Chapter 10

Clinical Studies of Disease Outcome

Just as some questions relating to disease occurrence and disease etiology are best answered by studying population groups, clinical problems often require the study of *groups* of patients. Many methods for studying patient groups are similar to the epidemiologic methods for studying populations, discussed in previous chapters.

The process by which healthy people become sick and the factors that determine who will become sick and who will stay healthy are the primary concern of epidemiology. Many clinical studies, on the other hand, aim at sick people and try to identify the factors that determine what the outcome of illness will be. This difference in focus between the two types of studies is illustrated in Fig. 10-1. Note that illness or disease can have several outcomes, including recovery, improvement, no change, worsening, complications, disability, and death.

The ultimate goal of epidemiology is to learn how to prevent

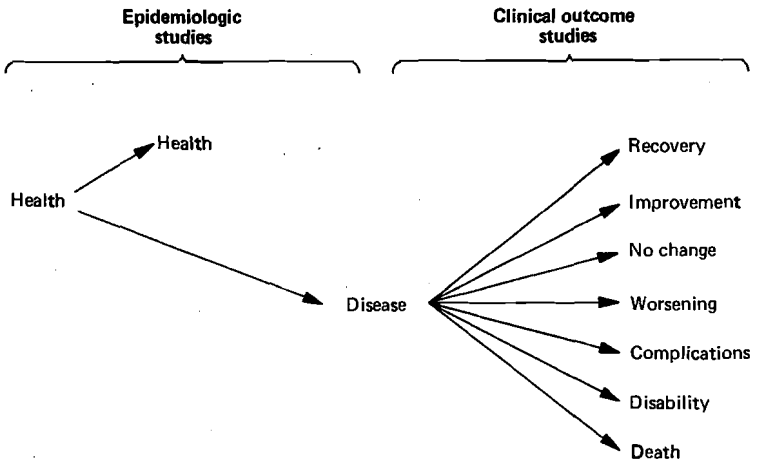


Figure 10-1 Areas of concern of epidemiologic studies and clinical outcome studies.

disease. The ultimate goal of clinical studies is to learn how to cure or successfully treat disease.

The purpose of this chapter is to demonstrate some of the parallels between clinical studies of disease outcome and epidemiologic studies and to describe the analytic methods commonly used to measure disease outcome.

Natural History of Disease

Studies of the *natural history of disease* are analogous to descriptive studies in epidemiology. The outcomes of a particular disease are observed and the proportions of the affected patients developing each outcome are measured. This information is the basis of *prognosis*, that is, predicting a patient's future. As in descriptive epidemiologic studies, disease outcomes are generally determined for major subgroups of patients such as males versus females, various age groups, and so on.

A good example of a study of the natural history of disease is Bland and Jones' (1951) 20-year study of 1,000 children and adoles-

cents with rheumatic fever or chorea. These patients, initially hospitalized at the House of the Good Samaritan in Boston, were carefully followed up into adulthood. Among the findings were that 65 percent of the children had signs of rheumatic heart disease when they recovered from their acute illness, but 16 percent of those with such signs had no evidence of heart disease 20 years later. On the other hand, 44 percent of those without apparent heart disease initially had valvular disease when they were examined as adults. Also described were the recurrence rates of acute rheumatic fever, the evolution of murmurs, and the frequency of deaths and other sequelae of the disease.

Analytic Studies

The clinical investigator usually wishes to go beyond general descriptions of prognosis and to determine what factors lead to improvement, worsening, death, and other outcomes. Such factors include patient characteristics and environmental influences. One of the main environmental factors that is investigated is, of course, therapy.

Analytic clinical investigations of prognostic factors may be carried out in a fashion quite analogous to prevalence, case-control, and incidence studies in epidemiology. A physician is conducting what amounts to an informal prevalence study when he makes rounds on two wards caring for paralyzed stroke patients and notices that in one ward, several patients have decubitus ulcers (bedsores) and on the other, the patients are ulcer-free. He will probably conclude that being on the first ward is conducive to the development of this complication of paralytic stroke and will make some appropriate comments to the nursing staff.

Analytic studies of factors affecting prognosis are usually similar to incidence studies. That is, attributes of the group of patients are assessed early in the course of the illness. Then, the patients are followed up to determine outcome.

The clinical investigator can adopt this prospective follow-up approach much more readily than can the epidemiologist. The rates of development of many disease outcomes are relatively high, compared to the incidence of most diseases in a population. Thus, a

relatively small patient group can be observed in one clinic or hospital until the various outcomes are noted.

Consider, for example, the follow-up study by Stahlman et al. (1967) to determine characteristics predicting the outcome of hyaline membrane disease in the newborn. Of 115 affected newborns studied, 33, or 29 percent, died in the neonatal period. A number of measurements taken within 12 hours of birth, such as arterial-blood oxygen tension, birth weight, and respiratory rate, all proved to be related to mortality, and statistical-significance tests showed that these relationships could not reasonably be attributed to chance. Thus, the predictive value of these measurements was demonstrable in this study of only several dozen patients.

Some analytic follow-up studies of prognosis deal with events that develop relatively slowly and infrequently, so that large numbers of patients must be followed for years. This is particularly true of chronic diseases. The Health Insurance Plan of Greater New York (HIP) has been investigating the prognosis of patients with angina pectoris and myocardial infarction. One such study demonstrated a relationship of blood pressure in these patients to the probability of subsequent myocardial infarction and cardiac death—the higher the blood pressure, the worse the prognosis. This study was based on 275 cases of angina pectoris and 881 cases of a first myocardial infarction found among 55,000 men during a 4-year case-finding period. The cases were followed up for 4.5 years (Frank et al., 1972).

When an analytic follow-up study cannot be carried out, it may be practical to use an approach analogous to the case-control method in epidemiology. That is, a group of patients with one particular outcome may be compared with a group showing another outcome, to see whether the two groups differ in any characteristic that might have affected or predicted the outcome. An example is Ellenberg's (1971) study of sexual impotence complicating diabetes mellitus. Forty-five impotent diabetic men ("cases") were compared with thirty male diabetics who were not impotent ("controls"). The potent diabetics were selected to match the impotent group with respect to age distribution and duration of diabetes. The striking difference between the two groups was in the percentage showing evidence of neuropathy affecting the autonomic system—82 percent of the impotent versus 10 percent of the potent. Thus it could be

concluded that most cases of impotence in diabetics were due to diabetic neuropathy rather than endocrine or other abnormalities.

Therapeutic Trials

The therapeutic trial is an experiment as applied to clinical medicine. In it, a drug, a surgical operation, or other therapy is applied to patients and the outcome is compared with that observed in a suitable control group.

It is essential that alternative therapies be evaluated in a well-controlled fashion using, whenever possible, the techniques of random allocation and blind assignment and assessment described in Chap. 9. The influence of the therapist's personality and the placebo effect (or tendency of patients to respond favorably even when a drug has no active ingredients) are potent determinants of outcome and should not be allowed to bias the experiment. Furthermore, because of wide variations in the way individual patients respond to treatment, large groups of patients are often required. Large groups will help ensure that an observed relationship between a treatment and an outcome is not due to chance and that the relationship has some general applicability.

The value of large patient series is apt to be forgotten by clinicians working with patients on an individual basis. A physician's use or avoidance of a particular therapy is often guided by his experience with a few patients. His view of the values or dangers of a particular treatment may be exaggerated just because, as luck would have it, the first two or three patients treated happened to respond unusually well or unusually poorly.

There is a widespread belief that the individual physician is the best judge of the value of a drug or other treatment. Through his knowledge of the patient, he may well be the best judge of what is most appropriate for that patient's particular problems. However, the average physician's limited experience with a few patients does not usually provide enough information to state a general principle or conclusion that one therapy is better than another. He may be able to detect dramatic effects such as the value of penicillin versus no antibiotic in treating lobar pneumococcal pneumonia. But conclusions as to less-striking differences between therapies should be

based on good-sized and representative series of patients with observations controlled as well as possible.

Medical history is full of examples of therapies which become accepted or popular in an epidemic of enthusiasm based on uncontrolled observations. Feeding this epidemic is the preference of authors and journals for reporting positive findings over negative findings. If the treatment is either not helpful or actually harmful, its use may eventually diminish or end, as its deficiencies become recognized. Unfortunately, during the period of general acceptance, withholding the treatment from some individuals, as is required in a well-controlled experiment, may be considered unethical. Thus it is important to perform a good therapeutic trial as early as possible after the therapy is developed.

Nevertheless, controlled trials are better carried out late than never. For example, the Boston Inter-Hospital Liver Group (BILG) recently completed a well-controlled therapeutic trial which failed to confirm the long-term value of a widely accepted surgical treatment (Resnick et al., 1969). Portacaval-shunt operations had been carried out as an elective prophylactic measure on patients with cirrhosis of the liver to relieve the excess pressure in esophageal varices and prevent serious bleeding episodes. Acceptance of the procedure by the medical profession was based on uncontrolled observations that cirrhotic patients who received this operation did better and lived longer than those who did not. What is often forgotten is that surgeons naturally prefer to operate on the relatively healthy or good-risk patients and reject the poor-risk patients as operative candidates.

In the BILG study, 93 cirrhotic patients with esophageal varices and no prior major bleeding episodes were randomly divided into a surgical and medical group. To avoid selection of the better-risk candidates for shunt surgery in this experiment, each patient was randomly assigned *after* the physicians and surgeons agreed that he or she was a candidate and *after* the patient had consented to have surgery. Both groups were followed up for several years.

The operation apparently did prevent bleeding episodes, as there were significantly more patients with subsequent hemorrhages in the medical group (12/45) than in the surgical group (1/48). However, the mortality of the surgical and medical patients was quite

similar. Although the surgical patients were less apt to die of bleeding, they were more apt to die of the hepatorenal syndrome. They were also more prone to develop hepatic encephalopathy.

Another recent controlled therapeutic trial did confirm the value of a much-used but still-debated treatment. For many years, even the individual practitioner could reliably observe that antihypertensive drug therapy brought about a dramatic improvement in the prognosis of severe and malignant hypertension. However the value of drugs for mild to moderate hypertension was less easy to recognize and, until quite recently, was subject to considerable debate. As a result, the Veterans Administration (1967, 1970) carried out a cooperative study in which 523 men with diastolic blood pressures of 90 to 129 mm Hg were assigned randomly to active drug therapy or placebo. Before random assignment there was a trial period during which the potentially uncooperative subjects—those who did not attend clinic regularly or take at least 90 percent of a marked placebo—could be eliminated. (Because most hypertensives feel well, there is little immediate gratification for them in following a regular therapeutic program.)

Therapeutic benefit to the drug-treated group was apparent after only 20 months of follow-up of those starting with diastolic levels of 115 to 129 mm Hg. Only 1 of 73 treated patients developed a major cardiovascular-renal complication, as compared to 27 of 70 control subjects, of whom 4 died. One other treated patient exhibited drug toxicity and had to be removed from the study therapy.

Longer follow-up of more subjects was required to demonstrate benefits of treating milder hypertension—90 to 114 mm Hg diastolic pressure. A total of 380 patients were followed up for an average of 3.3 years. Major complications were observed in 56 of 194 controls, as compared to only 22 of 186 treated subjects. Some complications, such as stroke, showed a markedly lower incidence among the treated group.

Concomitant with the reporting of controlled observations such as these has been a growing awareness that hypertension is serious, and that large numbers of persons in this country are hypertensive and not aware of it. Moreover, many persons who are aware of hypertension are not being treated adequately or consistently. Thus

the detection and sustained treatment of hypertension may become a major public health effort in the near future.

Commonly Used Measures of Disease Outcome

Rates Just as incidence rates are used in epidemiology to measure the development of disease in healthy persons, the outcomes of illness can be measured similarly in groups of sick persons. Thus one may speak of recovery rates, disability rates, death rates, and so on, referring to the proportion of the ill that recover, become disabled, or die per unit of time. Again, the proportion of the sick who manifest a particular outcome at one point in time is analogous to a point prevalence rate of disease in a general population.

Survival Measures of mortality outcome are often expressed in terms of *survival* rather than death. For comparative purposes, it is not particularly important whether one focuses on successes or failures. However, the data from clinical studies are so often analyzed and presented in terms of survival that it is desirable to be familiar with the approaches used. It should be remembered, also, that these measures need not be restricted to life and death. They can be applied to any mutually exclusive alternatives. Thus, in a study of the development of congestive heart failure in cardiac patients, remaining free of failure can be considered analogous to survival.

One of the most common measures of outcome is the proportion surviving for a particular duration. Any duration may be chosen—5 years is frequently used in studies of the surgical treatment of cancer, because for many types of cancer, if a patient survives for 5 years it is likely that he has been cured. Thus the "5-year survival rate" or "5-year cure rate" merely refers to that proportion of the original patient group still alive after 5 years of follow-up.

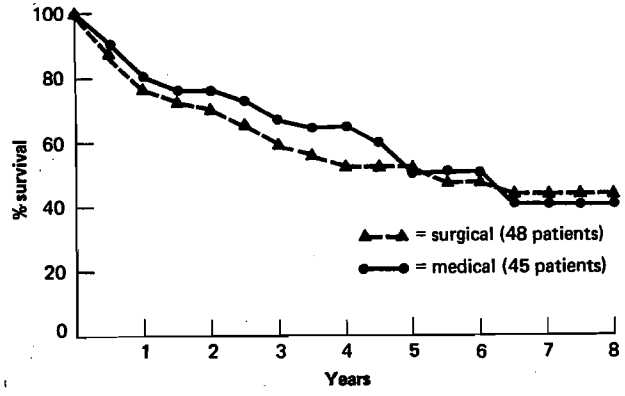
Another measure of survival that has been used is the "mean duration of survival." As mentioned in Chap. 2, page 19, the mean duration should be used for comparative purposes only when all

patients have died. When some are still living, it is preferable to compare *median duration of survival* or some other *quantile* of *survival durations* because once the stated percentage have died, their survival cannot change. For example, after 75 out of 100 patients have died, the survival duration of the seventy-fifth person becomes the 75th percentile of survival durations for the entire group. This cannot change no matter how much longer the other 25 live. The mean, on the other hand, is not finally determined until all 100 have died.

One of the most common and probably the most informative measures of survival is the survival curve. Starting initially at 100 percent, it shows the proportion still surviving at each subsequent point in time for as long as information is available. Fig. 10-2 shows the curves for the medical and surgical patients in the BILG study of portacaval shunt. The similarity in their survival experience is apparent.

Another graph, Fig. 10-3, shows marked differences in survival for several subgroups of patients with scleroderma, from the study by Medsger et al. (1971). The proportions of scleroderma patients surviving at the end of each year after entry into the study are shown by solid black circles. Those who had no involvement of their lungs,

Figure 10-2 Survival of surgical and medical patients in the Boston Inter-Hospital Liver Group's controlled therapeutic trial of portacaval shunt surgery for esophageal varices. *(Reproduced, by permission, from Resnick et al., 1969.)*



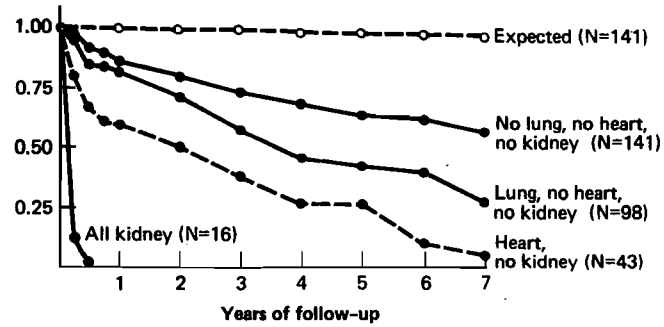


Figure 10-3 Survival of scleroderma patients according to organ involvement. Ordinate shows proportion surviving. *(Reproduced, by permission, from Medsger et al., 1971.)*

heart, or kidneys did the best, with 56 percent still alive after 7 years. Subgroups with poorer survival were next, those with lung involvement; then, those with heart involvement; and finally, those with kidney involvement, all of whom died within the first half year. For comparison, the expected survival curve is shown on top with clear circles. This is the survival that would have been expected for a group of this age, sex, and racial composition if the overall United States mortality rates for the study years had been applicable.

Construction of survival curves for a certain duration following a specific event or time does not require that all patients be observed for that entire duration. Consider an example in which persons are to be followed for 10 years starting at the time their disease was first diagnosed. The experience of a person who moves away and is lost to follow-up after 5 years is still useful in determining survival rates for the first 5 years. Similarly, someone who is diagnosed and enters the study 1 year before the date that follow-up observations are to be completed contributes to those persons observed during the first year after diagnosis.

Thus, all persons who are observed during each unit of time measured from the starting event can contribute their experience to the survival-rate computation for that time unit. The so-called *actuarial* or *life-table* method takes advantage of all these observations by computing survival rates for each time unit and combining these rates together into one composite survival curve. For details as

to methods, which are not difficult to carry out, see Berkson and Gage (1950), Cutler and Ederer (1958), or Hill (1971).

Importance of Starting Times When survival curves (or mortality rates) of two groups are to be compared, it is important that both have the same starting point. The starting time may be placed at the onset of symptoms, the first diagnosis, the beginning of therapy, discharge from a hospital, or some other landmark in the course of the disease.

Failure to follow this principle has led to many conflicting claims and erroneous conclusions as to benefits of therapy. For example, two equally good surgical treatments will appear to have different results if survival is measured from the hospital discharge date for one, and from the date of operation for the other. Measuring from date of discharge excludes operative and immediate postoperative mortality.

Although the inclusion or exclusion of operative mortality makes for an obvious error, more subtle and hard-to-recognize biases may result when follow-up of two groups does not begin at strictly comparable times. Consider a study to evaluate the efficacy of a new procedure for the early diagnosis of a disease. Even if detecting the disease early does not prolong life, it might appear to do so if survival is measured from the date of *early* diagnosis instead of from the usual diagnosis date resulting from traditional methods. Procedures for overcoming this bias are discussed by Feinleib and Zelen (1969).

Similarly, treatment measures for rapidly fatal diseases may appear more effective than they really are if they are initiated after a short delay. Part of the apparent improvement in in-hospital mortality from myocardial infarction, experienced by patients in coronary-care units, may be related to the fact that many heart attack victims die shortly after the onset of the attack. As noted by Kodlin, patients in coronary-care units have already survived the short delay between admission to the hospital and admission to the unit.

REFERENCES

Berkson, J., and R. P. Gage. 1950. Calculation of survival rates for cancer. *Proc. Staff Meet. Mayo Clinic*, **25**:270–286.

Bland, E. F., and T. D. Jones. 1951. Rheumatic fever and rheumatic heart disease: A twenty-year report on 1,000 patients followed since childhood. *Circulation*, **4**:836–843.

Cutler, S. J., and F. Ederer. 1958. Maximum utilization of the life table method in analyzing survival. *J. Chron. Dis.* **8**:699–712.

Ellenberg, M. 1971. Impotence in diabetes: The neurologic factor. *Ann. Intern. Med.*, **75**:213–219.

Feinleib, M., and M. Zelen. 1969. Some pitfalls in the evaluation of screening programs. *Arch. Environ. Health*, **19**:412–415.

Frank, C. W., E. Weinblatt, S. Shapiro, and R. Sager. 1972. Prognosis of men with coronary heart disease as related to blood pressure. *Circulation*, **38**:432–438.

Hill, A. B.: *Principles of Medical Statistics*, 9th ed. (London: Oxford University Press, 1971), pp. 228–236.

Kodlin, D. On the status of coronary care unit statistics. To be submitted for publication.

Medsger, T. A., A. T. Masi, G. P. Rodnan, T. G. Benedek, and H. Robinson. 1971. Survival with systemic sclerosis (Scleroderma): A life-table analysis of clinical and demographic factors in 309 patients. *Ann. Intern. Med.*, **75**:369–376.

Resnick, R. H., T. C. Chalmers, A. M. Ishihara, A. J. Garceau, A. D. Callow, E. M. Schimmel, E. T. O'Hara, and the Boston Inter-Hospital Liver Group. 1969. A controlled study of the prophylactic portacaval shunt: A final report. *Ann. Intern. Med.*, **70**:657–688.

Stahlman, M. T., E. J. Battersby, F. M. Shepard, and W. J. Blankenship: 1967. Prognosis in hyaline-membrane disease: use of a linear-discriminant. *New Engl. J. Med.*, **276**:303–309.

Veterans Administration Cooperative Study Group on Antihypertensive Agents. 1967. Effects of treatment on morbidity and mortality: results in patients with diastolic blood pressures averaging 115 through 129 mm. Hg. *J. Am. Med. Assoc.*, **202**:1028–1034.

Veterans Administration Cooperative Study Group on Antihypertensive Agents. 1970. II. Results in patients with diastolic blood pressure averaging 90 through 114 mm. Hg. *J. Am. Med. Assoc.*, **213**:1143–1152.

Chapter 11

Making Sense out of Statistical Associations

Positive findings of epidemiologic or clinical outcome studies are usually referred to as statistical associations. It is essential to have a proper perspective of the meaning and importance of statistical associations. All too frequently they are under- or overinterpreted. With regard to smoking, for example, those at one extreme discount the strong epidemiologic evidence relating cigarette smoking and lung cancer as being "only statistical." At the other extreme are those who quickly blame a whole host of health problems on cigarettes on the basis of weak epidemiologic evidence, without considering the possible role of other important characteristics of persons who smoke.

Statements and Measures of Statistical Association

In discussing the various types of epidemiologic and related studies in Chaps. 5 through 10, the usual methods of expressing the results of these studies have been mentioned several times. Typically, the

findings would be that persons having one characteristic or environmental exposure have a higher or lower incidence or prevalence of a disease than persons with a different characteristic or exposure. Or, the association may be expressed in terms of a greater or lesser proportion of the characteristic in the diseased as compared to the nondiseased. Similar statements may express the fact that there is an association between one characteristic and another, or between one disease and another.

In addition to these easily understood statements of association in terms of differences in rates or proportions, epidemiologists sometimes employ other statistical tools to measure and describe associations. For example, data may suggest that there is a linear relationship between two quantitative variables. In a perfect linear relationship, for every unit of increase in one variable the other increases or decreases proportionally. One useful measure of association, the correlation coefficient, indicates the degree to which a set of observations fits a linear relationship. (For method of computation and more discussion, see Hill, 1971, Chaps. 15 and 16 or Ipsen and Feigl, 1970, Chap. 9.) This coefficient, often represented by the letter r, can vary between +1 and −1. If $r = +1$, there is a perfect linear relationship in which one variable varies directly with the other. If $r = 0$, there is no association between the variables. If $r = -1$, there is again a perfect association, but one variable varies inversely with the other.

Plotted on a graph showing the relationship between two variables, data points would follow a slanted straight line if the correlation coefficient is +1 or −1. Where there is some, but not complete, correlation, the data points would not fall into line but would appear to cluster about a line. If there is no correlation at all, data points would form a regular or irregular clump with no underlying slanted line apparent. Note that the data points for the states in Fig. 11-1 show some degree of linear relationship between cigarettes sold per capita and coronary-heart-disease death rates. The correlation coefficient is +0.55.

Other methods of measuring associations are also used, but as mentioned, differences in rates or proportions are most commonly employed. Regardless of how a statistical association is measured or expressed, the same problems of interpretation apply.

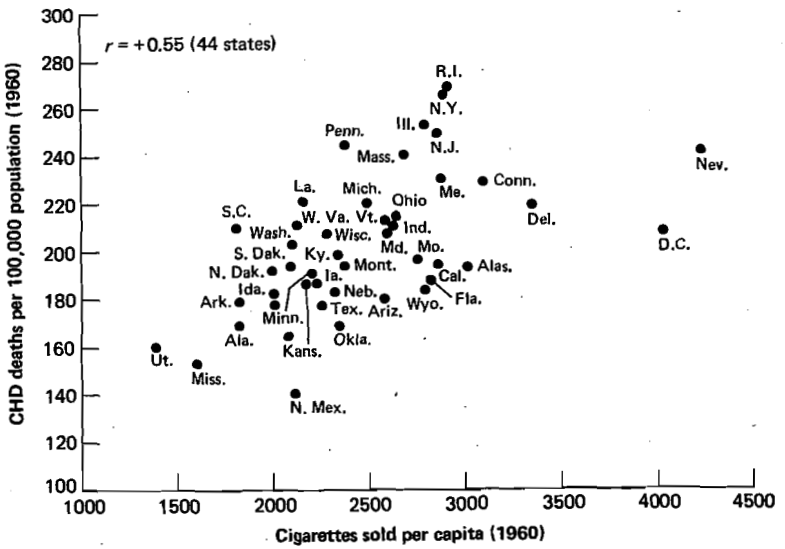


Figure 11-1 Relationship between the age-adjusted total death rate for coronary heart disease and per capita cigarette consumption in 44 states in 1960. *(Reproduced, by permission, from Friedman, 1967.)*

Associations Based on Groups of Groups

It has been emphasized in this book that, in epidemiology, the group is the unit of concern. Groups that provide the most useful and relevant information are *groups of individuals.* Nevertheless, it is also possible to study *groups of groups.* Statistical associations found in groups of groups may be useful, but they may also be quite misleading and not at all applicable to the individuals within the groups.

Consider, for example, the data shown in Fig. 11-1, relating per capita cigarette consumption to coronary-heart-disease mortality rates in 44 states in 1950. The statistical association shown graphically and by the correlation coefficient of +0.55, involves a group of states rather than a group of persons. Although the findings are suggestive of an association between cigarette smoking and coronary-heart-disease mortality in persons, we cannot be sure from

these data alone that the persons who smoked in these states truly experienced a higher coronary heart disease mortality rate. (Actually, the association between smoking and coronary heart disease death rates had already been shown in groups of individuals when the study yielding Fig. 11-1 was done. This study's purpose was to cast some light on the striking geographic variation in coronary mortality in the United States.)

The potential for drawing fallacious conclusions about groups of individuals from associations observed in groups of groups was emphasized by Robinson (1950), who termed the latter "ecological correlations." He noted, for example, that among *persons* age 10 and over in the United States there was a moderate *positive* association between being foreign-born and being illiterate. However, looked at on the basis of *geographic regions* (i.e., groups), there was a stronger *negative* correlation. That is, those regions with the lowest percentages of population foreign-born had the highest percentages who were illiterate. Thus a conclusion about the relationship of nativity to literacy based solely on a study of geographic units would have been quite misleading.

Most epidemiologic observations showing that geographic differences in disease rates parallel geographic differences in possible causative factors are associations involving groups of groups. The same may be said of parallel time trends. As such, these correlations in space and time are interesting clues, but their limitations should be recognized. Failure of investigators to respect the possible fallacies involved has contributed to the mistrust of statistics as exemplified by Disraeli's famous reference to "lies, damn lies, and statistics."

Evaluating Statistical Associations Involving Groups of Individuals

Fortunately, the main body of epidemiologic knowledge involves associations found in groups of individuals. When these associations emerge from a study, four basic questions usually require immediate attention:

1 Could the association have been observed just by chance?

2 Could other variables have accounted for the observed relationship?
3 To whom does the association apply?
4 Does the association represent a cause-and-effect relationship?

Evaluating the Possible Role of Chance

Regarding the first question, we have already mentioned in Chap. 3, page 25, that chance plays a role in determining the outcome of a study. The fewer the subjects, the more the observations may be influenced by chance sampling variation. Statistical significance tests are used to determine the probability that the observed association could have occurred by chance alone, if no association really exists. Selecting the appropriate test depends on the nature of the data and the method by which they are analyzed. For example, if the data analysis results in a fourfold table with subjects classified by presence or absence of a trait and of disease as illustrated by Table 3-2, page 39, the chi square test may be most appropriate. Comparing the mean level of a quantitative attribute in a disease group with the mean level in a control group may involve a "*t*" test of the difference between two means. The reader is referred to medical-statistics texts such as Hill (1971, Chaps, 11–14) or Ipsen and Feigl (1970, Chaps. 6, 8) for further details.

Unfortunately, the word "significant" in "statistically significant" is often misinterpreted as representing the medical or biological significance of an association. A slight difference in the mean hemoglobin concentration between two groups such as 0.1 gm/100 ml may be statistically significant if the two groups are large—that is, it is most unlikely to be due to chance. However this difference may be totally unimportant for health or longevity, or in relation to a disease under investigation. Thus, to say that one group's mean level is significantly lower than that of the other group has connotations that should be avoided by stressing the fact that *statistical* and not *biological* significance is being discussed.

Evaluating the Role of Other Variables

Ruling out chance as a likely explanation is only the first step in making sense out of an association. Equally, if not more, important, is to attempt to rule out other variables as possible explanations for

the association. To show in a very simple way how a third variable may account for part or all of a statistical association, an imaginary set of data is graphically plotted in Fig. 11-2. The figure shows, let us say, degree of coronary atherosclerosis measured by coronary angiography as related to hand-grip strength. Note that all eight data points form a pattern, showing an association between the two variables. That is, on the average, those with stronger grips tend to have more coronary atherosclerosis.

However, also note that four of the data points are shown by open circles and four by solid black circles. The open circles happen to represent four women and the black circles, four men. Looking at each sex group separately by covering the other four points, it can be seen that there is no relationship between grip strength and amount of atherosclerosis. It is only because the two sexes have been combined in one set of data that the association appears. Thus, sex difference constitutes a third underlying variable that completely explains the apparent association between grip strength and coronary atherosclerosis, which is therefore considered a *spurious* or *secondary* association.

Another set of fictitious data, shown in Table 11-1, again

Figure 11-2 Relationship between hand-grip strength and degree of coronary atherosclerosis. Fictitious data showing spurious correlation resulting from combining the data for men and women.

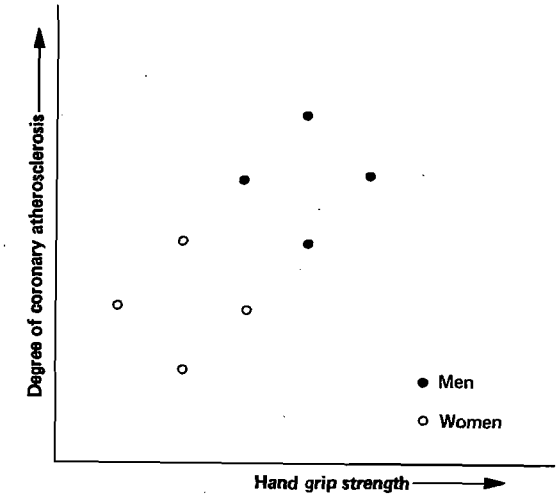


Table 11-1 Relationship between Parental Death and Low-Back Pain History. Fictitious Data Showing Spurious Association Due to Relation of Both Variables to Age.

Age	Total number	History of low-back pain	
		Number	Percent
30–39			
All subjects	200	20	10
Any parent dead	100	10	10
No parent dead	100	10	10
40–49			
All subjects	200	40	20
Any parent dead	140	28	20
No parent dead	60	12	20
50–59			
All subjects	200	60	30
Any parent dead	180	54	30
No parent dead	20	6	30
Total, all ages			
All subjects	600	120	20
Any parent dead	420	92	22
No parent dead	180	28	16

illustrates how an underlying variable, age, can result in an apparent association between two other variables when no real association exists. A total of 600 persons, ages 30–59, were asked whether they have ever been troubled by low-back pain and whether their parents were still living or whether either their mother or father had died.

The top section of the table shows the findings for the 200 subjects in their thirties. Twenty, or 10 percent, reported low-back pain. Also, half had both parents living and half reported at least one parent dead. Of the 100 in either parental-survival group, 10, or 10 percent, reported low-back pain. Thus, in this subgroup, death of a parent was not related to low-back pain.

The next section of the table shows the results for 200 subjects in their forties. At this later age a larger proportion had lost a parent ($^{140}/_{200}$), and a larger proportion reported low-back pain (20 percent), but parental death was again not related to low-back pain. In either parental survival group, 20 percent reported low-back pain.

The results for 200 subjects in their fifties also showed no relationship between the two study variables. The proportion with at least one dead parent was still higher ($^{180}/_{200}$), and the prevalence of a low-back-pain history was higher (30 percent) but again, the 30 percent low-back-pain prevalence held true for subjects both with and without a parent dead.

Now, look at what happens when the data for the three age groups are simply added together, as shown at the bottom of the table. A total of 22 percent of patients with a parent dead report low-back pain, whereas only 16 percent with both parents living have this complaint. The data for all ages combined appear to show that parental loss *is* related to low-back pain, whereas we know that in any age decade this is not the case.

The apparent relationship of low-back pain to parental loss in the total group is attributable to the difference in age distribution between those with and without a dead parent. Stated simply, those with a dead parent contain a higher proportion of older people and, therefore, are more apt to report low-back pain. Actually, 180, or 43 percent, of the 420 subjects with at least one parent dead were in their fifties, whereas only 20, or 11 percent, of the 180 subjects with no parents dead were in their fifties.

Handling Spurious Associations Due to Related Variables

Prevention Knowledge of previous epidemiologic findings or of the pathophysiology of the disease under investigation will often suggest related variables that may produce a spurious association. A study may be designed and carried out so as to prevent these related variables from producing misleading group differences. For example, cases and controls may be matched for age so that differences in age distribution will not lead to spurious associations such as the one described above.

It may not be possible to "control" all pertinent variables in this manner at the outset. Also, underlying variables may come to light or be thought of later, when the data are being analyzed. Fortunately, it is possible to analyze data in ways that take into account or control extraneous variables.

Specification The simplest method for controlling variables in the data analysis is *specification*. This involves examining the data separately for each subgroup of subjects who fall into one particular category or level of the variable to be controlled. In the above example involving a relationship between hand-grip strength and coronary atherosclerosis, the fact that the correlation is spurious and due to sex differences becomes obvious if we *specify* sex and look at the data separately for men and women. Similarly, if the parental-loss–back-pain association is examined in specific age groups, it is no longer apparent.

Actually, age and sex are so often related to disease occurrence and to other variables that it is customary to examine data in specific age-sex subgroups before combining them into an overall tabulation. This standard approach to data analysis in epidemiology is probably the reason that an epidemiologist has been defined as "a physician broken down by age and sex."

Just as specification can show associations to be spurious, it can also be used to show that suspected underlying variables are not explanations for an association. For example, in a study of smoking and the leukocyte count (Friedman et al., 1973), it was suspected that higher mean leukocyte counts in smokers than in nonsmokers might really be due to chronic bronchitis, which is related both to smoking and to the leukocyte count. The data were analyzed separately for persons with and without evidence of chronic bronchitis. When this was done, large smoker-nonsmoker differences in mean leukocyte count were still present in each subgroup and were, thus, not attributable to chronic bronchitis.

Adjustment Sometimes an investigator would like to compare two or more overall groups, knowing that they differ in a pertinent third variable. It is possible, by means of a procedure known as *adjustment*, to make such comparisons, controlling for differences

in an extraneous variable. For example, in evaluating the parental-loss–back-pain association, it is possible through *age-adjustment* to remove the effect of age as a "confounding" variable and compare subjects with and without parental loss to see if either group has a higher prevalence of a low-back-pain history.

Age adjustment by the *direct* method involves choosing a standard population and applying the rates observed for subjects in each specific age category to the corresponding members of the standard population. The choice of a standard population is fairly arbitrary. Often it is the population of a country at a particular time, such as the United States in 1960. Or, frequently, it is the total population involved in the study in question. Or, it may be one particular subgroup of that study population. In our low-back-pain study, for example, one might age-adjust the rates observed in the subgroup with no parental loss, to the subgroup with loss of a parent, or age-adjust the rates of both subgroups to the total study group.

To illustrate how this is accomplished, Table 11-2 shows the direct age adjustment of the rate of low-back pain in the subgroup without parental loss, to the total study population used as a standard. The rate observed in each age category of the subjects with no parental loss is multiplied by the number of subjects in the same age category in the standard population. This yields the number that would be observed in the standard population if the low-back pain rates in the group with no parental loss were applicable to the standard population. The numbers that would be observed in each age group of the standard population are then added together and the total is divided by the total number in the standard population, yielding the age-adjusted rate of 20 percent. In this example, the same age-specific rates of low-back pain were observed in the subjects with parental loss; therefore the age-adjusted rate for this subgroup would also be 20 percent. Thus, using age-adjusted rates, we would correctly conclude that parental loss was not related to low-back pain.

The *indirect* method of age-adjustment is somewhat different from the direct method. Instead of applying the study subgroup's age-specific rates to a standard population, the age-specific rates of the standard population are applied to the corresponding portions

Table 11-2 Example of Direct Age Adjustment: Observed Low-Back Pain Rates Applied to Standard Population Consisting of All Study Subjects

Age	Observed low-back-pain rate	×	Total number in age subgroup of standard population	=	Number that would be observed in standard population
30–39	10%		200		20
40–49	20%		200		40
50–59	30%		200		60
			Total 600		120
Age-adjusted rate = $^{120}/_{600}$ = 20%					

of the study subgroup. This procedure yields the numbers of cases that would be expected in the study subgroup if the age-specific rates in the standard population had been operative in the study subgroup. The overall expected rate in the study subgroup is then compared to the overall rate in the standard population. Any difference must be attributable to the difference between the age distribution of the subgroup and that of the standard population.

The study subgroup's overall observed rate is then corrected proportionally to make up for this difference in age distribution. For example, if the standard population's overall rate is 80 percent of the expected rate in the study subgroup, then the observed rate in the subgroup is reduced, by multiplying it by 80 percent. After the overall rates in various subgroups have been modified in this manner, they can then be compared fairly with one another. More detailed examples of age adjustment by the direct and indirect methods are given by Hill (1971, Chap. 17).

Indirect adjustment is preferable to direct when there are small numbers in age-specific groups. Rates used in direct adjustment would be based on these small numbers and would thus be subject to substantial sampling variation. With indirect adjustment the rates are more stable since they are based on a large standard population. Note that the expected rate or the expected number of cases, computed by the indirect method, is used in the ratio of observed/expected which constitutes the morbidity (or mortality) ratio described in Chap. 2.

It must be remembered that an age-adjusted rate is an artificial rather than an actual rate. Its value is that it permits one population to be compared with another, with age "controlled." It should not be used if what is wanted is not a comparison, but an accurate description of a population. The age-adjusted rate is a convenient summary of age-specific rates. The age-specific rates themselves are most informative and should be compared whenever possible.

This discussion of adjustment has focused on age adjustment because age is the variable that is most commonly controlled in this manner. However, direct or indirect adjustment may be applied to any variable suspected of playing a role in an association between two study variables.

Other methods More complex statistical procedures are also available for removing the effects of extraneous variables on statistical associations. These procedures involve the more traditional methods such as analysis of covariance, multiple correlation and multiple regression, and discriminant analysis. (The reader with some background in statistics may wish to refer to Morrison, 1967, for further discussion.) Newer methods of multivariate analysis have also been developed for epidemiologic studies of specific diseases.

These techniques are sometimes useful when it is apparent that several factors are not only associated with a disease but also with one another and one wishes to assess the relationship of each factor to the disease, independently of the other factors. An interesting example for the statistically minded reader is the multiple logistic method of Truett et al. (1967), as applied to coronary heart disease.

Although these methods appear to have definite value for certain epidemiologic studies, they all rest on assumptions. These assumptions must be understood by the user because they might or might not apply to the disease and other variables under investigation. Unfortunately, there has been a recent tendency to thoughtlessly throw some data into a computer together with a "canned" multivariate analysis program, expecting that the coefficients and other numbers that come out will somehow reveal a new secret of life. It must be stressed that no method of analysis, no matter how mathematically sophisticated, will substitute for careful evaluation of data based on good scientific judgment and knowledge of the disease process being studied.

General Applicability of an Association

In evaluating observed statistical associations one must always consider to whom they apply. The study in which the association is observed was conducted on a finite group of persons with certain characteristics. Would the association also hold true for other groups? Obviously, the more different groups that show the association, the more certain one can be that it is widely applicable. Where a variety of studies are lacking, it becomes a matter of judgment to determine whether an association observed in one group is applicable to another.

Questions of generality might be raised, for example, regarding the association between serum cholesterol level and coronary heart disease found in the Framingham Study. The study population is virtually all white. Thus it can legitimately be asked whether the same association holds true for blacks and Orientals. Fortunately, other studies provide a positive answer to this question.

More subtle is the fact the Framingham and other similar studies have as subjects volunteers or cooperative people. Does the cholesterol/coronary disease association apply also to uncooperative individuals? While volunteers do differ from others in certain characteristics, it is difficult to imagine that these characteristics would produce this observed relationship. Thus, one might reasonably judge that cholesterol is related to coronary heart disease in the uncooperative as well.

Statistical Associations and Cause-and-Effect Relationships

It is common knowledge that statistical associations do not necessarily imply causation. The "price of tea in China" is a frequently cited example of a variable which can be related statistically to some other variable but has no causal relation to it.

Statistical associations derived from well-controlled experimental studies can usually be interpreted to represent cause-and-effect relationships. Something is done and a result is observed. In epidemiology, however, most studies are observational, and an experiment to establish a cause-and-effect relationship may be

difficult or impossible to carry out. Vital decisions affecting public health and preventive medicine must be made on the basis of observational evidence. It is important, therefore, to have some basis for deciding whether or not a statistical association derived from an observational study represents a cause-and-effect relationship.

A number of authors have grappled with this philosophical problem. Certain criteria seem to be universally accepted, while others remain controversial. The reader wishing to explore this question in greater depth should refer to Chap. 2 of MacMahon and Pugh (1970), Chap. 24 of Hill (1971), Yerushalmy (1962), Larsen and Silvette (1968), and Susser (1973).

Strength of the Association In general, the stronger the association the more likely it represents a cause-and-effect relationship. Weak associations often turn out to be spurious and explainable by some known, or as yet unknown, third variable. In order for a strong association to be spurious, the underlying factor that explains it must have an even stronger relation to the disease (Bross, 1966). It is likely, although not certain, that the underlying variable with this even stronger relationship to the disease would be recognizable.

Strength of an association can be measured by the *relative risk*, or the ratio of the disease rate in those with the factor to the rate in those without. The relative risk of lung cancer in cigarette smokers as compared to nonsmokers is on the order of 10:1, whereas the relative risk of coronary heart disease is about 1.5:1. This difference suggests that cigarette smoking is more likely to be a causal factor for lung cancer than for coronary heart disease.

Time Sequence In a causal relationship the characteristic or event associated with the disease must *precede* the disease. This time relationship should be clear in incidence studies. In prevalence and case-control studies it may not always be obvious which came first.

Consistency with Other Knowledge If the association makes sense in terms of known biological mechanisms or other epidemiologic knowledge, it becomes more plausible as a cause-and-effect

relationship. Part of the attractiveness of the hypothesis that a high–saturated fat, high-cholesterol diet predisposes to atherosclerosis is the fact that a biologic mechanism can be invoked. Such a diet increases blood lipids which may in turn be deposited in arterial walls. A correlation between the number of telephone poles in a country and its coronary heart disease mortality rate lacks plausibility as a cause-and-effect relationship partly because it is difficult to imagine a biological mechanism whereby telephone poles result in atherosclerosis.

Failure to Find Other Explanations When a statistical association is observed, the thoughtful investigator will consider possible explanations for the relationship *other* than the observed variable's causing the disease. The data already collected may be used to learn whether these other possible explanations might hold true. Or, additional data may have to be obtained to answer such questions.

Failure to find an alternative to the cause-and-effect hypothesis despite conscientious searching does not prove that there is no alternative. But it does strengthen the evidence for a cause-and-effect relationship.

An interesting example of a search for other explanations comes from a case-control study showing an association between oral contraceptives and thromboembolic disease (Vessey and Doll, 1968). Since it is easy to overlook the diagnosis of deep-vein thrombosis or pulmonary embolism, the investigators considered the possibility that a history of oral-contraceptive use would alert the physician to these conditions, resulting in a spurious association. They reasoned that a spurious association of this type would be strongest among patients with the least evident disease, since this group would contain women whose condition was diagnosed only because they were known to have taken oral contraceptives. Cases were therefore classified by degree of certainty as to the presence of thromboembolism. It was found that the association with oral-contraceptive use was actually less marked among the less certain and milder cases than among the definite and severe cases. Thus, this alternative explanation could reasonably be rejected, lending greater credence to the idea that thromboembolism was actually caused by oral contraceptives.

Other Criteria The criteria listed below have been stressed by some authorities but to this author they seem less valuable as yardsticks for assessing a cause-and-effect relationship per se.

Gradient of Risk It has been stated that if there appears to be a dose-response relationship, this argues for a cause-and-effect relationship. For example, the fact that moderate cigarette smokers have a lung cancer death rate intermediate between nonsmokers and heavy smokers is considered evidence that cigarette smoking causes lung cancer.

This criterion would appear less satisfactory. Threshold phenomena are well known in nature, whereby no effect is seen until a causal stimulus reaches a certain level, above which a response is seen. In this situation a gradient of response might well be absent if two different dosages of the causal factor are well below the threshold level. Conversely, a spurious correlation could easily show a nice gradient. A spurious correlation of cigarette smoking with a disease caused by alcohol consumption might show an apparent dose-response relationship of disease incidence to amount smoked, due to a correlation between amount smoked and amount of alcohol consumed.

Consistency in Several Studies Finding the same association in several different studies provides assurance that the association *exists* and is not an artifact based on the way one particular study was carried out or based on an unusual group of study subjects. In this sense, consistency across studies is reassuring; but it does not argue strongly that an association is one of cause and effect.

Specificity By specificity is meant that the possible causal factor is observed to be associated with one or just a few diseases or effects, rather than a wide variety of diseases. One of the arguments that has been used against cigarette smoking as a cause of lung cancer is that in epidemiologic studies, smoking also appears to be associated with an assortment of seemingly unrelated diseases such as coronary heart disease, peptic ulcer, bladder cancer, and cirrhosis of the liver. It is argued either that smokers differ biologically from nonsmokers in a way that leads health to break down in a variety of ways or that these studies must have been affected by some kind of hidden bias or artifact that falsely incriminates smoking in so many ways.

Although it *is* reassuring when specificity is found, and an apparent lack of specificity *should* lead to some suspicion of an artifact, the importance of a lack of specificity as negative evidence has been overemphasized. This can be readily seen when one considers other recognized disease agents such as the tubercle bacillus and applies the lack of specificity argument to them. How, it might have been asked, can the tubercle bacillus cause an increased rate of lung lesions when it also has been associated with scrofula, meningitis, collapsed vertebrae, peritonitis, bleeding from the kidney, marked wasting, and so on. We now know that the tubercle bacillus can produce a variety of effects, and we have some understanding of the mechanisms by which these occur. Cigarette smoke has a variety of active constituents that get carried throughout the body, so that a lack of specificity is not surprising.

Statistical Associations between Diseases

Epidemiologic and clinical studies may reveal statistical associations between two or more diseases. Two diseases are associated in a population if the incidence or prevalence of one disease is higher when the other is present than when it is absent.

A true association between diseases may occur because one disease predisposes to another (e.g., diabetes mellitus and coronary heart disease) or because both diseases share a common etiologic factor (head injuries and cirrhosis of the liver, both due to alcoholism). Thus, discovery of disease associations may provide valuable information if the etiology of one disease is obscure.

Disease associations may be more apparent than real. Two diseases may produce similar signs, symptoms, or laboratory findings, thus leading to a greater chance of *diagnosis* of one disease if the other is present. Also, diseases are detected in the clinic, hospital, or at autopsy, and the presence of more than one disease may make it more likely for a person to show up at one of these diagnostic facilities. Due to this and other selective factors, diseases may appear to be associated at a medical facility even when they are not associated in the general population. Further discussion of disease associations and the potential fallacies involved may be found in Berkson (1946), Mainland (1953), Wijsman (1958), and Friedman (1968).

Even false associations due to selection may be useful to the clinician. For example, an association between inguinal hernia and colon cancer has been noted on the surgical ward (Terezis et al., 1963). Even if this association is not present in the general population, it still may be wise for surgeons to look for colon cancer in their patients with hernias.

REFERENCES

Berkson, J. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bull.*, **2**:47–53.

Bross, I. D. J. 1966. Spurious effects from an extraneous variable. *J. Chronic Dis.*, **19**:637–647.

Friedman, G. D. 1967. Cigarette smoking and geographic variation in coronary heart disease mortality in the United States. *J. Chron. Dis.*, **20**:769–779.

Friedman, G. D. 1968. The relationship between coronary heart disease and gallbladder disease: A critical review. *Ann. Intern. Med.*, **68**:222–235.

Friedman, G. D., A. B. Siegelaub, C. C. Seltzer, R. Feldman, and M. F. Collen. 1973. Smoking habits and the leukocyte count. *Arch. Environ. Health*, **26**:137–143.

Hill, A. B., *Principles of Medical Statistics*, 9th edition. (London: Oxford University Press, 1971).

Ipsen, J., and P. Feigl, *Bancroft's Introduction to Biostatistics.* (New York: Harper and Row, 1970).

Larsen, P. S., and H. Silvette, *Tobacco: Experimental and Clinical Studies* Supplement I. (Baltimore: Williams and Wilkins, 1968), pp. 346–362.

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

Mainland, D. 1953. The risk of fallacious conclusions from autopsy data of the incidence of diseases with applications to heart disease. *Am. Heart J.*, **45**:644–654.

Morrison, D. F., *Multivariate Statistical Methods.* (New York: McGraw-Hill Book Company, 1967).

Robinson, W. S. 1950. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.*, **15**:351–357.

Susser, M., *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* (New York: Oxford University Press, 1973).

Terezis, L. N., W. C. Davis, and F. C. Jackson. 1963. Carcinoma of the colon associated with inguinal hernia. *New Engl. J. Med.*, **268**:774.

Truett, J., J. Cornfield, and W. Kannel. 1967. A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.*, **20**:511–524.

Vessey, M. P., and R. Doll. 1968. Investigation of relation between use of oral contraceptives and thromboembolic disease. *Brit. Med. J.*, **2**:199–205.

Wijsman, R. A. 1958. Contribution to the study of the question of associations between two diseases. *Human Biology*, **30**:219–236.

Yerushalmy, J., Statistical considerations and evaluation of epidemiological evidence, in *Tobacco and Health* edited by G. James and T. Rosenthal. (Springfield, Ill.: Charles C Thomas, 1962), pp. 208–230.

Chapter 12

How to Carry Out a Study

Many health-care professionals wish to conduct a modest clinical or epidemiologic study. Hoping to answer one or more interesting questions, they find themselves in a good position to collect and analyze some appropriate data. However, to someone without previous research experience, the task often appears awesome, and it is not at all clear how to proceed.

This chapter is written as a general guide for the novice who wishes to carry out such a study. Obviously, each research project and each study setting presents unique problems which cannot be dealt with here. What will be presented is a general approach which emphasizes the practical difficulties that are frequently troublesome to the beginner.

Defining the Problem

The first step—and one of the most difficult ones—is defining the problem and choosing the question or questions to be answered.