adenocarcinoma of the colon or fracture of the femur). On the other hand, the disease name may focus on some real or supposed causative factor; e.g., pneumococcal pneumonia implies a pulmonary infection by the pneumococcus.

As knowledge about disease causation increases, the disease names are often switched from descriptive terms to terms implying a causal factor. Many ill persons who had been formerly named by a variety of descriptive terms become reclassified under a single causal heading. Similarly, a single descriptive heading may have contained patients with a variety of causally defined diseases. One of the former names for the condition we now call tuberculosis was *phthisis*, meaning "wasting away." Patients in whom wasting dominates the clinical picture constitute only a portion of persons with tuberculosis, and tuberculosis is only one of the causes of wasting.

Causal names for disease are useful in that they immediately imply means for prevention or therapy; in fact, they can drastically change the manner in which a particular health problem is handled. However, causal names can also lead to problems. When the focus on one causal factor such as an infectious agent is reflected in the disease name, we often forget that other factors are operating and tend to regard the infectious or other agent as the only cause.

In summary, disease names are important tools for thought and communication. However they must be viewed in proper perspective. They tend to mask differences among patients, and they have a way of influencing and narrowing our thinking. Disease names may even become "the thing itself," whereas the emphasis should be on the ill person. Furthermore, disease names are transitory. The naming and classifying of ill persons has changed markedly through history and will continue to change.

REFERENCE

MacMahon, B., and T. F. Pugh, *Epidemiology: Principles and Methods.* (Boston: Little, Brown, 1970), Chaps. 1, 2, and 4.

Chapter 2

Basic Measurements in Epidemiology

There is one thing I would be glad to ask you. When a mathematician engaged in investigating physical actions and results has arrived at his conclusions, may they not be expressed in common language as fully, clearly, and definitely as in mathematical formulae? If so, would it not be a great boon to such as I to express them so?

Michael Faraday,
Letter to James Clerk Maxwell

Epidemiology is a quantitative science. Its measured quantities and descriptive terms are used to describe *groups* of persons.

Counts The simplest and most frequently performed quantitative measurement in epidemiology is a count of the number of persons in the group studied who have a particular disease or a particular characteristic. For example, it may be noted that 10 people in a college dormitory developed infectious hepatitis or that 16 stomach cancer patients were foreign-born.

Proportions and Rates

In order for a count to be descriptive of a group it must be seen in proportion to it; i.e., it must be divided by the total number in the group. The 10 hepatitis cases would have quite a different significance for the dormitory if the dormitory housed 500 students than if it housed only 20. In the first case the proportion would be $^{10}/_{500}$, or 0.02, or 2 percent. (Percentage, or number per one hundred, is one of the most common ways of expressing proportions. Number per 1,000 or 1 million, or any other convenient base may be used.) In the second case the proportion would be $^{10}/_{20}$ or 0.50.

The use of denominators to convert counts into proportions seems almost too simple to mention. However, a proportion is one basic way to describe a group. *One of the central concerns of epidemiology is to find and enumerate appropriate denominators in order to describe and to compare groups in a meaningful and useful way.*

Certain kinds of proportions are used very frequently in epidemiology. These are referred to as *rates.* The various types of rates involve or imply some time relationship. The two most commonly used rates which every physician should understand and remember are the prevalence rate and the incidence rate:

Prevalence Rate

$$\text{Prevalence rate} = \frac{\text{number of persons with a disease}}{\text{total number in group}}$$

Prevalence describes a group at a certain point in time. It is like a snapshot of an existing situation. For example, *the prevalence of electrocardiographic abnormalities at our screening examination was 5 percent*; or, *the prevalence of diarrhea in the children's camp on July 13 was 33 percent.* Or, *the prevalence of significant hyperbilirubinemia in full-term infants on the third postpartum day is 20 percent.* As can be seen by the above examples the point in time is

not necessarily a true geometric point with no length, but is a relatively short time such as a day. Nor does the point have to be in calendar time. It can refer to an event which may happen to different persons at different times, such as an examination or the third postpartum day.

Incidence Rate

$$\text{Incidence rate} = \frac{\text{number of persons developing a disease}}{\text{total number at risk}} \text{ per unit of time}$$

Incidence describes the rate of development of a disease in a group over a period of time, which is included in the denominator. In contrast to a snapshot, incidence describes a continuing process over a given time period. For example, *the incidence of myocardial infarction is about 1 percent per year in men aged 55–59 in our community*; or, *at the height of the epidemic the incidence of chicken pox in the first grade children was 10 percent per day.*

Not everyone in a study population may be at risk for developing a disease. For example, some diseases are lifelong in duration, so that once you have it you cannot develop it again. Persons with such a disease are usually removed from the denominator population at risk.

In the medical literature the word "incidence" is often used to describe prevalence or simple proportion. For example, *the incidence of gallstones is 20 percent in middle-aged women*; or, *in our autopsy series the incidence of liver cirrhosis was 12 percent.* This imprecise use of "incidence" should be avoided, since the specific concept of incidence, defined as a rate of development, is a useful one.

Other Rates Some other rates, often used in epidemiology, are described below.

$$\text{Period prevalence} = \frac{\text{number of persons with a disease during a period of time}}{\text{total number in group}}$$

Sometimes one wishes to have a measure of all the diseases affecting a group during a period of time such as the year, 1970, rather than at a point in time. The period prevalence of a disease in 1970 turns out to be the prevalence at the beginning of 1970 plus the annual incidence during 1970.

$$\text{Mortality, or death, rate} = \frac{\text{number of persons dying (due to a particular cause or due to all causes)}}{\text{total number in group}} \text{ per unit of time}$$

Mortality rate is analogous to incidence but refers to the process of dying rather than the process of becoming ill.

Any rate may refer to a subgroup of a population. For example:

$$\text{Age–specific mortality rate} = \frac{\text{number of persons dying in a particular age group}}{\text{total number in the same age group}} \text{ per unit of time}$$

$$\text{Case fatality rate} = \frac{\text{number of persons dying due to a particular disease}}{\text{total number with the disease}}$$

Case fatality rate refers to the proportion of persons with a particular disease who die. The time element is usually not specified but may be, if desired, as with incidence.

A variety of other disease rates are described by Siegel (1967). In most rates the numerator must include only persons who are derived from the denominator population. The denominator is considered the total population at risk of being or becoming one of the numerator. Thus, these rates can be viewed as a statement of probability that a condition exists (prevalence) or will develop (incidence) in the population at risk.

Some rates depart somewhat from the ideal of having the numerator being derived from the denominator population at risk. This is done for convenience, because of the ready availability of

data that approximate the ideal. Consider the

$$\text{Maternal mortality rate} = \frac{\text{number of deaths from puerperal causes during a year}}{\text{number of live births during the same year}}$$

Actually, the true population of mothers at risk for puerperal death includes those that have had stillbirths as well as those that have had live births. Legally required registration and counting of live births makes this live-birth denominator much more accessible.

Handling Changing Denominators If a denominator population is growing or shrinking during the period of time for which a rate is to be computed, then it is customary to use the population size at the *midpoint* of the time interval as an estimate of the average population at risk. If an incidence rate is to be computed for the year 1973, then the population at risk as of July 1, 1973 is used for the denominator.

Comparison of Rates, Using Differences or Ratios

Differences It is often desired to compare a rate in one group with that in another. One may simply note both rates and observe that one is larger than the other. By subtracting the smaller from the larger, one may obtain the magnitude of the difference.

The difference between two incidence rates is sometimes called "attributable risk" if the two groups being compared differ in some other aspect that is believed to play a causal role in the disease. For example, in Hammond's (1966) study of smoking and mortality the lung cancer mortality rate in nonsmokers ages 55–69 was 19 per 100,000 persons per year as compared to 188 per 100,000 in cigarette smokers. The difference between the two lung cancer mortality rates was 169 per 100,000 per year. This is the lung cancer risk attributable to smoking, *if* smoking is the only important difference between the groups in factors affecting the development of lung cancer. Only the *excess* rate in smokers should be attributed to smoking—not the entire smokers' incidence rate—since nonsmokers develop some lung cancer, too.

Ratios Another way to compare two rates is by determining the ratio of one to the other, that is, dividing one by the other. In the smoking and lung cancer example, the ratio of the rate in smokers to that in nonsmokers was $^{188}/_{19}$ or 9.9. The smokers had a 9.9 times greater risk of dying from lung cancer than did the nonsmokers. The ratio of two rates is sometimes called the "relative risk," "risk ratio," "morbidity ratio," or, if mortality rates are under consideration, the "mortality ratio."

Ratio Comparisons of Several Groups to a Single Standard When one wishes to compare several different rates, it is often convenient to determine the ratio of all the different rates to a single standard. The standard of comparison may be an actual rate for a particular group that seems appropriate to use. In the study of smoking and lung cancer, smokers were divided according to the number of cigarettes currently smoked per day. Nonsmokers were again used as the standard of comparison, and their mortality rate was arbitrarily designated as 1.0. In comparison, the ratios for male smokers, ages 55–69, were 3.5 for smokers of 1 to 9 cigarettes per day, 8.8 for smokers of 10 to 19 cigarettes per day, 13.8 for smokers of 20 to 39 cigarettes per day, and 17.5 for smokers of 40 or more cigarettes per day.

It may be that the group to be used as a standard differs from the other groups in some important respect, resulting in a biased or unfair comparison. For example, suppose that the men in the different smoking categories not only had different smoking habits but were, on the average, of substantially different ages as well. Then it would not be fair to compare their lung cancer incidence as if differences in smoking were all that mattered, since we know that age is also important—the older one gets the higher the likelihood is of developing lung cancer. In order to eliminate this bias we have to determine as a standard of comparison an *expected rate* instead of an actual rate. To do this, we might calculate, for example, what lung cancer incidence rate would be expected in nonsmokers, as before, but now assuming that they were of the same age composition as

that of each group of smokers. The method for computing this expected rate involves what is called *age adjustment*, or *age standardization.* This will be discussed further in Chapter 11.

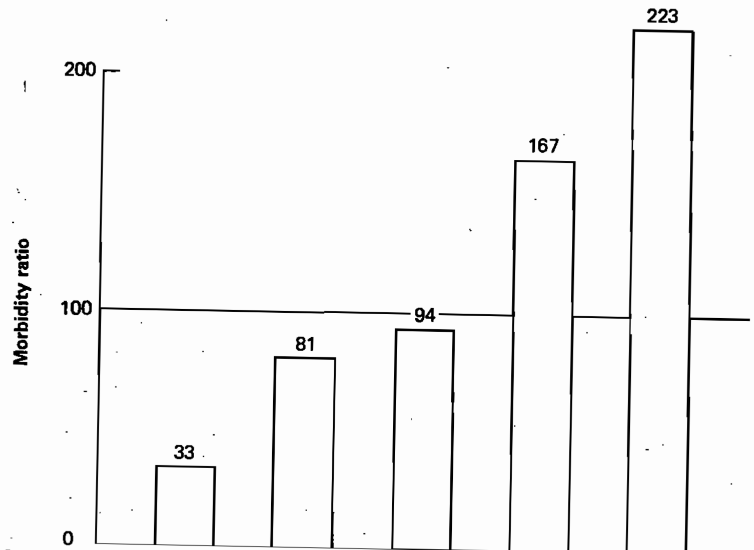
An example of a morbidity ratio comparison using an expected rate is shown in Fig. 2-1. In the Framingham Heart Study men and women in five different blood pressure level groups were compared with one another with respect to the subsequent incidence of coronary heart disease during 8 years. Morbidity ratios were used with an expected rate as a standard of comparison, set at 100 percent. The expected rate was that observed in the whole population, but age-adjusted so that it could be applied fairly to the particular blood pressure group under consideration.

In the figure it can be seen, for example, that for those persons with the lowest systolic blood pressure levels, less than 120 mm Hg, the observed incidence was $^{10}/_{629}$. The expected incidence, based on the experience of the whole population, was $^{30.7}/_{629}$. The ratio of these rates is $^{10}/_{30.7}$, or 33 percent. In contrast to the low incidence in the low blood pressure group, the incidence in the highest group, those with a systolic blood pressure of at least 180 mm Hg, was 223 percent of the expected incidence.

Quantitative Attributes

In considering counts, proportions, and rates we have been dealing with qualitative differences between people—presence or absence of disease, or possession of one versus another attribute. Other characteristics of groups that must be considered lie on a quantitative scale. These characteristics include such measures as height, weight, blood pressure, antibody titer, and diameter of tuberculin skin-test reaction. Epidemiology requires appropriate measures so that groups can be described and compared with respect to these quantitative attributes.

In discussing such measures, one must mention some concepts that are usually presented in books or courses on statistics or biostatistics (see Ipsen and Feigl, 1970). In this introduction to epidemiology it is not necessary to present statistical aspects in great detail, but certain basic measures do deserve mention. Parenthetically, it might be well to remark that one need not be highly



Systolic BP	< 120	120 – 139	140 – 159	160 – 179	180+
Pop. at risk	629	1617	1020	441	252
Obs. cases CHD	10	70	58	47	39
Exp. cases CHD	30.7	86.0	61.6	28.1	17.5

Figure 2-1 Risk of developing coronary heart disease (CHD) in 8 years according to initial systolic blood pressure level. Men and women, ages 30–59 years at entry. Framingham Heart Study. *(Reproduced, by permission, from Kagan et al., 1963.)*

talented in mathematics to understand or carry out epidemiologic studies. While some studies in epidemiology do require sophisticated statistical methods, most problems can be handled well by the simple quantitative measures described here.

Distributions The most complete summary of a quantitative measurement made on a group of persons is the *distribution*. The distribution tells either how many or what proportion of the group were found to have each value (or each small range of values) out of all the possible values that the quantitative measure can have. In addition, the counts or proportions (or percentages) may be cumulated by adding each successive amount to all those that preceded it.

A distribution of serum uric acid values for 1,734 nonsmoking white men, ages 40–49, is shown in Table 2-1. Note that both numbers and percentages are shown for both the distribution and cumulative distribution.

A distribution may be displayed graphically as a histogram, in which bars represent the numbers or proportions of subjects in each "class interval." The uric acid distribution in Table 2-1 is shown in Fig. 2-2 as a histogram. Note that in plotting a histogram the *area* of each bar communicates the number or proportion of subjects represented. If all bars represent class intervals of the same width, then the area is proportional to the height. If some class intervals or bars are wider, as are the extreme right and left bars in Fig. 2-2, their height must be scaled down proportionally.

Another way to display a distribution is by plotting a series of points. Each point shows the midpoint of an interval and the number or proportion of subjects falling into that interval. The points may be connected by straight lines, yielding a polygon, or they may be

Table 2-1 Distribution and Cumulative Distribution of Serum Uric Acid Concentrations: Nonsmoking Men, Ages 40–49

Serum uric acid (mg/100cc)	Distribution		Cumulative distribution	
	Number	Percent	Number	Percent
1.0–2.9	10	0.6	10	0.6
3.0–3.9	68	3.9	78	4.5
4.0–4.9	315	18.2	393	22.7
5.0–5.9	565	32.6	958	55.3
6.0–6.9	431	24.8	1,389	80.1
7.0–7.9	229	13.2	1,618	93.3
8.0–8.9	85	4.9	1,703	98.2
9.0–11.9	31	1.8	1,734	100.0
Total	1,734	100.0		

Mean = 5.93 mg/100 cc
Standard Deviation = 1.31 mg/100 cc
Range = 1.32 to 11.12, or 9.8 mg/100 cc
Median = 5.84 mg/100 cc
Interquartile Range = 5.07 to 6.79, or 1.72 mg/100 cc
Source: Kaiser-Permanente multiphasic examination data, 1964–1968, tabulated by A. B. Siegelaub, M.S.

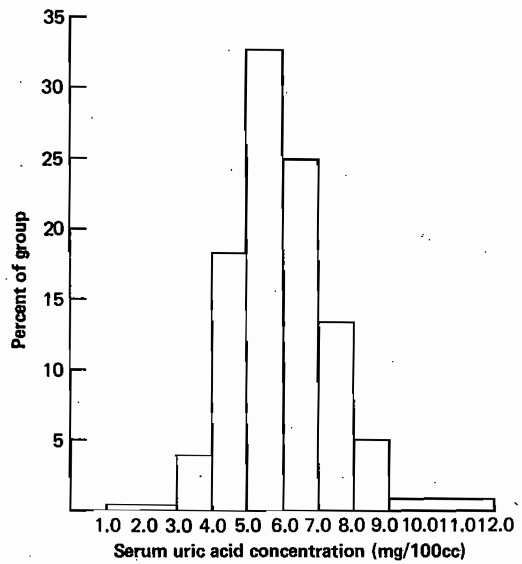


Figure 2-2 Percentage distribution of serum uric acid levels in Table 2-1, displayed as a histogram.

connected so as to form a smooth curve. The uric acid distribution in Table 2-1 is shown as a curve in Fig. 2-3.

Cumulative distributions are usually shown graphically by curves. Fig. 2-4 shows the cumulative distribution curve for the same uric acid data.

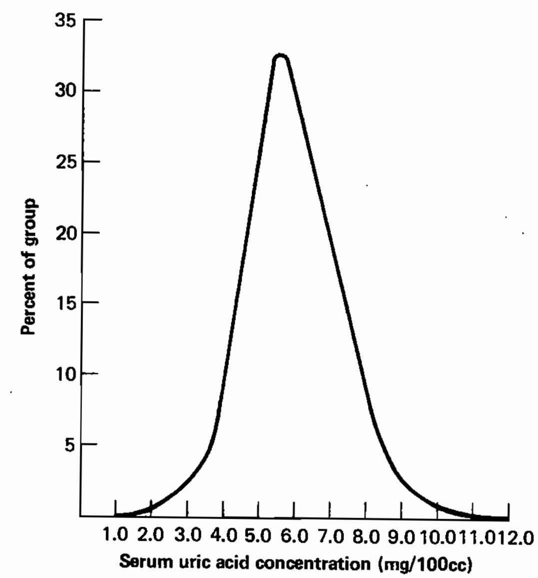
Means The *mean*, or arithmetic average, is one of the so-called measures of central tendency of the values for the whole group. It is computed by adding all the individual values together and dividing by the number in the group. When one wishes to compare two or more groups, it may be cumbersome to compare their entire distributions. Comparing means is much simpler. In many cases, for comparative purposes, the mean is a reasonably good representation of the group's values, and it can be expressed with just one number.

It should always be remembered though that the mean is only one feature of a distribution and that two differently shaped distribu-

tions may have the same mean. It is often important to know more about the distribution than just the mean. In some cases we may be most interested in knowing how many people are at one extreme of the distribution.

Standard Deviations A good supplement to the mean in describing a group is the *standard deviation*, which is a measure of dispersion or variation. One way to compute it is to (1) square the difference between each value and the mean, (2) add the squared differences, (3) divide that sum by the total number of values minus one, and (4) find the square root of the result of (3). The mean tells where the values for a group are centered. The standard deviation is a summary of how widely dispersed the values are around this center. The standard deviation is also needed in comparing means of different groups to see how likely it is that a difference between two means could have occurred by chance, using statistical significance tests.

Figure 2-3 Percentage distribution of serum uric acid levels in Table 2-1, displayed as a curve.



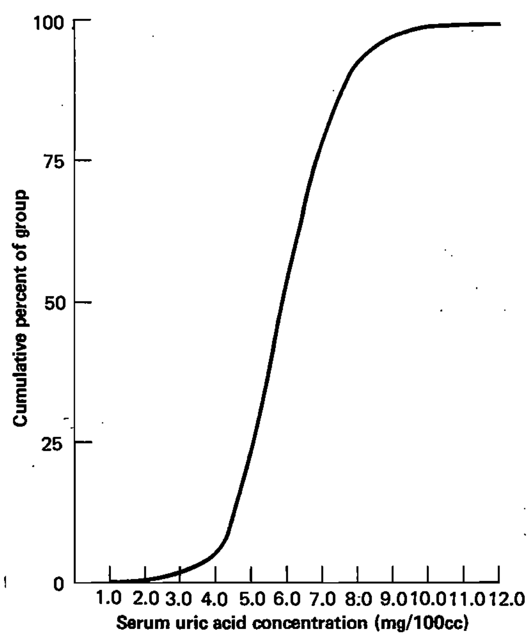


Figure 2-4 Cumulative percentage distribution of serum uric acid levels in Table 2-1, displayed as a curve.

Ranges The *range* of a distribution, the difference between the lowest value and the highest value observed, is, of course, another measure of dispersion. It is often less valuable than the standard deviation, however, since it only tells us about two members of a group. An extremely high or low value may be due to a measurement error.

Quantiles: Values That Divide a Group into Equal Parts Another way to describe a group on a quantitative scale or to classify each member of a group on such a scale is to divide the group into *quantiles*, or equal subgroups, along the scale. The simplest division is into two parts—the lower half and the upper half. The point on the scale that divides the group in this way is called the *median*. In the uric acid distribution shown in Table 2-1 the median value is 5.84 mg/100 cc. (When the median lies within an interval, e.g., between 5.0 and 6.0, we interpolate to estimate just where it lies). One-half of the group has values this high or higher and one-half has values this low or lower. Note that the median value can also be read from the cumulative distribution curve (Fig. 2-4) by seeing what uric acid value corresponds to the 50 percent point on the curve.

Just as one can compare two groups by their means, so one can also compare them by their medians. Medians are less often used than means but they have a few virtues that make them very useful in certain situations. One such situation is when a group has a few members with extreme values. The mean is substantially affected by these extreme values but the median is not. Suppose one wishes to summarize the weights of 22 women attending an obesity clinic. All but one are evenly distributed from 180 to 220 lb (i.e., 180, 182, 184, etc.). One is the fat lady in a traveling circus who weighs 420 lb. When she leaves, the mean weight of the clinic patients will drop by 10 lb, but the median will drop by only 1 lb. Medians are affected little by extreme values.

Another virtue of the median is its usefulness when some values are missing, but known to be above or below a certain level. Suppose one wishes to compare the age at death of two groups of fifty-year-old women exposed to different amounts of ionizing radiation. If one uses the mean age at death, then one must wait until all members of each group die. Conclusions cannot be drawn from the mean age of just some of the deaths, since an early difference between the two groups may be later counterbalanced by a difference in the opposite direction. By the time all the women have died, it is very probable that the investigator will also be dead or no longer interested in the study. Thus it is important to have an earlier answer. The median age at death is one such early measure, since it may be determined when only half the women in each group have died.

Groups may be divided into more than two parts. Three equal parts are known as *terciles*, four equal parts as *quartiles*, five as *quintiles*, ten as *deciles*. The finest division commonly used is into 100 parts, or *percentiles*. Percentiles are often useful for ranking individuals in relation to the total group. (Note that the borderlines between any divisions may be read from the cumulative distribution curve.)

Just as groups can be compared with respect to their medians, they can also be compared as to their borderlines between quartiles, and so on. Similarly, persons in the upper quartile of a value can be compared with those in each of the other quartiles. Also, one may wish to have a measure of dispersion in a group analogous to the standard deviation. The size of the interval between two percentiles, e.g., the 20th and 80th, can be used. One such measure of spread is the *interquartile range*, the interval between the top of the lowest quartile and the bottom of the highest quartile. Note that the interquartile range can easily be read off of a cumulative distribution curve as in Fig. 2-4.

Quantiles may prove very helpful in determining which of two quantitative variables has a stronger relationship to disease. In a particular population group the incidence of coronary heart disease may increase a certain amount with each 20-mm-Hg increase in systolic blood pressure and a different amount with each 20 mg/100 cc increase in serum cholesterol, but this tells us nothing of the relative importance of the two attributes since the units of measurement for blood pressure and cholesterol are completely different, and not at all comparable. A more appropriate contrast would be to note how much the incidence of coronary heart disease increases as one moves up the scale of each measurement by quantile divisions such as deciles or quartiles.

A good example of such a comparison is shown in Fig. 2-5. In the Framingham Heart Study two serum lipid measures, cholesterol and the cholesterol/phospholipid ratio, were compared to determine which was the better predictor of the subsequent development of coronary heart disease. The study population was divided into quartiles of each of the two lipid values. As shown by the morbidity ratios in the figure, the risk of coronary heart disease was clearly related to cholesterol, the incidence being distinctly higher in each successive quartile. In contrast, the increase in risk with increasing quartile of cholesterol/phospholipid ratio was slight, showing that the latter measure was a distinctly inferior predictor.

Epidemiologic Measurements in Perspective

In summary, epidemiology requires that groups of people be described and compared in a quantitative fashion. However, the

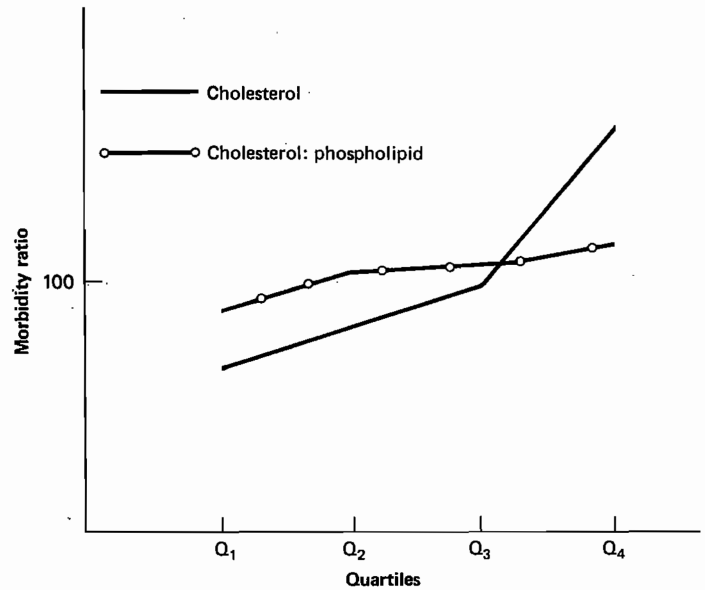


Figure 2-5 Risk of developing coronary heart disease in 10 years in subjects classified into quartiles of cholesterol and cholesterol/phospholipid ratio. Men, ages 30–59 years at entry. Framingham Heart Study. *(Reproduced in modified form, by permission, from Kannel et al., 1964.)*

particular characteristics of interest may be either qualitative or quantitative in nature.

When qualitative attributes are considered, persons with a particular attribute are counted, and the proportion of the total group studied that they constitute is determined. Since disease is the main concern of epidemiology, proportions of groups with disease or rates of disease are given primary attention. Disease rates are usually considered with respect to time. Disease present at one particular time is measured by a prevalence rate. Disease developing over a period of time is measured by an incidence rate.

Comparing disease rates among different groups is of primary importance. These comparisons are often expressed as differences between rates or as ratios of one rate to another.

Quantitative attributes are also important. It is often necessary to consider the entire distribution of the quantitative measure in a group. However, this distribution may be described in a summary fashion by such measures as the mean and standard deviation. Breaking the group into equal parts according to ranking on a quantitative scale (quantiles) serves many useful purposes.

Obviously, the measurements described in this chapter do not exhaust the repertory of the epidemiologist. Other measurements have been used, and new ones will be invented for specific purposes. The simple measures described are established, time-tested, and widely understood.

REFERENCES

Hammond, E. C. Smoking in relation to death rates of one million men and women, *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases*, edited by W. Haenszel, National Cancer Institute Monograph 19. U.S. Department of Health, Education, and Welfare, Bethesda, January 1966, pp. 127–204.

Ipsen, J., and P. Feigl, *Bancroft's Introduction to Biostatistics*, 2d ed. (New York: Harper and Row, 1970), Chaps. 2, 4, 5 and 11.

Kagan, A., W. B. Kannel, T. R. Dawber, and N. Revotskie. 1963. The coronary profile. *Ann. N.Y. Acad. Sci.*, **97**:883–894.

Kannel, W. B., T. R. Dawber, G. D. Friedman, W. E. Glennon, and P. M. McNamara. 1964. Risk factors in coronary heart disease: An evaluation of several serum lipids as predictors of coronary heart disease: the Framingham study. *Ann. Intern. Med.*, **61**:888–899.

Siegel, M., Indices of community health, in *Preventive Medicine* edited by D. Clark and B. MacMahon. (Boston: Little, Brown, 1967), pp. 67–79.

Chapter 3

Observations Used in Epidemiology

A wide variety of observations and measurements have been used by epidemiologists in their efforts to *describe* and *explain* the occurrence of disease in human populations. There are so many factors that influence human health and disease that almost any aspect of persons and their environments may be fair game for study. Depending upon what is being explored, epidemiologic studies may require the collaboration of scientists from other medical specialties and a variety of other disciplines. Ophthalmology, psychology, physical anthropology, bacteriology, and meteorology are just a few examples.

While we need not consider all varieties of data that may be used, certain types of observations recur frequently enough to deserve discussion. Health-care professionals must have some appreciation of the nature and limitations of these data sources. Not only are they used in scientific study, but they also provide the basis for vital decisions in day-to-day patient care.