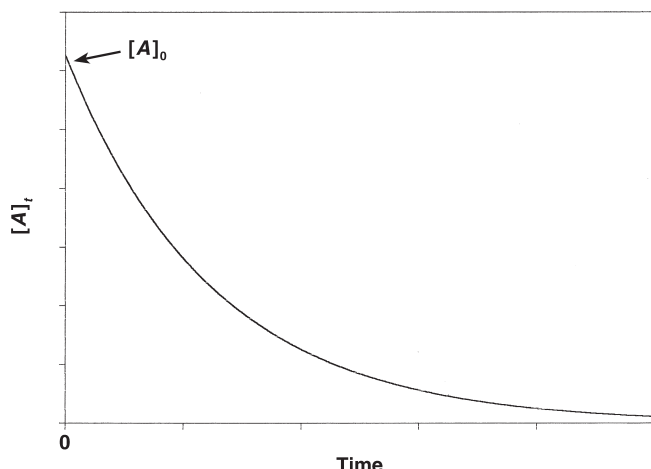


Figure 19-1. Plot of concentration of A versus time for a first-order reaction.

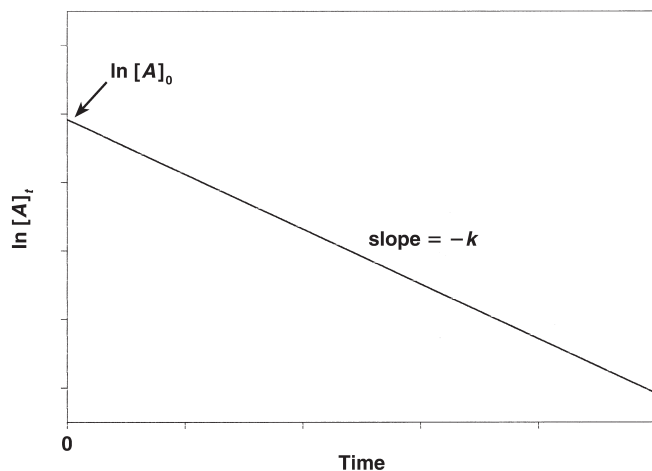


Figure 19-2. Plot of natural logarithm of the concentration of A versus time for a first-order reaction.

Equation 8 suggests that a plot of the natural logarithm of the concentration of the reactant as a function of time should give a linear plot with a slope equal to $-k$ and a y -intercept equal to the natural logarithm of the initial concentration of the reactant (Fig 19-2). Commonly a plot of the common logarithm of the concentration versus time is found in the literature for first-order reactions. In this case, according to Equation 9, the slope of this line would be equal to $-k/2.303$, and the y -intercept would be equal to the common logarithm of the initial concentration of the reactant.

$$\log [A]_t = -\frac{kt}{2.303} + \log [A]_0 \quad (9)$$

The rate constant, k , for a first-order reaction has a unit of reciprocal time (eg, s^{-1}).

Sometimes it may be necessary to determine the rate constant k from only two concentrations of the reactant, $[A]_1$ and $[A]_2$, obtained at two different times, t_1 and t_2 , in which case Equation 10 may be used.

$$k = \frac{1}{(t_2 - t_1)} \ln \frac{[A]_1}{[A]_2} \quad (10)$$

Another useful method for determining k is the fractional-life method, of which the half-life method is the most common. The *half-life method* involves measuring the time ($t = t_{1/2}$) that it takes for half of the initial concentration of the reactant to undergo reaction: $[A]_t = [A]_0/2$. Substituting these values into Equation 7 and rearranging to solve for k yields

$$k = \frac{\ln 2}{t_{1/2}} \quad (11)$$

It is apparent from Equation 11 that the half-life period for first-order reactions is constant and independent of the amount of reactant present. Thus, half of the initial concentration of the reactant undergoes reaction during the first half-life period, leaving 50% of the original concentration unreacted. During the second half-life period, which is identical to the time as the first half-life period for a first-order reaction, half of the remaining reactant reacts, leaving 25% of the initial concentration of the reactant unreacted. Similarly, after the third half-life period, 12.5% of the initial reactant would remain. After 10 first-order, half-life periods, only 0.098% of the original reactant remains unreacted. For precise studies, the rate of disappearance of a reactant should be followed over two or three half-life periods.

In some drug stability studies, it is necessary to determine the time that it takes for the loss of 10% of the drug, leaving 90% of the original drug concentration; that is, $[A]_t = 0.90[A]_0$

at $t = t_{0.90}$. This time can be determined with the knowledge of the rate constant by substituting these expressions into Equation 7 and rearranging to yield

$$t_{0.90} = \frac{\ln 0.90}{k} \quad (12)$$

First-order rate processes are not restricted to chemical reactions. The passive diffusion of drugs across biological membranes, and processes of drug absorption, distribution, metabolism, and excretion often can be shown to occur at rates proportional to the concentration of a drug, and thus can be described as first-order rate processes. The rate of growth of microorganisms and the rate of killing or inactivation of microorganisms by heat or chemical agents usually follow first-order kinetic processes. Radioactive decay always follows first-order kinetics.

SECOND-ORDER REACTIONS

There are two forms of second-order reactions. For the first case, it is assumed that the rate of reaction is proportional to the concentration of reactant A raised to the power of 2—that is, the reaction is second order with respect to A, in which case the rate equation takes the form

$$\frac{dX}{dt} = [A]_t^2 = ([A]_0 - aX)^2 \quad (13)$$

where a represents the stoichiometric coefficient of the reactant in the net equation. For the case where the stoichiometric coefficient of reactant A is 2, Equation 13 can be rearranged to give

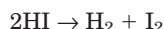
$$\int \frac{dX}{([A]_0 - 2X)^2} = k \int dt \quad (14)$$

When Equation 14 is integrated over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$), the following second-order integrated rate equation is obtained.

$$\frac{1}{[A]_t} = 2kt + \frac{1}{[A]_0} \quad (15)$$

It should be noted that since the stoichiometry was taken into account in this derivation, the stoichiometric coefficient, 2, has been incorporated into Equation 15. If the rate of reaction was determined solely on the disappearance of reactant A without considering the stoichiometry, then the rate constant for this reaction would be twice as large as the true rate constant. This occurs quite frequently in the literature, so the reader should be aware of this situation.

The decomposition of hydrogen iodide is a second-order reaction; in the gaseous state, hydrogen iodide forms hydrogen gas and molecular iodine according to the reaction



The integrated rate expression for this reaction follows the form given by Equation 15.

Equation 15 suggests that, for a second-order reaction, if the reciprocal of the concentration of reactant A is plotted as a function of time, the slope of the line is equal to the rate constant k , and the y -intercept is the reciprocal of the initial concentration of A (Fig 19-3). Rearranging Equation 15 and solving for k yields

$$k = \frac{1}{t} \frac{[A]_0 - [A]_t}{[A]_0[A]_t} \quad (16)$$

The rate constant for second-order reactions has units of reciprocal concentration and seconds (eg, $\text{M}^{-1}\text{s}^{-1}$).

The second type of a second-order reaction occurs if the rate of the reaction is proportional to the product of the concentrations of two reactants, each raised to the power of 1, that is, first order with respect to both reactants. Equation 17 shows the rate equation for such a reaction.

$$\frac{dX}{dt} = [A]_t[B]_t = ([A]_0 - aX)([B]_0 - bX) \quad (17)$$

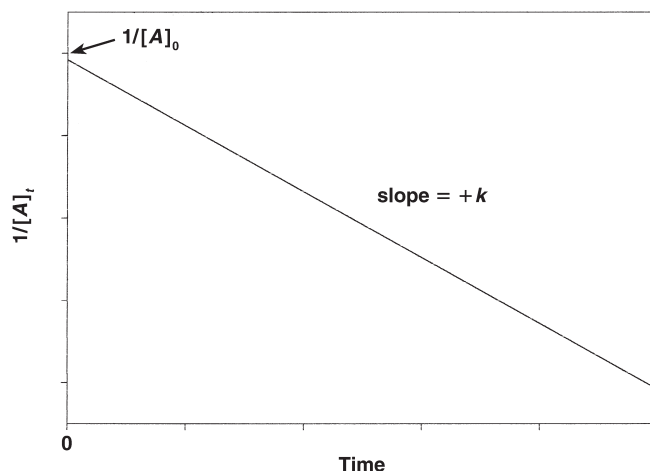


Figure 19-3. Plot of the reciprocal of the concentration of A versus time for a second-order reaction.

The stoichiometric coefficients of the reactants A and B are represented by a and b . For the case where a and b both equal 1, Equation 17 can be arranged to give the following:

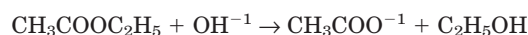
$$\int \frac{dX}{([A]_0 - X)([B]_0 - X)} = k \int dt \quad (18)$$

When Equation 18 is integrated over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$), the following second-order integrated rate equation is obtained:

$$\ln \frac{[A]_t}{[B]_t} = ([A]_0 - [B]_0)kt + \ln \frac{[A]_0}{[B]_0} \quad (19)$$

This suggests that if the left side of Equation 19 is plotted against time, the slope of the line would be equal to $([A]_0 - [B]_0)k$ and the y -intercept is equal to the natural logarithm of the ratio of the initial concentrations of reactants A and B. Equation 19 does not apply if the initial concentration of the two reactants are equal; in this case, Equation 18 reduces to Equation 14 and the integrated rate equation for this system reduces to Equation 15.

An example of a second-order reaction in which two reactants are involved is the saponification of an ester, such as ethyl acetate, in alkaline solution:



The course of this reaction may be followed by determining, by titration at specified times, the concentration of hydroxide ions remaining unreacted during the course of the reaction. This information and the initial concentrations of the ethyl acetate and hydroxide can be used to determine the rate constant in Equation 19.

Fractional-life methods can be applied readily to second-order reactions for the case when the order of the reaction with respect to one reactant is 2, or for the case when the initial concentrations of each of two reactants are equal when the order with respect to each reactant is 1. For example, the half-life of a second-order reaction is given by Equation 20.

$$t_{1/2} = \frac{1}{k[A]_0} \quad (20)$$

Unlike the half-life period for a first-order reaction, the half-life period for a second-order reaction is not constant, but rather is proportional to the reciprocal of the initial concentration of reactant. This means that the half-life period increases as a second-order reaction proceeds with time; thus, it takes twice as long to deplete a second-order reactant from 50 to 25% as it did to deplete the reactant from 100 to 50%.

THIRD-ORDER REACTIONS

Except for in the solution phase, third-order reactions are rare, as they require a simultaneous three-body collision of chemical species. There are a number of ways in which third-order reaction can occur—from a combination of three different chemical entities, for which the order of the reaction with respect to each of these is 1, to the simplest case in which three identical substances react, for which the order of the reaction with respect to that species is three. For the latter case, assuming the stoichiometric coefficient of the single reacting entity A is 3, then the rearranged rate equation is given by Equation 21.

$$\int \frac{dX}{([A]_0 - 3X)^2} = k \int dt \quad (21)$$

Upon integration of Equation 21 over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$), the following third-order integrated rate equation is obtained.

$$\frac{1}{[A]_t^2} = 6kt + \frac{1}{[A]_0^2} \quad (22)$$

Again, it should be noted, that if the stoichiometry was not taken into account and the *rate* was only determined by the rate of disappearance of reactant A , then Equation 22 would have 2 for the coefficient of kt instead of 6 and the value for the rate constant would be three times the value of the rate constant in Equation 22.

The equation for the half-life period for the case of Equation 22 is given by

$$t_{1/2} = \frac{1}{2k[A]_0^2} \quad (23)$$

Another type of a third-order reaction occurs when the rate of the reaction is proportional to the product of the concentrations of two reactants, one raised to the power of 1 and the other raised to the power of 2; it is first order with respect to one reactant and second order with respect to the other reactant. Equation 24 shows the rate equation for such a reaction.

$$\frac{dX}{dt} = [A]_t^2 [B]_t = ([A]_0 - aX)^2 ([B]_0 - bX) \quad (24)$$

If the stoichiometric coefficients, a and b , are both equal to 1, then Equation 24 can be rearranged and integrated over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$). The rate constant from this resulting equation is given by

$$k = \frac{1}{t} \frac{1}{[B]_0 - [A]_0} \frac{[A]_0 - [A]_t}{[A]_0 [A]_t} + \frac{1}{([B]_0 - [A]_0)^2} \ln \frac{[A]_t [B]_0}{[B]_t [A]_0} \quad (25)$$

However, if the stoichiometric coefficients are $a = 2$ and $b = 1$, then when Equation 24 is rearranged and integrated over the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$), the rate constant is determined by Equation 26.

$$k = \frac{1}{t} \frac{1}{2[B]_0 - [A]_0} \frac{[A]_0 - [A]_t}{[A]_0 [A]_t} + \frac{1}{(2[B]_0 - [A]_0)^2} \ln \frac{[A]_t [B]_0}{[B]_t [A]_0} \quad (26)$$

The rate constant for third-order reactions has units of reciprocal of the square of concentration per second (eg, $M^{-2}s^{-1}$).

Because of the rigors of the mathematics, when a third-order reaction is suspected, experimental conditions are often chosen so as to simplify the calculations. For example, for the third-order reaction in which the stoichiometric coefficients of the two reacting species are $a = 2$ and $b = 1$, such as that which led to the development of Equation 26, if the experimental conditions are set such that $[A]_0 = 2[B]_0$, it will lead to a much simpler integrated rate equation.

PSEUDO-ORDER REACTIONS

For some reactions, the rate of the reaction may be independent of the concentration of one or more of the reacting species over a wide range of concentrations. This may occur under these conditions:

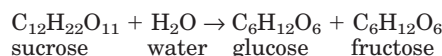
1. One or more of the reactants enters into the rate equation in great excess compared to the others.
2. One of the reactants is a catalyst.
3. One or more of the reactants is constantly replenished during the course of a reaction.

If this happens, the constant concentration term(s) in the rate equation is combined with the rate constant to give an *apparent rate constant*. For example, if the concentration of A in Equation 4 remains constant, then Equation 4 can be rewritten as

$$\text{Rate} = (k[A]^n)[B]^m \dots = k_{\text{app}}[B]^m \dots \quad (27)$$

where the apparent rate constant, k_{app} (sometimes referred to as the *pseudo-order rate constant*) now depends on the concentration of A raised to its power, n . Unfortunately, no information about n , the order of reaction with respect to A , can be determined from a single experiment. Rather, to gain an understanding of n , multiple experiments must be performed where the concentration of A is varied. A plot of the natural logarithm of k_{app} versus the natural logarithm of the concentration of A will give a slope that is equal to n .

In 1850, Wilhelmy performed the first quantitative kinetics study by following the rate of hydrolysis (inversion) of sucrose to glucose and fructose, according to the reaction



Wilhelmy found that this reaction followed the rate equation

$$-\frac{d[\text{C}_{12}\text{H}_{22}\text{O}_{11}]}{dt} = k_{\text{app}}[\text{C}_{12}\text{H}_{22}\text{O}_{11}] \quad (28)$$

which, upon rearrangement and integration, gives Equation 29.

$$\ln[\text{C}_{12}\text{H}_{22}\text{O}_{11}]_t = -k_{\text{app}}t + \ln[\text{C}_{12}\text{H}_{22}\text{O}_{11}]_0 \quad (29)$$

This reaction is now known to be a second-order reaction, as it is first order with respect to both sucrose and water. As for most typical aqueous solutions, the molar concentration of water (approximately 55.5 mol of water per liter) greatly exceeds the concentration of the solute sucrose. Therefore, even at moderate concentrations of sucrose, there is only a minor change in the molar concentration of water and the concentration of the solvent is practically constant over the course of the reaction. This allows the concentration of water to be incorporated into the apparent rate constant and the reaction appears to be first order.

As another example, if component A reacts in aqueous solution to go to product B , according to the first-order rate equation given by Equation 5 and the stoichiometric coefficient a is 1 (unity), then the concentration of A as a function of time should follow the exponential form of the integrated rate equation given by Equation 7. However, if this reaction occurs in a saturated solution of A (ie, $[A]_{\text{sat}}$) in the presence of excess solid A , and if the rate of converting solid A to aqueous A is greater than the rate of reaction in solution, then the rate of disappearance of A is given by

$$\frac{dX}{dt} = k[A]_{\text{sat}} = k_{\text{app}} \quad (30)$$

If Equation 30 is rearranged and integrated between the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$) and defining $[B]_t = X$, the following zero-order rate equation is obtained,

$$[B]_t = k_{\text{app}}t \quad (31)$$

which shows that as long as the solution remains saturated with A , the formation of B will occur at a constant rate. As an example, if a compound for which decomposition in solution is first order is present in excess of its maximum solubility (a sus-

pension), the concentration of the reactant in solution will be invariant so long as there is excess solid reactant present. The kinetics of such a system would then follow Equation 30.

First- and second-order reactions are by far the most common types of rate processes encountered regarding drug stability. If a reaction is of higher order than first order, it often is convenient to adjust experimental conditions so that the concentrations of all but one of the reactants remain constant throughout the experiment. If, for example, the concentration of hydroxide ion in the saponification of an ester is in great excess of the concentration of ester, or if a buffer system is employed to control hydroxide-ion concentration, then the concentration of hydroxide ion essentially is invariant throughout the course of the experiment. The observed rate of the reaction, therefore, depends only on the changing concentration of the ester, and the reaction is said to be *pseudo first-order*. The apparent first-order rate constant, k_{app} , thus obtained is $k[\text{OH}^{-1}]$ and, of course, is different for each hydroxide-ion concentration. The actual rate constant, k , can be obtained easily by dividing the experimentally determined apparent first-order rate constant, $k[\text{OH}^{-1}]$, by the concentration of hydroxide ion maintained throughout the study.

In the study of complex reactions, it is often desirable to use this approach of maintaining the concentration of all but one of the reactants constant to facilitate determining the dependency of the reaction rate on each of the reactants in turn.

MORE COMPLEX REACTIONS

Many chemical reactions do not follow the simple reaction kinetics listed above, but rather they often consist of two or more elementary processes that may lead to more complicated rate equations. For example, comparison of experimental measurements of the rate of the disappearance of the reactants and the appearance of the products may indicate that the reactants must be forming one or more intermediates before proceeding to form the products. Often chemical reactions proceed reversibly to form products before an equilibrium is established. There are many cases where the reactants simultaneously proceed through different mechanisms to form two or more products. These situations can lead to negative or noninteger orders of reactions with respect to reactants and products within the rate equation. Quite often, a series of experiments must be performed in which certain conditions are controlled in order to establish the order of the reaction of individual species involved in the chemical reaction before an overall rate equation can be established. The next several sections will look at some of the more common complex reactions.

Reversible Reactions

Many reactions are known to be reversible where the reactants go to form products but the products will reversibly revert back to reactants. The simplest example of this is in the case where reactant A follows a first-order kinetic process with a forward rate constant, k_f , to produce product B .



However, product B then follows a first-order rate process with a reverse rate constant, k_r , to reform reactant A .



Because, during the course of this reaction, reactant A is being simultaneously depleted and formed, the rate at which reactant A disappears is related to the forward and reverse rates according to Equations 32 and 33:

$$-\frac{d[A]}{dt} = \frac{d[A]_{\text{forward}}}{dt} - \frac{d[A]_{\text{reverse}}}{dt} \quad (32)$$

$$-\frac{d[A]_t}{dt} = k_f[A]_t - k_r[B]_t \quad (33)$$

If the initial concentration of B is zero, then at time $t = 0$ (initially) the rate equation is given solely by the forward rate equation. As the reaction proceeds, the reverse rate equation begins to contribute more and more substantially to the overall rate equation. Finally, a point will be reached at which the rate of the forward reaction is equal to the rate of reverse reaction and the overall rate is equal to 0. This is defined as a *dynamic equilibrium* and the concentration equilibrium constant, K_c , given by Equation 34, is equal to the ratio of the forward and reverse rate constants, where

$$K_c = \frac{[B]_{\text{eq}}}{[A]_{\text{eq}}} = \frac{k_f}{k_r} \quad (34)$$

$[B]_{\text{eq}}$ and $[A]_{\text{eq}}$ are the equilibrium concentrations of the product and reactant, respectively.

The rate equation expressed in Equation 33 can be rewritten to give

$$\frac{dX}{dt} = k_f([A]_0 - X) - k_rX \quad (35)$$

Rearrangement and integration of Equation 35 between the limits of $t = 0$ (at which $X = 0$) to $t = t$ (at which $X = X$) and defining $[B]_t = X$, the following expression for the concentration of A as a function of time is obtained:

$$[A]_t = \frac{k_f[A]_0 \exp[-(k_f + k_r)t] + k_r[A]_0}{(k_f + k_r)} \quad (36)$$

Simultaneous Reactions

Another very common reaction is when the reaction of one or more reactants lead to the formation of multiple products through different mechanistic pathways, each with characteristic rates:



and



For the case in which both reaction pathways are first order, then the rate of disappearance of reactant A is then given by Equation 37.

$$-\frac{d[A]_t}{dt} = k_1[A]_t + k_2[A]_t = (k_1 + k_2)[A]_t \quad (37)$$

Rearrangement and integration of Equation 37 gives

$$[A]_t = [A]_0 \exp[-(k_1 + k_2)t] \quad (38)$$

Since the rate of formation of product B is given by

$$\frac{d[B]}{dt} = k_1[A]_t \quad (39)$$

then, assuming that the initial concentration of B is 0, rearranging and integrating, and substituting Equation 38 into Equation 39 yields the following expression for the concentration of B as a function of time:

$$[B]_t = \frac{k_1[A]_0}{k_1 + k_2} (1 - \exp[-(k_1 + k_2)t]) \quad (40)$$

Using similar arguments, the concentration of C as a function of time is given by Equation 41.

$$[C]_t = \frac{k_2[A]_0}{k_1 + k_2} (1 - \exp[-(k_1 + k_2)t]) \quad (41)$$

It is of particular interest to note that if Equation 40 is divided by Equation 41, the ratio of the concentration of the products at any time is given by the ratio of the rate constants.

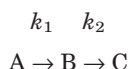
$$\frac{[B]_t}{[C]_t} = \frac{k_1}{k_2} \quad (42)$$

An example of this type of simultaneous reaction is the reaction of phenol with nitric acid to form both ortho- and para-nitrophenol through two simultaneous first-order reaction pathways. The relative concentrations of these two products is found to be given by Equation 42.

It is clear that if a kinetic experiment was performed without any a priori knowledge that the reaction is a simultaneous reaction, there is a danger that only the disappearance of a reactant or the appearance of only one of the products may lead to a faulty conclusion of the reaction mechanism. Care must be taken to attempt to identify and account for all of the chemical species in a chemical reaction to ensure that a proper rate mechanism is obtained.

Consecutive Reactions

One of the more common complex reactions is when a reactant decays through a series of consecutive reactions, forming one or more intermediates before forming a product. A simple case of a consecutive reaction is when reactant A proceeds through a first-order process to intermediate B which then decays to product C through another first-order process.



For cases such as this, it is often convenient to consider the situation in which the initial concentrations of B and C are 0 and the sum of the concentrations of A, B, and C at any time is equal to the initial concentration of the reactant A. In this case, the rate of disappearance of A is given by Equation 43 and the rate of appearance of product C is given by Equation 44.

$$-\frac{d[A]_t}{dt} = k_1[A]_t \quad (43)$$

$$\frac{d[C]_t}{dt} = k_2[B]_t \quad (44)$$

The derivative of the concentration of the intermediate B with respect to time consists of the rate of formation of B from the product A and the disappearance of B as it proceeds to product C, as shown by

$$\frac{d[B]_t}{dt} = k_1[A]_t - k_2[B]_t \quad (45)$$

Upon integration and rearrangement of Equation 43, the concentration of reactant A as a function of time can be expressed by

$$[A]_t = [A]_0 \exp(-k_1 t) \quad (46)$$

It should be noted that Equations 44 and 45 are not considered to be valid rate equations because, by convention, the concentration of an intermediate may not appear in a final rate equation. Therefore, an expression for the concentration of B as a function of time in terms of only the reactant or product must be developed. Substituting Equation 46 into Equation 45 and rearranging and integrating yields the following expression for the concentration of B as a function of time.

$$[B]_t = \frac{k_1[A]_0}{k_2 - k_1} (\exp[-k_1 t] - \exp[-k_2 t]) \quad (47)$$

Equation 47 can be substituted into Equations 44 and 45 to give appropriate rate expressions. Then Equation 45 can be rear-

ranged and integrated to give an expression for the concentration of C as a function of time.

$$[C]_t = \frac{[A]_0}{k_2 - k_1} (k_2 - k_2 \exp[-k_1 t]) - (k_1 - k_1 \exp[-k_2 t]) \quad (48)$$

EFFECTS ON REACTION RATE

Temperature

The application of heat to increase the rate of a chemical reaction is a common laboratory procedure. The rate of most solvolytic reactions of pharmaceuticals is increased roughly 2- to 3-fold by a 10° increase near room temperature. In 1889 Arrhenius noted that the variation with temperature of the rate constant of chemical reactions could be expressed by

$$k = A \exp[-E_a/RT] \quad (49)$$

where, according to collision theory, E_a is the Arrhenius activation energy (ie, the difference between the average energy of reactive molecules and the minimum energy required for reactants to proceed to products); $\exp[-E_a/RT]$ is the Boltzmann factor, which represents the fraction of molecules having energies greater than or equal to E_a ; the pre-exponential term A is a constant called the frequency factor; R is the gas constant (8.314 joule/mol-K or 1.987 cal/mol-K); and T is the absolute temperature. The Arrhenius equation can be expressed in a linear form according to Equation 50.

$$\ln k = \frac{E_a}{R} \frac{1}{T} + \ln A \quad (50)$$

Equations 49 and 50 are valid so long as the reaction mechanism does not change over the temperature range studied; a plot of the natural logarithm of the rate constant versus the reciprocal of the absolute temperature in which the rate constants are determined gives a negative slope that is equivalent to $-E_a/R$ (Fig 19-4). If a nonlinear plot is obtained, a thermally induced change in the reaction mechanism probably has occurred.

Differentiation of Equation 50 with respect to temperature, and then integrating between the limits of k_2 and k_1 at temperatures between T_2 and T_1 yields

$$\ln \frac{k_2}{k_1} = \frac{E_a}{R} \frac{T_2 - T_1}{T_2 T_1} \quad (51)$$

This equation allows E_a to be calculated for a reaction when the rate constants are known at two temperatures, or the rate con-

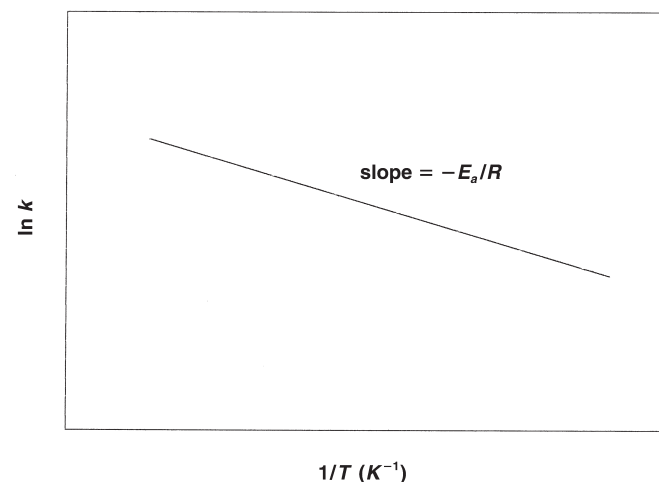


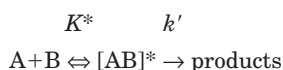
Figure 19-4. Variation of the rate constant with reciprocal absolute temperature, illustrating the Arrhenius equation.

stant at one temperature to be calculated if E_a and the rate constant at another temperature are known.

Most solvolytic reactions of pharmaceuticals exhibit activation energies in the range of 8 to 20 kcal/mol. Using Equation 50 and the appropriate activation energy, one readily can calculate that a reaction having an activation energy of 8 kcal/mol would show an approximately 1.5-fold increase in k for a temperature increase from 25° to 35°; a reaction having an activation energy of 20 kcal/mol would show a 3-fold increase in k for a similar temperature increase.

When two molecules undergo chemical interaction, it is reasonable to suppose that they first must collide and then, if conditions are right, undergo a rearrangement of certain electrons to form the bonds characteristic of the new molecules. However, not all collisions can cause a chemical change, or else chemical reactions would occur with great rapidity because collision frequencies are very high. While molecules or atoms must first collide if a reaction is to occur, the colliding molecules may not have an energy greater than or equal to the activation energy sufficient to overcome the mutual repulsion of the interacting molecules and enable them to approach close enough to each other to effect certain bond ruptures and/or establish new bonds characteristic of the products. The greater this energy requirement, the smaller the proportion of colliding molecules that will have the necessary energy, and the slower the reaction. In the Arrhenius equation, A is a factor related to frequency of collisions, and $\exp[-E_a/RT]$ is the probability that at temperature T a collision will occur with sufficient energy to provide a successful collision. The concept of energy of activation, in relationship to the energy of the reactants and of the products, is illustrated in Figure 19-5.

Eyring, in his transition state theory, proposed that reactants must proceed through an activated complex before proceeding to reactants. This is demonstrated by the reaction



where the reactants are considered to be in a rapid equilibrium with the activated complex or transition state, represented by $[AB]^*$, which then decays to products by a first-order process, according to the rate equation

$$\text{Rate} = k[AB]^* \quad (52)$$

However, as Equation 52 contains the concentration of the activated complex, an intermediate, it is not a valid rate equation and an expression in terms which include only the reactants or products must be substituted for this expression. Because the activated complex is in equilibrium with the reactants, the concentration of the activated complex can be given by

$$[AB]^* = K^*[A][B] \quad (53)$$

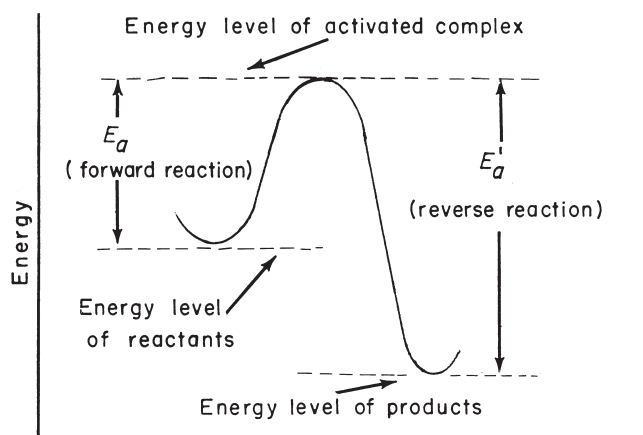


Figure 19-5. Relation between activation energy and energy levels of reactants, products and activated complex.

where K^* is the equilibrium constant. Substituting Equation 53 into Equation 52 yields

$$\text{Rate} = K^*k'[A][B] = k[A][B] \quad (54)$$

where k is equal to K^*k' and Equation 54 is a proper rate equation. Eyring was able to show that the rate constant, k' , of any reaction is given by the expression

$$k' = \frac{RT}{N_a h} K^* \quad (55)$$

where R is equal to 8.314 ergs/mol-K, N_a is Avogadro's number, and h is Planck's constant, which is equal to 6.625×10^{-27} erg-sec. K^* can be related to the thermodynamic parameters ΔG^* , ΔH^* , and ΔS^* through the equation

$$K^* = e^{-\Delta G^*/RT} = e^{(T\Delta S^* - \Delta H^*)/RT} \quad (56)$$

If Equation 56 is substituted into Equation 55, after it has been divided by the absolute temperature, the following linear equation is obtained.

$$\ln \frac{k'}{T} = \ln \frac{R}{N_a h} + \frac{\Delta S^*}{R} - \frac{\Delta H^*}{R} \frac{1}{T} \quad (57)$$

Thus, the thermodynamics of the formation of the activated complex can be determined from a plot of the natural logarithm of the ratio of the rate constant to absolute temperature versus the reciprocal absolute temperature.

CATALYSIS

In catalytic reactions, a molecule, called a catalyst, interacts with a reactant in a series of elementary processes in such a fashion as to lower the activation energy barrier (ie, E_a in Fig 19-5) of an uncatalyzed reaction. This change in mechanism causes the catalyzed reaction to run faster without changing the relative energy levels of either the reactants or products. During a catalytic reaction, the catalyst reacts with a reactant to form an intermediate that undergoes additional reaction(s) to form the product and the original catalytic molecule. Thus, while the catalyst is both consumed and produced in several elementary processes, there is no net change in the concentration of the catalyst during a catalyzed reaction but there is usually a substantial change in the rate in which the reaction occurs.

While there are several types of catalysis, such as homogeneous and heterogeneous catalysis, autocatalysis, etc., only a few general examples will be discussed here. Additional information about catalysis can be found in some of the references cited in the attached bibliography.

Specific Acid and Specific Base Catalysis

The terms *specific acid catalysis* and *specific base catalysis* refer to catalysis by the hydronium or hydrogen ion, and by the hydroxide ion, respectively. For example, if the rate of hydrolysis of an ester, such as ethyl acetate, is studied at a constant pH in a strongly buffered solution, the rate of disappearance of intact ester will be an apparent first-order reaction. If the reaction is studied in solutions buffered at several different pH values in a sufficiently acid pH region, a different apparent first-order rate constant will be observed for each pH value. The observed rate actually depends on the concentration of both the ester and hydrogen ion and, therefore, is a second-order reaction that appears to be a pseudo first-order reaction at the constant hydrogen-ion concentration in the buffer. Therefore, the observed first-order rate constant, k_{obs} , is proportional to the hydrogen ion concentration of the buffer system as shown by Equation 58.

$$k_{\text{obs}} = k_{\text{acid}} [H^+] \quad (58)$$

Taking the logarithm of Equation 58 yields

$$\log k_{\text{obs}} = \log k_{\text{acid}} + \log [\text{H}^+] \quad (59)$$

which upon applying the definition of pH yields

$$\log k_{\text{obs}} = \log k_{\text{acid}} - \text{pH} \quad (60)$$

Equation 60 suggests that a plot of $\log k_{\text{obs}}$ versus pH will be linear with a slope of -1 and a y -intercept of $\log k_{\text{acid}}$.

Similarly, if the same hydrolysis reaction is studied in buffered solutions at several pH values in a sufficiently alkaline region of the pH scale, the observed apparent first-order rate constants will be found to vary with hydroxide-ion concentration:

$$k_{\text{obs}} = k_{\text{base}} [\text{OH}^-] \quad (61)$$

and

$$\log k_{\text{obs}} = \log k_{\text{base}} + \log [\text{OH}^-] \quad (62)$$

However, the hydroxide ion concentration is related to the hydrogen ion concentration through the ionization constant of water, K_w , and Equation 60 becomes

$$\log k_{\text{obs}} = \log k_{\text{base}} + \log K_w + \text{pH} \quad (63)$$

Therefore, a plot of $\log k_{\text{obs}}$ versus pH in a heavily buffered alkaline solution would yield a straight line with a slope of $+1$ and a y -intercept equal to $\log k_{\text{base}} + \log K_w$.

Because of the equilibrium that exists between hydroxide and hydronium ions in aqueous solutions, each of these ions exists at all values of pH and the observed rate constant is actually given by the sum of Equations 58 and 61.

$$k_{\text{obs}} = k_{\text{acid}} [\text{H}^+] + k_{\text{base}} [\text{OH}^-] \quad (64)$$

The complete logarithm k_{obs} versus pH profile would be similar to that illustrated in Figure 19-6 for the hydrogen ion and hydroxide ion (specific acid and specific base) catalyzed hydrolysis of the ester atropine.² At relatively low values of pH the acid-catalyzed hydrolysis predominates; at relatively high values of pH the base-catalyzed hydrolysis predominates. The pH at which the minimum rate of hydrolysis is observed is a function of the relative magnitude of the specific rate constants k_{acid} and k_{base} . In the atropine example, the minimum rate of hydrolysis is at pH of 3.7, which indi-

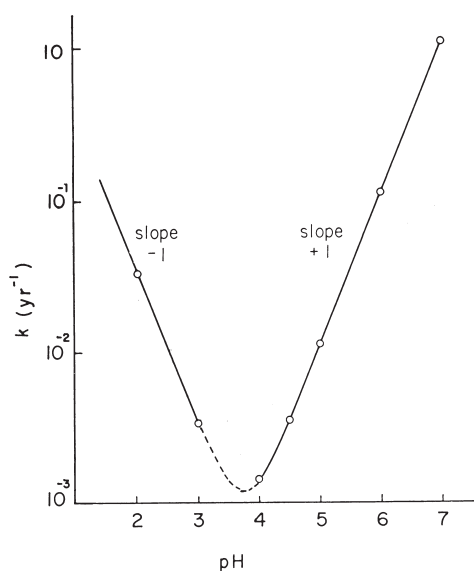


Figure 19-6. Apparent first-order rate of hydrolysis of atropine as a function of pH at 30°. The reaction is an illustration of specific hydrogen and hydroxide-ion catalysis. (From Kondritzer AA, Zvirblis P. *J APhA Sci Ed* 1957; 46: 531.)

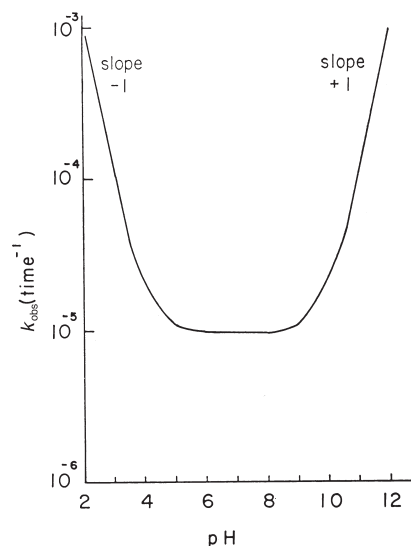


Figure 19-7. Apparent first-order rate of decomposition as a function of pH for a hypothetical case where $k_{\text{H}^+} = k_{\text{OH}^-} = 0.1$, $k_{\text{H}_2\text{O}} = 1 \times 10^{-5}$. The uncatalyzed reaction predominates in the pH region 5 to 9.

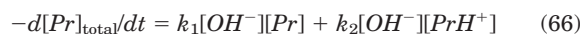
cates that $k_{\text{base}} > k_{\text{acid}}$. If k_{base} equals k_{acid} , then, at 25°, the expected minimum rate of the reaction would be expected to occur at pH 7.

A reaction may be catalyzed not only by hydrogen ion and hydroxide ion, but also by other Brønsted acids or bases such as the solvent water. This is referred to as general acid/base catalysis. In this case, the observed rate constant is given by

$$k_{\text{obs}} = k_{\text{water}} + k_{\text{acid}} [\text{H}^+] + k_{\text{base}} [\text{OH}^-] \quad (65)$$

where k_{water} is a pseudo-order rate constant that has the concentration of water, which is in large excess, incorporated into it. Figure 19-7 shows how a plot of the logarithm k_{obs} versus pH might appear in such a case. The flat region, where the rate of reaction apparently is not pH dependent, is the region where the solvent is much more important as a catalyst than either the hydrogen or hydroxide ions.

For compounds that are weak acids or weak bases, which can therefore exist in both ionized and nonionized species, the pH rate profiles become even more complex. Often, both the ionized and nonionized species are subject to decomposition and catalysis by hydrogen and hydroxide ion; but each of these species may react at a different rates. For example, the hydrolysis of the weakly basic drug procaine can be represented by¹



where Pr is the nonionized procaine molecule and PrH^+ is the protonated form. The concentration of each species can be related to the total procaine concentration by the relationships

$$[\text{Pr}] = \frac{[\text{OH}^-]}{K_b + [\text{OH}^-]} [\text{Pr}]_{\text{total}} \quad (67)$$

and

$$[\text{PrH}^+] = \frac{K_b}{K_b + [\text{OH}^-]} [\text{Pr}]_{\text{total}} \quad (68)$$

where K_b is the classical dissociation constant for the weak base procaine. The complete rate expression for procaine hydrolysis is given by Equation 69.

$$-\frac{d[\text{Pr}]_{\text{total}}}{dt} = \left[\frac{k_1[\text{OH}^-]^2}{K_b + [\text{OH}^-]} + \frac{k_2[\text{OH}^-]K_b}{K_b + [\text{OH}^-]} \right] [\text{Pr}]_{\text{total}} \quad (69)$$

The pH dependency of procaine hydrolysis is illustrated graphically in Figure 19-8³ by a plot of logarithm k_{obs} versus pOH for the pH region 7 to 13.

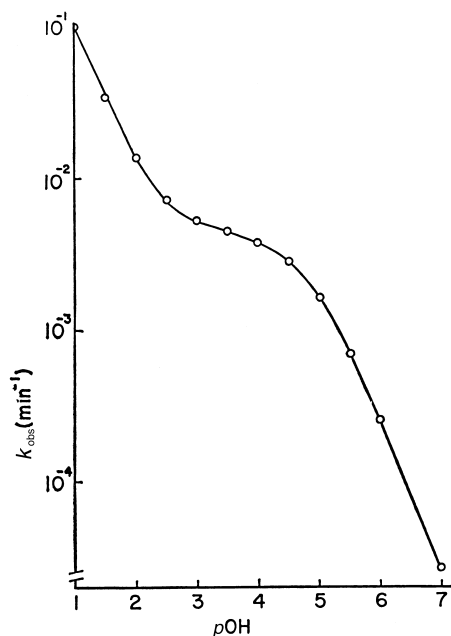


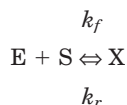
Figure 19-8. Apparent first-order rate of hydrolysis of procaine as a function of hydroxide-ion concentration at 40°. (From Higuchi T, Lachman L. *J APhA Sci Ed* 1955; 44: 52.)

General Acid or Base Catalysis

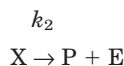
Acid or base catalysis is not restricted to the effect of hydrogen or hydroxide ion. Undissociated acids and bases often can be demonstrated to produce a catalytic effect, and in some instances metal ions and various anions can serve as catalysts. Mutarotation of glucose in acetate buffer is catalyzed by hydrogen ion, hydroxide ion, acetate ion, and undissociated acetic acid. Also, the rate of barbiturate hydrolysis in ammonia buffers is increased by increasing buffer concentration at constant pH as a result of catalysis by NH_3 . Hydrolysis of the amide function of chloramphenicol exhibits, in addition to solvent and specific acid–base catalysis, general acid–base catalysis in phosphate and citrate buffers. General acid–base catalysis is to be anticipated if there is evidence of a significant solvent catalysis, as illustrated in the pH–rate profile of Figure 19-7.

Enzyme Catalysis

In biological systems, catalytic molecules, called enzymes (E), reversibly bind to a substrate (S) to form an intermediate (X) which then decomposes to give a product (P) and the original enzyme.



and



The rate of this reaction, v , will be given by $d[\text{P}]/dt = k_2[\text{X}]$. However, this is an improper rate equation as X is an intermediate. Michaelis and Menten used a steady state approximation (ie, at sometime during the reaction the time rate of change of the intermediate will be zero) to calculate the concentration of X.

$$[\text{X}] = \frac{k_f[\text{E}]_0[\text{S}]_0}{k_f[\text{S}]_0 + k_r + k_2} \quad (70)$$

Upon substitution of Equation 70 into the rate equation, the initial rate, v_0 , (ie, The rate at time equal to zero) is given by Equation 71.

$$v_0 = \frac{k_f k_2 [\text{E}]_0}{k_f + \frac{k_r + k_2}{[\text{S}]_0}} = \frac{v_m}{1 + \frac{K_m}{[\text{S}]_0}} \quad (71)$$

where K_m , the Michaelis-Menten constant, is equal to $(k_r + k_2)/k_1$ and v_m , the maximum initial velocity of the reaction, is equal to $k_2 [\text{E}]_0$. The rate constant k_2 is often referred to as the turnover number which represents the number of molecules of product P created per second per mole of enzyme.

Equation 70 does not lend itself well to analysis as plots of v_0 vs. $[\text{S}]_0$ only asymptotically approaches the maximum velocity, v_m . Rearrangement of Equation 71 into Equation 72, known as the Lineweaver-Burk Equation, lends itself more readily to analysis as shown in Figure 19-9.

$$\frac{1}{v_0} = \frac{1}{v_m} + \frac{K_m}{v_m [\text{S}]_0} \quad (72)$$

An important area of study in enzyme kinetics is enzyme inhibition. There are two basic mechanisms in which an inhibitor, I, can inhibit and enzyme-catalyzed reaction. Competitive inhibition occurs when the inhibitor competes with the substrate, S, for the active binding site on the enzyme and blocks the catalytic action of the enzyme. In such a case, the formation of the enzyme-inhibitor complex is assumed to be in rapid equilibrium with the enzyme and inhibitor.



Equation 73 shows the resulting Lineweaver-Burk Equation for the case of a competitive inhibitor.

$$\frac{1}{v_0} = \frac{1}{v_m} + \left[1 + \frac{[\text{I}]}{K_I} \right] \frac{K_m}{v_m [\text{S}]_0} \quad (73)$$

K_I is the dissociation constant for the enzyme-inhibitor complex, $[\text{E}][\text{I}]/[\text{EI}]$. Figure 19-10 shows a typical Lineweaver-Burk graph for competitive inhibition. Each plot represents a different concentration of the competitive inhibitor. Note that all three plots intersect at the same point on the y-axis indicating that the reactions all have the same maximum velocity. However, since they do not intersect at the same point on the x-axis then they have different Michaelis-Menten constants, K_m .

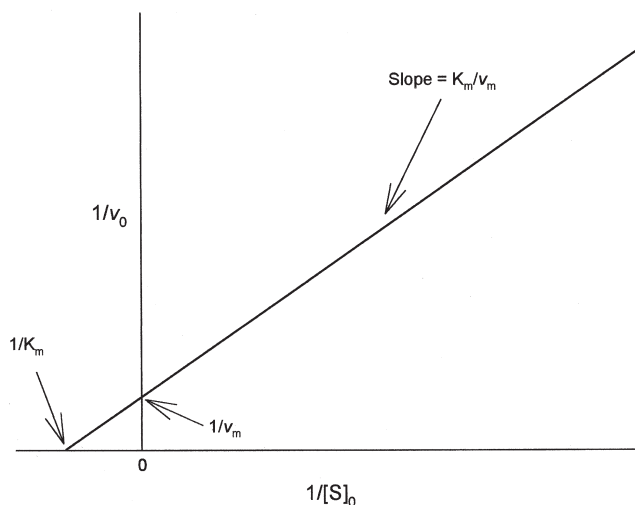


Figure 19-9. Lineweaver-Burk plot of modified Michaelis-Menten equation (Equation 72).

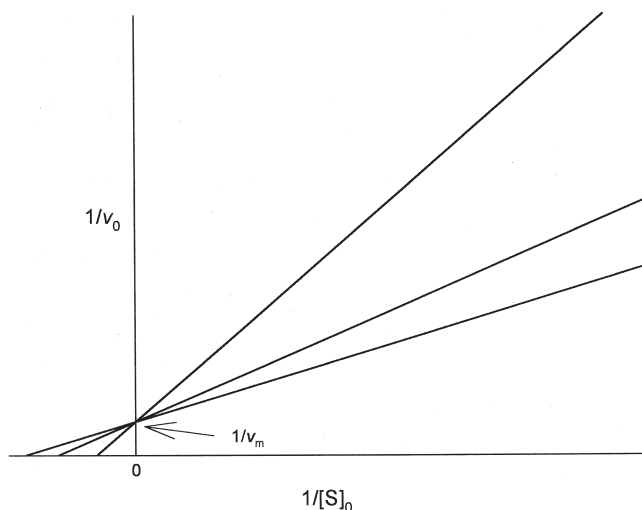
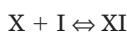


Figure 19-10. Lineweaver-Burk plot for competitive enzyme inhibition (Equation 73).

The other type of enzyme inhibition is noncompetitive inhibition. In this case the inhibitor does not bind to the active site of the enzyme but rather binds to another part of the enzyme or to the enzyme-substrate complex, X.



and



In this case, the equilibrium constant, K_I , is given by both $[E][I]/[EI]$ and $[X][I]/[XI]$. The resulting Lineweaver-Burk equation is given by Equation 74.

$$\frac{1}{v_0} = \left[\frac{1}{v_m} + \frac{K_m}{v_m[S]_0} \right] \left[1 + \frac{[I]}{K_I} \right] \quad (74)$$

Figure 19-11 shows a typical Lineweaver-Burk graph for noncompetitive inhibition. Each plot represents a different concentration of the competitive inhibitor. Note that all three plots intersect at the same point on the x-axis indicating that the reactions all have the same Michaelis-Menten constant, K_m .

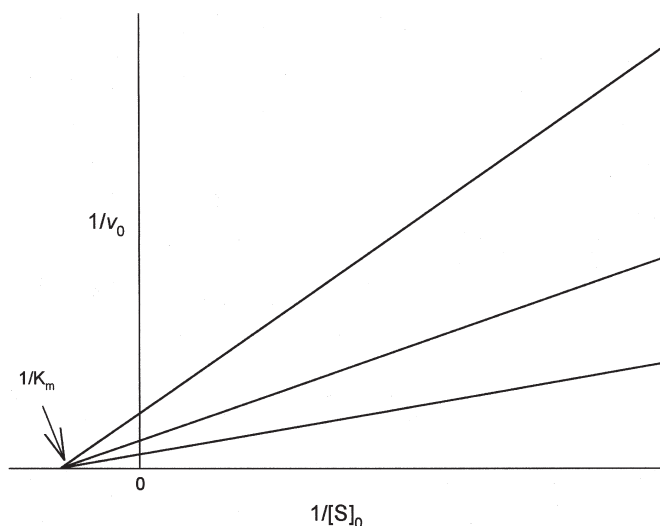


Figure 19-11. Lineweaver-Burk plot for noncompetitive enzyme inhibition (Equation 74).

However, since the plots do not intersect at the same point on the y-axis, then the reactions have different maximum velocities.

OTHER EFFECTS

Ionic Strength

In general, the effects of increasing concentrations of electrolytes on reaction rate can be predicted by consideration of the influence of ionic strength on interionic attraction. The Debye-Hückel equation may be used to demonstrate that increased ionic strength would be expected to decrease the rate of reaction between oppositely charged ions, and increase the rate of reaction between similarly charged ions. Thus, the hydrogen-ion catalyzed hydrolysis of sulfate esters is inhibited by increasing electrolyte concentration.



Reactions between ions and dipolar molecules, and reactions between neutral molecules generally are less sensitive to ionic strength effects than are reactions between ionic compounds. However, reactions that result in formation of oppositely charged ions as products may exhibit considerable increase in rate with increasing ionic strength.

Dielectric Constant of Solvent

Reactions involving ions of opposite charge are accelerated by solvents with low dielectric constants. For example, the rate of hydrogen ion-catalyzed hydrolysis of sulfate esters is much greater in low dielectric constant solvents, such as methylene chloride, than in water. Reaction between similarly charged species is favored by high dielectric constant solvents. Reaction between neutral molecules, which produce a highly polar transition state, such as the reaction of triethylamine with ethyl iodide to produce a quaternary ammonium salt, also will be enhanced by high dielectric constant solvents.

Hydrolysis (Solvolysis)

Hydrolysis of esters, such as procaine, aspirin, or atropine, represents one of the more common types of drug instability. *Ester hydrolysis* is either hydrogen- or hydroxide-ion catalyzed, although the catalysis that is important from the viewpoint of drug-product stability depends upon the specific compound and the pH of the solution. Amides generally are more stable than esters but are subject to catalysis by hydrogen and hydroxide ions, and often by general acids and bases. Some examples of the kinds of functional groups subject to hydrolytic cleavage

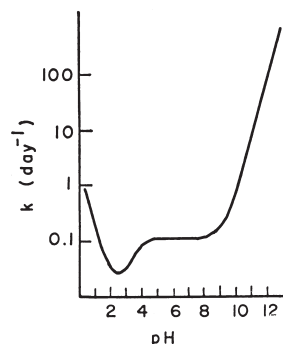
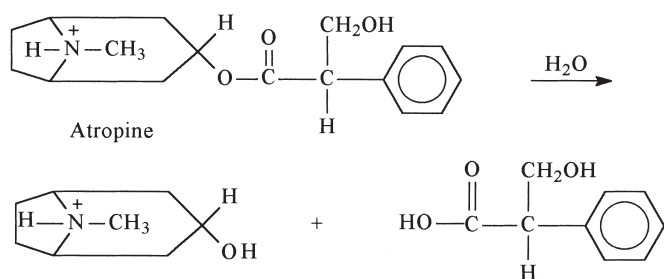


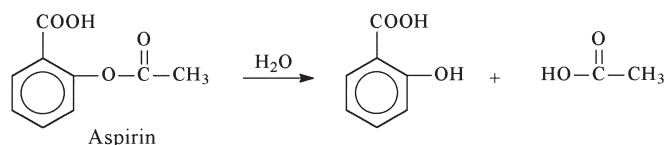
Figure 19-12. Apparent first-order rate of hydrolysis of aspirin as a function of pH at 17. (From Edwards LJ. *Trans Faraday Soc* 1950; 46: 723.)

and species shown to be catalysts for the reactions are presented below.



Hydrolysis of the ester function of atropine is typical of ester hydrolysis in that only catalysis by the hydrogen or hydroxide ions are important. Figure 19-6 illustrates a pH-rate profile which might be considered typical for such a reaction. Below pH 3, the principal reaction is hydrogen-ion catalyzed hydrolysis of the protonated form of atropine. Above pH 5, the principal reaction is hydroxide ion catalyzed hydrolysis of the same species. Maximum stability at 30° is at pH 3.7.

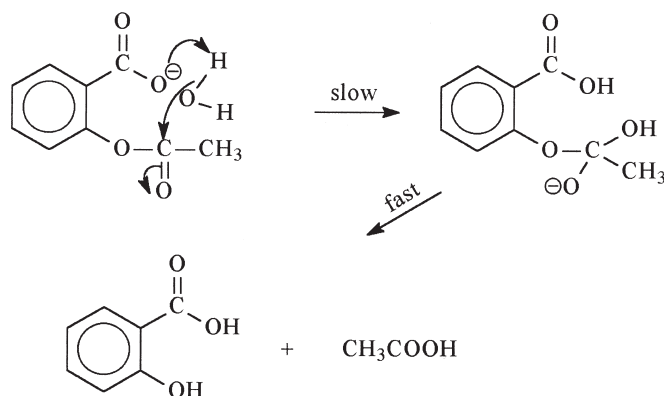
Hydrolytic cleavage of aspirin to salicylic acid and acetic acid was studied by Edwards.⁴



Edwards obtained the interesting pH-rate profile reproduced in Figure 19-12. The unusual pH-rate profile obtained for aspirin was attributed to a reaction of the form

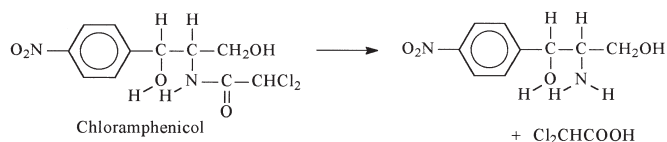
$$-d[\text{Aspirin}]_{\text{total}}/dt = k_1[\text{H}^+][\text{HA}] + k_2[\text{H}^+][\text{A}^-] + k_3[\text{OH}^-][\text{A}^-] + k_0[\text{A}^-] \quad (75)$$

where $[\text{HA}]$ represents undissociated aspirin and $[\text{A}^-]$ represents aspirin anion. The pH-independent anion hydrolysis indicated for the pH region 5 to 9 has been attributed to intramolecular catalysis by orthocarboxylate anion, rather than to general acid-base catalysis by water. It is principally this intramolecular catalysis that is responsible for the high instability of aqueous solutions of aspirin in the pharmaceutically useful pH range. Fersht and Kirby⁵ represented the intramolecular carboxylate ion reaction as a general base catalysis of attack by a water molecule. For nucleophiles such as ethanol, the terminal hydroxyl of polyethylene glycol (PEG) and the lysine ϵ -amino function in serum albumin also can participate in this reaction in the same manner as water. Thus, from aspirin in ethanol solution, ethyl acetate appears as a product; in polyethylene glycol, a polyethylene



glycol acetate is formed; and in a solution containing serum albumin (both *in vitro* and *in vivo*) aspirin produces an acetylated serum albumin. Whitworth et al.⁶ reasoned that an aspirin solution prepared in a PEG solvent containing no free hydroxyl groups would provide an aspirin solution of improved stability. They used acetylated PEG 400 as a solvent for aspirin and demonstrated that in such a solvent less than 1% aspirin loss occurred after 30 days at 45°.

Chloramphenicol decomposition below pH 7 proceeds primarily through hydrolytic cleavage of the amide function.



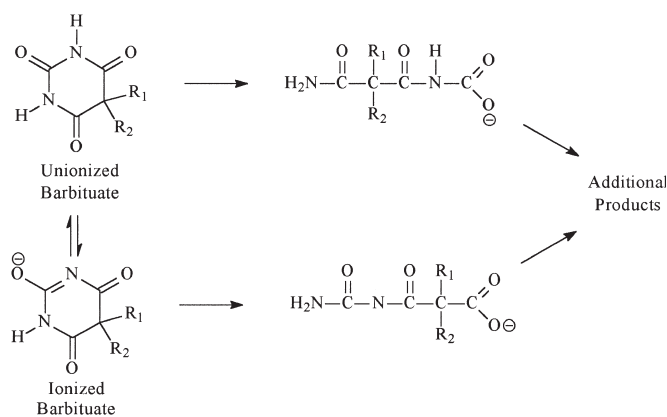
In the presence of a buffer, the reaction may be represented as

$$-d[\text{Camp}]/dt = (k_0 + k_1[\text{H}^+] + k_2[\text{OH}^-] + k_{\text{HB}}[\text{HB}] + k_{\text{B}}[\text{B}])(\text{Camp}) \quad (76)$$

In addition to hydrogen and hydroxide-ion catalysis there is an uncatalyzed (or water) reaction, and there may be general acid-base catalysis, represented above by the buffer species HB and B. In general, the rate of hydroxide-ion-catalyzed hydrolysis of amides is greater than rate of hydronium-ion-catalyzed hydrolysis.

Amides generally are much more stable than esters. Penicillins and cephalosporins are important exceptions to this rule because the amide bond is part of a strained four-membered ring (ie, a β -lactam). The decomposition of these compounds in aqueous solution is catalyzed by hydrogen ion, solvent, hydroxide ion, sugars, and many buffer species. Maximum stability occurs at about pH 7, but β -lactam antibiotics are too unstable to be formulated as solutions. For example, a buffered aqueous solution of penicillin G under refrigeration has a useful life of only about 1 week. Formation of the penicillanic acid by water-catalyzed rearrangement in acidic and neutral solutions is thought to be the first step in the degradation process.⁷

Barbiturate hydrolysis involves hydroxide-ion attack on both the undissociated acid, HP, and the ionized species, P^- .



Hydrogen ion catalyzed hydrolysis is not observed in the pH range of interest in pharmaceutical products.

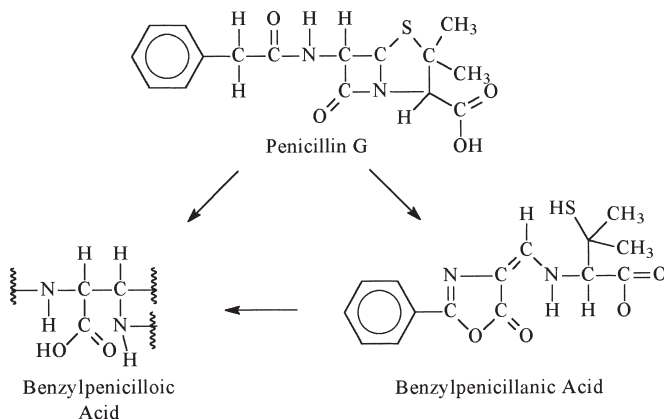
$$-d[\text{Barb}]/dt = k_1[\text{OH}^-][\text{HP}] + k_2[\text{OH}^-][\text{P}^-] \quad (77)$$

Hydrolysis of the amide (peptide) bond also occurs in protein and peptide drugs. This can occur by cleavage of the primary peptide linkage (R-NH-CO-R) between adjacent amino acids in the peptide chain. Hydrolysis of the free side-chain amide groups of asparagine and glutamine (deamidation) is another degradation

pathway for proteins. Insulin and recombinant human growth hormone undergo deamidation in solution.

Racemization

Many drugs are chiral and racemization is a common mechanism of degradation resulting in loss of biological activity. In proteins, a mixture of the D and L enantiomers is formed by base-catalyzed reaction of the natural L configuration. Acid-catalyzed racemization of epinephrine or base-catalyzed racemization of pilocarpine result in a loss of pharmacological activity.



Oxidation

Compounds such as phenols, aromatic amines, aldehydes, ethers, and unsaturated aliphatic compounds are subject to oxidation upon exposure to air or oxidizing chemicals. Epinephrine, ascorbic acid, phenothiazines, and vitamin A are examples of important pharmaceutical products that are oxidized readily. Proteins can undergo oxidative degradation by oxidation of methionine, a thioether, to its corresponding sulfoxide. Oxidation of the carbon-carbon double bonds in unsaturated fatty acids (eg, oleic acid) results in the fats and oils tasting rancid.

Of particular concern are oxidations that occur when solutions are exposed to atmospheric oxygen. Such reactions, termed *autoxidation* or *self-oxidation*, are complex reactions that proceed via a free-radical mechanism. A free radical is a highly unstable (highly reactive) species containing an unpaired electron. Autoxidation reactions are autocatalytic in that free-radical reactions generate additional free radicals, causing a chain reaction.

A technique used to protect pharmaceuticals susceptible to autoxidation is to include in the formulation agents that will react readily with free radicals, but that will terminate the chain propagation either by forming relatively stable, resonance-stabilized free radicals or by forming products that do not include additional free radicals.

Photochemical Decomposition

Numerous dyes and drugs are subject to photochemical decomposition. Light-catalyzed oxidations and reductions of photoexcited species are common and are often mechanistically complex reactions involving free-radical intermediates. Pharmaceuticals such as riboflavin, nifedipine, and the phenothiazines are examples of common drugs that are extremely light sensitive.

Interaction Between Components

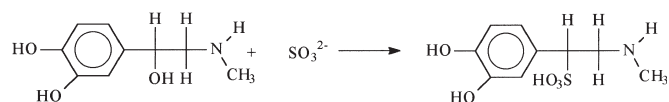
Because drugs are often combined in solution with buffers, antioxidants, flavoring agents, antimicrobial preservatives, and other drugs, potential interaction between the components of a formulation must be considered in pharmaceutical formulation

development. Some obvious interactions, such as the possibility of the reaction of a drug having a primary amino function with an aldehyde such as vanillin to produce a Schiff base, can be predicted; however, a number of interesting, less well-recognized reactions have been encountered.

In addition to buffer species acting as general acid-base catalysts, as previously indicated, some buffer species undergo specific interactions with drug molecules to form new chemical compounds. The formation of amides in aqueous solution from amines such as procaine and buffers such as citric acid has been observed.

The aromatic function of procaine reacts with glucose to form procaine *N*-glycoside; also, phenylethylamine reacts with dehydroacetic acid to form a Schiff base-type compound. Catechols have been shown to catalyze penicillin hydrolysis.

It has been demonstrated that bisulfite, an agent commonly employed to protect epinephrine against oxidative decomposition, is capable of inducing epinephrine degradation through attack on the chiral side chain.



Although a solution of folic acid alone is stable to light, a combination of riboflavin and folic acid showed a rapid loss of folic acid through formation of a coupled oxidation-reduction system in which riboflavin was photoreduced, with folic acid being used as a reducing substrate and being itself irreversibly oxidized. In the dark and in the presence of oxygen, the riboflavin was regenerated, and when the solution was again irradiated, the cycle was repeated with further destruction of folic acid. In this case, the riboflavin acts as a photosensitizer in this reaction and would cause the decomposition not only of folic acid, but also of ascorbic acid or any other easily oxidized substrate.

The presence of micellar surfactants and certain high-molecular-weight polymers commonly employed in pharmaceuticals also have been shown to lead to decreased drug stability in some cases. Both nonionic and anionic surfactants, as well as polymers such as polyvinylpyrrolidone, accelerate the photodecomposition of riboflavin in aqueous solution. Nonionic surfactants also are capable of increasing the rate of hydrolysis of sulfate esters which may be incorporated in or on the micellar surface.

Physical Instability

The introduction of an increasing number of drugs derived from developments in biotechnology necessitates greater awareness of instability occurring as a result of loss of drug activity through structural changes unrelated to disruption of covalent bonds. Protein-based drugs may lose activity as a result of a change in superstructure (secondary, tertiary, quaternary) that is independent of chemical modification. Superstructure changes, which may alter protein drug activity, include denaturation (unfolding), aggregation, surface adsorption, and precipitation. Treatments with potential for inducing such changes include temperature changes, pH extremes, and agitation or foaming resulting from shaking. High shear encountered in manufacturing or in drug delivery systems also may denature protein drugs. Detection of instability of a physical nature generally requires one or more biological assays, or physical assay methods that are sensitive to the critical superstructure change.

DRUG STABILIZATION

Some drug decomposition reactions, such as photolytic and oxidative reactions, are relatively easy to avoid by protecting the components from light (photodecomposition) or exclusion of oxygen and by use of chain-terminating reagents or free-radical scavengers to minimize free-radical-mediated reactions. Solvol-

ysis reactions, however, cannot be stopped by such procedures, but several techniques may be employed to retard reactions sufficiently to permit the formulation of a suitable drug product. The following approaches may be useful in attempts to retard solvolytic reactions.

Selection of Optimum pH, Buffer, and Solvent

Consideration of the mechanism of the reaction and the way in which the reaction rate is influenced by pH, buffer species, and solvent permits the selection of the optimum conditions for drug stability. Often, however, ideal conditions for maximum stability may be unacceptable from the viewpoint of pharmaceutically acceptable formulation or therapeutic efficacy; thus, it may be necessary to prepare a formulation with conditions less than optimum for drug stability. If a suitable compromise between conditions for maximum stability and conditions for a pharmaceutically acceptable formulation cannot be achieved, techniques such as those described below may be useful in retarding solvolysis reactions.

Specific Complexing Agents

The technique of stabilization by forming complexes in solution was introduced by Higuchi and Lachman,⁸ who demonstrated that the rate of hydrolysis of the ester function of benzocaine was retarded significantly in the presence of caffeine, a reagent with which the benzocaine formed a soluble complex. It was demonstrated further that, in these systems, the complexed drug did not hydrolyze at all, and that the observed rate of hydrolysis could be ascribed to the concentration of the free or uncomplexed drug that was in equilibrium with the drug complex.

Boric acid chelation of the catechol function of epinephrine stabilizes epinephrine against attack by bisulfite and sulfite. The complex of povidone (polyvinylpyrrolidone) and iodine has been used for many years as a topical antiseptic because of its higher iodine concentration, slow release of iodine from the complex, and lower toxicity.

Surfactants

It has been demonstrated that the incorporation of benzocaine into surfactant micelles could retard significantly the rate of ester hydrolysis. Nonionic and anionic surfactants retarded the hydroxide-ion-catalyzed hydrolysis, but cationic surfactants somewhat increased the rate of hydroxide-ion-catalyzed hydrolysis. Similar observations have been reported for a number of drugs that are sufficiently lipophilic to be solubilized by surfactant micelles.

Suspensions

If the solubility of a labile drug is reduced and the drug is prepared in a suspension form, the rate at which the drug degrades will be related only to the concentration of dissolved drug rather than to the total concentration of drug in the product. Thus it has been demonstrated that penicillin G procaine suspensions degraded at a rate proportional to the low concentration of penicillin in solution. Because the penicillin in solution was in equilibrium with excess solid penicillin G procaine, the penicillin concentration in solution was constant and the observed order of reaction was apparently zero order.

Refrigeration

Storage below room temperature usually will retard solvolytic reactions. Storage in the frozen state generally is an effective means of retarding degradative reactions. Several antibiotics

are sold as frozen solutions in flexible plastic bags. An exception is sodium ampicillin dissolved in 5% dextrose solution, which showed approximately 10% decomposition after 4 hr of storage at 5° and more than 13% loss after storage for the same period in the frozen state at -20°.

Stability Testing of Pharmaceutical Products

If a product is to be marketed, it must be stable over relatively long storage times at room temperature or at the actual temperature at which it will be shipped and stored prior to its ultimate use. Thus, the rate of degradation may have to be studied over an undesirably long period of time in order to determine the product's stability under normal storage conditions.

To avoid this undesirable delay in evaluating possible formulations, the manufacturer attempts to predict stability under conditions of room temperature or actual storage conditions by using data for the rate of decomposition obtained at several elevated temperatures. This is accomplished using an Arrhenius plot to predict, from high-temperature data, the rate of product breakdown to be expected at actual lower temperature storage conditions.

Prediction based on data obtained at elevated temperatures generally is satisfactory for solution dosage forms. Success is more uncertain when nonhomogeneous products are involved. Suspensions of drugs may not provide linear Arrhenius plots because often there is the possibility that the solid phase, which exists at elevated temperature, may not be the same solid phase that exists at room temperature. Such differences in the solubility of the several solid phases may invalidate the usual Arrhenius plots. These difficulties should be anticipated when polymorphic crystal forms or several different solvates are known to exist for a specific solute. Also, when solid dosage forms (eg, tablets) are subjected to high temperatures, changes in the quantity of moisture in the product may greatly influence the stability of the product.

Arrhenius plots also suffer limitations when applied to reactions that have relatively low activation energies and, therefore, are not accelerated greatly by an increase in temperature. Where usually it is desirable to determine drug stability by analyzing samples for the amount of intact drug remaining—in instances where there is very little drug decomposition and particularly when it is not convenient to accelerate the reaction by increasing temperature—it sometimes is advantageous to determine initial reaction rates from the determination of the amount of reaction product formed.

Using modern methods of analysis, such as high-performance liquid chromatography (HPLC), it is often possible to measure the rate of formation of a degradation product. By using this technique, very small amounts of degradation (less than 1% loss of parent compound) can be detected, resulting in a more sensitive indication of product stability than can be obtained by analyzing potency.

Since manufacturers are interested primarily in the time required to produce just a few-percent breakdown in their product, it is not uncommon to employ terminology such as $t_{0.90}$ or $t_{0.95}$, which is the time required for the drug to decompose to 90 or 95%, respectively, of original potency.

An Arrhenius-type plot, analogous to that illustrated in Figure 19-4, can be obtained by plotting the logarithm of the time required for the specified fractional decomposition versus the reciprocal of absolute temperature. The time required for the product to decrease in potency to 90% of original potency at room temperature then can be obtained directly from the plot.

REFERENCES

- Garrett ER. In: Bean HS, et al, eds. *Advances in Pharmaceutical Sciences*, vol 2. New York: Academic Press, 1967, Chap 2.
- Kondritzer AA, Zvirblis P. *J APhA Sci Ed* 1957; 46: 531.

3. Higuchi T, et al. *J APhA Sci Ed* 1950; 39: 405.
4. Edwards LJ. *Trans Faraday Soc* 1950; 46: 723.
5. Fersht AR, Kirby AJ. *J Am Chem Soc* 1967; 89: 4857.
6. Whitworth CA, et al. *J Pharm Sci* 1973; 62: 1184.
7. Yamana T, et al. *J Pharm Sci* 1977; 66: 861.
8. Higuchi T, Lachman L. *J APhA Sci Ed* 1955; 44: 52.

BIBLIOGRAPHY

Carstensen JT. *Drug Stability, Principles and Practices*. New York: Dekker, 1990.

- Connors KA, et al. *Chemical Stability of Pharmaceuticals*, 2nd ed. New York: Wiley, 1986.
- Espenson JH. *Chemical Kinetics and Reaction Mechanisms*, 2nd ed. New York: McGraw-Hill, 1995.
- Fung HL. In: Banker GS, Rhodes CT, eds. *Modern Pharmaceuticals*, 2nd ed. New York: Dekker, 1990, Chap 6.
- House JE. *Principles of Chemical Kinetics*. Dubuque, IA: WC Brown, 1997.
- Houston, PL. *Chemical Kinetics and Reaction Dynamics*, Boston: McGraw-Hill, Inc., 2001.
- Lachman L, DeLuca P, Akers M. In: Lachman L, et al, eds. *The Theory and Practice of Industrial Pharmacy*, 3rd ed. Philadelphia: Lea & Febiger, 1986, Chap 6.

Interfacial Phenomena

Paul M Bummer, PhD



Very often it is desirable or necessary in the development of pharmaceutical dosage forms to produce multiphase dispersions by mixing together two or more ingredients that are not mutually miscible and capable of forming homogeneous solutions. Examples of such dispersions include:

- Suspensions (solid in liquid)
- Emulsions (liquid in liquid)
- Foams (vapor in liquids)

Because these systems are not homogeneous and thermodynamically stable, over time they will show some tendency to separate on standing to produce the minimum possible surface area of contact between phases. Thus, suspended particles agglomerate and sediment, emulsified droplets cream and coalesce, and the bubbles dispersed in foams collapse to produce unstable and nonuniform dosage forms. One way to prevent or slow down this natural tendency for further phase separation is to add materials that can accumulate at the interface to provide some type of energy barrier to aggregation and coalescence. Such materials are said to exhibit *surface activity* or to act as *surface-active agents*.

In this chapter the fundamental physical chemical properties of molecules situated at interfaces will be discussed so that the reader can gain a better understanding of how problems involving interfaces can be resolved in designing pharmaceutical dosage forms by the use of surface-active agents.

INTERFACIAL FORCES AND ENERGETICS

In the bulk portion of each phase, molecules are attracted to each other equally in all directions, such that no resultant forces are acting on any one molecule. The strength of these forces determines whether a substance exists as a vapor, liquid, or solid at a particular temperature and pressure.

At the boundary between phases, however, molecules are acted upon unequally because they are in contact with other molecules exhibiting different forces of attraction. For example, the primary intermolecular forces in water are due to hydrogen bonds, whereas those responsible for intermolecular bonding in hydrocarbon liquids, such as mineral oil, are due to London dispersion forces.

Thus, molecules situated at the interface experience interaction forces dissimilar to those experienced in each bulk phase. In liquid systems such unbalanced forces can be satisfied by spontaneous movement of molecules from the interface into the bulk phase. This leaves fewer molecules per unit area at the interface (greater intermolecular distance) and reduces the actual contact area between dissimilar molecules.

Any attempt to reverse this process by increasing the area of contact between phases—that is, bringing more molecules into the interface—causes the interface to resist expansion and behave as though it is under a tension everywhere in a tangential direction. The force of this tension per unit length of interface generally is called the *interfacial tension*, except when dealing with the air–liquid interface, where the terms *surface* and *surface tension* are used.

To illustrate the presence of a tension in the interface, consider an experiment where a circular metal frame, with a looped piece of thread loosely tied to it, is dipped into a liquid. When the frame is removed and exposed to the air, a film of liquid will be stretched entirely across the circular frame, as when one uses such a frame to blow soap bubbles. Under these conditions (Fig 20-1A), the thread will remain collapsed. If a heated needle is used to puncture and remove the liquid film from within the loop (Fig 20-1B), the loop will stretch spontaneously into a circular shape.

The result of this experiment demonstrates the spontaneous reduction of interfacial contact between air and the liquid remaining; indeed, it illustrates that a tension causing the loop to remain extended exists parallel to the interface. The circular shape of the loop indicates that the tension in the plane of the interface exists at right angles or normal to every part of the looped thread. The total force on the entire loop divided by the circumference of the circle, therefore, represents the tension per unit distance of surface, or the surface tension.

Just as work is required to extend a spring under tension, work should be required to reverse the process seen in Figure 20-1A and B, thus bringing more molecules to the interface. This may be seen quantitatively by considering an experiment where tension and work may be measured directly. Assume that we have a rectangular wire with one movable side (Fig 20-2). Assume further that by dipping this wire into a liquid, a film of liquid will form within the frame when it is removed and exposed to the air. As seen earlier in Figure 20-1, when it comes in contact with air, the liquid surface will tend to contract with a force, F , as molecules leave the surface for the bulk. To keep the movable side in equilibrium, an equal force must be applied to oppose this tension in the surface. The surface tension, γ , of the liquid may be defined as $F/2l$, where $2l$ is the distance of surface over which F is operating. The factor 2 arises out of considering two surfaces, top and bottom. Upon expansion of the surface by a very small distance, Δx , the work done (W) is

$$W = F\Delta x \quad (1)$$

and therefore,

$$W = \gamma 2l\Delta x \quad (2)$$

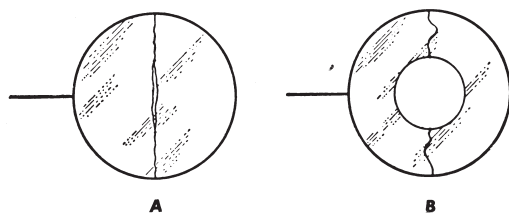


Figure 20-1. A circular wire frame with a loop of thread loosely tied to it: (A) a liquid film on the wire frame with a loop in it; (B) the film inside the loop is broken. (From Semat H. *Fundamentals of Physics*, 3rd ed. New York: Holt Reinhart Winston, 1957.)

Since

$$\Delta A = 2l\Delta x \quad (3)$$

where ΔA is the change in area due to the expansion of the surface, it may be concluded that

$$W = \gamma\Delta A \quad (4)$$

Thus, the work required to create a unit area of surface, known as the *surface free energy/unit area*, is equivalent to the surface tension of a liquid system—the greater the area of interfacial contact between phases, the greater the free-energy increase for the total system. Because a prime requisite for equilibrium is that the free energy of a system be at a minimum, it is not surprising to observe that phases in contact tend to reduce area of contact spontaneously.

Liquids, being mobile, may assume spherical shapes (smallest interfacial area for a given volume), as when ejected from an orifice into air or when dispersed into another immiscible liquid. If a large number of drops are formed, further reduction in area can occur by having the drops coalesce, as when a foam collapses or when the liquid phases making up an emulsion separate.

In the centigrade-gram-second (cgs) system, surface tension is expressed in units of dynes per centimeter (dyne/cm), while surface free energy is expressed in erg/cm². As an erg is a dyne-cm, both sets of units are equivalent. In the SI (international units) system, surface tension is expressed in mN/m and surface free energy in mJ/m².

Values for the surface tension of a variety of liquids are given in Table 20-1, and interfacial tension values for various liquids against water are given in Table 20-2. Other combinations of immiscible phases could be given, but most heterogeneous systems encountered in pharmacy usually contain water. Values for these tensions are expressed for a particular temperature. Because an increased temperature increases the thermal energy of molecules, the work required to bring molecules to the interface should be less, and thus the surface and interfacial tension will be reduced. For example, the surface tension of water is 76.5 dynes/cm at 0° and 63.5 dynes/cm at 75°.

As would be expected from the discussion so far, the relative values for surface tension should reflect the nature of intermolecular forces present, hence the relatively large values for

Table 20-1. Surface Tension of Various Liquids at 20°

SUBSTANCE	SURFACE TENSION (dyne/cm)
Mercury	476
Water	72.8
Glycerin	63.4
Oleic acid	32.5
Benzene	28.9
Chloroform	27.1
Carbon tetrachloride	26.8
1-Octanol	26.5
Hexadecane	27.4
Dodecane	25.4
Decane	23.9
Octane	21.8
Heptane	19.7
Hexane	18.0
Perfluoroheptane	11.0
Nitrogen (at 75 K)	9.4

mercury (metallic bonds) and water (hydrogen bonds), and the lower values for benzene, chloroform, carbon tetrachloride, and the *n*-alkanes.

Benzene, with π electrons, exhibits a higher surface tension than the alkanes of comparable molecular weight, but increasing the molecular weight of the alkanes (and hence intermolecular attraction) increases their surface tension closer to that of benzene. The lower values for the more nonpolar substances, perfluoroheptane and liquid nitrogen, demonstrate this point even more strongly.

Values of interfacial tension should reflect the differences in chemical structure of the two phases involved—the greater the tendency to interact, the less the interfacial tension. The 20-dyne/cm difference between air–water tension and that at the octane–water interface reflects the small but significant interaction between octane molecules and water molecules at the interface. This is seen also in Table 20-2 by comparing the values for octane and octanol, oleic acid and the alkanes, or chloroform and carbon tetrachloride. In each case the presence of chemical groups capable of hydrogen bonding with water markedly reduces the interfacial tension, presumably by satisfying the unbalanced forces at the interface. These observations strongly suggest that molecules at an interface arrange themselves or orient so as to minimize differences between bulk phases.

That this phenomenon occurs even at the air–liquid interface is seen when one notes the relatively low surface-tension values of very different chemical structures such as the *n*-alkanes, octanol, oleic acid, benzene, and chloroform. Presumably, in each case the similar nonpolar groups are oriented toward the air with any polar groups oriented away toward the bulk phase. This tendency for molecules to orient at an interface is a basic factor in interfacial phenomena and will be discussed more fully in succeeding sections.

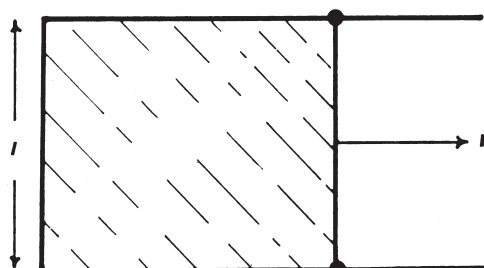


Figure 20-2. A movable wire frame containing a film of liquid being expanded with a force, F .

Table 20-2. Interfacial Tension of Various Liquids Against Water at 20°

SUBSTANCE	INTERFACIAL TENSION (dyne/cm)
Decane	52.3
Octane	51.7
Hexane	50.8
Carbon tetrachloride	45.0
Chloroform	32.8
Benzene	35.0
Mercury	428
Oleic acid	15.6
1-Octanol	8.51

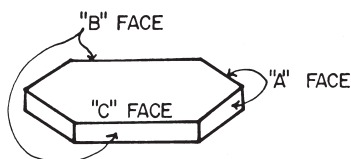


Figure 20-3. Adipic acid crystal showing various faces. (From Michaels AS. *J Phys Chem* 1961; 65: 1730.)

Solid substances such as metals, metal oxides, silicates, and salts, all containing polar groups exposed at their surface, may be classified as *high-energy solids*, whereas nonpolar solids such as carbon, sulfur, glyceryl tristearate, polyethylene, and polytetrafluoroethylene (Teflon) may be classified as *low-energy solids*. It is of interest to measure the surface free energy of solids; however, the lack of mobility of molecules at the surface of solids prevents the observation and direct measurement of a surface tension. It is possible to measure the work required to create new solid surface by cleaving a crystal and measuring the work involved. However, this work not only represents free energy due to exposed groups but also takes into account the mechanical energy associated with crystal fracture (ie, plastic and elastic deformation and strain energies due to crystal structure and imperfections in that structure).

Also contributing to the complexity of a solid surface is the heterogeneous behavior as a result of the exposure of different crystal faces, each having a different surface free energy/unit area. For example, adipic acid, $\text{HOOC}(\text{CH}_2)_4\text{COOH}$, crystallizes from water as thin hexagonal plates with three different faces, as shown in Figure 20-3. Each unit cell of such a crystal contains adipic acid molecules oriented such that the hexagonal planes (faces) contain exposed carboxyl groups, while the sides and edges (A and B faces) represent the side view of the carboxyl and alkyl groups and thus are quite nonpolar. Indeed, interactions involving these different faces reflect the differing surface free energies.²

Other complexities of solid surfaces include roughness and porosity.³ Even in the absence of chemical contamination, such as that occurring during recrystallization, surface energy changes in a solid can be induced by unit operations such as milling, resulting in an altered pattern of drug dissolution.^{4,5} In view of all these potential complications that are difficult to quantify, surface free energy values for solids, when reported, should be regarded as average values, often dependent on the method used and not necessarily the same for other samples of the same substance.

Table 20-3 lists some average values of γ_{sv} for a variety of solids, ranging in polarity from Teflon to copper, obtained by various indirect techniques.

ADHESIONAL AND COHESIONAL FORCES

Of prime importance to those dealing with heterogeneous systems is the question of how two phases will behave when

brought in contact with each other. It is well known, for instance, that some liquids, when placed in contact with other liquid or solid surfaces, will remain retracted in the form of a drop (known as a *lens*), while other liquids may exhibit a tendency to spread and cover the surface of this liquid or solid.

Based upon concepts developed to this point, it is apparent that the individual phases will exhibit a tendency to minimize the area of contact with other phases, thus leading to phase separation. On the other hand, the tendency for interaction between molecules at the new interface will offset this to some extent and give rise to the spontaneous spreading of one substance over the other.

In essence, therefore, phase affinity is increased as the forces of attraction between different phases (*adhesional forces*) become greater than the forces of attraction between molecules of the same phase (*cohesional forces*). If these adhesional forces become great enough, miscibility will occur and the interface will disappear. The present discussion is concerned only with systems of limited phase affinity, where an interface still exists.

A convenient approach used to express these forces quantitatively is work of adhesion and work of cohesion. The *work of adhesion*, W_a , is defined as the free energy/cm² required to separate two phases at their boundary and is equal but opposite in sign to the free energy/cm² released when the interface is formed. In an analogous manner the *work of cohesion* for a pure substance, W_c , is the work/cm² required to produce two new surfaces, as when separating different phases, but now both surfaces contain the same molecules. This is equal and opposite in sign to the free energy/cm² released when the same two pure liquid surfaces are brought together and eliminated.

By convention, when the work of adhesion between two substances, A and B, exceeds the work of cohesion for one substance (eg, B), spontaneous spreading of B over the surface of A should occur with a net loss of free energy equal to the difference between W_a and W_c . If W_c exceeds W_a , no spontaneous spreading of B over A can occur. The difference between W_a and W_c is known as the *spreading coefficient*, S . Only when S is positive will spreading occur.

The values for W_a and W_c (and hence S) may be expressed in terms of surface and interfacial tensions, when one considers that upon separation of two phases, A and B, γ_{AB} ergs of interfacial free energy/cm² (interfacial tension) are lost, but that γ_A and γ_B erg/cm² of energy (surface tensions of A and B) are gained; upon separation of bulk-phase molecules in an analogous manner, $2\gamma_A$ or $2\gamma_B$ erg/cm² will be gained. Thus,

$$W_a = \gamma_A + \gamma_B - \gamma_{AB} \quad (5)$$

and

$$W_c = 2\gamma_A \text{ or } 2\gamma_B \quad (6)$$

for B spreading on the surface of A. Therefore,

$$S_B = \gamma_A + \gamma_B - \gamma_{AB} - 2\gamma_B \quad (7)$$

or

$$S_B = \gamma_A - (\gamma_B + \gamma_{AB}) \quad (8)$$

Using Equation 8 and the values of surface and interfacial tension given in Tables 20-1 and 20-2, the spreading coefficient can be calculated for three representative substances—decane, benzene, and oleic acid—on water at 20°.

$$\text{Decane: } S = 72.8 - (23.9 + 52.3) = -3.4$$

$$\text{Benzene: } S = 72.8 - (28.9 + 35.0) = 8.9$$

$$\text{Oleic Acid: } S = 72.8 - (32.5 + 15.6) = 24.7$$

As expected, relatively nonpolar substances such as decane exhibit negative values of spreading coefficient, whereas the more-polar materials yield positive values—the greater the polarity of the molecule, the more positive the value of S .

Table 20-3. Values of γ_{sv} for Solids of Varying Polarity

SOLID	γ_{sv} (dyne/cm)
Teflon	19.0
Paraffin	25.5
Polyethylene	37.6
Polymethyl methacrylate	45.4
Nylon	50.8
Indomethacin	61.8
Griseofulvin	62.2
Hydrocortisone	68.7
Sodium Chloride	155
Copper	1300

The importance of the cohesive energy of the spreading liquid may be noted also by comparing the spreading coefficients for hexane on water and water on hexane.

$$S_{H/W} = 72.8 - (18.0 + 50.8) = 10.0$$

$$S_{W/H} = 18.0 - (72.8 + 50.8) = -105.6$$

Here, despite the fact that both liquids are the same, the high cohesion and air–liquid tension of water prevents spreading on the low-energy hexane surface, while the very low value for hexane allows spreading on the water surface. This also is seen when comparing the positive spreading coefficient of hexane to the negative value for decane on water.

To see whether spreading does or does not occur, a powder such as talc or charcoal can be sprinkled over the surface of water such that it floats; then, a drop of each liquid is placed on this surface. As predicted, decane will remain as an intact drop, while hexane, benzene, and oleic acid will spread out, as shown by the rapid movement of solid particles away from the point where the liquid drop was placed originally.

An apparent contradiction to these observations may be noted for hexane, benzene, and oleic acid when more of each substance is added: lenses now appear to form even though initial spreading occurred. Thus, in effect a substance does not appear to spread over itself.

It is now established that the spreading substance forms a monomolecular film that creates a new surface that has a lower surface free energy than pure water. This arises because of the apparent orientation of the molecules in such a film so that their most hydrophobic portion is oriented toward the spreading phase. It is the lack of affinity between this exposed portion of the spread molecules and the polar portion of the remaining molecules that prevents further spreading. This may be seen by calculating a final spreading coefficient where the new surface tension of water plus monomolecular film is used. For example, the presence of benzene reduces the surface tension of water to 62.2 dyne/cm so that the final spreading coefficient, is

$$S = 62.2 - (28.9 + 35.0) = -1.7$$

The lack of spreading exhibited by oleic acid should be reflected in an even more negative final spreading coefficient, as the very polar carboxyl groups should have very little affinity for the exposed alkyl chain of the oleic acid film. Spreading so as to form a second layer with polar groups exposed to the air also would seem very unlikely, thus leading to the formation of a lens.

WETTING PHENOMENA

In the experiment described above it was shown that talc or charcoal sprinkled onto the surface of water float despite the fact that their densities are much greater than that of water. In order for immersion of the solid to occur, the liquid must displace air and spread over the surface of the solid; when liquids cannot spread over a solid surface spontaneously, and, therefore, S , the spreading coefficient, is negative, we say that the solid is not wetted.

An important parameter reflecting the degree of wetting is the angle made by the liquid with the solid surface at the point of contact (Fig 20-4). By convention, when wetting is complete, the contact angle is 0° ; in nonwetting situations it theoretically can increase to a value of 180° , where a spherical droplet makes contact with solid at only one point.

To express contact angle in terms of solid–liquid–air equilibria, one can balance forces parallel to the solid surface at the point of contact between all three phases (see Fig 20-4), as expressed in

$$\gamma_{SV} = \gamma_{SL} + \gamma_{LV} \cos \theta \quad (9)$$

where γ_{SV} , γ_{SL} , and γ_{LV} represent the surface free energy/unit area of the solid–air, solid–liquid and liquid–air interfaces, re-

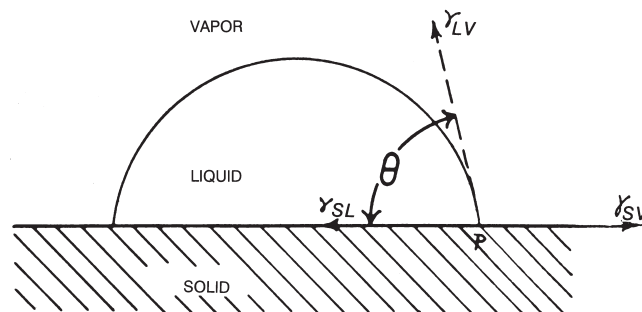


Figure 20-4. Forces acting on a nonwetting liquid drop exhibiting a contact angle of θ . (From Zisman WA. *Adv Chem Ser* 1964; 43: 1.)

spectively. Although difficult to use quantitatively because of uncertainties with γ_{SV} and γ_{SL} measurements, conceptually the equation, known as the Young equation, is useful because it shows that the loss of free energy due to elimination of the air–solid interface by wetting is offset by the increased solid–liquid and liquid–air area of contact as the drop spreads out.

The $\gamma_{LV} \cos \theta$ term arises as the horizontal vectorial component of the force acting along the surface of the drop, as represented by γ_{LV} . Factors tending to reduce γ_{LV} and γ_{SL} , therefore, will favor wetting, while the greater the value of γ_{SV} , the greater the chance for wetting to occur. This is seen in Table 20-4 for the wetting of a low-energy surface, paraffin (hydrocarbon), and a higher energy surface, nylon (polyhexamethylene adipamide). Here, the lower the surface tension of a liquid, the smaller the contact angle on a given solid, and the more polar the solid, the smaller the contact angle with the same liquid.

With Equation 9 in mind and looking at Figure 20-5, it is now possible to understand how the forces acting at the solid–liquid–air interface can cause a dense nonwetted solid to float if γ_{SL} and γ_{LV} are large enough relative to γ_{SV} .

The significance of reducing γ_{LV} was first developed empirically by Zisman⁶ when he plotted $\cos \theta$ versus the surface tension of a series of liquids and found that a linear relationship, dependent on the solid, often was obtained. When such plots are extrapolated to $\cos \theta$ equal to 1, or 0° contact angle, a value of surface tension required to just cause complete wetting is obtained. Doing this for a number of solids, it was shown that this surface tension (known as the critical surface tension, γ_c) parallels expected solid surface energy γ_{SV} —the lower γ_c , the more nonpolar the surface.

Table 20-5 indicates some of these γ_c values for different surface groups, indicating such a trend. Thus, water with a surface tension of about 72 dyne/cm will not wet polyethylene ($\gamma_c = 31$ dyne/cm) but heptane, with a surface tension of about 20 dyne/cm, will. Likewise, Teflon (polytetrafluoroethylene) ($\gamma_c = 19$) is not wetted by heptane but is wetted by perfluoroheptane with a surface tension of 11 dyne/cm.

Table 20-4. Contact Angle on Paraffin and Nylon for Various Liquids of Differing Surface Tension

SUBSTANCE	SURFACE TENSION (dyne/cm)	CONTACT ANGLE ($^\circ$)	
		PARAFFIN	NYLON
Water	72.8	105	70
Glycerin	63.4	96	60
Formamide	58.2	91	50
Methylene iodide	50.8	66	41
α -Bromonaphthalene	44.6	47	16
tert-Butylnaphthalene	33.7	38	spreads
Benzene	28.9	24	spreads
Dodecane	25.4	17	spreads
Decane	23.9	7	spreads
Nonane	22.9	spreads	spreads

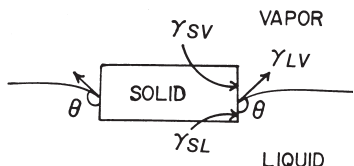


Figure 20-5. Forces acting on a nonwetting solid at the air+liquid+solid interface: contact angle θ greater than 90° .

One complication associated with the wetting of high-energy surfaces is the lack of wetting after the initial formation of a monomolecular film caused by the spreading substance. As in the case of oleic acid spreading on the surface of water, the remaining liquid retracts because of the low-energy surface produced by the oriented film. This phenomenon, often called *autophobic behavior*, is an important factor in many systems of pharmaceutical interest because many solids expected to be wetted easily by water may be rendered hydrophobic if other molecules dissolved in the water can form these monomolecular films at the solid surface.

CAPILLARITY

Because water shows a strong tendency to spread out over a polar surface such as clean glass (contact angle equal to 0°), one would expect to observe a meniscus forming when water is contained in a glass vessel such as a pipet or buret. This behavior is accentuated dramatically if a fine-bore capillary tube is placed into the liquid (Fig 20-6). Not only will the wetting of the glass produce a more highly curved meniscus, but the level of the liquid in the tube will be appreciably higher than the level of the water in the beaker.

The spontaneous movement of a liquid into a capillary or narrow tube due to surface forces is defined as *capillarity* and is responsible for a number of important processes involving the penetration of liquids into porous solids. In contrast to water in contact with glass, if the same capillary is placed into mercury (contact angle on glass: 130°), not only will the meniscus be inverted (Fig 20-7), but the level of the mercury in the capillary will be lower than in the beaker. In this case one does not expect mercury or other *nonwetting* liquids to penetrate pores easily unless external forces are applied.

To examine more closely the factors giving rise to the phenomenon of capillarity, consider the case of a liquid that rises to a height, h , above the bulk liquid in a capillary having a radius, r . As shown in Figure 20-6, if the contact angle of water on glass is 0° , a force, F , will act upward and vertically

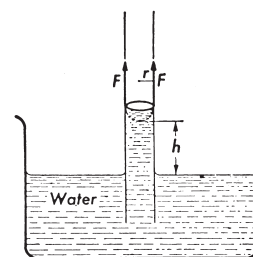


Figure 20-6. Capillary rise for a liquid exhibiting 0° contact angle. (From Semat H. *Fundamentals of Physics*, 3rd ed. New York: Holt Reinhart Winston, 1957.)

along the circle of liquid–glass contact. Based upon the definition of surface tension, this force will be equal to the surface tension, γ , multiplied by the circumference of the circle, $2\pi r$. Thus,

$$F = \gamma 2\pi r \tag{10}$$

This force upward must support the column of water, and because the mass, m , of the column is equal to the density, d , multiplied by the volume of the column, $\pi r^2 h$, the force W opposing the movement upward will be

$$W = mg = \pi r^2 dgh \tag{11}$$

where g is the gravity constant.

Equating the two forces at equilibrium gives

$$\pi r^2 dgh = \gamma 2\pi r \tag{12}$$

so that

$$h = \frac{2\gamma}{rdg} \tag{13}$$

Thus, the greater the surface tension and the finer the capillary radius, the greater the rise of liquid in the capillary.

If the contact angle of liquid is not 0° (Fig 20-8), the same relationship may be developed, except the vertical component of F which opposes the weight of the column is $F \cos \theta$ and, therefore

$$h = \frac{2\gamma \cos \theta}{rdg} \tag{14}$$

This indicates the very important fact that if θ is less than 90° , but greater than 0° , the value of h will decrease with increasing contact angle until at 90° ($\cos \theta = 0^\circ$), $h = 0$. Above 90° , values of h will be negative, as indicated in Figure 20-7 for mercury. Thus, based on these equations it may be concluded that capillarity will occur spontaneously in a cylindrical pore even if the contact angle is greater than 0° , but it will not occur at all if the contact angle becomes 90° or more. In solids with irregularly shaped pores the relationships between parameters in Equation 14 will be the same, but they will be more difficult to quantitate because of nonuniform changes in pore radius throughout the porous structure.

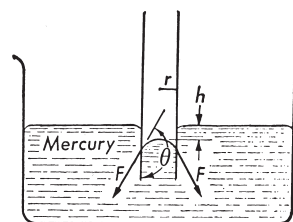


Figure 20-7. Capillary fall for a liquid exhibiting a contact angle, θ , that is greater than 90° . (From Semat H. *Fundamentals of Physics*, 3rd ed. New York: Holt Reinhart Winston, 1957.)

Table 20-5. Critical Surface Tensions of Various Polymeric Solids

POLYMERIC SOLID	Γ_c (dyne/cm AT 20°C)
Polymethacrylic ester of ϕ' -octanol	10.6
Polyhexafluoropropylene	16.2
Polytetrafluoroethylene	19
Polytrifluoroethylene	22
Poly(vinylidene fluoride)	25
Poly(vinyl fluoride)	28
Polyethylene	31
Polytrifluorochloroethylene	31
Polystyrene	33
Poly(vinyl alcohol)	37
Poly(methyl methacrylate)	39
Poly(vinyl chloride)	39
Poly(vinylidene chloride)	40
Poly(ethylene terephthalate)	43
Poly(hexamethylene adipamide)	46

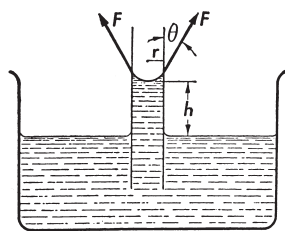


Figure 20-8. Capillary rise for a liquid exhibiting a contact angle, θ , that is greater than 0° but less than 90° . (From Semat H. *Fundamentals of Physics*, 3rd ed. New York: Holt Reinhart Winston, 1957.)

PRESSURE DIFFERENCES ACROSS CURVED SURFACES

From the preceding discussion of capillarity another important concept follows. In order for the liquid in a capillary to rise spontaneously it must develop a higher pressure than the lower level of the liquid in the beaker. However, because the system is open to the atmosphere, both surfaces are in equilibrium with the atmospheric pressure. To be raised above the level of liquid in the beaker and produce a hydrostatic pressure equal to hgd , the pressure just below the liquid meniscus, in the capillary, P_1 , must be less than that just below the flat liquid surface, P_0 , by hgd , and therefore

$$P_0 - P_1 = hgd \quad (15)$$

Because, according to Equation 14,

$$h = \frac{2\gamma \cos \theta}{rgd}$$

then

$$P_0 - P_1 = \frac{2\gamma \cos \theta}{r} \quad (16)$$

For a contact angle of 0° , where the radius of the capillary is the radius of the hemisphere making up the meniscus,

$$P_0 - P_1 = \frac{2\gamma}{r} \quad (17)$$

The consequences of this relationship (known as the Laplace equation) are important for any curved surface when r becomes very small and γ is relatively significant. For example, a spherical droplet of air formed in a bulk liquid and having a radius r will have a greater pressure on the inner concave surface than on the convex side, as expressed in Equation 17. Direct measurement of the pressure difference, $(P_0 - P_1)$, for an air bubble of known radius allows the determination of the surface tension of either a pure liquid or a solution of surface active substance. Both static (constant radius) and dynamic (radius changing in a cyclic fashion as a function of time) measurements have been employed. The latter treatment, known as the pulsating bubble method, has been very useful in the study of some of the biophysical properties and associated disease states of pulmonary surfactant, a mixture of surface active materials lining the small airways of the mammalian lung.⁷ One of the less appreciated advantages of this method for measuring surface tension is the need for only a very small sample size, typically on the order of $50 \mu\text{L}$.

Another direct consequence of what Equation 17 expresses is the fact that very small droplets of liquid, having highly curved surfaces, will exhibit a higher vapor pressure, VP , than observed are over a flat surface of the same liquid at VP' . Equation 18, called the *Kelvin equation*, expresses the ratio of VP/VP' to droplet radius r , and surface tension γ :

$$\log \frac{P}{P'} = \frac{2\gamma M}{2.303RT\rho r} \quad (18)$$

where M is the molecular weight, R is the gas constant in erg/mol/degree, T is temperature, and ρ is the density in g/cm^3 . Values for the ratio of vapor pressures are given in Table 20-6 for water droplets of varying size. Such ratios indicate why it is possible for very fine water droplets in clouds to remain uncondensed despite their close proximity to one another.

This same behavior may be seen when measuring the solubility of very fine solid particles, as both vapor pressure and solubility are measures of the escaping tendency of molecules from a surface. Indeed, the equilibrium solubility of extremely small particles has been shown to be greater than the usual value noted for coarser particles; the greater the surface energy and smaller the particles, the greater this effect.

ADSORPTION

Vapor Adsorption on Solid Surfaces

It was suggested earlier that a high surface or interfacial free energy may exist at a solid surface if the unbalanced forces at the surface and the area of exposed groups are quite great.

Substances such as metals, metal oxides, silicates, and salts—all containing exposed polar groups—may be classified as high-energy or hydrophilic solids; nonpolar solids such as carbon, sulfur, polyethylene, or Teflon (polytetrafluoroethylene) may be classified as low-energy or hydrophobic solids (see Table 20-3). Whereas liquids satisfy their unbalanced surface forces by changes in shape, pure solids (which exhibit negligible surface mobility) must rely on reaction with molecules either in the vapor state or in a solution that comes in contact with the solid surface to accomplish this.

Vapor adsorption is the simplest model demonstrating how solids reduce their surface free energy in this manner. Depending on the chemical nature of the adsorbent (solid) and the adsorbate (vapor), the strength of interaction between the two species may vary from strong specific chemical bonding to interactions produced by the weaker, more nonspecific London dispersion forces. Ordinarily, these latter forces are those responsible for the condensation of relatively nonpolar substances such as N_2 , O_2 , CO_2 , or hydrocarbons.

When chemical reaction occurs, the process is called *chemisorption*; when dispersion forces predominate, the term *physisorption* is used. Physisorption occurs at temperatures approaching the liquefaction temperature of the vapor; for chemisorption, temperatures depend on the particular reaction involved. Water-vapor adsorption to various polar solids can occur at room temperature through hydrogen-bonding, with binding energies intermediate to physisorption and chemisorption.

To study the adsorption of vapors onto solid surfaces, one must measure the amount of gas adsorbed/unit area or unit mass of solid, at different pressures of gas. Because such studies usually are conducted at constant temperature, plots of volume adsorbed versus pressure are known as *adsorption isotherms*. If the physical or chemical adsorption process is

Table 20-6. Ratio of Observed Vapor Pressure (P) to Expected Vapor Pressure (P') of Water at 25°C With Varying Droplet Size

P/P'	DROPLET SIZE (μm)
1.001	1
1.01	0.1
1.1	0.01
2.0	0.005
3.0	0.001
4.2	0.00065
5.2	0.00060

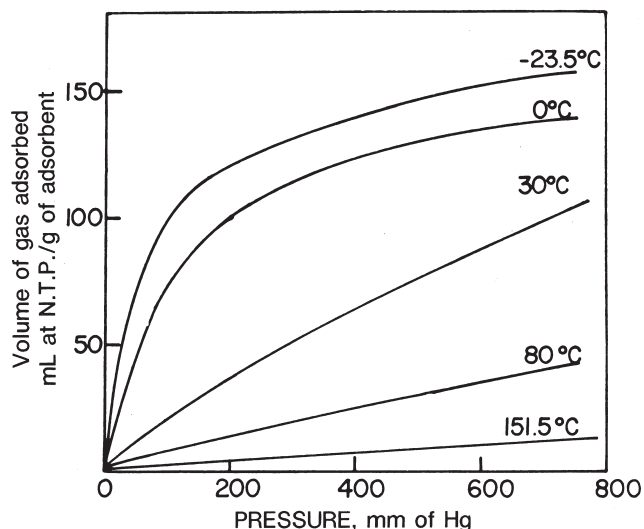


Figure 20-9. Adsorption isotherms for ammonia on charcoal. (From Titoff Z. *Z Phys Chem Leipzig* 1910; 74: 641.)

monomolecular, the adsorption isotherm should appear similar to those shown in Figure 20-9. Adsorption significantly increases with increasing pressure, followed by a leveling off, which is due either to a saturation of available specific chemical groups, as in chemisorption, or to the entire available surface being covered by physically adsorbed molecules. Adsorption reduction with increasing temperature occurs because the adsorption process is exothermic. In the case of physical adsorption at low temperatures after adsorption levels off, often a marked increase in adsorption occurs, presumably due to multilayered adsorption. In this case vapor molecules essentially condense upon themselves as the liquefaction pressure of the vapor is approached. Figure 20-10 illustrates one type of isotherm generally seen with multilayered physisorption.

To have a quantitative understanding of the adsorption process and to be able to compare different systems, two factors must be evaluated. It is important to know the capacity of the solid or the maximum amount of adsorption under a given set of conditions and the affinity of a given substance for the solid surface—how readily does it adsorb for a given amount of pressure? In effect, the second term is the equilibrium constant for the process. For many systems vapor-adsorption data may fit a

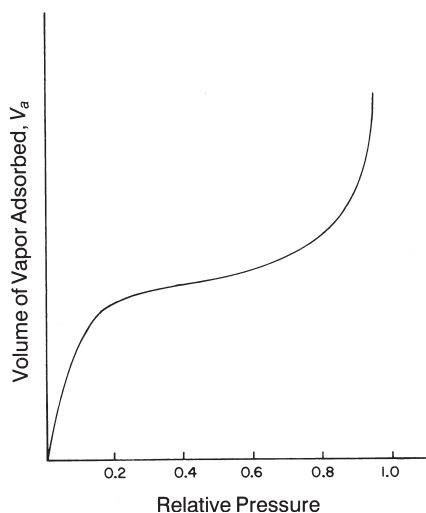


Figure 20-10. Typical plot for multilayer physical adsorption of a vapor on a solid surface.

very general, but somewhat empirical equation, the Freundlich equation:

$$V_a = kp^n \quad (19)$$

where V_a is the volume of gas adsorbed, p is the gas pressure, and k and n are constants reflecting adsorption affinity and capacity.

A significant theoretical improvement along these lines was the theory of *monomolecular adsorption* proposed by Langmuir. He postulated that for adsorption to occur a solid must contain uniform adsorption sites, each capable of holding a single gas molecule. Molecules colliding with the surface may bounce off elastically or they may remain in contact for a period of time. It is this contact over a period of time that Langmuir termed *adsorption*.

Two major assumptions were made in deriving the adsorption equation:

1. Only those molecules striking an empty site can be adsorbed; hence, only monomolecular adsorption occurs.
2. The forces of interaction between adsorbed molecules are negligible and, therefore, the probability of a molecule adsorbing onto or desorbing from any site is independent of the surrounding sites.

With these assumptions and applying the kinetic theory of gases, it can be shown that

$$V_a = (V_m k' p) / (1 + k' p) \quad (20)$$

where V_m is the volume of gas covering all of the adsorption sites with a single layer of molecules and k' is a constant that reflects the affinity of the gas for the solid.

A test of fit to this equation can be made by expressing it in linear form.

$$\frac{p}{V_a} = \frac{1}{V_m k'} + \frac{p}{V_m} \quad (21)$$

The value of k' is, in effect, the equilibrium constant and may be used to compare affinities of different substances for the solid surface. The value of V_m is valuable because it indicates the maximum number of sites available for adsorption. In the case of physisorption the maximum number of sites is actually the total surface area of the solid; therefore, the value of V_m can be used to estimate surface area if the volume and area/molecule of vapor are known.

Since physisorption most often involves some multilayered adsorption, an equation based on the Langmuir equation, the B.E.T. equation, normally is used to determine V_m and solid surface areas. Equation 22 is the B.E.T. equation:

$$V_a = \frac{V_m c p}{(p_0 - p)[1 + (c - 1)(p/p_0)]} \quad (22)$$

where c is a constant and p_0 is the vapor pressure of the adsorbing substance.⁹ Experimentally, the most widely used vapor for this purpose is nitrogen, which adsorbs nonspecifically on most solids near its boiling point at -195° and appears to occupy about $16 \text{ \AA}^2/\text{molecule}$ on a solid surface.

Adsorption from Solution

By far one of the most important aspects of interfacial phenomena encountered in pharmaceutical systems is the tendency for substances dissolved in a liquid to adsorb to various interfaces. Adsorption from solution is generally more complex than that from the vapor state because of the influence of the solvent and any other solutes dissolved in the solvent. Although such adsorption generally is limited to one or two molecular layers at most, the presence of other molecules often makes the interpretation of adsorption mechanisms much more difficult than for chemisorption or physisorption of a vapor. Because monomolecular adsorption from solution is so widespread at all interfaces, we will first discuss the nature of monomolecular films and then return to a discussion of adsorption from solution.

Insoluble Monomolecular Films

It was suggested above that molecules exhibiting a tendency to spread out at an interface might be expected to orient so as to reduce the interfacial free energy produced by the presence of the interface. Direct evidence for molecular orientation has been obtained from studies dealing with the spreading on water of insoluble polar substances containing long hydrocarbon chains, such as fatty acids.

In the late 19th century Pockels and Rayleigh showed that a very small amount of olive or castor oil, when placed on the surface of water, spreads out, as discussed above. If the amount of material was less than could physically cover the entire surface, only a slight reduction in the surface tension of water was noted. However, if the surface was compressed between barriers, as shown in Figure 20-11, the surface tension was reduced considerably.

Devaux extended the use of this technique by dissolving small amounts of solid in volatile solvents and dropping the solution onto a water surface. After assisting the water-insoluble molecules to spread, the solvent evaporated, leaving a surface film containing a known amount of solute.

Compression and measurement of surface tension indicated that a maximum reduction of surface was reached when the number of molecules/unit area was reduced to a value corresponding to complete coverage of the surface. This suggested that a monomolecular film forms and that surface tension is reduced upon compression because contact between air and water is reduced by the presence of the film molecules. Beyond the point of closest packing, the film apparently collapses very much as a layer of corks floating on water would be disrupted when laterally compressed beyond the point of initial physical contact.

Using a refined quantitative technique based on these studies, Langmuir¹¹ spread films of pure fatty acids, alcohols, and esters on the surface of water. Comparing a series of saturated fatty acids, differing only in chain length, he found that the area/molecule at collapse was independent of chain length, corresponding to the cross-sectional area of a molecule oriented in a vertical position (see Fig 20-11). He further concluded that this molecular orientation involved association of the polar carboxyl group with the water phase and the nonpolar alkyl chain out toward the vapor phase.

In addition to the evidence for molecular orientation, Langmuir's work with surface films revealed that each substance exhibits film properties which reflect the interactions between molecules in the surface film. This is seen best by plotting the difference in surface tension of the clean surface γ_0 , and that of the surface covered with the film γ , versus the area/molecule A produced by film compression (total area/the number of molecules). The difference in surface tension is called the surface pressure, π , and thus

$$\pi = \gamma_0 - \gamma \quad (23)$$

Figure 20-12 depicts such a plot for a typical fatty acid monomolecular film. At areas greater than $50 \text{ \AA}^2/\text{molecule}$ the molecules are far apart and do not cover enough surface to reduce the surface tension of the clean surface to any extent and thus the lack of appreciable surface pressure. Because the molecules in the film are quite free to move laterally in the sur-

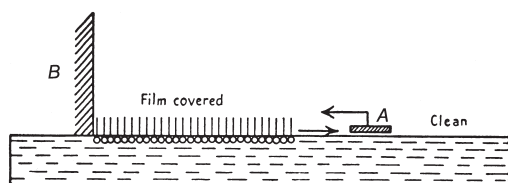


Figure 20-11. Insoluble monomolecular film compressed between a fixed barrier B , and a movable barrier A . (Osipow *LI. Surface Chemistry: Theory and Applications*. New York: Reinhold, 1962.)

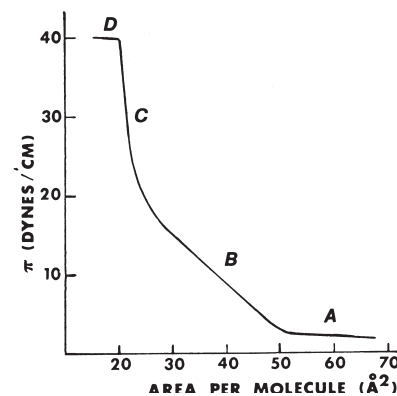


Figure 20-12. A surface pressure–area curve for an insoluble monomolecular film: Region A , gaseous film; Region B , liquid film; Region C , solid film; Region D , film collapse.

face, they are said to be in a two-dimensional *gaseous* or *vapor* state.

As the intermolecular distance is reduced upon compression, the surface pressure rises because the air–water surface is being covered to a greater extent. The rate of change in π with A , however, will depend on the extent of interaction between film molecules—the greater the rate of change, the more “condensed” the state of the film.

In Figure 20-12, from 50 to $30 \text{ \AA}^2/\text{molecule}$, the curve shows a steady increase in π , representative of a two-dimensional “liquid” film, where the molecules become more restricted in their freedom of movement because of interactions. Below $30 \text{ \AA}^2/\text{molecule}$, the increase in π occurs over a narrow range of A , characteristic of closest packing and a two-dimensional “solid” film.

Any factor tending to increase polarity or bulkiness of the molecule—such as increased charge, number of polar groups, reduction in chain length, or the introduction of aromatic rings, side chains, and double bonds—should reduce molecular interactions. On the other hand, the longer the alkyl chain and the less bulky the polar group, the closer the molecules can approach and the stronger the extent of interaction in the film.

Soluble Films and Adsorption from Solution

If a fatty acid exhibits highly gaseous film behavior on an aqueous surface, a relatively small change in π with A over a considerable range of compression should be expected. Indeed, for short-chain compounds such as lauric acid (12 carbons) or decanoic acid, not only is the change in π small with decreasing A , but at a point just before the expected closest packing area, the surface pressure becomes constant without any collapse.

If lauric acid is converted to the laurate ion, or if a shorter chain acid such as octanoic acid is used, spreading on water and compression of the surface produces no increase in π . These results illustrate that the more polar the molecule (hence, the more *gaseous* the film), the higher the area/molecule where a constant surface pressure occurs. This behavior may be explained by assuming that polar molecules form monomolecular films when spread on water but that, upon compression, they are caused to enter the aqueous bulk solution rather than to remain as an intact insoluble film. The constant surface pressure with increased compression arises because a constant number of molecules/unit area remain at the surface in equilibrium with dissolved molecules. The extent of such behavior will be greater for substances exhibiting weaker intermolecular interaction and greater water solubility.

Starting from the other direction, it can be shown that short-chain acids and alcohols (when dissolved in water) reduce the surface tension of water, thus producing a surface

pressure, just as with insoluble films (see Equation 23). That dissolved molecules are accumulating at the interface in the form of a monomolecular film is suggested from the similarity in behavior to systems where lightly soluble molecules are spread on the surface. For example, compressing the surface of a solution containing "surface-active" molecules has no effect on the initial surface pressure, whereas increasing bulk-solution concentration tends to increase surface pressure, presumably by shifting the equilibrium between surface and bulk molecules.

At this point one may ask, why should water-soluble molecules leave an aqueous phase and accumulate or *adsorb* at an air-solution interface? Because any process will occur spontaneously if it results in a net loss in free energy, such must be the case for the process of adsorption. A number of factors will produce such a favorable change in free energy:

- The presence of the oriented monomolecular film reduces the surface free energy of the air-water interface.
- The hydrophobic group on the molecule is in a lower state of energy at the interface, where it no longer is as surrounded by water molecules, than when it is in the bulk-solution phase.
- Increased interaction between film molecules also will contribute to this process.

A further reduction in free energy occurs upon adsorption because of the gain in entropy associated with a change in water structure. Water molecules, in the presence of dissolved alkyl chains are more highly organized or *ice-like* than they are as a pure bulk phase; hence, the entropy of such structured water is lower than that of bulk water.

The process of adsorption requires that the ice-like structure *melt* as the chains go to the interface, and thus an increase in the entropy of water occurs. The adsorption of molecules dissolved in oil can occur but it is not influenced by water structure changes, and hence, only the first factors mentioned are important here.

It is very rare that significant adsorption can occur at the hydrocarbon-air interface as little loss in free energy comes about by bringing hydrocarbon chains with polar groups attached to this interface. On the other hand, at oil-water interfaces, the polar portions of the molecule can interact with water at the interface, leading to significant adsorption.

Thus, whereas water-soluble fatty acid salts are adsorbed from water to air-water and oil-water interfaces, their undissociated counterparts, the free fatty acids, which are water insoluble, form insoluble films at the air-water interface, are not adsorbed from oil solution to an oil-air interface, but show significant adsorption at the oil-water interface when dissolved in oil.

From this discussion it is possible also to conclude that adsorption from aqueous solution requires a lower solute concentration to obtain the same level of adsorption if the hydrophobic chain length is increased or if the polar portion of the molecule is less hydrophilic. On the other hand, adsorption from nonpolar solvents is favored when the solute is quite polar.

Because soluble or adsorbed films cannot be compressed, there is no simple, direct way to estimate the number of molecules/unit area coming to the surface under a given set of conditions. For relatively simple systems it is possible to estimate this value by application of the Gibbs equation, which relates surface concentration to the surface-tension change produced at different solute activities. The derivation of this equation is beyond the scope of this discussion, but it arises from a classical thermodynamic treatment of the change in free energy when molecules concentrate at the boundary between two phases. The equation may be expressed as

$$\Gamma = - \frac{a}{RT} \frac{d\gamma}{da} \quad (24)$$

where Γ is the moles of solute adsorbed/unit area, R is the gas constant, T is the absolute temperature and $d\gamma$ is the change in surface tension with a change in solute activity, da , at activity a .

For dilute solutions of nonelectrolytes, or for electrolytes when the Debye-Hückel equation for activity coefficient is applicable, the value of a may be replaced by solute concentration, c . Because the term dc/c is equal to $d \ln c$, the Gibbs equation is often written as

$$\Gamma = - \frac{1}{RT} \frac{d\gamma}{d \ln c} \quad (25)$$

In this way the slope of a plot of γ versus $\ln c$ multiplied by $1/RT$ should give Γ at a particular value of c .

Figure 20-13 depicts typical plots for a series of water-soluble surface-active agents differing only in the alkyl chain length. A greater reduction of surface tension occurs at lower concentrations for longer chain-length compounds. In addition, there are greater slopes with increasing concentration, indicating more adsorption (Equation 25), and an abrupt leveling of surface tension at higher concentrations. This latter behavior reflects the self-association of surface-active agent to form micelles which exhibit no further tendency to reduce surface tension. The topic of micelles will be discussed later in Chapter 21.

If one plots the values of surface concentration, Γ versus concentration c , for substances adsorbing to the vapor-liquid and liquid-liquid interfaces, using data such as those given in Figure 20-13, one generally obtains an adsorption isotherm shaped like those in Figure 20-9 for vapor adsorption. Indeed, it can be shown that the Langmuir equation (Equation 20) can be fitted to such data when written in the form

$$\Gamma = \frac{\Gamma_{\max} k'c}{1 + k'c} \quad (26)$$

where Γ_{\max} is the maximum surface concentration attained with increasing concentration and k' is related to k in Equation 20. Combining Equations 24 and 26 leads to a widely used relationship between surface-tension change Π (see Equation 23), and solute concentration c , known as the Syszkowski equation.

$$\Pi = \Gamma_{\max} RT \ln(1 + k'c) \quad (27)$$

Mixed Films

It would seem reasonable to expect that the properties of a surface film could be varied greatly if a mixture of surface-active agents were in the film. As an example, consider that a mixture of short- and long-chain fatty acids would be expected to show a degree of *condensation* varying from the gaseous state when the short-chain substance is used in high amount to a highly condensed state when the longer chain substance predomi-

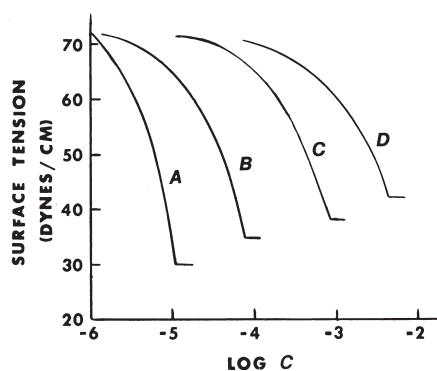


Figure 20-13. The effect of increasing chain length on the surface activity of a surfactant at the air-aqueous solution interface (each curve differs from the preceding or succeeding by two methylene groups with A, the longest chain, and D, the shortest).

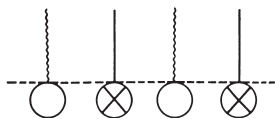


Figure 20-14. A mixed monomolecular film. \equiv , a long-chain ion; \circ , a long-chain nonionic compound.

nates. Thus, each component in such a case would operate independently by bringing a proportional amount of film behavior to the system.

More often, the ingredients of a surface film do not behave independently, but rather interact to produce a new surface film. An obvious example would be the combination of organic amines and acids which are charged oppositely and would be expected to interact strongly. In addition to such polar-group interactions, chain-chain interactions strongly favor mixed condensed films. An important example of such a case occurs when a long-chain alcohol is introduced along with an ionized long-chain substance. Together the molecules form a highly condensed film despite the presence of a high number of like charges. Presumably this occurs as seen in Figure 20-14, by arranging the molecules so that ionic groups alternate with alcohol groups; however, if chain-chain interactions are not strong, the ionic species often will be displaced by the more nonpolar unionized species and will “desorb” into the bulk solution.

On the other hand, sometimes the more soluble surface-active agent produces surface pressures in excess of the collapse pressure of the insoluble film and displaces it from the surface. This is an important concept because it is the underlying principle behind cell lysis by surface-active agents and some drugs, and behind the important process of detergency.

Adsorption from Solution on to Solid Surfaces

Adsorption to solid surfaces from solution may occur if the dissolved molecules and the solid surface have chemical groups capable of interacting. Nonspecific adsorption also will occur if the solute is surface active and if the surface area of the solid is high. This latter case would be the same as occurs at the vapor-liquid and liquid-liquid interfaces. As with adsorption to liquid interfaces, adsorption to solid surfaces from solution generally leads to a monomolecular layer, often described by the Langmuir equation in the form:

$$x/M = [(x/M)m k^* c] / (1 + k^* c) \quad (28)$$

where x is the amount of adsorbed solute, M is the total weight of solid, x/M is the amount of solute adsorbed per unit weight of solid at concentration c , k^* is a constant, and $(x/M)m$ is the amount of solute per unit weight covering the surface with a complete monolayer. However, as Giles¹² has pointed out, the variety of combinations of solutes and solids, and hence the variety of possible mechanisms of adsorption, can lead to a number of more complex isotherms. In particular, adsorption of surfactants and polymers, of great importance in a number of pharmaceutical systems, still is not understood well on a fundamental level, and may in some situations even be multilayered.

Adsorption from solution may be measured by separating solid and solution and either estimating the amount of adsorbate adhering to the solid or the loss in concentration of adsorbate from solution. In view of the possibility of solvent adsorption, the latter approach really only gives an apparent adsorption. For example, if solvent adsorption is great enough, it is possible to end up with an increased concentration of solute after contact with the solid; here, the term *negative adsorption* is used.

Solvent not only influences adsorption by competing for the surface, but as discussed in connection with adsorption at liquid surfaces, the solvent will determine the escaping tendency of a solute; for example, the more polar the molecule, the less the ad-

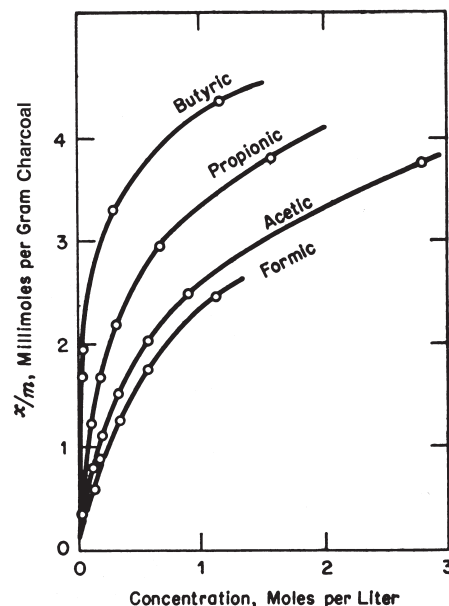


Figure 20-15. The relation between adsorption and molecular weight of fatty acids. (From Weiser HB. *A Textbook of Colloid Chemistry*. New York: Elsevier, 1949.)

sorption that occurs from water. This is seen in Figure 20-15, where adsorption of various fatty acids from water onto charcoal increases with increasing alkyl chain length or nonpolarity. It is difficult to predict these effects, but, in general, the more chemically unlike the solute and solvent and the more alike the solid surface groups and solute, the greater the extent of adsorption. Another factor that must be kept in mind is that charged solid surfaces, such as polyelectrolytes, will strongly adsorb oppositely charged solutes. This is similar to the strong specific binding seen in gas chemisorption, and it is characterized by significant monolayer adsorption at very low concentrations of solute. See Figure 20-16 for an example of such adsorption.

Adsorption onto activated charcoal has been shown to be extremely useful in the emergency treatment of acute overdosage of a variety of drugs taken by the oral route.¹⁵ Overall effectiveness of commercially available activated-charcoal suspen-

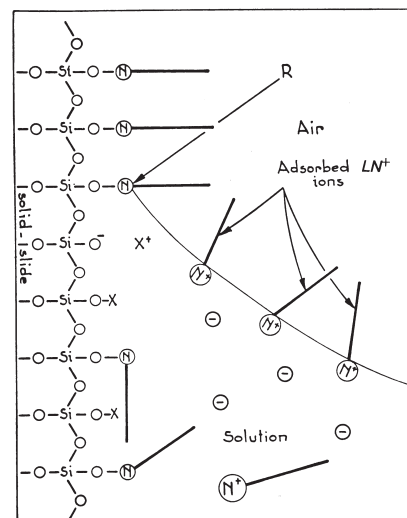


Figure 20-16. The adsorption of a cationic surfactant, LN^+ , onto a negatively charged silica or glass surface, exposing a hydrophobic surface as the solid is exposed to air. (From Ter-Minassian-Saraga L. *Adv Chem Ser* 1964; 43: 232.)

sions as an antidote in oral poisonings appears to be directly related to the total charcoal surface area.¹⁶ Drug adsorption to charcoal tends to follow both the Langmuir model as well as the Freundlich model. In addition, a drug that is un-ionized at gastric pH will adsorb to charcoal to a greater extent than will the ionized form of the drug, probably because of less repulsive interactions in the adsorbed state of neutral molecules. Great care must be exercised in the formulation of activated-charcoal suspensions because pharmaceutical adjuvants employed in suspensions have the potential to adsorb to the charcoal and block sites for drug adsorption.

SURFACE-ACTIVE AGENTS

Throughout the discussion so far, examples of surface-active agents (surfactants) have been restricted primarily to fatty acids and their salts. It has been shown that both a hydrophobic portion (alkyl chain) and a hydrophilic portion (carboxyl and carboxylate groups) are required for their surface activity, the relative degree of polarity determining the tendency to accumulate at interfaces. It now becomes important to look at some of the specific types of surfactants available and to see what structural features are required for different pharmaceutical applications.

The classification of surfactants is quite arbitrary, but one based on chemical structure appears best as a means of introducing the topic. It is generally convenient to categorize surfactants according to their polar portions because the nonpolar portion usually is made up of alkyl or aryl groups. The major polar groups found in most surfactants may be divided as follows: anionic, cationic, amphoteric, and nonionic. As shall be seen, the last group is the largest and most widely used for pharmaceutical systems, so that it will be emphasized in the discussion that follows.

Types

ANIONIC AGENTS—The most commonly used anionic surfactants are those containing carboxylate, sulfonate, and sulfate ions. Those containing carboxylate ions are known as soaps and generally are prepared by the saponification of natural fatty acid glycerides in alkaline solution. The most common cations associated with soaps are sodium, potassium, ammonium, and triethanolamine; the chain length of the fatty acids ranges from 12 to 18.

The extent of solubility in water is influenced greatly by the length of the alkyl chain and the presence of double bonds. For example, sodium stearate is quite insoluble in water at room temperature, whereas sodium oleate under the same conditions is quite water soluble.

Multivalent ions, such as calcium and magnesium, produce marked water insolubility, even at lower alkyl chain lengths; thus, soaps are not useful in hard water that is high in content of these ions. Soaps, being salts of weak acids, are subject also to hydrolysis and the formation of free acid plus hydroxide ion, particularly when in more concentrated solution.

To offset some of the disadvantages of soaps, a number of long alkyl chain sulfonates, as well as alkyl aryl sulfonates such as sodium dodecylbenzene sulfonate, may be used; the sulfonate ion is less subject to hydrolysis and precipitation in the presence of multivalent ions. A popular group of sulfonates, widely used in pharmaceutical systems, are the dialkyl sodium sulfosuccinates, particularly sodium bis-(2-ethylhexyl)sulfosuccinate, best known as Aerosol OT or docusate sodium. This compound is unique in that it is soluble both in oil and water, and hence forms micelles in both phases. It reduces surface and interfacial tension to low values and acts as an excellent wetting agent in many types of solid dosage forms (Table 20-7).

A number of alkyl sulfates are available as surfactants, but by far the most popular member of this group is sodium lauryl

Table 20-7. Effect of Aerosol OT Concentration on the Surface Tension of Water and the Contact Angle of Water with Magnesium Stearate

CONCENTRATION (M $\times 10^6$)	Γ_{sv}	θ ($^\circ$)
1.0	60.1	120
3.0	49.8	113
5.0	45.1	104
8.0	40.6	89
10.0	38.6	80
12.0	37.9	71
15.0	35.0	63
20.0	32.4	54
25.0	29.5	50

sulfate, which is used widely as an emulsifier and solubilizer in pharmaceutical systems. Unlike the sulfonates, sulfates are susceptible to pH-dependent hydrolysis leading to the formation of the long-chain alcohol.

CATIONIC AGENTS—A number of long-chain cations, such as amine salts and quaternary ammonium salts, often are used as surface-active agents when dissolved in water; however, their use in pharmaceutical preparations is limited to that of antimicrobial preservation rather than as surfactants. This arises because the cations adsorb so readily at cell membrane structures in a nonspecific manner, leading to cell lysis (eg, hemolysis), as do anionics to a lesser extent. It is in this way that they act to destroy bacteria and fungi.

Since anionic and nonionic agents are not as effective as preservatives, one must conclude that the positive charge of these compounds is important; however, the extent of surface activity has been shown to determine the amount of material needed for a given amount of preservation. Quaternary ammonium salts are preferable to free amine salts as they are not subject to effect by pH in any way; however, the presence of organic anions such as dyes and natural polyelectrolytes is an important source of incompatibility and such a combination should be avoided.

AMPHOTERIC AGENTS—The major groups of molecules falling into the amphoteric category are those containing carboxylate or phosphate groups as the anion, and amino or quaternary ammonium groups as the cation. The former group is represented by various polypeptides, proteins, and the alkyl betaines; the latter group consists of natural phospholipids such as the lecithins and cephalins. In general, long-chain amphoterics, which exist in solution in zwitterionic form, are more surface active than are ionic surfactants having the same hydrophobic group, because in effect the oppositely charged ions are neutralized. However, when compared to nonionics, they appear somewhere between ionic and nonionic.

PROTEINS—Considering the rapidly growing importance of proteins as therapeutic agents, the unique surface characteristics of these biological macromolecules deserve some special attention. Therapeutic proteins have been shown to be extremely surface active, and they adsorb to clinically important surfaces such as glass bottles and syringes, sterile filters, and plastic IV bags and administration sets; the result is treatment failures. In general, proteins can adsorb to a whole variety of surfaces, both hydrophobic and hydrophilic. From the standpoint of the surface, protein adsorption appears to be maximized when the electrical charge of the surface is opposite that of the protein or when the surface is extremely hydrophobic. From the standpoint of the protein, the extent of adsorption depends on the molecular weight, the number of hydrophobic side chains, and the relative distribution of cationic and anionic side chains. The effect of ionic strength is usually to enhance adsorption by shielding adjacent proteins from repulsive electrical interactions. Adsorption is also maximized when the pH of

the protein solution is equal to the pI (isoelectric point) of the molecule, again due to minimized electrical repulsion.

When different proteins compete for adsorption sites on a single surface, the effect of molecular weight becomes most striking. Early in the adsorption process the protein with the smaller molecular weight, which can diffuse to the surface more rapidly, initially occupies the interface. After some time, it is found that the larger molecular weight protein has displaced the smaller protein since the larger molecule has more possible interaction points with the surface and thus greater total energy of interaction.

The most important consequence of therapeutic protein adsorption is the loss of bioactivity, the reasons for which include loss of therapeutic agent by irreversible adsorption to the surface, possible structural changes in the protein induced by the interface, and surface-associated aggregation and precipitation of the protein. Each of these consequences is related to the structure adopted by the protein in the interfacial region. The native three-dimensional structure of a protein in solution is the result of a complex balance between attractive and repulsive forces. Surface can easily disrupt the balance of forces in proteins residing in the interfacial region and cause the molecule to undergo a change, unfolding from the native to the extended configuration. As it is unlikely that the extended configuration will refold back to the native state upon release from the interface, the protein is considered to be denatured. Like other polymers, the unfolding of the protein at the interface is thought to minimize the contact of apolar amino acid side chains with water.

In addition, electrical interactions, both within the protein and between the protein and the surface, strongly modulate the configuration assumed at the interface. Motion of the interface, such as comes about during shaking of a solution, appears to accelerate the surface-associated denaturation. Some proteins appear to be rather vulnerable to surface-induced structural alterations, whereas others are very resistant. Algorithms for predicting those proteins most vulnerable to the structure-damaging effects of interfaces are not yet available. Empirical observations suggest that those proteins easily denatured in solution by elevated temperatures may also be most sensitive to interfacial denaturation.

The best defense against untoward effects on the structure of proteins induced by surfaces appears to be prevention of adsorption. Research in the field of biomaterials has shown that surfaces that are highly hydrophilic are less likely to serve as sites for protein adsorption. Steric hindrance of adsorption by bonding hydrophilic polymers, such as polyethylene oxide, to a surface also appears to be successful in minimizing adsorption. Formulations of proteins intended for parenteral administration frequently contain synthetic surfactants to preserve bioactivity. The specific molecular mechanism of protection is not understood and can involve specific blocking of adsorption to the interface or enhanced removal from the interface before protein unfolding can occur. In support of the former mechanism is the observation that surfactants most successful at protecting proteins from interfacial denaturation contain long polyethylene oxide chains capable of blocking access of the protein to the surface.

PHOSPHOLIPIDS—All lecithins contain the L- α -glycerophosphocholine skeleton esterified to two long-chain fatty acids (often oleic, palmitic, stearic, and linoleic). Typically, for pharmaceutical use, lecithins are derived from egg yolk or soybean. Although possessing a polar zwitterionic *head* group, the twin hydrocarbon *tails* result in a surfactant with very low water solubility in the monomer state. With the exception of the skin, phospholipids make up a vast majority of the lipid component of cell membranes throughout the body. As a result, the biocompatibility of lecithin is high, accounting for the increasing popularity of use in formulations intended for oral, topical, and intravenous use. Egg yolk lecithins are used extensively as the main emulsifying agent in the fat emulsions intended for intravenous use.

The ability of the lecithins to form a tough but flexible film between the oil and water phases is responsible for the excellent physical stability shown the IV fat emulsions. In aqueous media, phospholipids are capable of assembling into concentric bilayer structures known as liposomes. The therapeutic advantage of such a lipid assembly for drug delivery depends upon the encapsulation of the active ingredient either within the interior aqueous environment or within the hydrophobic region of the bilayer. Deposition of the liposome within the body appears to be dependent upon a number of factors, including the composition of the phospholipids employed in the bilayer and the diameter of the liposome.

The unique surface properties of phospholipids are critical to the function of the pulmonary system. Pulmonary surfactant is a mixture of phospholipids and other associated molecules secreted by type II pneumocytes. In the absence of pulmonary surfactant (as in a neonate born prematurely), the high surface energy of the pulmonary alveoli and airways can be diminished only by physical collapse of these structures and resulting elimination of the air-water interface. As a consequence of airway collapse, the lung fails to act as an organ of gas exchange. Pulmonary surfactant maintains the morphology and function of the alveoli and airways by markedly decreasing surface energy through decreasing the surface tension of the air-water interface.

The most prevalent component of pulmonary surfactant, dipalmitoylphosphatidylcholine (DPPC), is uniquely responsible for forming the very rigid surface film necessary to reduce the surface tension of the interface to a value near 0. Such an extreme reduction in surface tension is most critical during the process of exhalation of the lung where the air-water interfacial area is decreasing. Although DPPC does form the rigid film, in the absence of additives it is unable to respread over an expanding interface typical of a lung during the inhalation phase. An anionic phospholipid, phosphatidylglycerol, in conjunction with a surfactant-associated protein, SP-C, appears to aid the respreading of DPPC and to maintain mechanical stability of the interface. A truly remarkable feature is that pulmonary surfactant is able to carry out the cycle of reducing surface tension to near 0 during exhalation and then reexpanding over the interface during inhalation at whatever rate is necessary by the respiratory pattern.

Commercially available pulmonary surfactant replacement preparations contain DPPC as the primary ingredient. Agents that aid in the respreading of DPPC may differ depending upon the source of the surface-active material.

NONIONIC AGENTS—The major class of compounds used in pharmaceutical systems are the nonionic surfactants, as their advantages with respect to compatibility, stability, and potential toxicity are quite significant. It is convenient to divide these compounds into those that are relatively water insoluble and those that are quite water soluble. The major types of compounds making up this first group are the long-chain fatty acids and their water-insoluble derivatives. These include:

- Fatty alcohols such as lauryl, cetyl (16 carbons), and stearyl alcohols
- Glycerol esters such as the naturally occurring mono-, di-, and triglycerides
- Fatty acid esters of fatty alcohols and other alcohols such as propylene glycol, polyethylene glycol, sorbitan, sucrose, and cholesterol. Included also in this general class of nonionic water-insoluble compounds are the free steroidal alcohols such as cholesterol.

To increase the water solubility of these compounds and to form the second group of nonionic agents, polyoxyethylene groups are added through an ether linkage with one of their alcohol groups. The list of derivatives available is much too long to cover completely, but a few general categories will be given.

The most widely used compounds are the polyoxyethylene sorbitan fatty acid esters, found in pharmaceutical formulations that are to be used both internally and externally. Closely related compounds include polyoxyethylene glyceryl and

steroidal esters, as well as the comparable polyoxypropylene esters. It is also possible to have a direct ether linkage with the hydrophobic group, as with a polyoxyethylene–stearyl ether or a polyoxyethylene–alkyl phenol. These ethers offer advantages because, unlike the esters, they are quite resistant to acidic or alkaline hydrolysis.

Besides the classification of surfactants according to their polar portion, it is useful to have a method that categorizes them in a manner that reflects their interfacial activity and their ability to function as wetting agents, emulsifiers, and solubilizers. Variation in the relative polarity or nonpolarity of a surfactant significantly influences its interfacial behavior, so some measure of polarity or nonpolarity should be useful as a means of classification.

One such approach assigns a hydrophile–lipophile balance (HLB) number for each surfactant; although the method was developed by a commercial supplier of one group of surfactants, it has received widespread application.

The HLB value, as originally conceived for nonionic surfactants, is merely the percentage weight of the hydrophilic group divided by 5 in order to reduce the range of values. On a molar basis, therefore, a 100% hydrophilic molecule (polyethylene glycol) would have a value of 20. Thus, an increase in polyoxyethylene chain length increases polarity, and hence the HLB value; at constant polar chain length, an increase in alkyl chain length or number of fatty acid groups decreases polarity and the HLB value. One immediate advantage of this system is that to a first approximation one can compare any chemical type of surfactant to another type when both polar and nonpolar groups are different.

Values of HLB for nonionics are calculable on the basis of the proportion of polyoxyethylene chain present; however, to determine values for other types of surfactants, it is necessary to compare physical chemical properties reflecting polarity with those surfactants having known HLB values.

Relationships between HLB and phenomena such as water solubility, interfacial tension, and dielectric constant have been used. Those surfactants exhibiting values greater than 20 (eg, sodium lauryl sulfate) demonstrate hydrophilic behavior in ex-

cess of the polyoxyethylene groups alone. Refer to Chapter 22 for further information.

Acknowledgment—The author is grateful to Professor George Zografi for his continuing mentorship and support.

REFERENCES

1. Semat H. *Fundamentals of Physics*, 3rd ed. New York: Holt Rinehart Winston, 1957.
2. Michaels AS. *J Phys Chem* 1961; 65:1730.
3. Ring TA. *Powder Tech* 1991; 65:195.
4. Elamin AA, et al. *Int J Pharmaceut* 1994; 111:159.
5. Dirkson JA, Ring TA. *Chem Eng Sci* 1991; 46:2389.
6. Zisman WA. *Adv Chem Ser* 1964; 43:1.
7. Putz G, et al. *J Appl Physiol* 1994; 76:1425.
8. Titoff Z. *Z Phys Chem Leipzig* 1910; 74:641.
9. Brittain HG. *Physical Characterization of Pharmaceutical Solids*. New York: Dekker, 1995.
10. Osipow LI. *Surface Chemistry: Theory and Applications*. New York: Reinhold, 1962.
11. Langmuir I. *J Am Chem Soc* 1917; 39:1848.
12. Giles CH. In: EH Lucassen-Reynders, ed. *Anionic Surfactants*. New York: Dekker, 1981, Chap 4.
13. Weiser HB. *A Textbook of Colloid Chemistry*. New York: Elsevier, 1949.
14. Ter-Minassian-Saraga L. *Adv Chem Ser* 1964; 43:232.
15. Cooney DO. *Activated Charcoal in Medical Applications*, Dekker, New York, 1995.
16. Modi NB, et al. *Pharm Res* 1994; 11:318.

BIBLIOGRAPHY

- Adamson AW. *Physical Chemistry of Surfaces*, 5th ed. New York: Wiley Interscience, 1990.
- David JT, Rideal EK. *Interfacial Phenomena*, 2nd ed. New York: Academic Press, 1963.
- Hiemenz PC. *Principles of Colloid and Surface Chemistry*, 2nd ed. New York: Dekker, 1986.
- MacRitchie F. *Chemistry at Interfaces*. San Diego: Academic Press, 1990.
- Shaw DJ. *Introduction to Colloid and Surface Chemistry*, 4th ed. London: Butterworths, 1992.