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.

Measures of Data Quality: Validity and Reliability

Observations or measurements, whether made by man or machine, involve some degree of error. Errors affect two important aspects of data quality—*validity* and *reliability.*

Validity Validity, or accuracy, is a measure of how closely the observations correspond to the actual state of affairs. As a clinical illustration, consider a patient with a rapid irregular heartbeat due to atrial fibrillation. Measurement of his heart rate by the radial pulse is considered inaccurate or lacking in validity because some heart beats produce a pulse too weak to be felt at the wrist. Compared to the true heart rate the radial pulse rate is *biased* toward lower values, resulting in what is commonly known as a "pulse deficit."

Reliability Reliability or reproducibility is a measure of how closely a series of observations of exactly the same thing match one another. If the cholesterol concentration of two portions of the same serum specimen is measured in an automated chemical analyzer, the two results should ideally be exactly the same. To the extent that they are not, the analyzer is said to lack reliability.

Effects of Lack of Validity and Reliability

Observations may be highly reliable but invalid. The cholesterol concentration on duplicate specimens may always agree within 5 mg/100 cc. Yet the readings may consistently be about 30 mg/100 cc too high.

This lack of validity does not necessarily rule out the use of the data. In some instances, knowing a person's absolute level of cholesterol may not be as important as knowing how that person ranks in his group. If all the group's values are 30 mg/100 cc too high, each person in the group will still be properly ranked in relation to the others. However, if one wishes to compare the mean cholesterol for that entire group with the mean of another group, for whom serum cholesterol has been measured accurately, the comparison will be unfair or biased.

Now consider the effects of unreliability. If a group of observations is unreliable, most will also be invalid due to departures from

the true values. However, if the unreliability is due to fluctuations that center around the true value, then the average or mean of a large series of observations may be quite a valid measure of the true average or mean. In this case many individuals will be improperly ranked relative to one another, if the ranking is based on one measurement for each. However, a comparison of the mean cholesterol of one large group with that of another may be quite fair and unbiased.

Usual Sources of Variation in Measurements

Not all the fluctuations in measurements or observations are attributable to lack of validity or reliability. The attributes themselves usually vary in a variety of ways.

Consider the distribution of blood pressures found in a community survey in which each subject has two measurements made. The major components of variation in the distribution are as follows:

Differences among subgroups—e.g., blacks have higher blood pressures, on the average, than whites; older persons have higher blood pressures than younger ones.

Differences among individuals within a subgroup—e.g., among black men, age 50, some individuals have higher blood pressures than others.

Differences within each individual—due to a variety of influences, each individual's blood pressure varies from one moment to the next. Some of these intraindividual differences may follow a regular pattern, e.g., diurnal variation.

Measurement errors—even if all blood pressures measured were exactly the same, they would appear to vary because of the failings of the observer, be it human or a mechanical device.

Sampling Variation

Another source of error or variation in data, known as *sampling variation*, is due to chance. It can be overcome by studying groups that are sufficiently large.

When we study the occurrence of a disease in a group of men, aged 50–59, in a community, we would like to think that our findings are applicable to all men of that age decade in that community. The

findings would undoubtedly be true of all 50-to-59-year-old men in the community if we studied all of them, but we usually have to take a sample. If the sample is selected in such a way that all men have an equal chance of being chosen, then we have what is called a *random sample*.

Experience and the laws of probability tell us that the larger the sample that is studied, the more likely are the findings to be representative of the total population. Conversely, the smaller the sample, the more likely we are to be misled. If repeated samples are drawn from a population, the findings in each sample will differ from one another—thus the term, "sampling variation." The larger the sample size, the less the variation, and the less chance of error.

This fact may be readily seen in the classic example of a large bag full of an equal mixture of black and white marbles. If an observer tries to determine what proportion of the marbles are white by pulling out only two marbles, he has a 25 percent probability of picking out two white marbles and concluding erroneously that all the marbles are white. If he pulls out four marbles instead, his chances of getting all white marbles are much less, only 1 in 16, or about 6 percent. One may apply the laws of probability to compute the likelihood of this false conclusion with any size sample; the result corresponds with our intuitive feeling that the more marbles one looks at, the less the chance of concluding that those in the bag are all white.

Thus, the larger the sample or group studied, the less the probability that chance error may occur. Statistical significance tests (such as "*t*" or chi square tests, and a variety of others described in statistics texts) are used to measure the probability of chance errors, given the size and characteristics of the study population and the question that is being asked. The result of a test of statistical significance is a probability level or "*p*" value, as frequently seen in medical journal articles. The expression "$p < 0.05$" means that there is less than a 5 percent probability that the observed result could have occurred by chance error.

Clinical Observations

Clinical observations are the primary basis for decisions as to the presence or absence of a particular disease. The most basic clinical

observations constitute the clinical history and physical examination. These are usually obtained by physicians, nurses, and other specially trained physicians' assistants.

The means for obtaining a history and physical examination need not be described here, but some comment about their limitations is in order. Many physicians have had memorable experiences in the unreliability of the medical history interview when they were medical students. Consider this all-too-familiar example. In preparation for rounds with the professor of cardiology the student devotes 10 minutes to careful questioning of the patient concerning nocturnal dyspnea and convinces himself that the patient indeed becomes short of breath at night and must sit up in bed in order to breathe more easily. After presenting the history during rounds the next day, the embarrassed student hears the patient tell the professor that he has never been short of breath at night.

The physical examination is no more reliable. If the patient is examined by half a dozen physicians, there will often be one or two who will hear (or not hear) a faint diastolic murmur not heard (or heard) by the others. The same degree of disagreement may be expected concerning the palpability of an elusive spleen. Differences in observer skill cannot be denied. Yet the murmur-hearers and spleen-feelers hold the psychological advantage, and objectivity probably suffers as a result.

Blood pressure, measured with a sphygmomanometer, has been a convenient measurement for the study of observer error in clinical medicine. It is a very sobering experience to be among a group viewing a movie prepared by Wilcox (1961), which shows a series of 14 views of a descending column of mercury in a sphygmomanometer accompanied by Korotkov's sounds amplified from a stethoscope. The group is asked to record the systolic and diastolic pressure for each measurement displayed. Even though all observers are seeing the same column of mercury and hearing the same sounds, the differences in the recorded results are striking. The greatest surprise comes when the viewers, learning that some of the early and late scenes are exactly the same, find discrepancies in their own readings for duplicate measurements.

When the results of a series of blood pressure measurements are tabulated, one human source of error that usually comes to light is *digit preference*. Physicians may tend to record values rounded off

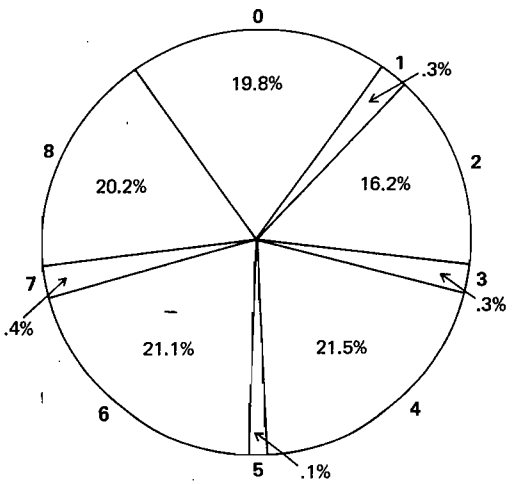
to a last digit of 5 or 0, or a preference for even over odd numbers becomes apparent (Fig. 3-1). Also noted has been a tendency to slant borderline values downward to avoid making unpleasant diagnoses.

Observations of Medical Specialists

Physicians in certain medical specialties make particular observations that are supposed to provide highly objective evidence as to the presence or absence of disease. Radiologists have the x-ray, cardiologists have the electrocardiogram, and pathologists have their stained microscopic sections. Implicit in giving a pathologist the last word in a clinicopathologic conference is perhaps the feeling that his observations will not only shed additional light on difficult problems, but that they are more reliable and valid than those of a bedside clinician.

A few of these specialists have made important contributions to our knowledge of the extent of observer variation in medicine. They

Figure 3-1 Percentage distribution of terminal digits on both systolic and diastolic blood pressure readings by an examining physician in the Los Angeles Heart Study. *(Reproduced, by permission, from Chapman, Clark, and Coulson, 1966.)*



have had the interest and courage to participate in studies to compare observations of the same visual object by different members of the same specialty or to compare duplicate observations by the same individual. The lack of reliability, even in these so-called objective measurements, has been striking.

Perhaps the classic series of studies in this area was carried out by Yerushalmy (1969) and his associates in the field of radiology. In one such study 14,541 entering college students received 70-mm chest photofluorograms. Each film was interpreted twice by two physicians and once by six others. Follow-up study of students with films read as "positive" by more than one reader, was accomplished by 14- by 17-in. chest film interpreted by a group of radiologists. The final interpretation regarding the presence of pulmonary tuberculosis was that 177 students had films that were "roentgenologically positive," 61 were "roentgenologically urgent," and 13 were "clinically active." Each of these cases, of course, had initial films that had been read by eight different readers. The percentages of original readings that were falsely read as negative were as follows:

	False negatives
Roentgenologically positive	26.9%
Roentgenologically urgent	25.4%
Clinically active	25.0%

Thus about one-quarter of all these nontrivial cases were missed the first time by competent x-ray readers.

Another series of 1,256 14- by 17-in. films were interpreted by a group of five competent radiologists and tuberculosis specialists. The number of films read as positive for tuberculosis by each reader was 56, 59, 62, 70, and 109, respectively. "Moreover, the radiologist who selected 109 did not include all those selected by the one who selected only 56." Similarly each reader read a different number as being positive when he read the films a second time. In each case some of the films read as positive once were read as negative by the same reader on another occasion.

The presence or absence of significant disease was not the only subject of inter- and intraobserver disagreement. Commonly accepted descriptive terms for pulmonary lesions such as "active,"

"inactive," "fibrotic," "soft," "hard," and "cavity" showed great differences among readers. After 2 years of work in trying to develop a reliable classification scheme to describe pulmonary lesions, the group of radiologists concluded that they had failed. "It was disappointing to find that many conferences and much practice, together and apart failed to increase reliability and agreement to a useful degree."

Interpretation of serial roentgenograms, the basis for many clinical decisions about tuberculosis patients, was also found to be grossly inconsistent. In making a judgment as to whether two x-ray films taken at different times showed progression, regression, or stability of disease, two readers disagreed with each other in about one-third of cases and a single reader disagreed with himself in about one-fifth of cases.

Clinical Diagnoses

Diagnoses are inferences or conclusions based on clinical and laboratory observations. Not only may these observations be incorrect, but the reasoning leading to the conclusions may also be in error. Yet even if the observations are complete and accurate and the reasoning is sound, differently trained physicians use different criteria for making the same diagnosis. Also leading to observer variations is the fact that one physician may have access to more laboratory tests or other specialized data. Furthermore, different terms may be used to refer to the same clinical condition, and a single term (e.g., "arteriosclerotic heart disease") may have different meanings to different physicians.

Thus, clinical diagnoses by themselves are indeed undependable indicators of disease for scientific study. Whenever possible, specific criteria should be established for making diagnoses. These criteria should be adhered to carefully and described clearly so that the work may be repeated or evaluated by others.

Medical Chart Review

Both epidemiologic studies and patient care frequently rely upon the review and abstracting of information from medical records. Just as

is found for other types of observations, the reading of charts involves substantial amounts of error. Even if the information is relatively complete and the various handwritings are legible, two chart readers will extract differing information. Usually, however, matters are much worse, with missing information and cryptic or illegible physicians' notes.

Disease Reporting

Physicians are legally required to report certain diseases to the local public health authorities at the time of diagnosis. The primary purpose of this is to detect the onset of epidemics of certain serious diseases and to provide information so that appropriate community-wide control measures can be undertaken. In addition to their usefulness in disease control these data may also be used to measure disease incidence in the community.

Despite official requirements, many diseases are underreported. For example it has been estimated that, despite the mounting concern over the recent epidemic of venereal disease, only one-fourth of all cases are reported. Desire to avoid social stigma for patients, the pressures of other work, and laxity are among the reasons that have been given for underreporting.

This is not only the case with regard to certain infectious diseases. In the 1950's and 1960's two large agencies, the American Medical Association and the U.S. Food and Drug Administration carried out special programs to encourage physicians to report instances of suspected adverse drug reactions. The purposes of these programs were to obtain some measure of the frequency of occurrence of various drug reactions and to provide a means of receiving early warnings of as yet unsuspected side effects of drugs. The response of physicians was quite disappointing. By and large, busy physicians do not wish to take the time to fill out the reporting forms. In a study of various approaches to detecting adverse drug reactions in a hospital, Cluff et al. (1964) judged a system whereby physicians were to fill out a drug reaction card at the time of discharge to be "completely unsatisfactory," since intensive daily surveillance of just one service yielded four times as many reactions as were listed by report card from the entire hospital.

Death Certificates and Mortality Statistics

Mortality data for nations, states, and communities, as obtained from death certificates, have played an important role in epidemiologic research for more than a century. Many major problems and inaccuracies are associated with death certificates. (See Feinstein, 1968, for detailed discussion.) Nevertheless, they constitute a widely implemented collection of data about fatal illnesses that can be used to study disease occurrence on a local, national, or international scale.

Death certificate diagnoses are usually clinical diagnoses and are thus subject to all the vagaries described above. In addition, the patient may have had several diseases contributing to his death, but under current procedures, only one underlying cause is to be selected. Before 1949 in the United States, coding rules automatically led to the choice of one underlying cause out of several possibilities. For example, if both diabetes mellitus and heart disease were listed on the death certificate, diabetes was coded as the underlying cause even if the doctor felt that heart disease was more to blame. Starting in 1949, the physician was asked to indicate the underlying cause. While this may have been an improvement, it resulted in some sudden changes in apparent mortality rates (e.g., a drop in diabetes mortality, as would be expected); it also forced physicians to oversimplify many complex situations where multiple causes might have been involved. For this reason, many authorities have urged the adoption of a multiple-cause coding system for death certificates. If mortality statistics are to become more meaningful, it would be helpful if physicians were trained in uniform and proper procedures for filling out death certificates.

Other changes in diagnostic classifications have been made in the *International Classification of Diseases*, now in its eighth revision, leading to abrupt changes in reported mortality rates for the diseases affected. Studies of time trends in disease mortality must take into account these coding changes as well as the technological advances that lead to increased diagnoses of particular conditions and changes in the fashion of allocating deaths to one disease instead of another. Fig. 3-2, from a study by Reid and Evans (1970), shows time trends in mortality rates for nephritis, hypertension, and

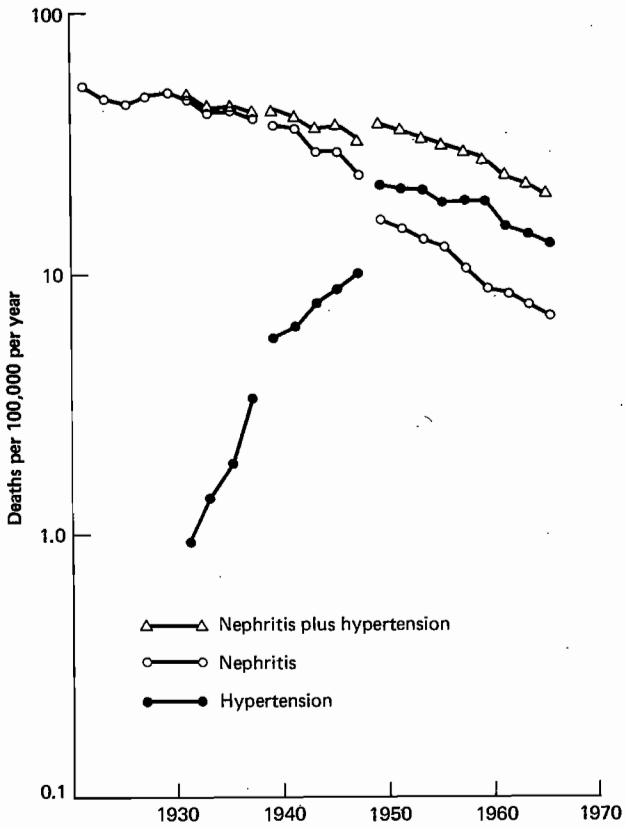


Figure 3-2 Mortality rates from nephritis, hypertension, and both combined, among males aged 45–54 years in England and Wales, from 1931–1966. *(Reproduced, by permission, from Reid and Evans, 1970.)*

both combined, among men ages 45–54 in England and Wales, and illustrates several of these factors. The gaps in the curves reflect changes in disease classification. The sharp rise in the death rate for hypertension between 1931 and 1950 probably reflects the increased use of the sphygmomanometer and an increased awareness of the importance of hypertension. The reciprocal changes in hypertension and nephritis deaths may represent an increasing tendency to

attribute uremia to kidney damage produced by hypertension rather than by inflammation.

Responses to Questionnaires

The clinical history is only one of many kinds of data that may be obtained by questionnaires. Data relating to social status or exposures to environmental hazards can also be obtained in this manner.

It does not seem necessary to belabor the frailties of human observations and their written or oral communications any further, except to encourage again a reasonably skeptical attitude toward the results of questionnaire studies and to show some examples of problems commonly encountered.

Nonresponse If given a choice, a substantial proportion of individuals will not answer questions. In 1971 a questionnaire was mailed to 8,250 Kaiser Foundation Health Plan members participating in a study to evaluate periodic checkups involving multiphasic screening. Because of the incomplete response to one mailing, four subsequent mailings were sent out to nonrespondents. The percentage of the total group responding to each mailing is shown in Table 3-1. The final nonresponse rate of 20.4 percent (100 percent minus 79.6 percent) is not at all unusual for a mailed questionnaire.

Nonresponse can also occur under more controlled or supervised conditions. As part of a multiphasic examination at Kaiser-Permanente, patients are given a self-administered questionnaire containing a series of questions about their smoking habits. The answers to the questions about smoking were used to classify examinees in a study of the characteristics of smokers and nonsmokers (Friedman et al., 1972). In doing so it was found that about 12.7 percent of 111,024 persons did not answer at least one of the crucial questions about present or past smoking habits.

Nonresponse would not constitute a serious problem if it merely reduced the number of subjects available for study; however, it may also lead to a biased study sample if the respondents and nonrespondents differ with respect to health or some other characteristic being studied. Unfortunately, this is frequently the case.

Table 3-1 Response to Five Mailings of a Questionnaire by 8,250 Kaiser Health Plan Members*

Mailing	Percentage of total study group responding
First	43.4
Second	15.4
Third	8.6
Fourth	7.0
Fifth	5.2
Total	79.6

*Data tabulated by Barbara A. Campbell, M.A.

Inconsistent or Otherwise Unusable Responses It is surprising how often persons will answer both "yes" and "no" to the same questionnaire item or provide otherwise inconsistent responses. In the study of smoking just referred to 2.3 percent of subjects did not indicate that they smoked cigarettes, but then gave a positive response to some duration of smoking or quantity of cigarettes smoked. Because of this and other serious inconsistencies, plus the omissions described above, 16.5 percent, or about one-sixth of the total subjects, had to be eliminated.

Overreporting of Disease Symptoms Patients who either deny or exaggerate disease symptoms are well known to physicians. In a study of the reliability of a self-administered questionnaire (Collen et al., 1969) it was found that on the average, one-fifth of persons who answered "yes" to a symptom question the first time, denied the symptom when the questionnaire was administered again at the same examination. Physicians who perform follow-up examinations after patients have answered a symptom questionnaire often find that positive responses to questions about serious symptoms either cannot be substantiated or appear nonsignificant upon careful history-taking. As an example of the likely overreporting of symptoms on a self-administered questionnaire, 15.9 percent of

1,950 girls, ages 15–19, taking Kaiser-Permanente multiphasic examinations, answered "yes" to the following question describing symptoms almost pathognomonic of angina pectoris: "In the past year have you had repeated pain (or pressure or tight feeling) in your chest when you walked fast or uphill and that left after a few minutes rest?"

Presenting Oneself in a Favorable Light This is such a universal trait that it hardly needs to be mentioned except that it can introduce systematic biases into epidemiologic studies. Persons tend to deny venereal disease and drug abuse and underestimate their alcohol consumption.

Household Surveys

Information about medical conditions and other pertinent social and personal characteristics is frequently obtained by household interview survey. Assessments of health and social problems by survey may be the basis of determining priorities for community or national policy. Thus, the limitations of this study method should be well understood.

Problems associated with survey data and the techniques of obtaining representative samples of individuals for questioning have been a major concern of social scientists. In the health field, the *National Health Survey* (NHS) has been authorized by the United States Congress to carry out surveys by the household interview method since 1957. In the course of this work NHS scientists have carried out important methodologic studies to determine the accuracy of interview-acquired information.

In one such study (Madow, 1967), patients' reporting of chronic conditions was compared to the chronic conditions recorded by physicians in their medical records during a 1-year period. Overall, 45.3 percent or almost half the chronic conditions recorded by the physicians were not reported by the patients despite the fact that patients were given a fairly comprehensive checklist of conditions to jog their memories.

Thus, interview data about illness are apt to be incomplete. As might be expected, conditions for which the patient made more

frequent doctor visits were more apt to be reported, as were those for which a doctor was seen more recently. Furthermore, conditions were more likely to be reported in an interview if they affected the person's way of life, for example, by causing pain or worry or limitations in his work or in what he could eat or drink.

Laboratory Data

Mechanical, electrical, and chemical measurements are also subject to error. Well-run clinical laboratories maintain continuing quality control programs to monitor the validity and reliability of their measurements. When significant errors occur, monitoring permits institution of prompt corrective action.

Yet, even with the most careful quality control, significant errors occur, due both to known and unknown factors. Many of these factors cannot be controlled within the laboratory. Only in the past decade, for example, did it become generally known that exposure of a blood specimen to light would cause a breakdown of bilirubin and significantly lower the serum bilirubin concentration measured in the laboratory. Similarly, ingestion of a variety of drugs can affect the measurement of important blood constituents. A well-known example is the effect of iodide-containing drugs on the protein-bound iodine test of thyroid function.

STUDYING RELATIONSHIPS IN IMPERFECT DATA: THE VALUE OF INVESTIGATING LARGE GROUPS

This section is, to the author, one of the most important in this book. It will attempt to bridge a serious gap in understanding and communication between the scientifically minded clinician and the epidemiologist.

As will be developed more fully in the next chapter, one of the primary concerns of the epidemiologist, like other scientists, is the study of *relationships*. The epidemiologist focuses on relationships between diseases and other human or environmental attributes by studying population groups.

The clinician focuses on the individual patient and strives to obtain complete and accurate information, in order to provide the

best possible diagnosis and treatment. In his appropriate concern for the patient's welfare, he can tolerate few avoidable errors in this information. Accustomed to high standards in his pursuit of information and the expenditure, if necessary, of hundreds of dollars per patient in laboratory tests and specialized diagnostic procedures, he becomes intolerant of the use of relatively low-quality data such as questionnaires or death certificates in epidemiologic studies.

A case in point is the difficulty in convincing some neurologists of the validity of epidemiologic studies of stroke that do not include an evaluation of all study subjects by a neurologist. Neurologists spend years learning the subtleties of the neurological examination and the fine points of differentiating strokes from a variety of other neurological conditions (many of which are quite rare). To many physicians with such a background it is inconceivable that one would undertake a scientific study of stroke based, say, on identification of cases simply by asking, "Have you ever had a stroke?"

Yet in a study of a large population, the human and financial resources to provide a neurologist's examination for all subjects are not available now, nor will they be in the foreseeable future. So let's compromise and have any ill persons in whom the attending physician suspects a stroke evaluated by a neurologist. This approach is more workable and can be employed in special intensive population studies such as the Framingham Study (described in Chap. 8). Yet even there, practical difficulties arise; if a person has a stroke which is rapidly fatal or which occurs out of town, he will probably not be seen by a neurologist.

The epidemiologist is not in favor of bad data. He wants the best he can get. But experience has shown that he can discern important relationships, even in data of relatively poor quality because studying large groups provides power to overcome error. With some validity to the data and large enough numbers of study subjects to minimize sampling error, one may still derive some valuable information from poor quality data.

Consider the following numerical example. Suppose that we wish to determine whether there is a relationship of stroke to hypertension and we can only use a questionnaire which asks, "Have you ever had a stroke?" and "Have you ever had high blood pressure?" The questionnaire is administered to 10,000 persons,

ages 65–74. Let us postulate that the true state of affairs for this population happens to be that 200 persons have had a stroke and 2,000 have had high blood pressure. Of the stroke cases, 150 had high blood pressure. The *true* population breakdown is shown in Table 3-2.

A slight digression here may be of value to the reader who is unfamiliar with the presentation of data in a "two-by-two" or "fourfold" table, frequently used in epidemiology and exemplified by Table 3-2. These tables show the relationship of one "yes-or-no," or dichotomous, variable to another. The presence or absence of one disease or characteristic is indicated at the left and the presence or absence of the other is shown at the top.

Table 3-2 shows how the population is divided in the four possible ways according to each of the two characteristics. The number, 150, in the upper left corner indicates that there are 150 persons with both a history of stroke and of hypertension. The number, 1,850, to the right of the 150, represents 1,850 persons with a history of hypertension but no history of stroke. The sum of 150 and 1,850, or 2,000, is shown at the far upper right and represents all persons with a history of hypertension. The 8,000 persons without a history of hypertension are shown in the second row. Fifty of the 8,000, on the left, have a history of stroke. The 7,950, next to them, do not have a history of stroke. The total of 50 plus 7,950, or 8,000, is shown to the right. Totals of the columns are shown below and represent the 200 persons with a history of stroke and the 9,800

Table 3-2 "True" Breakdown of a Population of 10,000 Persons, Ages 65–74, According to the Presence or Absence of a History of Hypertension and a History of Stroke (Fictitious Data)

		Stroke history (True)		
		Present	Absent	Total
Hypertension history (True)	Present	150	1,850	2,000
	Absent	50	7,950	8,000
	Total	200	9,800	10,000

without. The grand total of the population, or 10,000, is shown at the lower right-hand corner.

Returning now to the argument at hand, the prevalence of a history of stroke in those with a history of high blood pressure is $^{150}/_{2,000}$ or 7.5 percent. The prevalence of a stroke history in those without a hypertension history is $^{50}/_{8,000}$ or 0.625 percent. Thus, if one could only know the true situation, one would find that those with high blood pressure in the past had $^{7.5}/_{0.625}$, or 12 times, the likelihood of the nonhypertensives, of having a history of stroke.

Now let us estimate that our questionnaire only elicits positive responses to the stroke question from 160, or four-fifths, of the stroke cases and, in addition, 196, or 2 percent, of the 9,800 nonstroke cases answered "yes" to the stroke question by mistake. Let us also assume that only one-half of hypertensives were aware of, and reported, their elevated blood pressure and that 5 percent of nonhypertensives erroneously reported that they were hypertensive.

As a result of these errors, some of the persons from each "true" category will be moved to each of the four "reported" categories. For example, consider the 150 persons with *true* strokes and *true* hypertension. Only half report their hypertension. Of the 75 reporting either hypertension or nonhypertension one-fifth do not report their stroke. So the 150 "true" stroke cases with hypertension will be distributed into the four "reported" categories as shown in Table 3-3.

Table 3-3 Parceling Out the 150 Persons with a "True" History of Both Stroke and Hypertension into Four Categories According to What They Will Report on the Questionnaire (Fictitious data)

		Stroke history (reported)	
		Present	Absent
Hypertension history (reported)	Present	60	15
	Absent	60	15

One may go through this exercise with each of the other three "true" categories and divide each into the four "reported" categories. If one then adds all the persons in each of the "reported" categories, the (rounded) result is as shown in Table 3-4.

Now the observed prevalence of a history of stroke in prior hypertensives is $^{88}/_{1,400}$, or 6.3 percent. This is about twice the 3.1 percent prevalence ($^{268}/_{8,600}$) in prior normotensives. *Despite the poor quality of the data, the relationship between hypertension and stroke, while not as strong as in reality, may still be perceived.* Thus, the study of relationships in groups of people can, to some degree, overcome certain kinds of error.

This is not an argument for using poor data when better are obtainable. One must always be aware of the limitations of his data and how inaccuracies and biases may affect his results. In the example it was assumed that the failure to report hypertension was equally true of persons with and without stroke. If stroke affected memory so as to further diminish the reporting of hypertension in the stroke case group, then the study might have missed the stroke-hypertension relationship completely, or might even have led to the opposite conclusion. Thus, data can be, and often are so bad as to be unrevealing or even misleading, despite large numbers.

The example given illustrates another epidemiologic principle. Where relationships are observed in data with an appreciable number of misclassified subjects (e.g., persons with a disease classified as not having it), the results are conservative. That is, the

Table 3-4 Findings in the Total Population Based upon What They Report on the Questionnaire (Fictitious Data)

		Stroke history (reported)		
		Present	Absent	Total
Hypertension history (reported)	Present	88	1,312	1,400
	Absent	268	8,332	8,600
	Total	356	9,644	10,000

relationship in real life is greater than is revealed by the data. In the above example the misclassifications of patients regarding their blood pressure or stroke status reduced an actual twelvefold increase of stroke in hypertensives to an observed twofold increase.

Nevertheless, the study of large groups allows one to detect important relationships, using poor data that are intolerable in conscientious patient care. This, then, is the explanation to the clinician of the seeming tolerance of epidemiology for inadequate data.

REFERENCES

Chapman, J. M., V. A. Clark, and A. H. Coulson. 1966. Problems of measurement in blood pressure surveys: Inter-observer differences in blood pressure determinations. *Am. J. Epidemiology*, **84**:483–494.

Cluff, L. E., G. F. Thornton, and L. G. Seidl. 1964. Studies on the epidemiology of adverse drug reactions: I. Methods of surveillance. *J. Am. Med. Assoc.*, **188**:976–983.

Collen, M. F., J. L. Cutler, A. B. Siegelaub, R. L. Cella. 1969. Reliability of a self-administered questionnaire. *Arch. Intern. Med.*, **123**:664–681.

Feinstein, A. R. 1968. Clinical epidemiology. II. The identification rates of disease. *Ann. Intern. Med.*, **69**:1037–1061.

Friedman, G. D., C. C. Seltzer, A. B. Siegelaub, R. Feldman, and M. F. Collen. 1972. Smoking among white, black and yellow men and women: Kaiser-Permanente Multiphasic Health examination data, 1964–1968. *Am. J. Epidemiology*, **96**:23–35.

Madow, W. G., Interview data on chronic conditions compared with information derived from medical records. National Center for Health Statistics Report, Ser. 2, No. 23, U.S. Department of Health, Education, and Welfare, 1967.

Reid, D. D., and J. G. Evans. 1970. New drugs and changing mortality from non-infectious disease in England and Wales. *Brit. Med. Bull.*, **26**:191–196.

Wilcox, J. 1961. Observer factors in the measurement of blood pressure. *Nursing Research*, **10**:4–20.

Yerushalmy, J. 1969. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. *Radiol. Clin. N. Amer.*, **7**:381–392.

Chapter 4

Basic Methods of Study

In the two preceding chapters the reader has been introduced to the data employed in epidemiology and the basic measurements that are used to describe groups of persons. It is now appropriate to consider the major types of epidemiological investigation. Each type of study uses these tools in a particular way and has a unique logical framework. In addition, each type of study is especially appropriate for the unique circumstances surrounding any particular investigation—the aims of the investigation, the populations available for study, and the human and financial resources that can be brought to bear on the problem.

Relationships

Much of the effort of medical scientists in understanding the etiology of disease and developing appropriate therapies involves a study of