5

Types of Epidemiologic Studies

Kenneth J. Rothman and Sander Greenland

The scientific experiment is emblematic of scientific activity. What constitutes an experiment? In common parlance, an *experiment* refers to any trial or test. For example, a professor might introduce new teaching methods as an experiment. For many scientists, the term has a more specific meaning: An experiment is a set of observations, conducted under controlled circumstances, in which the scientist manipulates the conditions to ascertain what effect such manipulation has on the observations. Some might enlarge this definition to include controlled observations without manipulation of the conditions. Thus, the astrometric observations during the solar eclipse of 1919 that corroborated Einstein's general theory of relativity have often been referred to as an experiment. For epidemiologists, however, the word *experiment* usually connotes that the investigator manipulates the conditions studied.

The ideal experiment would create sets of circumstances across which only one factor affecting the outcome of interest would vary. To achieve this objective would require control of all the relevant conditions that might affect the outcome under study. Unfortunately, in the biologic sciences, the conditions affecting most outcomes are so complex and occult that they cannot be made uniform. In the study of the causes of cancer, for example, it is impossible to create conditions that will invariably give rise to cancer after a fixed time interval, even if the population is a group of cloned laboratory mice. Inevitably, there will be what is called "biologic variation," which refers to variation in the set of conditions that produces the effect.

In biologic experimentation, then, the ideal of creating duplicate sets of circumstances in which only one relevant factor varies is unrealistic (many would argue that this objective is unrealistic in other branches of science also). Instead, the experimenter settles for creating circumstances in which the amount of variation of factors that might affect the outcome is small in comparison with the variation of the key factor under study. Thus, it may be impossible to make all animals in an experimental group eat exactly the same amount of food. Variability in food consumption could pose a problem if it affected the outcome under study. If it were kept small, however, variability in food consumption may not affect the experiment very much. The investigator would usually be satisfied if the

variability of extraneous factors (i.e., those factors other than the key study variables) was too small to affect the outcome under study to an important extent.

Epidemiologic study types have their roots in the concepts of scientific experimentation. When epidemiologic experiments are feasible, their design is guided by principles that reduce variation by extraneous factors in comparison with the study factors. Epidemiologic experiments include *clinical trials* (with patients as subjects), *field trials* (with interventions assigned to individual community members), and *community intervention trials* (with interventions assigned to whole communities). When experiments are not feasible, epidemiologists design nonexperimental studies to simulate what might have been learned if an experiment had been conducted. Nonexperimental studies include *cohort* studies, in which subjects are classified (and possibly selected) according to their exposure status and followed over time to ascertain disease incidence; *case-control* studies, in which subjects are selected according to their disease status and further classified according to their exposure status; *proportional mortality* studies, which are best viewed as a type of case-control study; *cross-sectional* studies, including *prevalence* studies; and *ecologic* studies, in which the units of observation are groups of people.

EXPERIMENTAL STUDIES

In an experiment, those who are exposed to the agent or putative cause are exposed only because the investigator has assigned the exposure to the subject. Furthermore, the reason for assigning the specific exposure to the particular subject must be simply the pursuit of the study protocol—that is, the only reason for the assignment must be to conform to the protocol rather than to meet the needs of the subject. For example, suppose that a physician treating headache had prescribed a patented drug to her wealthy patients and a generic counterpart to her indigent patients, because the presumed greater reliability of the patented version was in her judgment not worth the greater cost for those of modest means. Should the physician later want to compare the effects of the two medications, she could not consider herself to be conducting an experiment, despite the fact that the investigator herself had assigned the exposures. To conduct a proper experiment, she would have to assign the drugs according to a protocol that would reduce variation between the treatment groups with respect to other potential causes of headache. The assignment of exposure in experiments is designed to help the study rather than the individual subject. If it is done to help the subject, then a nonexperimental study is still possible, but it should not be called an experiment. Sometimes the term *quasi-experiment* is used to refer to controlled studies in which exposure was assigned but not according to a randomized experimental protocol (Cook and Campbell, 1979).

Because the goals of the study rather than the subject's needs determine the exposure assignment, ethical constraints limit the circumstances in which epidemiologic experiments are feasible. Experiments are ethically permissible only when adherence to the scientific protocol does not conflict with the subject's best interests. Specifically, there should be reasonable assurance that no participating subject could be treated better than the two or more treatment possibilities that the protocol provides. From this requirement comes the obvious constraint that any exposures or treatments given to subjects should be limited to potential preventives of disease or disease consequences. This limitation alone confines most etiologic research to the nonexperimental variety.

A second constraint is that all of the treatment alternatives should be equally acceptable under present knowledge. A third constraint is that subjects admitted to the study should not be thereby deprived of some preferable form of treatment or preventive that is

not included in the study. For example, it is unethical to include a placebo therapy as one of the arms of a clinical trial if an accepted remedy or preventive of the outcome already exists. The best available therapy should be the comparison for any new treatment. Additionally, subjects must be fully informed of their participation in an experiment and of the possible consequences.

Even with these limitations, many epidemiologic experiments are conducted. Most fall into the specialized area of *clinical trials,* which are epidemiologic studies of different treatments for patients who already have some disease (*trial* is used as a synonym for *experiment).* Epidemiologic experiments that aim to evaluate primary preventives (agents intended to prevent disease onset in the first place) are less common than clinical trials; these studies are usually *field trials* or *community intervention trials.*

Clinical Trials

A clinical trial is an experiment with patients as subjects. The goal of a clinical trial is either to evaluate a potential cure for a disease or to find a preventive of disease sequelae such as death or disability. The exposures in a clinical trial are not primary preventives, since they do not prevent occurrence of the initial disease, but they are preventives of the sequelae of the initial disease. For example, a modified diet after an individual suffers a myocardial infarction may prevent reinfarction and subsequent death, or chemotherapeutic agents given to cancer patients may prevent recurrence of cancer.

Subjects in clinical trials must be diagnosed as having the disease in question and must be admitted to the study soon enough following diagnosis to permit the treatment assignment to occur in a timely fashion. Subjects whose illness is too mild or too severe to permit the form of treatment or alternative treatment being studied must be excluded. Treatment assignment should be designed to minimize variation of extraneous factors that might affect the comparison. For example, if some physicians participating in the study favored the new therapy, they could conceivably influence the assignment of, say, their own patients or perhaps the more seriously afflicted patients to the new treatment. If the more seriously afflicted patients tended to get the new treatment, then valid evaluation of the new treatment would be compromised.

To avoid this and related problems, it is customary to assign treatments in clinical trials in a way that promotes comparability among treatment groups with respect to unmeasured "baseline" characteristics, deters manipulation of assignments by study personnel, and permits causal inferences with a minimum of assumptions. It is almost universally agreed that a random assignment scheme is the best way to accomplish these objectives (Byar et al., 1976; Peto et al., 1976; Gelman et al., 1995). The validity of the trial ultimately depends on the extent to which the random process achieves similarity of the treatment groups with respect to the baseline distribution of unmeasured risk factors.

Whenever feasible, clinical trials should attempt to employ *blinding* with respect to the treatment assignment. Ideally, the individual who makes the assignment, the patient, and the assessor of the outcome should all be ignorant of the treatment assignment. Blinding prevents certain biases that could affect assignment, assessment, or compliance. Most important is to keep the assessor blind, especially if the outcome assessment is subjective, such as a clinical diagnosis (some outcomes, such as death, will be relatively insusceptible to bias in assessment). Patient knowledge of treatment assignment can affect compliance with the treatment regime and can bias perceptions of symptoms that might affect the outcome assessment. Studies in which both the assessor and the patient are blinded as to the treatment assignment are known as *double-blind* studies. A study in which the individual who makes

the assignment is unaware which treatment is which (such as might occur if the treatments are coded pills and the assigner does not know the code) may be described as *triple-blind.*

Depending on the nature of the treatment, it may not be possible or practical to keep knowledge of the treatment assignment from some or all of these three parties. For example, a treatment may have well known side effects that allow the patients to identify the treatment. The investigator needs to be aware of and report these possibilities.

If there is no accepted treatment for the condition being studied, it may be useful to employ a *placebo* as the comparison treatment. Placebos are inert treatments intended to have no effect other than the psychologic benefit of offering treatment, which itself can be a powerful effect. By employing a placebo, an investigator can control for the psychologic component of offering treatment and study the nonpsychologic benefits of a new intervention. In addition, employing a placebo facilitates blinding if there would otherwise be no comparison treatment. Placebos, however, may be considered unethical in some settings, especially when an effective treatment is available; in that case, the best available treatment should be used as a comparison (Rothman and Michels, 1994). Placebos are also not necessary when the objective of the trial is solely to compare different treatments.

Whenever noncompliance with the assigned treatment is possible, it will be important to measure its extent. Investigators may directly query subjects about their compliance. Occasionally, biochemical measures of compliance may be available and acceptable. Compliance measures can be used to improve estimates of treatment effects (Angrist et al., 1996).

Field Trials

Field trials differ from clinical trials in that they deal with subjects who have not yet gotten disease and therefore are not patients. Whereas the patients in a clinical trial may face the complications of their disease with high probability during a relatively short time, typically the risk of contracting a given disease for the first time is comparatively small. Consequently, field trials usually require a greater number of subjects than clinical trials and therefore are usually much more expensive. Furthermore, since the subjects are not patients, who usually come to a central location for treatment, a field trial often necessitates visiting subjects in the field (at work, home, or school) or establishing centers from which the study can be conducted and to which subjects are urged to report. These design features add to cost.

The expense of field trials limits their use to the study of preventives of either extremely common or extremely serious diseases. Several field trials were conducted to determine the efficacy of large doses of vitamin C in preventing the common cold (Karlowski et al., 1975; Dykes and Meier, 1975). Poliomyelitis, a rare but serious illness, was a sufficient public health concern to warrant what may have been the largest formal human experiment ever attempted, the Salk vaccine trial, in which the vaccine or a placebo was administered to hundreds of thousands of school children (Francis et al., 1955). When the disease outcome occurs rarely, it is more efficient to study subjects thought to be at higher risk. Thus, the trial of hepatitis B vaccine was carried out in a population of New York City male homosexuals, among whom hepatitis B infection occurs with much greater frequency than is usual among New Yorkers (Szmuness, 1980).

Similar reasoning is often applied to clinical trials, which may concentrate on patients at high risk of adverse outcomes. Several clinical trials of the effect of lowering serum cholesterol levels on the risk of myocardial infarction have been undertaken with subjects who have already experienced a myocardial infarction because such patients are at high risk for a second infarction (Leren, 1966; Detre and Shaw, 1974). It is much more costly

to conduct a trial designed to study the effect of lowering serum cholesterol on the first occurrence of a myocardial infarction because many more subjects must be included to provide a reasonable number of outcome events to study. The Multiple Risk Factor Intervention Trial (MRFIT) was a field trial of several primary preventives of myocardial infarction, including diet; although it admitted only high-risk individuals and endeavored to reduce risk through several simultaneous interventions, the study involved 12,866 subjects and cost $115 million (nearly half a billion dollars in present-day dollars) (Kolata, 1982).

As in clinical trials, exposures in field trials should be assigned in a way that promotes comparability of groups and removes any discretion in assignment from the study's staff. A random assignment scheme is again an ideal choice, but the difficulties of implementing such a scheme in a large-scale field trial can outweigh the advantages. For example, it may be convenient to distribute vaccinations to groups in batches that are handled identically, especially if storage and transport of the vaccine is difficult. Because such choices can seriously affect the interpretation of experimental findings, the advantages and disadvantages need to be carefully weighed.

Community Intervention and Cluster Randomized Trials

The community intervention trial is an extension of the field trial that involves intervention on a community-wide basis. Conceptually, the distinction hinges on whether or not the intervention is implemented separately for each individual. Whereas a vaccine is ordinarily administered singly to individual people, water fluoridation to prevent dental caries is ordinarily administered to individual water supplies. Consequently, water fluoridation was evaluated by community intervention trials in which entire communities were selected and exposure (water treatment) was assigned on a community basis. Other examples of preventives that might be implemented on a community-wide basis include fast-response emergency resuscitation programs and educational programs conducted using mass media, such as Project Burn Prevention in Massachusetts (MacKay and Rothman, 1982).

Some interventions are implemented most conveniently with groups of subjects smaller than entire communities. Dietary intervention may be made most conveniently by family or household; environmental interventions may affect an entire office, plant, or residential building. Protective sports equipment may have to be assigned to an entire team or league. Intervention groups may be army units, classrooms, vehicle occupants, or any other group whose members are simultaneously exposed to the intervention. The scientific foundation of experiments using such interventions is identical to that of community intervention trials. What sets all these studies apart from ordinary field trials is that the intervention is more easily assigned to groups than to individuals.

Field trials in which the treatment is assigned randomly to groups of participants are said to be *cluster randomized.* The larger the size of the group to be randomized relative to the total study size, the less that is accomplished by random assignment. If only two communities are involved in a study, one of which will receive the intervention and the other of which will not, such as in the Newburgh-Kingston water fluoridation trial (Ast et al, 1956), it cannot matter whether the community that receives the fluoride is assigned randomly or not; differences in baseline characteristics will have the same magnitude whatever the method of assignment—only the direction of the differences will be affected. Only if the numbers of groups randomized to each intervention are large will it be likely that randomization leads to similar distributions of baseline characteristics among the intervention groups.

NONEXPERIMENTAL STUDIES

The limitations imposed by ethics and cost restrict epidemiologic research to nonexperimental studies in most circumstances. While it is unethical for an investigator to expose a person to a potential cause of disease simply to learn about etiology, people often willingly or unwillingly expose themselves to many potentially harmful factors. The extent of such exposures has been eloquently described by MacMahon (1979):

> They choose a broad range of dosages of a variety of potentially toxic substances. Consider the cigarette habit to which hundreds of millions of persons have exposed themselves at levels ranging from almost zero (for those exposed only through smoking by others) to the addict's three or four cigarettes per waking hour, and the consequent two million or more deaths from lung cancer in the last half century in this country alone. Consider the fact that fewer than half of American women pass through menopause without either having their uterus surgically removed, being liberally dosed with hormones that are known to increase cancer risk in animals, or both. Consider the implications of the fact that more than fifty million women worldwide take regularly for contraceptive purposes a combination of hormones that essentially cuts off the function of their own ovaries.

The goal of all research is to obtain valid evidence regarding the hypothesis under study. Ideally, we would want the quality of evidence from nonexperimental research to be as high as that obtainable from a well designed experiment, had one been possible. In an experiment, however, the investigator has the power to assign exposures in a way that enhances the validity of the study, whereas in nonexperimental research the investigator cannot control the circumstances of exposure. If those who happen to be exposed have a greater or lesser risk for the disease than those who are not exposed, a simple comparison between exposed and unexposed will not reflect accurately the effect of the exposure.

Since the investigator cannot assign exposure in nonexperimental studies, he or she must rely heavily on the primary source of discretion that remains, the selection of subjects. If the paradigm of scientific observation is the experiment, then the paradigm of nonexperimental epidemiologic research is the "natural experiment," in which nature emulates an experiment. By far the most renowned example, the prototype of all natural experiments, is the elegant study of cholera in London conducted by John Snow. In London during the mid-nineteenth century, there were several water companies that piped drinking water to residents. Snow's natural experiment consisted of comparing the cholera mortality rates for residents subscribing to two of the major water companies: the Southwark and Vauxhall Company, which piped impure Thames river water contaminated with sewage, and the Lambeth Company, which in 1852 changed its collection from opposite Hungerford Market to Thames Ditton, thus obtaining a supply of water free of the sewage of London. As Snow (1860) described it,

> . . . the intermixing of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over an extensive part of London, admitted of the subject being sifted in such a way as to yield the most incontrovertible proof on one side or the other. In the subdistricts...supplied by both companies, the mixing of the supply is of the most intimate kind. The pipes of each company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one company and a few by the other, according to the decision of the owner or occupier at the time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in either the condition or occupation of the persons receiving the water of the different companies...it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this.

> The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentle folks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from impurity.
>
> To turn this experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur....

There are two primary types of nonexperimental studies in epidemiology. The first, the *cohort study* (also called the *follow-up study* or *incidence study),* is a direct analogue of the experiment; different exposure groups are compared, but (as in Snow's study) the investigator does not assign the exposure. The other, the incident case-control study, or simply the *case-control study*, employs an extra step of sampling according to the outcome of individuals in the population. This extra sampling step can make a case-control study much more efficient than a cohort study of the entire population, but it introduces a number of subtleties and avenues for bias that are absent in typical cohort studies.

Cohort Studies

In the classic cohort study, the investigator defines two or more groups of people that are free of disease and that differ according to the extent of their exposure to a potential cause of the disease. These groups are referred to as the study *cohorts* (from the Latin word for one of the ten divisions of a Roman legion). In such studies, there is at least one cohort thought of as the exposed cohort—those individuals who have experienced the putative causal event or condition—and another cohort thought of as the unexposed, or reference cohort. There may be more than just two cohorts, but each cohort would represent a group with a different level or type of exposure. For example, an occupational cohort study of chemical workers might comprise cohorts of workers in a plant who work in different departments of the plant, with each cohort being exposed to a different set of chemicals. The investigator measures and compares the incidence rate of the disease in each of the study cohorts.

In Snow's natural experiment, the study cohorts were residents of London who consumed water from either the Lambeth Company or the Southwark and Vauxhall Company and who lived in districts where the pipes of the two water companies were intermixed. Snow was able to estimate the frequency of cholera deaths, using households as the denominator, separately for people in each of the two cohorts (Snow, 1860):

> According to a return which was made to Parliament, the Southwark and Vauxhall Company supplied 40,046 houses from January 1 to December 31, 1853, and the Lambeth Company supplied 26,107 houses during the same period; consequently, as 286 fatal attacks of cholera took place, in the first four weeks of the epidemic, in houses supplied by the former company, and only 14 in houses supplied by the latter, the proportion of fatal attacks to each 10,000 houses was as follows: Southwark and Vauxhall 71, Lambeth 5. The cholera was therefore fourteen times as fatal at this period, amongst persons having the impure water of the Southwark and Vauxhall Company, as amongst those having the purer water from Thames Ditton.

Many cohort studies begin with but a single cohort that is heterogeneous with respect to exposure history. Comparisons of disease experience are made within the cohort across subgroups defined by one or more exposures. Examples include studies of cohorts defined from membership lists of administrative or social units, such as cohorts of doctors or nurses, or cohorts defined from employment records, such as cohorts of factory workers.

Case-Control Studies

Case-control studies are best understood by defining a source population, which represents a hypothetical study population in which a cohort study might have been conducted. If a cohort study were undertaken, the primary tasks would be to identify the exposed and unexposed denominator experience, measured in person-time units of experience or as the number of people in each study cohort, and then to identify the number of cases occurring in each person-time category or study cohort. In a case-control study, the cases are identified and their exposure status is determined just as in a cohort study, but denominators from which rates could be calculated are not measured. Instead, a control group of study subjects is sampled from the entire source population that gives rise to the cases.

The purpose of the control group is to determine the relative (as opposed to absolute) size of the exposed and unexposed denominators within the source population. From the relative size of the denominators, the relative size of the incidence rates (or incidence proportions, depending on the nature of the data) can be estimated. Thus, case-control studies yield estimates of relative effect measures. Because the control group is used to estimate the distribution of exposure in the source population, the cardinal requirement of control selection is that the controls must be sampled independently of their exposure status.

In sum, case-control studies of incident cases differ from cohort studies according to how subjects are initially selected. A cohort study identifies and follows a population or populations to observe disease experience; a case-control study involves an additional step of selecting cases and controls from this population. More detailed discussions of cohort and case-control studies will be presented in Chapters 6 and 7.

Prospective Versus Retrospective Studies

Studies can be classified further as either prospective or retrospective. We define a prospective study as one in which exposure and covariate measurements are made before the cases of illness occur. In a retrospective study these measurements are made after the cases have already occurred.

The distinction between the classification as cohort or case-control and prospective or retrospective should be firmly drawn, because these two axes for classifying epidemiologic studies have often been confused: Early writers referred to cohort studies as prospective studies and to case-control studies as retrospective studies because cohort studies usually begin with identification of the exposure status and then measure disease occurrence, whereas case-control studies usually begin by identifying cases and controls and then measure exposure status. The terms *prospective* and *retrospective,* however, are more usefully employed to describe the timing of disease occurrence with respect to exposure measurement. For example, case-control studies can be either prospective or retrospective. A prospective case-control study uses exposure measurements taken before disease, whereas a retrospective case-control study uses measurements taken after disease. Both cohort and case-control studies may employ a mixture of prospective and retrospective measurements, using data collected before and after disease occurred.

The prospective/retrospective distinction is sometimes used to refer to the timing of subject identification, rather than measurement of exposure and covariates. With this usage, a retrospective (or historical) cohort study involves the identification and follow-up of subjects, but the subjects are identified only after the follow-up period under study has

ended. The identification of the subjects, their exposure, and their outcome must be based on existing records or memories.

Experiments are always prospective cohort studies, because the investigator first assigns the exposure and then must wait until disease events occur. On the other hand, many occupational cohort studies are retrospective, in the sense that subjects are selected after the disease occurred. The advantages and drawbacks of prospective and retrospective measurement and selection will be discussed in Chapters 8 and 9.

Cross-Sectional Studies

A study that includes as subjects all persons in the population at the time of ascertainment or a representative sample of all such persons, including those who have the disease, and that has an objective limited to describing the population at that time, is usually referred to as a *cross-sectional study.* A cross-sectional study conducted to estimate prevalence is called a *prevalence study.* Usually, the exposure information is ascertained simultaneously with the disease information, so that different exposure subpopulations may be compared with respect to their disease prevalence.

Cross-sectional studies need not have etiologic objectives. For example, delivery of health services often requires knowledge only of how many items will be needed (such as number of hospital beds), without reference to the causes of the disease. Nevertheless, prevalence data are so often used for etiologic inferences that a thorough understanding of their limitations is essential.

One problem, often discussed under the topic of length-biased sampling, is that the cases in a cross-sectional study will overrepresent cases with long duration and underrepresent those with short duration of illness. To see this, consider two extreme situations involving a disease with a highly variable duration. A person contracting this disease at age 20 and living until age 70 can be included in any cross-sectional study during the person's 50 years of disease. A person contracting the disease at age 40 and dying within a day has almost no chance of inclusion. Thus, if the exposure does not alter disease risk but causes the disease to be very mild if contracted (so that exposure is positively associated with duration), the prevalence of exposure will be elevated among cases; as a result, a very positive exposure–disease association will be observed in a cross-sectional study, even though exposure has no effect on disease risk. If exposure does not alter disease risk but causes the disease to be rapidly fatal if contracted (so that exposure is negatively associated with duration), prevalence of exposure will be very low among cases; as a result, the exposure–disease association observed in the cross-sectional study will be very negative, even though exposure has no effect on disease risk.

There are analytic methods for dealing with the potential relation of exposure to duration (e.g., Simon, 1980b). These methods require that we obtain either the diagnosis dates of the study cases or information on the distribution of durations for the study disease at different exposure levels. Even with no relation of exposure to duration, however, one still faces the problem that current exposure may have little relation to exposure during the time etiologically relevant to current disease. Such a time is removed from the present by two historical spans: (1) the induction time between relevant exposure and disease occurrence (which remains hypothetical until good data on induction time are obtained) and (2) the time from disease occurrence to the time of the study (which can and should be measured, preferably from medical records).

Cross-sectional studies often deal with exposures that cannot change, such as blood type or other invariable personal characteristics. For such exposures, current information

is as useful as any. For variable exposures, however, current information is less desirable than etiologically more relevant information from before the case occurred. In a study of the etiology of respiratory cancer that compares smoking information on cases and noncases, the current smoking habits of subjects are not nearly as relevant as their smoking histories before the cancer developed. The cross-sectional approach to such a question could well be viewed as a case-control study with an excessively large control group (because few people in a population would have respiratory cancer), with smoking information from an inappropriate time period, and with biased case ascertainment (short-duration cases are much less likely to be seen than long-duration cases). Of course, the time-period problem could be addressed by asking subjects about their smoking history, rather than about current smoking.

Although current information is often too recent to be etiologically relevant, occasionally there is adequate justification for its use. If there is reason to believe that current exposure closely corresponds with the relevant past exposure and if recall of previous exposure is likely to be unreliable, it may be reasonable to use current exposure status as a proxy for the relevant exposure. Studies on dietary preferences, for example, often extract detailed current information because precise information on food consumption can thus be obtained, whereas recall of dietary information is likely to be vague and unreliable. Studies dependent on current exposure when past exposure is relevant suffer in validity to the extent that the previous exposure of subjects differs from current exposure.

Cross-sectional studies may involve sampling subjects differentially with respect to disease status. Such studies are sometimes called *prevalent case-control studies*, since their relation to prevalence studies is analogous to the relation of incident case-control studies to cohort studies (Morgenstern and Thomas, 1993).

Proportional Mortality Studies

A proportional mortality study includes only dead subjects. The proportion of dead exposed subjects assigned to one or more specific index causes of death is compared with the proportion of dead unexposed subjects assigned to the index causes. The resulting *proportional mortality ratio* (often abbreviated PMR) is the traditional measure of the effect of the exposure on the index causes of death. Superficially, the comparison of proportions of subjects dying from a specific cause for an exposed and an unexposed group resembles a cohort study measuring incidence. The resemblance is deceiving, however, because a proportional mortality study does not involve the identification and follow-up of cohorts. All subjects are dead at the time of entry into the study.

The premise of a proportional mortality study is that if the exposure causes (or prevents) a specific fatal illness, there should be proportionately more (or fewer) deaths from that illness among dead people who had been exposed than among dead people who had not been exposed. It is well recognized that this reasoning suffers two important flaws. First, a PMR comparison cannot distinguish whether exposure causes the index causes of death or prevents the reference (nonindex) causes of death; exposure could also have some mixture of these effects (McDowall, 1983). For example, a proportional mortality study could find a proportional excess of cancer deaths among heavy aspirin users compared with nonusers of aspirin, but this finding might be attributable to a preventive effect of aspirin on cardiovascular deaths, which compose the great majority of noncancer deaths. An implicit assumption of a proportional mortality study is that the overall death rate for categories other than the ones under study is not related to the exposure.

The second major problem in mortality comparisons is that they cannot determine the extent to which exposure causes the index causes of death or worsens the prognosis of the illnesses corresponding to the index causes. For example, an association of aspirin use with stroke deaths among all deaths could be due to an aspirin effect on the incidence of strokes, the severity of strokes, or some combination of these effects.

The ambiguities in interpreting a PMR are not necessarily a fatal flaw, since the measure will often provide leads worth pursuing about causal relations. In many situations, it may be only one or a few narrow causes of death that are of interest, and it may be judged implausible that an exposure would substantially affect prognosis or reference deaths. Nonetheless, many of the difficulties in interpreting proportional mortality studies can be mitigated by considering a proportional mortality study as a variant of the case-control study. To do so requires conceptualizing a combined population of exposed and unexposed individuals in which the cases occurred. The cases are those deaths, both exposed and unexposed, in the specific category or categories of interest; the controls are other deaths (Miettinen and Wang, 1981).

The principle of control series selection is to choose individuals representing the source population from which the cases came, to learn the distribution of exposure within that population. Instead of sampling controls directly from the source population, we can sample reference deaths occurring in the source population, provided that the exposure distribution among the deaths sampled is the same as the distribution in the source population; that is, the exposure should not be related to the control causes of death (McLaughlin et al., 1985). If we keep the objectives of control selection in mind, it becomes clear that we are not bound to select as controls all causes of death other than index cases. We can instead select as controls a limited set of reference causes of death, chosen on the basis of a presumed lack of association with the exposure. In this way, other causes of death for which a relation with exposure is known, suspected, or merely plausible can be excluded.

The principle behind selecting the control causes of death for inclusion in the study is identical to the principle of selecting a control series for any case-control study: The control series should be selected independently of exposure, with the aim of estimating the proportion of the source population experience that is exposed. Deaths from causes not included as part of the control series may be excluded from the study or may be studied as alternative case groups.

Treating a proportional mortality study as a case-control study can thus enhance study validity. It also provides a basis for estimating the usual epidemiologic measures of effect that can be derived from such studies (Wang and Miettinen, 1982). The conceptual clarity that results from considering proportional mortality studies as case-control studies warrants dropping the term *proportional mortality study* from the epidemiologist's lexicon, and ending use of the misleading PMR.

Ecologic Studies

All the study types described thus far share the characteristic that the observations made pertain to individuals. It is possible to conduct research in which the unit of observation is a group of people rather than an individual; such studies are called *ecologic* or *aggregate studies*. The groups may be classes in a school, factories, cities, counties, or nations. The only requirement is that information on the populations studied is available to measure the exposure and disease distributions in each group. Incidence and mortality are commonly used to quantify disease occurrence in groups. Exposure is also measured

by some overall index; for example, county alcohol consumption may be estimated from alcohol tax data, information on socioeconomic status is available for census tracts from the decennial census, and environmental data (temperature, air quality, and so on) may be available locally or regionally.

Because the data in ecologic studies are measurements averaged over individuals, the degree of association between exposure and disease need not reflect individual-level associations (Morgenstern, 1982; Richardson et al., 1987; Greenland and Robins, 1994). In addition, use of proxy measures for exposure (e.g., alcohol tax data rather than consumption data) and disease (mortality rather than incidence) further distort the associations (Brenner et al., 1992b). Finally, ecologic studies usually suffer from unavailability of data necessary for adequate control of confounding in the analysis (Greenland and Robins, 1994). All of these problems can combine to produce results that may be of questionable validity. Despite such problems, ecologic studies can be useful for detecting associations of exposure distributions with disease occurrence. Even if confounded by unknown or uncontrollable factors, such associations may signal the presence of effects worthy of further investigation. A detailed discussion of ecologic studies is presented in Chapter 23.

Hypothesis Generation Versus Hypothesis Screening

Studies in which validity is less secure have sometimes been referred to as "hypothesis-generating" studies to distinguish them from so-called "analytic studies," in which validity may be better. Ecologic studies have often been considered as hypothesis-generating studies because of concern about various biases. The distinction, however, between hypothesis-generating and analytic studies is not conceptually accurate. It is the investigator, not the study, that generates hypotheses, and any type of data may be used as a way of testing a hypothesis. For example, international comparisons indicate that Japanese women have a much lower breast cancer rate than women in the United States. These data are ecologic and subject to the usual concerns about the many differences that exist between cultures. Nevertheless, the finding corroborates a number of hypotheses, including the theories that early menarche, high-fat diets, and large breast size (all more frequent among U.S. women than Japanese women) may be important determinants of breast cancer risk (e.g., see Trichopoulos and Lipman, 1992). The international difference in breast cancer rates is neither hypothesis generating nor analytic, for the hypotheses arose independently of this finding. Thus, the distinction between hypothesis-generating and analytic studies is one that is best replaced by a more accurate distinction.

A proposal that we view favorably is to refer to preliminary studies of limited validity or precision as *hypothesis-screening studies*. In analogy with screening of individuals for disease, such studies represent a relatively easy and inexpensive test for the presence of an association between exposure and disease. If such an association is detected, it is subject to more rigorous and costly tests using a more valid study design. While the screening analogy should not be taken to an extreme, it does better describe the progression of studies than the hypothesis-generating/analytic study distinction.