Field Epidemiology

Second Edition

Edited by
MICHAEL B. GREGG

Selected Chapters

Chapter 7

DESIGNING STUDIES IN THE FIELD

Richard C. Dicker

For all but the most straightforward field investigations, you are likely to design and conduct an epidemiologic study of some sort. This is sometimes called an analytic study, to distinguish it from a descriptive study. The basic ingredients of epidemiologic studies consist of two groups: the observed group, such as a group of ill or exposed persons, and a comparison group, which provides baseline or "expected" data. Using these groups is an efficient way to evaluate hypotheses about causes of disease and risk factors that have been raised in earlier phases of the investigation. By comparing the observed data with the expected data from the comparison group, you can quantify the relationship between possible risk factors and disease, and can test the statistical significance of the various hypotheses that have been raised.

The gold standard for an epidemiologic study is an experimental study such as a therapeutic trial, in which study participants are enrolled, randomly assigned into intervention or nonintervention (placebo) exposure groups, and then monitored over time. In public health practice, however, epidemiologists rarely conduct such experiments, because they are seldom in a position to assign exposures—exposures have generally already occurred through genetics, circumstance, or choice. As a result, almost all studies conducted by field epidemiologists are observational studies, in which the epidemiologists document rather than determine exposures.

You will likely conduct two types of epidemiologic studies. In a *cohort* or *follow-up study*, enrollment of the study group is based on exposure characteristics or membership in a particular group. The occurrence of health-related outcomes (like diseases) is then determined and the frequency of those occurrences is compared among exposure groups. In a *case-control study*, enrollment is based on the presence ("case") or absence ("control") of disease, and the frequency of exposures is compared between the cases and controls. Each type of study has its strengths and limitations, but each has an important place in field investigations.

This chapter provides an overview of these two study designs, emphasizing methodologic considerations in the field. For more in-depth discussion of the theory and other features of study design, the reader is referred to other epidemiology texts.[1-3]

DEFINING EXPOSURE GROUPS

Since both cohort and case-control studies are used to quantify the relationship between exposure and disease, defining what is meant by "exposure" and "disease" is critical. In general, exposure is used quite broadly, meaning demographic characteristics, genetic or immunologic makeup, behaviors, environmental exposures, and other factors that might influence one's risk of disease.

Since precise exposure information is essential to accurate estimation of an exposure's effect on disease, exposure measures should be as objective and standard as possible. An exposure may be a relatively discrete event or characteristic, and developing a measure of exposure is conceptually straightforward, for example, whether a person ate the shrimp appetizer at Restaurant A or whether a person had received influenza vaccine this year. While these exposures may be straightforward in theory, they are subject to the whims of memory. Memory aids, such as Restaurant A's menu, and exposure documentation, such as a vaccination card or medical record, may help in these situations.

Some exposures can be subdivided by dose or duration (number of glasses of apple cider, number of years working in a coal mine). A pathogen may require a minimum (threshold) level of exposure to cause disease and may be more likely to cause disease with increasing exposures. The disease may require prolonged exposure or have a long latency or incubation period. These relationships may be missed by characterizing exposure simply as "yes" or "no." Similarly, the vehicle of infection, for example, may be a component or ingredient of other measured exposures. One could then create a composite measure, such as whether a person ate any item with mayonnaise as an ingredient.

Some exposures are subtle or difficult to quantify. Surrogate measures may be used (census track or level of education as a surrogate for socioeconomic status, which in turn may be a surrogate for access to health care, adequacy of housing, nutritional status, etc.), but should be interpreted with caution.

DEFINING OUTCOMES ("CASE DEFINITION")

A case definition is a set of standard criteria for deciding whether an individual should be classified as having the health condition of interest. A case definition consists of clinical criteria and, particularly in the setting of an outbreak investigation, certain restrictions on time, place, and person. The clinical criteria may include confirmatory laboratory tests, if available, or combinations of symptoms, signs, and other findings, but in general they should be kept simple and objective, for example, the presence of elevated antibody titers, three or more loose bowel movements per day, illness severe enough to require hospitalization, or primary hospital discharge diagnosis of ICD-9 code 480-486). The case definition may be restricted by time (e.g., to persons with onset of illness within the past 2 months), by place (e.g., to employees at a particular manufacturing plant or to residents of a town), and by person (e.g., to persons who had previously tested negative for chlamydia or to children at least 9 months old). Whatever the criteria, they must be

applied consistently and without bias to all persons under investigation to ensure that persons with illness are characterized consistently over time, locale, and clinical practice.

A case definition can have degrees of certainty, for example, a suspect case (usually based on clinical and sometimes epidemiologic criteria) versus a confirmed case (based on laboratory confirmation). For example, during an outbreak of measles, a person with fever and rash may be categorized as having a suspect, probable, or confirmed case of measles, depending on the strength of the additional laboratory and epidemiologic evidence. Sometimes a case is temporarily classified as suspect or probable while awaiting laboratory results. Depending on the lab results, the case will be reclassified as either confirmed or "not a case." Sometimes a case is permanently classified as suspect or probable, because, in the midst of a large outbreak of a known agent, investigators need not use precious time and resources to identify the same agent from every person with consistent clinical findings and history of exposure.

The case definition may also vary depending on the purpose. For case finding in a local area, the case definition should be relatively sensitive to capture as many potential cases as possible, that is, throw the net wide. However, for enrolling persons into an epidemiologic study to identify risk factors, a relatively specific or narrow case definition will minimize misclassification and bias.

For an epidemiologic study, a definition for controls may be just as important as the definition for cases. That is, since misclassification and bias may result if some controls actually have the disease under study, you may wish to adopt a control definition to exclude persons with mild or asymptomatic cases of the disease. In a study of a cluster of thyrotoxicosis, a surprise finding was that 75% of asymptomatic family members of cases had elevated thyroid function tests.[4] Had these family members been enrolled as controls, the epidemiologic study would not have identified a difference in exposure between cases and controls, and the association with consumption of locally produced ground beef that inadvertently included bits of thyroid gland would have been missed.

COHORT STUDIES

In concept, a cohort study, like an experimental study, begins with a group of persons without the disease under study but with different exposure experiences, and follows them over time to find out if they develop disease or a health condition of interest. In a cohort study, though, each person's exposure status is merely recorded rather than assigned randomly by the investigator. Then, the occurrence of disease among persons with different exposures is compared to assess whether the exposures are associated with increased risk of disease.

A cohort study sometimes begins by enrolling everyone in a population regardless of exposure status, then characterizing each person's exposure status after enrollment. Alternatively, a sample rather than the whole population could be enrolled. The enrollees are then followed over time for occurrence of the disease(s) of interest. Examples of large cohort studies that span many years include the Framingham Study, a study of cardiovascular disease among residents of Framingham, Massachusetts,[5] and the Nurses' Health Study, a study of the effects of oral contraceptives, diet, and lifestyle risk factors

among over 100,000 nurses.[6] Another example of this type of cohort study is one that enrolls all employees of a manufacturing plant before ascertaining each person's job type or exposure to a manufacturing process or chemical. A third example is a study that enrolls all persons who attended a banquet, then elicits food consumption histories to determine exposure. Note that in cohort studies that enroll all or a sample of a population without regard to exposure status, a wide variety of exposures as well as a wide variety of outcomes can be examined.

A cohort study can also begin with the enrollment of persons based on their exposure status. In this type of cohort study, two or more groups defined by their exposure status are enrolled. For example, an investigator may decide to enroll 100 persons exposed to some agent and 100 persons who were not exposed but are otherwise comparable. In this type of cohort study, while a wide variety of outcomes can be examined, assessment of exposure may be restricted to the one used to define the enrollment groups.

In a *prospective cohort study*, enrollment takes place before the occurrence of disease. In fact, any potential subject who is found to have the disease at enrollment will be excluded. Thus each subsequently identified case is an *incident* case. Incidence may be quantified as the number of cases over the sum of time that person was followed (*incidence rate*), or as the number of cases over the number of persons being followed (*attack rate* or *risk*). A major challenge for prospective studies is to maintain follow-up that is as complete as possible and comparable for each exposure group.

Note that, for a prospective study, disease should not have already occurred. Therefore, in field epidemiology, a prospective study is only likely to be conducted after a known exposure and a long incubation or latency period before illness. One example is the follow-up study of persons exposed to nuclear tests in Utah.[7] More commonly, cohort studies are conducted by field epidemiologists in response to a noted cluster or outbreak of disease in a well-defined population. A cohort study in which persons are enrolled after disease has already occurred is called a *retrospective cohort study*. In the typical "church picnic" outbreak where all or a representative sample of participants provide information on both their food exposures and whether they became ill, the investigator can calculate attack rates of disease in those who did or did not eat each food, and compare those attack rates to identify the food associated with the greatest increase in risk (see Appendix at end of book). This retrospective cohort type of study is the technique of choice for an acute outbreak in a well-defined population, particularly one for which a roster of names and contact information such as telephone numbers are available. Examples include not only the church picnic for which membership lists are available but also weddings and other gatherings, cruise ships, nursing homes, and schools. Retrospective cohort studies can also be used in a noninfectious disease context and are popular in occupational epidemiology. For example, a group of persons exposed to a worksite hazard years ago or over many years (e.g., workers exposed to vinyl chloride during the manufacturing process) and a comparable group not exposed (e.g., workers in a different part of the same plant) are constructed from available employment records, and the morbidity or mortality of the two groups is determined and compared[8] (see Chapter 17). However, when the population at risk is not known (e.g., as with

nationwide epidemics), the only expedient and scientifically sound way to analyze the problem is to use the case-control method.

CASE-CONTROL STUDIES

Whereas a cohort study proceeds conceptually from exposure to disease, a case-control study begins conceptually with disease and looks backward at prior exposures. Specifically, in a case-control study, a group of people with the disease of interest (cases or case-patients) and an appropriate group of people without disease are enrolled, and their prior exposures are ascertained. Differences in exposure between the two groups indicate an association between the exposure and disease under study.

Selection of Subjects

The case-control study begins with the identification of cases and the selection of controls. The case group represents the "observed" exposure experience, while the control group is needed to provide the "expected" level of exposure.

The cases in a case-control study must meet the case definition, that is, they must have the disease in question. The case definition must be independent of the exposure(s) under study. Ideally, the cases will be limited to new or incident cases rather than prevalent cases, so that the study does not confuse factors associated with disease occurrence with those associated with survival. Because field investigations rarely find all the cases, because you often are under strong pressure to find an answer, and because having only 70% to 80% of all cases is usually enough to perform an adequate study, you will usually attempt to enroll all persons who are eligible and meet the case definition. Since one goal of an analytic study is to quantify the relationship between exposure and disease, you should use a relatively narrow or specific case definition to ensure that cases truly have the disease—minimizing one source of misclassification bias.

A comparable group of controls must be identified and enrolled. While this statement is simple, debates about the selection of controls can be among the most complex in epidemiology.[9] The controls should not have the disease in question and, like the cases, should be identified independently of exposure. As a general rule, the controls should be representative of the population from which the cases arose, so that if a control had developed the disease, he or she would have been included as a case in the study. Suppose the cases are persons with community-acquired pneumonia admitted to a single hospital. The controls should be persons who would be admitted to the same hospital if they had the disease. This condition helps ensure comparability between cases and controls, since persons admitted to a different hospital may reflect a different population with a variety of different host characteristics and other exposures that may affect risk of disease. Commonly, controls for hospital-based cases are selected from the group of patients admitted to the same hospital, but with diagnoses other than the case-defining illness. Similarly, cases diagnosed in the outpatient setting may be compared to controls from the same clinical practices. Cases scattered through a community are often compared with community-based controls.

Controls should be free of the disease under study. This underscores the

importance of both the case definition and the control definition in distinguishing persons who have the disease from those who do not. In some studies, controls are required to have laboratory or other confirmation that they are disease free. In other studies, lack of symptoms and signs of illness are presumed to indicate absence of disease. However, the stricter the definition of the controls, the less opportunity for misclassification (enrolling someone with mild or asymptomatic disease as a control) and bias.

Consider the thyrotoxicosis outbreak mentioned earlier, with about 75% of asymptomatic family members with elevated thyroid function tests because they ate the same contaminated ground beef as the cases. Had the investigators not tested the family members, and had the family members thus been included in the control group, they would have had exposures similar to the cases, making the exposure-disease association harder to identify.

In general, controls should be at risk for the disease. While this can be challenged on academic grounds, the assertion has face validity and needs little justification. For example, in a case-control study of risk factors for uterine cancer, most epidemiologists would not include men in the control group. While men might adequately represent the distribution of A-B-O blood groups in the population, they surely would represent an inappropriate estimate of the "expected" levels of sexual activity, contraceptive choices, and the like.

Sometimes the choice of a control group is so vexing that investigators decide to use more than one type of control group. For example, in a study where the cases are persons hospitalized with West Nile encephalitis, you might want to select a hospital-based control group (since only a minority of persons with West Nile infection require hospitalization and are the cases most easily found) and a community-based control group. If the two control groups provide similar results and conclusions about risk factors for West Nile infection, then the credibility of the findings is increased. On the other hand, if the two control groups yield conflicting results, then you must struggle to develop plausible explanations.

Types of Controls

Controls come from a variety of sources, each with potential strengths and weaknesses. As noted previously, two of the guiding principles in selecting a control group are whether they represent the population from which the cases came, and whether they will provide a good estimate of the level of exposure one would expect in that population. Some common sources of controls include persons served by the same health-care institutions or providers as the cases; members of the same institution or organization; relatives, friends, or neighbors; or a random sample of the community from which the cases came.

For outbreaks in hospitals or nursing homes, the source of controls is usually other patients or residents of the facility. For example, in the investigation of postoperative surgical site infections the epidemiologist might select as controls persons who had similar surgery but who did not develop postoperative infections. The advantages of using such controls are that they come from the same catchment area as the cases, have similar access to medical care, have comparable medical records, have time

on their hands, and are usually cooperative. The disadvantage is that they may have conditions that are associated either positively or negatively with the disease or risk factors of interest. For example, hospitalized patients are more likely to be current or former smokers than the general population. Depending on the disease and risk factors under study, the best strategy may be to select controls with only a limited number of diagnoses known to be independent of the exposures and disease, or, alternatively, to select controls with as broad a range of diagnoses as possible, so that no one diagnosis has undue influence.

In other settings with a well-defined or easily enumerated population, controls generally come from lists of persons in that population who did not become ill. For example, controls for an outbreak of nausea, lightheadedness, and fainting among seventh graders at a middle school might be seventh grade students at the same school who did not experience those symptoms. Similarly, on a cruise ship, controls might be selected as a random sample of well passengers or perhaps cabin mates of cases who ate together but remained well. These population- based controls have advantages similar to those listed for hospital-based controls, but without the disadvantage of having another disease.

When an outbreak occurs in a community at large, controls may be randomly selected from that community. However, the epidemiologist is not likely to have an available list of all persons from which to choose. Therefore, he or she must enlist controls either by telephoning a randomly or systematically selected set of telephone numbers, or by mailings to residents, or by conducting a door-to-door neighborhood survey. Each approach has its relative strengths and weaknesses, and associated potential biases. For example, both telephone dialing and door-to-door canvassing are labor intensive and are best done in the evenings when people are likely to be home. Even so, the public has become wary of telephone solicitations and even more so of strangers, however well intentioned, knocking on their doors. Mailings require far less labor but have notoriously low response rates, and those who respond may be a skewed rather than representative group (see Chapter 11).

When an investigation is not limited to a specific location but, for instance, involves the entire United States (e.g., toxic shock syndrome and tampon use or HIV infection and sexual practices), the selection of an appropriate control group is not as straightforward. In such circumstances epidemiologists have successfully used friends, relatives, or neighbors as controls. Typically, the investigator interviews a case, then asks for the names and telephone numbers of perhaps three friends to call as possible controls. One advantage is that the friends of an ill person are usually quite willing to participate, knowing that their cooperation may help solve the puzzle. On the other hand, they may be too similar to the cases, sharing personal habits and other exposures. The consequence of enrolling controls who are too similar to the cases—called "overmatching"—makes it harder to identify exposure-disease associations.

Sampling Methods for Selecting Controls

A variety of approaches can be used to select controls, depending on the hypotheses to be evaluated, the urgency of the investigation, the resources available, and the setting.

All persons at risk

Occasionally, an outbreak occurs in a well-defined, relatively small population. Examples include a food-borne outbreak among persons who attended a wedding, or a nosocomial outbreak among patients in the intensive care unit of a hospital. All persons with the disease under study could be called cases, and all persons who did not become ill could be called controls. However, since the entire population is available for study, you could and should analyze the data as a cohort study, computing and comparing rates of disease among exposed and unexposed groups, rather than analyzing the data in case-control fashion.

Random or systematic sampling

When an outbreak occurs in a population with a large number of potential controls, you can choose a random or systematic sample of the population. If a roster is available, you could choose a random sample by using a table or computer-generated list of random numbers to select individuals. For a systematic sample, you would select every tenth or thirtieth (or other appropriate interval) person on the list. When no roster is available you might resort to the technique called "random digit dialing," dialing random telephone numbers with the same area codes and exchanges as the cases. Whichever strategy is used, potential controls with symptoms and signs similar to the cases should either be excluded, or if they meet the definition for a case, be evaluated and included as cases, if appropriate.

Pair matching

Pair matching is the selection of one or more controls for each case, who have the same or similar specified characteristics as that case. For example, if the criteria for pair matching were same gender, school, and grade as the case, and the control-to-case ratio were one-to-one, then a female ninth grade case at Lincoln High School would need to be matched to a female Lincoln High School ninth grade control. Although the term *pair matching* implies one case and one control, the term may also refer to two, three, or even four controls matched to each case.

In field epidemiology, pair matching is used in two circumstances—to control for potential confounding, or for logistical ease. In the first circumstance, one or more factors may be suspected to confound the relationship between exposure and disease; that is, the factor may be linked to the exposure and, independently, be a risk factor for the disease. To help eliminate the intertwining of the effect of the confounder with the effect of the other exposures of interest, the epidemiologist may choose to match on the confounder. The result is that the cases and controls are the same in terms of the confounding factor, and, when analyzed properly, any apparent association between the exposure and the disease cannot be due to confounding by the matching factor. Note that matching in the design of the study, that is, choosing controls matched to the cases, requires the use of matched analysis methods (see Chapter 8).

A second reason for pair matching is simple expedience. As noted earlier, sometimes the quickest and most convenient method of selecting controls is to ask the cases for the names of friends, or to walk next door to a neighbor's home. This is pair matching because Jane's friend or neighbor in Seattle is not the friend or neighbor of Mary in Chicago. While such pair matching may be done for expedience, the net result is that cases and controls generally do wind up being matched for such difficult-to-measure factors such as socioeconomic status, cultural influence, exposure to local advertising, and the like.

Frequency matching

Frequency matching, also called category matching, is an alternative to pair matching. Frequency matching involves the selection of controls in proportion to the distribution of certain characteristics of the cases. For example, if 70% of the cases were ninth graders, 20% were eighth graders, and 10% were seventh graders, then the same proportion of controls would be selected from those grades. Frequency matching works best when all the cases have been identified before control selection begins.

Matching has several advantages. Matching is conceptually simple. It can save time and resources, as noted above with friend controls, and it can control for confounding by numerous social factors that are difficult to quantify and, hence, otherwise difficult to control for in an analysis. Finally, if the matching factor would have been a strong confounder, then matching improves the precision or power of the analysis.

However, matching has important disadvantages as well. First and foremost, if you match on a factor, you can no longer evaluate its effect on disease in your study, because you have made the controls and cases alike on that factor. For example, if infants with nosocomial infections in a neonatal intensive care unit were matched by birth weight to other newborns, then investigators would not be able to study birth weight itself as a risk factor for infection. Therefore, you should only match on factors that you do not need to evaluate. Second, if you use too many or too rigid matching criteria, it may be hard to find appropriate controls, and you may have to toss out cases if appropriately matched controls cannot be found. For example, one disadvantage of using sibling controls is that cases who are only children in a family have no eligible controls and cannot be included in the study. Finally, matching on factors that are not confounders tends to decrease a study's precision.

Size of Control Group

The size of the control group may be determined by circumstances, resources, or power considerations. Circumstance, for example, the number of eligible controls, sometimes is a limiting factor. At other times, time and resources may limit the number of controls that can be enrolled. However, when the size of the population from which the cases arose is large and resources are adequate, power calculations may be performed to determine the optimal number of controls needed to identify an important association. Most case-control studies use a control-to-case ratio of either 1:1, 2:1, or 3:1. In general, little power is gained with control-to-case ratios in excess of 3:1 or 4:1.

COMPARISONS OF COHORT AND CASE-CONTROL STUDIES

Some outbreaks occur in settings that are amenable to either a retrospective cohort or case-control study design. Others are better suited to one study type or the other. The advantages and disadvantages of these two approaches are listed in Table 7-1.

Risk Measurement

One of the most important advantages of the cohort design is that you can directly measure the disease risk (attack rate) of disease. This information is particularly important if the exposure is at the discretion of the individual. Only a cohort study can fill in the blank of "What is my risk of developing [name of disease] if I choose to [be exposed]?" The case-control study, with a set number of cases and an arbitrary number of controls, does not permit calculation of disease risk for a given exposure group.

Rare Exposure

Cohort studies are better suited than case-control studies for examining health effects following a relatively rare exposure. With a cohort approach, all persons

Table 7–1. Features of Case-Control and Retrospective Cohort Studies

FEATURE	CASE-CONTROL STUDY	RETROSPECTIVE COHORT STUDY
Sample size	Smaller	Larger
Costs	Less	More because of size
Study time	Short	Short
Rare disease	Efficient	Inefficient
Rare exposure	Inefficient	Efficient
Multiple exposures	Can examine	Often can examine
Multiple outcomes	Cannot examine	Can examine
Natural history	Cannot ascertain	Can ascertain
Disease risk	Cannot measure	Can measure
Recall bias	Potential problem	Potential problem
Loss to follow-up	Not an issue	Potential problem
Selection bias	Potential problem	Potential problem

with the exposure can be enrolled and monitored, as well as a sample of comparable persons who were not exposed. This rationale explains the popularity of retrospective cohort studies in occupational epidemiology, where a group of workers with an exposure common among that group but relatively rare in the community at large can be followed over time.

Rare Disease

Case-control studies are the design of choice for sporadic occurrences of an otherwise rare disease in a population. All cases and an appropriate number of controls can be enrolled and exposures evaluated for association with disease. In contrast, a cohort study

would have to enroll an extremely large number of persons to have enough with the outcome of interest.

POTENTIAL PITFALLS IN THE DESIGN AND CONDUCT OF EPIDEMIOLOGIC STUDIES

Designing and conducting a good epidemiologic study in the field is not easy. In designing a study you must make many choices. Many of these choices have no right answer but involve trade-offs or compromises between theory and practical issues such as time constraints and resources. Other choices involve deciding between two less-than-perfect options, such as two different control groups each with potential flaws. Some of the pitfalls that result from less-than-ideal study design and conduct are described below.

Selection Bias

Selection bias is a systematic error in choosing the study groups to be enrolled (e.g., cases and controls in a case-control study, exposed and unexposed groups in a cohort study) or in the enrollment of study participants that results in a mistaken estimate of an exposure-disease association. Consider, for example, a disease with low pathogenicity, that is, one with many asymptomatic cases. If a case-control study were conducted but controls were not tested for evidence of asymptomatic infection, then at least some of the controls may have the infection under study. The exposures among these mislabeled controls will be the same as the cases, resulting in an underestimate of the exposure-disease relationship. Another source of selection bias is diagnostic bias, in which knowledge of the exposure-disease hypothesis may prompt a clinician to make a diagnosis. For example, a physician may be more likely to diagnose pulmonary embolism in a woman he knows to be taking oral contraceptives—any subsequent analyses will show an association between oral contraceptives and pulmonary embolism! A third source of selection bias is nonresponse bias, in which persons who choose to participate may differ in important ways from persons who choose not to participate or cannot be found. In occupational epidemiology a well-known source of selection bias is called the *healthy worker effect*, wherein workers who remain on the job are, in general, more healthy and fit than the population at large, and comparisons between workers and the general population may not be appropriate. The list of types of selection bias is lengthy, so investigators must be careful to use an objective and consistent case definition; select controls that represent the population from which the cases arose, using objective and consistent control criteria; and work hard to promote high response rates among all groups.

Information Bias

Information bias is a systematic error in the collection of exposure or outcome data about the study participants that results in a mistaken estimate of an exposure's effect on the risk of disease. One of the most common types of information bias is recall bias, in which one group is more likely than the other to remember and report an exposure. For example, persons who developed severe diarrhea are very likely to have thought about all

the preceding meals and foods they had eaten, while healthy controls are not. Interviewer bias occurs when interviewers are more probing about exposures with the cases than with the controls. To minimize information bias, good studies use standard and pretested questionnaires or data collection forms, and interviewers or abstractors who are trained in the objective use of the forms. Memory aids, such as calendars, menus, or photographs of medications, can often aid participants' recall.

Confounding

Confounding is the distortion of an exposure-disease association by a third factor that is related to both exposure and disease. Consider, for example, a study of an investigational cancer drug versus "usual treatment." Suppose that most people who received the drug had early-stage disease, and most people who received usual treatment had later-stage disease. Then even if the investigational drug had no beneficial effect, it might look efficacious because its effect was intertwined with that of disease stage. For a factor to be a confounder it must be an independent risk factor for the disease, and it must be unequally distributed among the exposure groups. Since age is independently associated with almost every health condition imaginable, age automatically fulfills one of the two criteria for confounding, so it must always be considered a potential confounder.

In observational studies, confounding can be addressed through restriction, matching, stratified analysis, or modeling. Restriction means, simply, that the study population is limited to a narrowly defined population. In the investigational drug example above, if the study had been limited to persons with early-stage disease, the disease stage could not confound the results. Similarly, if age is a suspected confounder, the study could be limited to a narrow age range. Matching in the study design has been addressed previously in this chapter. Matching in the analysis, as well as stratified analysis and modeling, is addressed in Chapter 8.

Small Sample Size

Sample size and power calculations can provide estimates of the number of subjects needed to find an association that is statistically significant and that you consider important. In practice, the size of a study is sometimes limited by the number of cases, time, and resources available. While the two most popular measures of effect—risk ratio and odds ratio—are not influenced by the size of the study, their measures of precision—confidence intervals—and measures of statistical significance, such as chi-square tests and *p* values, are all affected by study size. Many an investigator has wished for a larger study after calculating a large and potentially important risk ratio or odds ratio that, alas, is not statistically significant and has a wide confidence interval. Would a larger study confirm the association statistically different from the null, or would it show that the apparent association was indeed just chance variation from the null? Often, the investigator will never know. Determination of an adequate sample size in advance could avoid this situation.

SUMMARY

Cohort and case-control studies are the two types of analytic studies used most commonly by field epidemiologists. They are effective mechanisms for evaluating—quantifying and testing—hypotheses suggested in earlier phases of the investigation. Cohort studies, which are oriented conceptually from exposure to disease, are appropriate in settings in which an entire population is well defined and available for enrollment, such as invited guests at a wedding reception. Cohort studies are also appropriate when you can define and enroll groups by exposure, such as employees working in different parts of a manufacturing plant. Case-control studies, on the other hand, are quite useful when the population is less clearly defined. Case-control studies, oriented from disease to exposure, identify persons with disease ("cases") through, say, surveillance, and a comparable group of persons without disease ("controls"), then the exposure experiences of the two groups are compared. While conceptually straightforward, the design of a good epidemiologic study requires many decisions including who make up an appropriate comparison group, whether or not to match, and how best to avoid potential biases.

REFERENCES

1. Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, Boston.
2. Schlesselman, J.J. (1982), *Case-Control Studies*. Oxford University Press, New York.
3. Breslow, N.E., Day, N.E. (1997). *Statistical Methods in Cancer Research*, vol 2. The *design and analysis of cohort studies*. IARC Scientific Publications, Lyon, France.
4. Hedberg, C.W., Fishbein, D.B., Janssen, R.S., et al. (1987). An outbreak of thyrotoxicosis caused by the consumption of bovine thyroid gland in ground beef. *N Engl J Med* 316, 993-98.
5. Dawber, T.R., Kannel, W.B., Lyell, L.P. (1963). An approach to longitudinal studies in a community: the Framingham study. *Ann NY Acad Sci* 107, 539-56.
6. Colditz, G.A. (1995). The Nurses' Health Study: a cohort of US women followed since 1976. *J Am Med Womens Assoc* 50, 40-44, 63.
7. Caldwell, G.G., Kelley, D.B., Zack, M., et al. (1983). Mortality and cancer frequency among military nuclear test (smoky) participants, 1957-1979. *J Am Med Assoc* 250, 620-24.
8. Waxweiler, R.J., Stringer, W., Wagoner, J.K., et al. (1976). Neoplastic risk among workers exposed to vinyl chloride. *Ann NY Acad Sci* 271, 40-48.
9. Wacholder, S., McLaughlin, J.K., Silverman, D.T., et al. (1992). Selection of controls in case-control studies: I. Principles. *Am J Epidemiol* 135, 1019-28.
10. Morse, L.J., Bryan, J.A., Hurley, J.P., et al. (1972). The Holy Cross College football team hepatitis outbreak. *J Am Med Assoc* 219, 706-8.

Chapter 8

ANALYZING AND INTERPRETING DATA

Richard C. Dicker

The purpose of many field investigations is to identify causes, risk factors, sources, vehicles, routes of transmission, or other factors that put some members of the population at greater risk than others of having an adverse health event. In some field investigations, identifying a "culprit" is sufficient; if the culprit can be eliminated, the problem is solved. In other field settings, the goal may be to quantify the relationship between exposure (or any population characteristic) and an adverse health event. Quantifying this relationship may lead not only to appropriate interventions but also to advances in knowledge about disease causation. Both types of field investigation require appropriate but not necessarily sophisticated analytic methods. This chapter describes the strategy for planning an analysis, methods for conducting the analysis, and guidelines for interpreting the results.

PREANALYSIS PLANNING

What to Analyze

The first step of a successful analysis is to lay out an analytic strategy in advance. A thoughtfully planned and carefully executed analysis is just as critical for a field investigation as it is for a protocol-based study. Planning is necessary to assure that the appropriate hypotheses will be considered and that the relevant data will be appropriately collected, recorded, managed, analyzed, and interpreted to evaluate those hypotheses. Therefore, the time to decide on what (and how) to analyze the data is before you design your questionnaire, *not* after you have collected the data. As illustrated in Figure 8-1, the hypotheses that you wish to evaluate drive the analysis. (These hypotheses are usually developed by considering the common causes and modes of transmission of the condition under investigation; talking with patients and with local medical and public health staff; observing the dominant patterns in the descriptive epidemiologic data; and scrutinizing the outliers in these data.) Depending on the health condition being investigated, the hypotheses should address the source of the agent, the mode (and vehicle or vector) of transmission, and the exposures that caused disease. They should obviously be testable, since the role of the analysis will be to evaluate them.

Once you have determined the hypotheses to be evaluated, you must decide which data to collect in order to test the hypotheses. (You will also need to determine the best study design to use, as describe in the previous chapter.) There is a saying in clinical medicine that "If you don't take a temperature, you can't find a fever".[1] Similarly, in field epidemiology, if you neglect to ask about a potentially important risk factor in the questionnaire, you cannot evaluate its role in the outbreak. Since the hypotheses to be

tested dictate the data you need to collect, the time to plan the analysis is before you design the questionnaire.

Questionnaires and other data collection instruments are not limited to risk factors, however. They should also include identifying information, clinical information, and descriptive factors. Identifying information (or ID codes linked to identifying information stored elsewhere) allows you to recontact the respondent to ask additional questions or provide follow-up information. Sufficient clinical information should be collected to determine whether a patient truly meets the case definition. Clinical data on spectrum and severity of illness, hospitalization, and sequelae may also be useful. Descriptive factors related to time, place, and person should be collected to adequately characterize the population, assess comparability between groups (cases and controls in a case-control study; exposed and unexposed groups in a cohort study), and help you generate hypotheses about causal relationships.

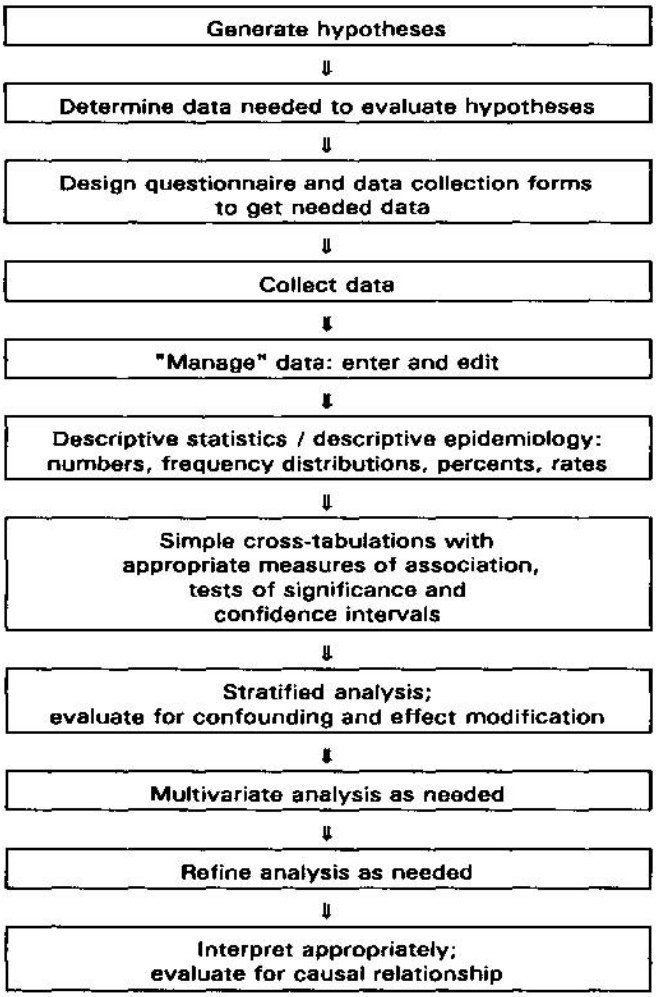

Figure 8-1. Steps in an analysis.

Data Editing

Usually, data for an analytic study are collected on paper questionnaires. These data are then entered into a computer. Increasingly, data are entered directly into a computer as they are obtained. In either situation, good data management practices will facilitate the analysis. These practices include, at the very least,

- Ensuring that you have the right number of records, with no duplicates
- Performing quality-control checks on each data field

Check that the number of records in the computerized database matches the number of questionnaires. Then check for duplicate records. It is not uncommon for questionnaires to be skipped or entered twice, particularly if they are not all entered at one sitting.

Two types of quality-control checks should be performed before beginning the analysis: range checks and logic (or consistency) checks. A range check identifies values for each variable that are "out of range" (i.e., not allowed, or at least highly suspicious). If, for the variable "gender," "male" is coded as 1 and "female" as 2, the range check should flag all records with any value other than 1 or 2. If 3's, F's, or blanks are found, review the original questionnaire, recontact the respondent, or recode those values to "known missing." For the variable "weight (in pounds)," an allowable range for adults might be 90 to 250. It is quite possible that some respondents will weigh more or less than this range, but it is also possible that values outside that range represent coding errors. Again, you must decide whether to attempt to verify the information or leave it as entered. The effort needed to confirm and complete the information should be weighed against the effect of lost data in the analysis—for a small study, you can ill afford missing data for the key variables but can tolerate it for less important variables. Under no circumstances should you change a value just because "it doesn't seem right."

A logic check compares responses to two different questions and flags those that are inconsistent. For example, a record in which "gender" is coded as "male" and "hysterectomy" is coded as "yes" should probably be flagged! Dates can also be compared—date of onset of illness should usually precede date of hospitalization (except in outbreaks of nosocomial infection, when date of hospitalization *precedes* date of onset) and date of onset should precede date of report. Again you must decide how to handle inconsistencies.

Two additional principles should guide data management. First, document everything, particularly your decisions. Take a blank copy of the questionnaire and write the name of each variable next to the corresponding question on the questionnaire. If, for the variable "gender," you decide to recode F's as 2's and recode 3's and blanks as 9's for "known missing," write those decisions down as well, so that you and others will know how to recode unacceptable values for gender in the future.

Note that you cannot create logic checks in advance for every possible contingency. Many inconsistencies in a database come to light during the analysis. Treat these inconsistencies the same way—decide how best to resolve the inconsistency (short of making up better data!) and then document your decision.

The second principle is, "Never let an error age." Deal with the problem as soon as you find it. Under the pressures of a field investigation, it is all too common to forget

about a data error, analyze the data as they are, and then be embarrassed during a presentation when calculations or values in a table do not seem to make sense.

Developing the Analysis Strategy

After the data have been edited, they are ready to be analyzed. But before you sit down to analyze the data, first develop an analysis strategy (Table 8-1). The analysis strategy is comparable to the outline you would develop before sitting down to write a term paper. It lays out the key components of the analysis in a logical sequence and provides a guide to follow during the actual analysis. An analytic strategy that is well planned in advance will expedite the analysis once the data are collected.

Table 8–1. Sequence of an Epidemiologic Analysis Strategy

1. Establish how the data were collected and plan to analyze accordingly.
2. Identify and list the most important variables in light of what you know about the subject matter, biologically plausible hypotheses, and the manner in which the study will be (or was) conducted:
   - Exposures of interest
   - Outcomes of interest
   - Potential confounders
   - Variables for subgroup analysis
3. To become familiar with the data, plan to perform frequency distributions and descriptive statistics on the variables identified in step 2.
4. To characterize the study population, create tables of clinical features and descriptive epidemiology (table shells should be created in advance).
5. To assess exposure-disease associations, create two-way tables based on study design, prior knowledge, and hypotheses (table shells should be created in advance).
6. Create additional two-way tables based on interesting findings in the data.
7. Create three-way tables, refinements (e.g., dose-response; sensitivity analysis) and subgroup analysis based on design, prior knowledge, hypotheses, or interesting findings in the data.

The first step in developing the analysis strategy is recognizing how the data were collected. For example, if you have data from a cohort study, think in terms of exposure groups and plan to calculate rates. If you have data from a case-control study, think in terms of cases and controls. If the cases and controls were matched, plan to do a matched analysis. If you have survey data, review the sampling scheme—you may need to account for the survey's design effect in your analysis.

The next step is deciding which variables are most important. Include the exposures and outcomes of interest, other known risk factors, study design factors such as variables you matched on, any other variables you think may have an impact on the analysis, and variables you are simply interested in. In a small questionnaire, perhaps all variables will be deemed important. Plan to review the frequency of responses and descriptive statistics for each variable. This is the best way to become familiar with the

data. What are the minimum, maximum, and average values for each variable? Are there any variables that have many missing responses? If you hope to do a stratified or subgroup analysis by, say, race, is there a sufficient number of responses in each race category?

The next step in the analysis strategy is sketching out table shells. A table shell (sometimes called a "dummy table") is a table such as a frequency distribution or two-way table that is titled and fully labeled but contains no data. The numbers will be filled in as the analysis progresses.

You should sketch out the series of table shells as a guide for the analysis. The table shells should proceed in a logical order from simple (e.g., descriptive epidemiology) to complex (e.g., analytic epidemiology). The table shells should also indicate which measures (e.g., odds ratio) and statistics (e.g., chi square) you will calculate for each table. Measures and statistics are described later in this chapter.

One way to think about the types and sequence of table shells is to consider what tables you would want to show in a report. One common sequence is as follows:

Table 1: Clinical features (e.g., signs and symptoms, percent lab-confirmed, percent hospitalized, percent died, etc.)
Table 2: Descriptive epidemiology
Time: usually graphed as line graph (for secular trends) or epidemic curve
Place: (county of residence or occurrence, spot or shaded map)
Person: "Who is in the study?" (age, race, gender, etc.)

For analytic studies,

Table 3: Primary tables of association (i.e., risk factors by outcome status)
Table 4: Stratification of Table 3 to separate effects and to assess confounding and effect modification
Table 5: Refinements of Table 3 (e.g., dose-response, latency, use of more sensitive or more specific case definition, etc.)
Table 6: Specific subgroup analyses

The following sequence of table shells (A through I) was designed before conducting a case-control study of Kawasaki syndrome (a pediatric disease of unknown cause that occasionally occurs in clusters). Since there is no definitive diagnostic test for this syndrome, the case definition requires that the patient have fever plus at least four of five other clinical findings listed in Table Shell A. Three hypotheses to be tested by the case-control study were the syndrome's purported association with antecedent viral illness, recent exposure to carpet shampoo, and increasing household income.

Since descriptive epidemiology has been covered in Chapter 5, the remainder of this chapter addresses the analytic techniques most commonly used in field investigations.

Table Shell A. Diagnostic Criteria for Kawasaki Syndrome Cases with Onset October–December

CRITERION	NUMBER	PERCENT
1. Fever ≥ 5 days	—	(%)
2. Bilateral conjunctival injection	—	(%)
3. Oral changes	—	(%)
Injected lips	—	(%)
Injected pharynx	—	(%)
Dry, fissured lips	—	(%)
Strawberry tongue	—	(%)
4. Peripheral extremity changes	—	(%)
Edema	—	(%)
Erythema	—	(%)
Periungual desquamation	—	(%)
5. Rash	—	(%)
6. Cervical lymphadenopathy > 1.5 cm	—	(%)

Table Shell B. Days of Hospitalization, Kawasaki Syndrome Cases with Onset October–December

DAYS OF HOSPITALIZATION	FREQUENCY
0	—
1	—
2	—
3	—
4	—
5	—
6	—
7	—
8	—
9	—
and so on to maximum	—
Unknown	—
Range:	—
Mean:	—
Median:	—

Table Shell C. Frequency Distribution of Serious Complications among Kawasaki Syndrome Cases with Onset October–December

CRITERION	NUMBER	PERCENT
Arthritis	__	(%)
Coronary artery aneurysm	__	(%)
Other complications (list:)	__	(%)
Death	__	(%)

Table Shell D. Demographic Characteristics of Kawasaki Syndrome Cases with Onset October–December

DEMOGRAPHIC CHARACTERISTIC		NUMBER	PERCENT
Age	< 1 yr	__	(%)
	1 yr	__	(%)
	2 yr	__	(%)
	3 yr	__	(%)
	4 yr	__	(%)
	5 yr	__	(%)
	≥ 6 yr	__	(%)
Gender	Male	__	(%)
	Female	__	(%)
Race	White	__	(%)
	Black	__	(%)
	Asian	__	(%)
	Other	__	(%)

Table Shell E. Frequency Distribution by County of Residence, Kawasaki Syndrome Cases, October–December

COUNTY	NUMBER	PERCENT	POPULATION	ATTACK RATE
County A	__	(%)	—	—
County B	__	(%)	—	—
County C	__	(%)	—	—
County D	__	(%)	—	—
County E	__	(%)	—	—
County F	__	(%)	—	—

Table Shell F. Frequency Distribution by Household Income, Kawasaki Syndrome Cases, October–December

ANNUAL HOUSEHOLD INCOME[a]	NUMBER	PERCENT
< $15,000	—	(%)
$15,000–$29,999	—	(%)
$30,000–$44,999	—	(%)
≥ $45,000	—	(%)

[a]May need to revise categories of household income to portray range.

Table Shell G. Kawasaki Syndrome and Antecedent Illness, Case Control Study

		CASES	CONTROLS	TOTAL	
ANTECEDENT ILLNESS	YES	—	—	—	Odds ratio = ___ 95% CI = (,)
	NO	—	—	—	x^2 = ____, *P* value = ____
	TOTAL	—	—	—	

Table Shell H. Kawasaki Syndrome and Carpet Shampoo, Case Control Study

		CASES	CONTROLS	TOTAL	
CARPET SHAMPOO	YES	—	—	—	Odds ratio = ___ 95% CI = (,)
	NO	—	—	—	x^2 = ____, *P* value = ____
	TOTAL	—	—	—	

Table Shell I. Kawasaki Syndrome and Carpet Shampoo, Case Control Study

		CASES	CONTROLS	TOTAL	
HOUSEHOLD INCOME (IN THOUSANDS OF DOLLARS)	<15	—	—	—	
	15–30	—	—	—	x^2 = ____,
	30–45	—	—	—	*P* value = ____
	45+	—	—	—	
	TOTAL	—	—	—	

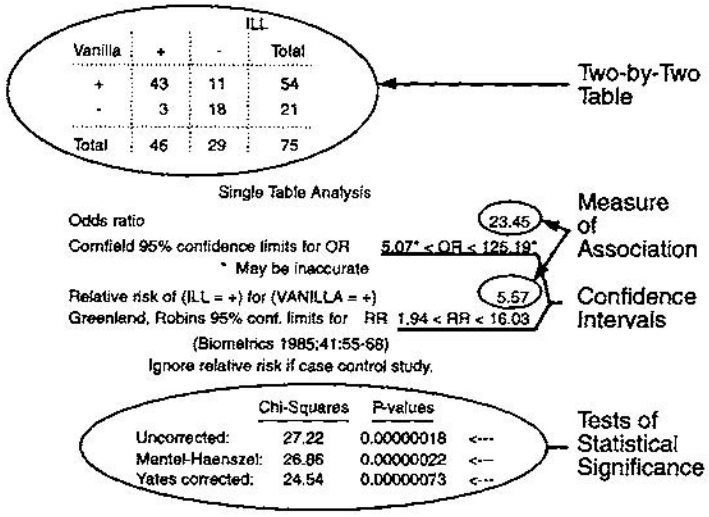

Figure 8-2. Typical *Epi Info* output from the analysis module, using the tables command. [Source: Dean, A.G. et al. 1994.[2]]

Figure 8-2 depicts a screen from *Epi Info*'s Analysis module (see Chapter 12). It shows the output from the "Tables" command for data from a typical field investigation. Note the four elements of the output: (1) a two-by-two table, (2) measures of association, (3) tests of statistical significance, and (4) confidence intervals. Each of these elements is discussed below.

The Two-by-Two Table

> "Every epidemiologic study can be summarized in a two-by-two table."
>
> —H. Ory

In many epidemiologic studies, exposure and the health event being studied can be characterized as binary variables (e.g., "yes" or "no"). The relationship between exposure and disease can then be cross-tabulated in a *two-by-two table*, so named because both the exposure and disease have just two categories (Table 8-2). One can put disease status

Table 8–2. Data layout and notation for standard two-by-two table.

	ILL	WELL	TOTAL	ATTACK RATE
EXPOSED	a	b	h_1	a/h_1
UNEXPOSED	c	d	h_0	c/h_0
TOTAL	v_1	v_0	t	v_1/t

(e.g., ill vs. well) along the top and exposure status along the side. (*Epi Info,* a microcomputer program written for field use, also follows this convention, although some epidemiologic textbooks do not [Chapter 12].) The intersection of a row and a column in which a count is recorded is known as a *cell.* The letters a, b, c, and d within the four cells of the two-by-two table refer to the number of persons with the disease status indicated in the column heading and the exposure status indicated to the left. For example, c is the number of unexposed ill/case subjects in the study. The *horizontal* row totals are labeled h_1 and h_0 (or h_2 and the *vertical* column totals are labeled v_1 and v_0 (or v_2). The total number of subjects included in the two-by-two table is written in the lower right corner and is represented by the letter t or n. Attack rates (the proportion of a group of people who develop disease during a specified time interval) are sometimes provided to the right of the row totals.

Data from an outbreak investigation in South Carolina are presented in Table 8-3. The table provides a cross-tabulation of turkey consumption (exposure) by presence or absence of *Salmonella* gastroenteritis (outcome). Attack rates (56.4 percent for those who ate turkey; 12.2 percent for those who did not) are given to the right of the table.

Table 8–3. Turkey Consumption and Gastrointestinal Illness, *Salmonella* Outbreak, South Carolina, 1990

		ILL	WELL	TOTAL	ATTACK RATE
ATE TURKEY?	YES	115	89	204	56.4%
	NO	5	36	41	12.2%
	TOTAL	120	125	245	49.0%

Source: Luby et al. (1993) [3]

MEASURES OF ASSOCIATION

A measure of association quantifies the strength or magnitude of the statistical association between the exposure and the health problem of interest. Measures of association are sometimes called measures of effect because—if the exposure is causally related to the disease—the measures quantify the effect of having the exposure on the incidence of disease. In cohort studies, the measure of association most commonly used is the relative risk. In case-control studies, the odds ratio is the most commonly used measure of association. In cross-sectional studies, either a prevalence ratio or a prevalence odds ratio may be calculated.

Relative Risk (Risk Ratio)

The relative risk is the risk in the exposed group divided by the risk in the unexposed group:

$$\text{Relative risk (RR)} = \text{risk}_{\text{exposed}} / \text{risk}_{\text{unexposed}} = (a/h_1) / (c/h_0)$$

The relative risk reflects the excess risk in the exposed group compared with the unexposed (background, expected) group. The excess is expressed as a ratio. In acute outbreak settings, risk is represented by the attack rate. The data presented in Table 8-3 show that the relative risk of illness, given turkey consumption, was 0.564/0.122 = 4.6. That is, persons who ate turkey were 4.6 times more likely to become ill than those who did not eat turkey. Note that the relative risk will be greater than 1.0 when the risk is greater in the exposed group than in the unexposed group. The relative risk will be less than 1.0 when the risk in the exposed group is less than the risk in the unexposed group, as is usually the case when the exposure under study is vaccination.

Odds Ratio (Cross-Product Ratio, Relative Odds)

In most case-control studies, because you do not know the true size of the exposed and unexposed groups, you do not have a denominator with which to calculate an attack rate or risk. However, using case control data, the relative risk can be approximated by an odds ratio. The odds ratio is calculated as

$$\text{Odds ratio (OR)} = ad/bc$$

In an outbreak of group A *Streptococcus* (GAS) surgical wound infections in a community hospital, 10 cases had occurred during a 17-month period. Investigators used a table of random numbers to select controls from the 2,600 surgical procedures performed during the epidemic period. Since many clusters of GAS surgical wound infections can be traced to a GAS carrier among operating room personnel, investigators studied all hospital staff associated with each patient. They drew a two-by-two table for exposure to each staff member and calculated odds ratios. The two-by-two table for exposure to nurse A is shown in Table 8-4. The odds ratio is calculated as 8 x 49/2 x 5 = 39.2. Strictly speaking, this means that the *odds* of being exposed to nurse A were 39 times higher among cases than among controls. It is also reasonable to say that the odds of developing a GAS surgical wound infection were 39 times higher among those exposed to nurse A than among those not exposed. For a rare disease (say, less than 5 percent), the odds ratio approximates the relative risk. So in this setting, with only 10 cases out of 2,600 procedures, the odds ratio could be interpreted as indicating that the *risk* of developing a GAS surgical wound infection was 39 times higher among those exposed to nurse A than among those not exposed.

Table 8–4. Surgical Wound Infection and Exposure to Nurse A, Hospital M, Michigan, 1980

		CASE	CONTROL	TOTAL
EXPOSED TO NURSE A?	YES	8	5	13
	NO	2	49	51
	TOTAL	10	54	64

Source: Berkelman et al. (1982). [4]

The odds ratio is a very useful measure of association in epidemiology for a variety of reasons. As noted above, when the disease is rare, a case-control study can yield an odds ratio that closely approximates the relative risk from a cohort study. From a theoretical statistical perspective (beyond the scope of this book), the odds ratio also has some desirable statistical properties and is easily derived from multivariate modeling techniques.

Prevalence Ratio and Prevalence Odds Ratio

Cross-sectional studies or surveys generally measure the prevalence (existing cases) of a health condition in a population rather than the incidence (new cases). Prevalence is a function of both incidence (risk) and duration of illness, so measures of association based on prevalent cases reflect both the exposure's effect on incidence and its effect on duration or survival.

The prevalence measures of association analogous to the relative risk and the odds ratio are the *prevalence ratio* and the *prevalence odds* ratio, respectively.

In the two-by-two table (Table 8-5), the prevalence ratio = 0.20/0.05 = 4.0. That is, exposed subjects are four times as likely as are unexposed subjects to have the condition. In the example above, the prevalence odds ratio = (20) (380) / (80) (20) = 4.75. The *odds* of having disease is 4.75 times higher for the exposed than the unexposed group. Note that when the prevalence is low, the values of the prevalence ratio and the prevalence odds ratio will be similar.

Table 8–5. Data from a Hypothetical Cross-Sectional Survey

		HAVE CONDITION?			
		YES	NO	TOTAL	PREVALENCE
EXPOSED?	YES	20	80	100	0.20
	NO	20	380	400	0.05
	TOTAL	40	460	500	

MEASURES OF PUBLIC HEALTH IMPACT

A measure of public health impact places the exposure-disease association in a public health perspective. It reflects the apparent contribution of an exposure to the frequency of disease in a particular population. For example, for an exposure associated with an increased risk of disease (e.g., smoking and lung cancer), the attributable risk percent represents the expected reduction in disease load if the exposure could be removed (or never existed). The population attributable risk percent represents the proportion of disease in a population attributable to an exposure. For an exposure associated with a decreased risk of disease (e.g., vaccination), a prevented fraction could be calculated that

represents the actual reduction in disease load attributable to the current level of exposure in the population.

Attributable Risk Percent (Attributable Fraction [or Proportion] among the Exposed, Etiologic Fraction)

The attributable risk percent is the proportion of cases in the exposed group presumably attributable to the exposure. This measure assumes that the level of risk in the unexposed group (assumed to be the baseline or background risk of disease) also applies to the exposed group, so that only the *excess* risk should be attributed to the exposure. The attributable risk percent can be calculated with either of the following formulas (which are algebraically equivalent):

$$\text{Attributable risk percent} = (\text{risk}_{\text{exposed}} - \text{risk}_{\text{unexposed}}) / \text{risk}_{\text{exposed}}$$
$$= (RR - 1) / RR$$

The attributable risk percent can be reported as a fraction or can be multiplied by 100 and reported as a percent. Using the turkey consumption data in Table 8-3, the attributable risk percent is (0.564 - 0.122) / 0.564 = 78.4 percent. Therefore, over three-fourths of the gastroenteritis that occurred among persons who ate turkey may be attributable to turkey consumption. The other 21.6 percent is attributed to the baseline occurrence of gastroenteritis in that population.

In a case-control study, if the odds ratio is thought to be a reasonable approximation of the relative risk, you can calculate the attributable risk percent as

$$\text{Attributable risk percent} = (OR-1) / OR$$

Population Attributable Risk Percent (Population Attributable Fraction)

The population attributable risk percent is the proportion of cases in the entire population (both exposed and unexposed groups) presumably attributable to the exposure. Algebraically equivalent formulas include

$$\text{Population attributable risk percent} = (\text{risk}_{\text{overall}} - \text{risk}_{\text{unexposed}}) / \text{risk}_{\text{overall}}$$
$$= P(RR-1) / [P(RR-1) + 1]$$

where P = proportion of population exposed = h_1/t

Applying the first formula to the turkey consumption data, the population attributable risk percent is (0.490 - 0.122) / 0.490 = 75.1 percent. In situations in which most of the cases are exposed, the attributable risk percent and population attributable risk percent will be close. For diseases with multiple causes (e.g., many chronic diseases) and uncommon exposures, the population attributable risk percent may be considerably less than the attributable risk percent.

The population attributable risk percent can be estimated from a population-based case-control study by using the OR to approximate the RR and by using the

proportion of controls exposed to approximate P; that is, $P = b/v_0$ (assuming that the controls are representative of the entire population).

Prevented Fraction in the Exposed Group (Vaccine Efficacy)

If the risk ratio is less than 1.0, you can calculate the prevented fraction, which is the proportion of potential new cases that would have occurred in the absence of the exposure. In other words, the prevented fraction is the proportion of potential cases prevented by some beneficial exposure, such as vaccination. The prevented fraction in the exposed group is calculated as

$$\text{Prevented fraction among the exposed} = (\text{risk}_{\text{unexposed}} - \text{risk}_{\text{exposed}} / \text{risk}_{\text{unexposed}}$$
$$= 1 - RR$$

Table 8-5 presents data from a 1970 measles outbreak along the Texas- Arkansas border. Because some cases had occurred among children vaccinated against measles, the public questioned the effectiveness of the measles vaccine. As shown in Table 8-6, the risk of measles among vaccinated children was about 4 percent of the risk among unvaccinated children. Vaccine efficacy was calculated to be 96 percent indicating that vaccination prevented 96 percent of the cases that might have otherwise occurred among vaccinated children had they not been vaccinated.

Note that the terms "attributable" and "prevented" convey much more than statistical association. They imply a cause-and-effect relationship between the exposure and disease. Therefore, these measures should not be presented routinely but only after thoughtful inference of causality.

Table 8–6. Vaccination Status and Occurrence of Measles, Texarkana, 1970

	MEASLES	NO MEASLES	TOTAL	RISK PER 1,000	RELATIVE RISK
VACCINATED	27	6,323	6,350	4.2	0.04
NOT VACCINATED	512	4,323	4,835	105.9	(REFERENCE)
TOTAL	539	10,646	11,185	48.2	

Vaccine efficacy = (105.9 – 4.2) / 105.9 = 0.96

Source: Landrigan (1972). 5

TESTS OF STATISTICAL SIGNIFICANCE

Tests of statistical significance are used to determine how likely it is that the observed results could have occurred by chance alone, if exposure was not actually related to disease. In the paragraphs below, we describe the key features of the tests most commonly used with two-by-two tables. For discussion of theory, derivations, and other topics beyond the scope of this book, we suggest that you consult one of the many biostatistics textbooks, which cover these subjects well.

In statistical testing, you assume that the study population is a sample from some large "source population." Then assume that, in the source population, incidence of disease is the same for exposed and unexposed groups. In other words, assume that, in the source population, exposure is not related to disease. This assumption is known as the *null hypothesis*. (The *alternative hypothesis*, which may be adopted if the null hypothesis proves to be implausible, is that exposure *is* associated with disease.) Next, compute a measure of association, such as a relative risk or odds ratio. Then, calculate the test of statistical significance such as a chi square (described below). This test tells you the probability of finding an association as strong as (or stronger than) the one you have observed if the null hypothesis were really true. This probability is called the *P value*. A very small *P* value means that you would be very unlikely to observe such an association if the null hypothesis were true. In other words, a small *P* value indicates that the null hypothesis is implausible, given the data at hand. If this *P* value is smaller than some predetermined cutoff (usually 0.05 or 5 percent), you can discard ("reject") the null hypothesis in favor of the alternative hypothesis. The association is then said to be "statistically significant."

In reaching a decision about the null hypothesis, be alert to two types of error. In a *type I error* (also called *alpha error*), the null hypothesis is rejected when in fact it is true. In a *type II error* (also called *beta error*), the null hypothesis is not rejected when in fact it is false.

Both the null hypothesis and the alternative hypothesis should be specified in advance. When little is known about the association being tested, you should specify a null hypothesis that the exposure is not related to disease (e.g., RR = 1 or OR = 1). The corresponding alternative hypothesis states that exposure and disease are associated (e.g., RR ? 1 or OR ? 1). Note that this alternative hypothesis includes the possibilities that exposure may either increase or decrease the risk of disease.

When you know more about the association between a given exposure and disease, you may specify a narrower ("directional") hypothesis. For example, if it is well established that an exposure increases the risk of developing a particular health problem (e.g., smoking and lung cancer), you can specify a null hypothesis that the exposure does not increase risk of that condition (e.g., RR = 1 or OR = 1) and an alternative hypothesis that exposure does increase the risk (e.g., RR > 1 or OR > 1). Similarly, if you were studying a well-established protective relationship [measles-mumps-rubella (MMR) vaccine and measles], you could specify a null hypothesis that RR = 1 and an alternative hypothesis that RR < 1.

A nondirectional hypothesis is tested by a "two-tailed" test. A directional hypothesis is tested with a "one-tailed" test. In general, the cutoff for a one-tailed test is twice the cutoff of a two-tailed test (i.e., 0.10 rather than 0.05). Since raising the cutoff for rejecting the null hypothesis increases the likelihood of making a type I error, epidemiologists in field situations generally use a two-tailed test.

Two different tests, each with some variations, are used for testing data in a two-by-two table. These two tests, described below, are the Fisher exact test and the chi-square test. These tests are not specific to any particular measure of association. The same test can be used regardless of whether you are interested in risk ratio, odds ratio, or attributable risk.

Fisher Exact Test

The Fisher exact test is considered the "gold standard" for a two-by-two table and is the test of choice when the numbers in a two-by-two table are small. Assume that the null hypothesis is true in the source population and that the values in the four cells but not the row and column totals of the two-by-two table could change. The Fisher exact test involves computing the probability of observing an association in a sample equal to or greater than the one observed. The technique for deriving this probability is outlined in Appendix 8-1.

As a rule of thumb, the Fisher exact test is the test of choice when the *expected* value in any cell of the two-by-two table is less than 5. The expected value is calculated by multiplying the row total by the column total and dividing by the table total. However, calculating the Fisher exact test, which is tedious at best for small numbers, becomes virtually impossible when the numbers get large. Fortunately, with large numbers, the chi-square test provides a reasonable approximation to the Fisher exact test.

Chi-Square Test

When you have at least 30 subjects and the expected value in each cell of the two-by-two table is at least 5, the chi-square test provides a reasonable approximation to the Fisher exact test. Plugging the appropriate numbers into the chi-square formula, you get a value for the Chi-square. Then look up its corresponding two-tailed *P* value in a chi-square table (see Appendix 8-2). A two-by-two table has one degree of freedom,* and a chi-square larger than 3.84 corresponds to a two-tailed *P* value smaller than 0.05.

At least three different formulas of the chi-square for a two-by-two table are in common use; *Epi Info* presents all three.

$$\text{Pearson uncorrected } \chi^2 = \frac{t\,(ad-bc)^2}{(v_1)\,(v_0)\,(h_1)\,(h_0)}$$

$$\text{Yates corrected } \chi^2 = \frac{t\left(\left|ad-bc\right| - \left(\frac{t}{2}\right)\right)^2}{(v_1)\,(v_0)\,(h_1)\,(h_0)}$$

$$\text{Mantel-Haenzel } \chi^2 = \frac{(t-1)\,(ad-bc)^2}{(v_1)\,(v_0)\,(h_1)\,(h_0)}$$

*Degrees of freedom equals the number of rows in the table minus1 times the number of columns in the table minus 1. So for a two-by-two table, degrees of freedom = (2 - 1) x (2 - 1) =1.

For a given set of data in a two-by-two table, the Pearson chi-square formula gives the largest chi-square value and hence the smallest *P* value. This *P* value is often somewhat smaller than the "gold standard" *P* value calculated by the Fisher exact method. So the Pearson chi-square is more apt to lead to a type I error (concluding that there is an association when there is not). The Yates corrected chi square gives the largest *P* value of the three formulas, sometimes even larger than the corresponding Fisher exact *P* value. The Yates correction is preferred by those epidemiologists who want to minimize their likelihood of making a type I error, but it increases the likelihood of making a type II error. The Mantel-Haenszel formula, popular in stratified analysis,

yields a *P* value which is slightly larger than that from the Pearson chi square but often smaller than the *P* value from the Yates corrected chi square and Fisher exact *P* value. Table 8-7 shows the data for macaroni consumption and risk of gastroenteritis from the South Carolina *Salmonella* outbreak. For these data, the Pearson and Mantel-Haenszel chi-square formulas yield *P* values smaller than 0.05 (the usual cutoff for rejecting the null hypothesis). In contrast, the corrected chi-square formula yields a *P* value closer to but slightly larger than the Fisher exact *P* value (the "gold standard"). Both *P* values are larger than 0.05, indicating that the null hypothesis should *not* be rejected. Fortunately, for most analyses the three chi-square formulas provide similar enough *P* values to make the same decision regarding the null hypothesis based on all three.

Table 8–7. Macaroni Consumption and Gastroenteritis, *Salmonella* Outbreak, South Carolina, 1990

	ILL	WELL	TOTAL	RISK	
EXPOSED	76	63	39	54.7%	Relative risk = 1.3
UNEXPOSED	44	62	106	41.5%	Odds ratio = 1.7
TOTAL	120	125	245		

$$\text{Uncorrected } \chi^2 = \frac{(245)\,(76 \times 62 - 63 \times 44)^2}{(120)\,(125)\,(139)\,(106)} = 4.17$$

$$\text{Mantel-Haenszel } \chi^2 = \frac{(245 - 1)\,(76 \times 62 - 63 \times 44)^2}{(120)\,(125)\,(139)\,(106)} = 4.16$$

$$\text{Corrected } \chi^2 = \frac{(245)\,[\left|76 \times 62 - 63 \times 44\right| - (\frac{245}{2})]^2}{(120)\,(125)\,(139)\,(106)} = 3.66$$

The corresponding two-tailed *P* values are as follows:

Uncorrected $\chi^2 = 4.17$, *P* value = 0.041
Mantel-Haenszel $\chi^2 = 4.16$, *P* value = 0.042
Corrected $\chi^2 = 3.66$, *P* value = 0.056

Fisher exact *P* value (two-tail) = 0.053

Source: Luby et al. (1993) [3]

Which Test to Use?

The Fisher exact test should be used if the expected value in any cell is less than 5. Remember that the expected value for any cell can be determined by multiplying the row total by the column total and dividing by the table total.

If all expected values in the two-by-two table are 5 or greater, then you can choose among the chi-square tests. Each of the three formulas shown above has its advocates among epidemiologists, and *Epi Info* provides all three. Many field epidemiologists prefer the Yates corrected formula because they are least likely to make type I error (but most likely to make a type II error). Epidemiologists who frequently perform stratified analyses are accustomed to using the Mantel-Haenszel formula, so they tend to use this formula even for simple two-by-two tables.

Measure of Association versus Test of Significance

The measures of association, such as relative risk and odds ratio, reflect the strength of the relationship between an exposure and a disease. These measures are generally independent of the size of the study and may be thought of as the "best guess" of the true degree of association in the source population. However, the measure gives no indication of its reliability (i.e., how much faith to put in it).

In contrast, a test of significance provides an indication of how likely it is that the observed association may be due to chance. Although the chi-square test statistic is influenced both by the magnitude of the association and the study size, it does not distinguish the contribution of each one. Thus the measure of association and the test of significance (or a confidence interval, see below) provide complementary information.

Interpreting Statistical Test Results

"Not significant" does not necessarily mean "no association." The measure of association (relative risk, odds ratio) indicates the direction and strength of the association. The statistical test indicates how likely it is that the observed association may have occurred by chance alone. Nonsignificance may reflect no association in the source population but may also reflect a study size too small to detect a true association in the source population.

Statistical significance does not by itself indicate a cause-effect relationship. An observed association may indeed represent a causal relationship, but it may also be due to chance, selection bias, information bias, confounding, and other sources of error in the design, execution, and analysis of the study. Statistical testing relates only to the role of chance in explaining an observed association, and statistical significance indicates only that chance is an unlikely (though not impossible) explanation of the association. You must rely on your epidemiologic judgment in considering these factors as well as consistency of the findings with those from other studies, the temporal relationship between exposure and disease, biological plausibility, and other criteria for inferring causation. These issues are discussed at greater length in the last section of this chapter.

Finally, statistical significance does not necessarily mean public health significance. With a large study, a weak association with little public health (or clinical) relevance many nonetheless be "statistically significant." More commonly, relationships of public health and/or clinical importance fail to be "statistically significant" because the studies are too small.

CONFIDENCE INTERVALS FOR MEASURES OF ASSOCIATION

We have just described the use of a statistical test to determine how likely the difference between an observed association and the null state is consistent with chance variation. Another index of the statistical variability of the association is the *confidence interval*. Statisticians define a 95% confidence interval as the interval that, given repeated sampling of the source population, will include or "cover" the true association value 95 percent of the time. The confidence interval from a single study may be roughly interpreted as the range of values that, given the data at hand and in the absence of bias,

has a 95 percent chance of including the "true" value. Even more loosely, the confidence interval may be thought of as the range in which the "true" value of an association is likely to be found, or the range of values that is consistent with the data in your study.

The chi-square test and the confidence interval are closely related. The chi-square test uses the observed data to determine the probability (*P* value) under the null hypothesis, and you "reject" the null hypothesis if the probability is less than some preselected value, called alpha, such as 5 percent. The confidence interval uses a preselected probability value, alpha, to determine the limits of the interval, and you can reject the null hypothesis if the interval does not include the null association value. Both indicate the precision of the observed association; both are influenced by the magnitude of the association and the size of the study group. While both measure precision, neither addresses validity (lack of bias).

You must select a probability level (alpha) to determine limiting values of the confidence interval. As with the chi-square test, epidemiologists traditionally choose an alpha level of 0.05 or 0.01. The "confidence" is then 100 x (1 - alpha) percent (e.g., 95 percent or 99 percent).

Unlike the calculation of a chi square, the calculation of a confidence interval is a function of the particular measure of association. That is, each association measure has its own formula for calculating confidence intervals. In fact, each measure has several formulas. There are "exact" confidence intervals and a variety of approximations.

Interpreting the Confidence Interval

As noted above, a confidence interval is sometimes loosely regarded as the range of values consistent with the data in a study. Suppose that you conducted a study in your area in which the relative risk for smoking and disease X was 4.0, and the 95 percent confidence interval was 3.0 to 5.3. Your single best guess of the association in the general population is 4.0, but your data are consistent with values anywhere from 3.0 to 5.3. Note that your data are *not* consistent with a relative risk of 1.0; that is, your data are *not* consistent with the null hypothesis. Thus, the values that are included in the confidence interval and values that are excluded both provide important information.

The width of a confidence interval (i.e., the values included) reflects the precision with which a study can pinpoint an association such as a relative risk. A wide confidence interval reflects a large amount of variability or imprecision. A narrow confidence interval reflects little variability and high precision. Usually, the larger the number of subjects or observations in a study, the greater the precision and the narrower the confidence interval.

As stated earlier, the measure of association provides the "best guess" of our estimate of the true association. If we were in a casino, that "best guess" would be the number to bet on. The confidence interval provides a measure of the confidence we should have in that "best guess," that is, it tells us how much to bet! A wide confidence interval indicates a fair amount of imprecision in our best guess, so we should not bet too much on that one number. A narrow confidence interval indicates a more precise estimate, so we might want to bet more on that number.

Since a confidence interval reflects the range of values consistent with the data in a study, one can use the confidence interval to determine whether the data are consistent with the null hypothesis. Since the null hypothesis specifies that the relative risk (or odds ratio) equals 1.0, a confidence interval that includes 1.0 is consistent with the null hypothesis. This is equivalent to deciding that the null hypothesis cannot be rejected. On the other hand, a confidence interval that does not include 1.0 indicates that the null hypothesis should be rejected, since it is inconsistent with the study results. Thus the confidence interval can be used as a test of statistical significance.

SUMMARY EXPOSURE TABLES

If the goal of the field investigation is to identify one or more vehicles or risk factors for disease, it may be helpful to summarize the exposures of interest in a single table, such as Table 8-8. For a food-borne outbreak, the table typically includes each food item served, numbers of ill and well persons by food consumption history, food-specific attack rates (if a cohort study was done), relative risk (or odds ratio), chi square and/or *P* value, and, sometimes, a confidence interval. To identify a culprit, you should look for a food item with two features:

1. An elevated relative risk, odds ratio, or chi square (small *P* value), reflecting a substantial difference in attack rates among those exposed to the item and those not exposed.
2. Most of the ill persons had been exposed, so that the exposure could "explain" most if not all of the cases.

In Table 8-8, turkey has the highest relative risk (and smallest *P* value) and can account for 115 of the 120 cases.

Table 8–8. Food-Specific Attack Rates for Persons Who Ate Sunday Lunch, *Salmonella* Outbreak, South Carolina, 1990[a]

	ATE			DID NOT EAT					
FOOD	# CASES	TOTAL	AR %	# CASES	TOTAL	AR %	RR	(95% CI)	P VALUE
Turkey	115	204	56	5	41	12	4.6	(2.0, 10.6)	<0.001
Ham	65	121	54	54	122	44	1.2	(0.9, 1.6)	0.178
Dressing	99	186	53	21	59	36	1.5	(1.0, 2.2)	0.027
Gravy	85	159	53	35	85	41	1.3	(1.0, 1.7)	0.090
Macaroni	76	139	55	44	106	42	1.3	(1.0, 1.7)	0.056
Beans	96	183	52	23	61	38	1.4	(1.0, 2.0)	0.065
Corn	80	153	52	40	92	43	1.2	(0.9, 1.6)	0.229
Rolls	78	158	49	41	84	49	1.0	(0.8, 1.3)	0.958
Butter	47	88	53	73	157	46	1.2	(0.9, 1.5)	0.365
Tea	102	203	50	18	42	43	1.2	(0.8, 1.7)	0.482
Coffee	9	28	32	111	217	51	0.6	(0.4, 1.1)	0.090
Cranberries	42	74	57	78	171	46	1.2	(1.0, 1.6)	0.144

[a]AR indicates attack rate; RR, relative risk, and CI, confidence interval.
Source: Luby et al. (1993). [3]

STRATIFIED ANALYSIS

Although it has been said that every epidemiologic study can be summarized in a two-by-two table, many such studies require more sophisticated analyses than those described so far in this chapter. For example, two different exposures may appear to be associated with disease. How do you analyze both at the same time? Even when you are only interested in the association of one particular exposure and one particular outcome, a third factor may complicate the association. The two principal types of complications are *confounding* and *effect modification*. Stratified analysis, which involves examining the exposure-disease association within different categories of a third factor, is one method for dealing with these complications.

Stratified analysis is an effective method for looking at the effects of two different exposures on the disease. Consider a hypothetical outbreak of hepatitis A among junior high school students. The investigators, not knowing the vehicle, administered a food consumption questionnaire to 50 students with hepatitis A and to 50 well controls. Two exposures had elevated odds ratios and statistically significant *P* values: milk and donuts (Table 8-9). Donuts were often consumed with milk, so many people were exposed to both or neither. How do you tease apart the effect of each item?

Stratification is one way to tease apart the effects of the two foods. First, decide which food will be the exposure of interest and which will be the stratification variable. Since donuts has the larger odds ratio, you might choose donuts as the primary exposure and milk as the stratification variable. The results are shown in Table 8-10. The odds ratio for donuts is 6.0, whether milk was consumed or not. Now, what if you had decided to look at the milk-illness association, stratified by donuts? Those results are shown in Table 8-11. Clearly, from Table 8-10, consumption of donuts remains strongly associated with disease, regardless of milk consumption. On the other hand, from Table 8-11, milk consumption is not independently associated with disease, with an odds ratio of 1.0 among those who did and did not eat donuts. Milk only *appeared* to be associated with illness because so many milk drinkers also ate donuts.

Table 8–9. Hepatitis A and Consumption of Milk and Donuts

MILK	CASES	CONTROLS	TOTAL	
EXPOSED	37	21	58	Odds ratio = 3.9
UNEXPOSED	13	29	42	Yates-corrected $\chi^2 = 9.24$
TOTAL	50	50	100	*P* value = 0.0002
DONUTS	**CASES**	**CONTROLS**	**TOTAL**	
EXPOSED	40	20	60	Odds ratio = 6.0
UNEXPOSED	10	30	40	Yates-corrected $\chi^2 = 15.04$
	50	50	100	*P* value = 0.0001

Table 8–10. Hepatitis A and Donut Consumption, Stratified by Milk

		DRANK MILK		DID NOT DRINK MILK			
		CASES	CONTROLS			CASES	CONTROLS
ATE DONUT?	YES	36	18	ATE DONUT?	YES	4	2
	NO	1	3		NO	9	27
		Odds ratio = 6.0				Odds ratio = 6.0	

Table 8–11. Hepatitis A and Milk Consumption, Stratified by Donuts

		ATE DONUT		DID NOT EAT DONUT			
		CASES	CONTROLS			CASES	CONTROLS
DRANK MILK?	YES	36	18	DRANK MILK?	YES	1	3
	NO	4	2		NO	9	27
		Odds ratio = 1.0				Odds ratio = 1.0	

An alternative method for analyzing two exposures is with a two-by-four table, as shown in Table 8-12. In that table, exposure 1 is labeled "EXP 1"; exposure 2 is labeled "EXP 2." To calculate the risk ratio for each row, divide the attack rate ("risk") for that row by the attack rate for the group not exposed to either exposure (bottom row in Table 8-12). To calculate the odds ratio for each row, use that row's values for a and b in the usual formula, ad/bc.

Table 8–12. Data Layout for Two-by-Four Table, Analyzing Two Exposures at Once

EXP 1	EXP 2	ILL	WELL	TOTAL	RISK	RISK RATIO	ODDS RATIO
Yes	Yes	a_{YY}	b_{YY}	h_{YY}	a_{YY}/h_{YY}	$Risk_{YY}/Risk_{NN}$	$a_{YY}d/b_{YY}c$
No	Yes	a_{NY}	b_{NY}	h_{NY}	a_{NY}/h_{NY}	$Risk_{NY}/Risk_{NN}$	$a_{NY}d/b_{NY}c$
Yes	No	a_{YN}	b_{YN}	h_{YN}	a_{YN}/h_{YN}	$Risk_{YN}/Risk_{NN}$	$a_{YN}d/b_{YN}c$
No	No	c	d	h_{NN}	c/h_{NN}	1.0 (Ref)	1.0 (Ref)

With this presentation, it is easy to see the effect of exposure 1 alone (row 3) compared with the unexposed group (row 4), exposure 2 alone (row 2) compared with the unexposed group (row 4), and exposure 1 and 2 together (row 1) compared with the unexposed group (row 4). Thus the separate and joint effects can be assessed.
From Table 8-13, you can see that donuts alone had an odds ratio of 6.0, whereas milk alone had an odds ratio of 1.0. Together, donuts and milk had an odds ratio of 6.0, the same as donuts alone. In other words, donuts, but not milk, were associated with illness. The two-by-four table summarizes the stratified tables in one and eliminates the need to designate one of the foods as the primary exposure and the other as the stratification variable.

Table 8–13. Hepatitis A and Consumption of Milk and Donuts, in Two-by-Four Table Layout

DONUT	MILK	CASE	CONTROL	ODDS RATIO
Yes	Yes	36	18	6.0
No	Yes	1	3	1.0
Yes	No	4	2	6.0
No	No	9	27	1.0 (Ref)

Confounding

Stratification also helps in the identification and handling of confounding. *Confounding is the distortion of an exposure-disease association by the effect of some third factor (a "confounder").* A third factor may be a confounder and distort the exposure-disease association if it is

- Associated with the outcome independent of the exposure—that is, even in the nonexposed group (In other words, it must be an independent "risk factor.")
- Associated with the exposure but not a consequence of it

To separate out the effect of the exposure from the effect of the confounder, stratify by the confounder.

Consider the mortality rates in Alaska versus Arizona. In 1988, the crude mortality rate in Arizona was 7.9 deaths per 1,000 population, over twice as high as the crude mortality rate in Alaska (3.9 deaths per 1,000 population). Is living in Arizona more hazardous to one's health? The answer is no. In fact, for most age groups, the mortality rate in Arizona is about equal to or slightly lower than the mortality rate in Alaska. The population of Arizona is older than the population of Alaska, and death rates rise with age. Age is a confounder that wholly accounts for Arizona's apparently elevated death rate—the age-adjusted mortality rates for Arizona and Alaska are 7.5/1000 and 8.4/1000, respectively. Note that age satisfies the two criteria described above: increasing age is associated with increased mortality, regardless of where one lives; and age is associated with state of residence (Arizona's population is older than Alaska's).

Return to the sequence in which an analysis should be conducted (Figure 8-1). After you have assessed the basic exposure-disease relationships using two-by-two tables, you should stratify the data by "third variables"—variables that are cofactors, potential confounders, or effect modifiers (described below). If your simple two-by-two table analysis has identified two or more possible risk factors, each should be stratified by the other or others. In addition, you should develop a list of other variables to be assessed. The list should include the known risk factors for the disease (one of the two criteria for a confounder) and matching variables. Then stratify or separate the data by categories of relevant third variables. For each stratum, compute a stratum-specific measure of association. Age is so often a real confounder that it is reasonable to consider it a potential confounder in almost any data set. Using age as an example, you could

separate the data by 10-year age groups (strata), create a separate two-by-two table of exposure and outcome for each stratum, and calculate a measure of association for each stratum.

The result of this type of analysis is that, within each stratum, "like is compared with like." If the stratification variable is gender, then in one stratum the exposure-disease relationship can be assessed for women and in the other the same relationship can be assessed for men. Gender can no longer be a confounder in these strata, since women are compared with women and men are compared with men.

To look for confounding, first look at the smallest and largest values of the stratum-specific measures of association and compare them with the crude value. If the crude value does not fall within the range between the smallest and largest stratum-specific values, confounding is surely present.

Often, confounding is not quite that obvious. So the next step is to calculate a summary "adjusted" measure of association as a weighted average of the stratum-specific values. The most common method of controlling for confounding is by stratifying the data and then computing measures that represent weighted averages of the stratum-specific data. One popular technique was developed by Mantel and Haenszel. This and other methods are described in Reference 6. After calculating a summary value, compare the summary value to the crude value to see if the two are "appreciably different." Unfortunately, there are no hard-and-fast rules or statistical tests to determine what constitutes "appreciably different." In practice, we assume that the summary adjusted value is more accurate. The question then becomes, "Does the crude value adequately approximate the adjusted value, or would the crude value be misleading to a reader?" If the crude and adjusted values are close, you can use the crude because it is not misleading and it is easier to explain. If the two values are appreciably different (10 percent? 20 percent?), use the adjusted value.

After deciding whether the crude or adjusted or stratum-specific measures of association are appropriate, you can then perform hypothesis testing and calculate confidence intervals for the chosen measures.

Effect Modification

The third use of stratification is in assessing effect modification. *Effect modification* means, simply, that the degree of association between an exposure and an outcome differs in different subgroups of the population. For example, a measles vaccine (exposure) may be highly effective (strong association) in preventing disease (outcome) if given after a child is 15 months of age (stratification variable = age at vaccination, stratum 1 = = 15 months), but less effective (weaker association) if given before 15 months (age stratum 2 = < 15 months). As a second example, tetracyc line (exposure) may cause (strong association) tooth mottling (outcome) among children (stratifier = age, stratum 1 = children), but tetracycline does not cause tooth mottling among adults. In both examples, the association or effect is a function of, or is modified by, some third variable. Effect modification is enlightening because it raises questions for further research. Why does the effect vary? In what way is one group different from the other? Studying these and related questions can lead to insights into pathophysiology, natural history of disease, and genetic or acquired host characteristics that influence risk.

Basically, evaluation for effect modification involves determining whether the stratum-specific measures of association differ from one another. Identification of effect modification is really a two-part process involving these questions:

1. Is the range of associations wide enough to be of public health or scientific importance? (A credo of field epidemiology is that "a difference, to be a difference, has to make a difference.")
2. Is the range of associations likely to represent normal sampling variation? Evaluation can be done either qualitatively ("eyeballing the results") or quantitatively (done with multivariate analysis such as logistic regression or with statistical tests of heterogeneity).

Another difference is important to note: confounding is extremely common because it is just an artifact of the data. True effect modification, on the other hand, usually represents a biological phenomenon and hence is much less common.

ADDITIONAL ANALYSES

Two additional areas are worth mentioning, although technical discussions are beyond the scope of this book. These two areas are the assessment of dose-response relationships and modeling.

Dose Response

In epidemiology, *dose-response* means increased risk of disease with increasing (or, for a protective exposure, decreasing) amount of exposure. Amount of exposure may reflect intensity of exposure (e.g., milligrams of L-tryptophan or number of cigarettes per day) or duration of exposure (e.g., number of months or years of exposure) or both.

If an association between an exposure and a health problem has been established, epidemiologists often take the next step to look for a dose-response effect. Indeed, the presence of a dose-response effect is one of the well-recognized criteria for inferring causation. Statistical techniques are available for assessing such relationships, even when confounders must be taken into account.

The first step, as always, is organizing your data. One convenient format is a 2-by-H table, where H represents the categories or doses of exposure.

As shown in Table 8-14, an odds ratio (or a risk ratio for a cohort study) can be calculated for each dose relative to the lowest dose or the unexposed group. You can calculate confidence intervals for each dose as well.

Merely eyeballing the data in this format can give you a sense of whether a dose-response relationship is present. If the odds ratios increase or decrease monotonically, a statistically significant dose-response relationship may be present. The Mantel extension test is one method of assessing the statistical significance of a dose-response effect. The mechanics of this test are described in Reference 7. The test yields a chi-square statistic with one degree of freedom.

Table 8-14. Data Layout and Notation for Dose-Response Table

	ILL	WELL		*Odds ratio*
Dose 5	a_5	b_5	h_5	a_5d/b_5c
Dose 4	a_4	b_4	h_4	a_4d/b_4c
Dose 3	a_3	b_3	h_3	a_3d/b_3c
Dose 2	a_2	b_2	h_2	a_2d/b_2c
Dose 1	a_1	b_1	h_1	a_1d/b_1c
Dose 0	c	d	h_0	1.0 (reference)
	v_1	v_0	t	

Modeling

There comes a time in the life of many epidemiologists when neither simple nor stratified analysis can do justice to the data. At such times, epidemiologists may turn to modeling. Modeling is a technique of fitting the data to particular statistical equations. One group of models are regression models, where the outcome is a function of exposure variables, confounders, and interaction terms (effect modifiers). The types of data usually dictate the type of regression model that is most appropriate. For example, logistic regression is the model most epidemiologists choose for binary outcome variables (ill/well, case/control, alive/dead, etc.).

In logistic regression, a binary outcome (dependent) variable is modeled as a function of a series of independent variables. The independent variables should include the exposure or exposures of primary interest and may include confounders and more complex interaction terms. Software packages provide beta coefficients for each independent term. If the model includes only the outcome variable and the primary exposure variable coded as (0,1), then $e^{ß}$ should equal the odds ratio you could calculate from the two-by-two table. If other terms are included in the model, then $e^{ß}$ equals the odds ratio adjusted for all the other terms. Logistic regression can also be used to assess dose-response relationships, effect modification, and more complex relationships. A variant of logistic regression called conditional logistic regression is particularly appropriate for pair-matched data.

Other types of models used in epidemiology include Cox proportional hazards models for life-table analysis, binomial regression for risk ratio analysis, and Poisson regression for analysis of rare-event data.

Keep in mind that *sophisticated analytic techniques cannot atone for sloppy data.* Analytic techniques such as those described in this chapter are only as good as the data to which they are applied. Analytic techniques, whether they be simple, stratified, or multivariate, use the information at hand. They do not ask or assess whether the proper comparison group was selected, whether the response rate was adequate, whether exposure and disease were properly defined, or whether the data coding and entry were free of errors. Analytic techniques are merely tools; as the analyst, you are responsible for knowing the quality of the data and interpreting the results appropriately.

MATCHING IN CASE-CONTROL STUDIES

Early in this chapter we noted that different study designs require different analytic methods. Matching is one design that requires methods different from those described so far. Because matching is so common in field studies, this section addresses this important topic.

Matching generally refers to a case-control study design in which controls are intentionally selected to be similar to case-subjects on one or more specified characteristics (other than the exposure or exposures of interest). The goal of matching, like that of stratified analysis, is to "compare like with like." The characteristics most appropriately specified for matching are those that are potential confounders of the exposure-disease associations of interest. By matching cases and controls on factors such as age, gender, or geographic area, the distribution of those factors among cases and controls will be identical. In other words, the matching variable will not be associated with case-control status in the study. As a result, if the analysis is properly done, the matching variable will not confound the association of primary interest.

Two types of matching schemes are commonly used in epidemiology. One type is *pair matching*, where each control is selected according to its similarity to a *particular* case. This method is most appropriate when each case is unique in terms of the matching factor, for example, 50 cases widely scattered geographically. Each case could be matched to a friend or neighborhood control. That control is suitably matched to that particular case-subject, but not to any other case-subject in the study. The matching by design into these unique pairs must be maintained in the analysis.

The term "pair matching" is sometimes generalized to include not only matched pairs (case and one control), but matched triplets (case and two controls), quadruplets, and so on. The term also refers to studies in which the number of matched controls per case varies, so long as the controls are matched to a specific case.

The other type of matching is *category matching*, also called *frequency matching*. Category matching is a form of stratified sampling of controls, wherein controls are selected in proportion to the number of cases in each category of a matching variable. For example, in a study of 70 male and 30 female case-subjects, if 100 controls were also desired, you would select 70 male controls at random from the pool of all non-ill males and 30 female controls from the female pool. The pairs are not unique; any male control is a suitable match to any male case-subject. Data collected by category matching in the study design must be analyzed using stratified analysis.

Matching has several advantages. Matching on factors such as neighborhood, friendship, or sibship may control for confounding by numerous social factors that would be otherwise impossible to measure and control. Matching may be cost- and time-efficient, facilitating enrollment of controls. For example, matched friend controls may be identified while interviewing each case-subject, and these friends are more likely to cooperate than controls randomly selected from the general population. And finally, matching on a confounder increases the statistical efficiency of an analysis and thus provides narrower confidence intervals.

Matching has disadvantages, too. The primary disadvantage is that matching on a factor prevents you from examining its association with disease. If the age and gender

distribution of case-subjects and controls are identical because you matched on those two factors, you cannot use your data to evaluate age and gender as risk factors themselves. Matching may be both cost- and time-inefficient, if considerable work must be performed to identify appropriately matched controls. The more variables to be matched on, the more difficult it will be to find suitably matched controls. In addition, matching on a factor that is not a confounder or having to discard cases because suitable controls could not be found decreases statistical efficiency and results in wider confidence intervals. Finally, matching complicates the analysis, particularly if other confounders are present.

In summary, matching is desirable and beneficial when you know beforehand that (1) you do not wish to examine the relationship between the matching factor and disease, (2) the factor is related to risk of disease so it is a potential confounder, and (3) matching is convenient or at least worth the potential extra costs to you. When in doubt, do not match, or match only on a strong risk factor that is likely to be distributed differently between exposed and unexposed groups and that is not a risk factor you are interested in assessing.

Matched Pairs

The basic data layout for a matched pair analysis appears at first glance to resemble the simple unmatched two-by-two tables presented earlier in this chapter, but in reality the two are quite different. In the matched-pair two-by-two table, each cell represents the number of matched pairs who meet the row and column criteria. In the unmatched two-by-two table, each cell represents the number of individuals who meet the criteria.

In Table 8-15, E+ denotes "exposed" and E- denotes "unexposed." Cell f thus represents the number of pairs made up of an exposed case and an unexposed control. Cells e and h are called *concordant pairs* because the case and control are in the same exposure category. Cells f and g are called *discordant pairs*.

In a matched-pair analysis, only the discordant pairs are informative. The odds ratio is computed as

$$\text{Odds ratio} = f / g$$

The test of significance for a matched pair analysis is the McNemar chi-square test. Both uncorrected and corrected formulas are commonly used.

$$\text{Uncorrected McNemar test} = \frac{(f - g)^2}{(f + g)}$$

$$\text{Corrected McNemar test} = \frac{(|f - g| - 1)^2}{(f + g)}$$

Table 8-16 presents the data from a pair-matched case-control study conducted in 1980 to assess the association between tampon use and toxic shock syndrome.[6]

Table 8–15. Data Layout and Notation for Matched-Pair Two-by-Two Table

		CONTROLS E+	CONTROLS E-	Total
CASES	E+	e	f	e + f
	E-	g	h	g + h
	TOTAL	e + g	f + h	e + f + g + h

Table 8–16. Continual Tampon Use during Index Menstrual Period in Case-Control Pairs, Toxic Shock Syndrome Study, 1980

		CONTROLS YES	CONTROLS NO	TOTAL
CASES	YES	33	9	42
	NO	1	1	2
	TOTAL	34	10	44 pairs

Odds ratio = 9 / 1 = 9.0

McNemar uncorrected chi-square test = $(9 - 1)^2 / (9 + 1) = 6.40$ ($P = 0.01$)

McNemar corrected chi-square test = $(|9 - 1| - 1)^2 / (9 + 1) = 4.90$ ($P = 0.03$)

Source: Shands et al. (1980). [8]

Matched Triplets

The data layout for a study in which two controls are matched to each case is shown in Table 8-17. Each cell is named f_{ij} where i is the number of exposed cases (1 if the case is exposed, 0 if the case is unexposed), and j is the number of exposed controls in the triplet. Thus cell f_{02} contains the number of triplets in which the case is unexposed but both controls are exposed.

A formula for calculating an odds ratio with *any* number of controls per case is

$$\text{OR} = \frac{\text{Number of unexposed controls matched with exposed cases}}{\text{Number of exposed controls matched with unexposed cases}}$$

For matched triplets, this formula reduces to

$$\text{Odds ratio} = \frac{2f_{10} + f_{11}}{2f_{02} + f_{01}}$$

Table 8-17 shows data from a case-control study of Kawasaki syndrome in Washington State.[9] For each of 16 cases-subjects, two age- and neighborhood-matched controls were identified. Although the study found no association with carpet cleaning, it did find the usual association with high household income (Table 8-18).

Table 8–17. Data Layout and Notation for a Matched Case-Control Study with Two Controls per Case

		PERCENT EXPOSED CONTROLS		
		2 of 2	1 of 2	0 of 2
CASES	E+	f_{12}	f_{11}	f_{10}
	E–	f_{02}	f_{01}	f_{00}

Table 8–18. Kawasaki Syndrome and Annual Household Income > $40,000, Washington State, 1986

		NUMBER OF EXPOSED CONTROLS		
		2 of 2	1 of 2	0 of 2
CASES	E+	0	1	7
	E–	0	4	4

Odds ratio = $(2 \times 7 + 1) / (2 \times 0 + 4) = 3.8$

Source: Dicker (1986). [9]

Larger Matched Sets and Variable Matching

Analogous analytic methods are available for matched sets of any fixed size and for sets with variable numbers of controls per case.[10] Such data are best analyzed with appropriate computer software, such as *Epi Info.*

Does a matched design require a matched analysis?

Does a matched design require a matched analysis? Usually, yes. In a pair- matched study, if the pairs are unique (siblings, friends, etc.), then pair-matched analysis is needed. If the pairs were based on a nonunique characteristic such as gender or race, stratified analysis is preferred. In a frequency matched study, stratified analysis is necessary.

In practice, some epidemiologists perform the appropriate matched analysis, then "break the match" and perform an unmatched analysis on the same data. If the results are similar, they may opt to present the data in unmatched fashion. In most instances, the unmatched odds ratio will be closer to 1.0 than the matched odds ratio ("bias toward the null"). Less frequently, the "broken" or unmatched odds ratio will be further from the null. These differences, which are related to confounding, may be trivial or substantial. The chi-square test result from unmatched data may be particularly misleading, usually being larger than the McNemar test result from the matched data. The decision to use a matched analysis or unmatched analysis is analogous to the decision to present crude or adjusted results. You must use your epidemiologic judgment in deciding whether the unmatched results are misleading to your audience or, worse, to yourself!

INTERPRETING FIELD DATA

> "Skepticism is the chastity of the intellect….
> Don't give it away to the first attractive hypothesis that comes along."
>
> M. B. Gregg,
> after George Santayana

Does an elevated relative risk or odds ratio or a statistically significant chi-square test mean that the exposure is a true cause of disease? Certainly not. Although the association may indeed be causal, flaws in study design, execution, and analysis can result in apparent associations that are actually artifacts. Chance, selection bias, information bias, confounding, and investigator error should all be evaluated as possible explanations for an observed association.

One possible explanation for an observed association is chance. Under the null hypothesis, you assume that your study population is a sample from some source population and that incidence of disease is not associated with exposure in the source population. The role of chance is assessed through the use of tests of statistical significance. (As noted above, confidence intervals can be used as well.) A very small *P* value indicates that the null hypothesis is an *unlikely* explanation of the result you found. Keep in mind that chance can never be ruled out entirely— even if the *P* value is small, say 0.01. Yours may be the one sample in a hundred in which the null hypothesis is true and chance *is* the explanation! Note that tests of significance only evaluate the role of chance. They do not say anything about the roles of selection bias, information bias, confounding, or investigator error, discussed below.

Another explanation for the observed explanation is selection bias. *Selection bias* is a systematic error in the study groups or in the enrollment of study participants that results in a mistaken estimate of an exposure's effect on the risk of disease. In more simplistic terms, selection bias may be thought of as a problem arising from who gets into the study. Selection bias may arise either in the design or in the execution of the study. Selection bias may arise from the faulty design of a case-control study if, for example, too loose a case definition is used (so some persons in the case group do not actually have the disease being studied), asymptomatic cases go undetected among the controls, or an inappropriate control group is used. In the execution phase, selection bias may result if eligible subjects with certain exposure and disease characteristics choose not to participate or cannot be located. For example, if ill persons with the exposure of interest know the hypothesis of the study and are more willing to participate than other ill persons, then cell a in the two-by-two table will be artificially inflated compared to cell c, and the odds ratio will also be inflated. So to evaluate the possible role of selection bias, you must look at how cases and controls were specified and how they were enrolled.

Another possible explanation of an observed association is information bias. *Information bias* is a systematic error in the collection of exposure or outcome data about the study participants that results in a mistaken estimate of an exposure's effect on the

risk of disease. Again, in more simplistic terms, information bias is a problem with the information you collect from the people in the study. Information bias may arise in a number of ways, including poor wording or understanding of a question on a questionnaire, poor recall (what did YOU have for lunch a week ago Tuesday?), or inconsistent interviewing technique. Information bias may also arise if a subject knowingly provides false information, either to hide the truth or, as is common in some cultures, in an attempt to please the interviewer.

As discussed earlier in this chapter, confounding can also distort an association. To evaluate the role of confounding, ensure that a list of potential confounders has been drawn up, that they have been evaluated for confounding, and that they have been controlled for as necessary.

Finally, investigator error has been known to be the explanation for some apparent associations. A missed button on a calculator, an erroneous transcription of a value, or use of the wrong formula can all yield artifactual associations! Check your work, or have someone else try to replicate it.

So before considering whether an association may be causal, consider whether the association may be explained by chance, selection bias, information bias, confounding, or investigator error. Now suppose that an elevated risk ratio or odds ratio has a small *P* value and narrow confidence interval, so chance is an unlikely explanation. Specification of cases and controls is reasonable and participation was good, so selection bias is an unlikely explanation. Information was collected using a standard questionnaire by an experienced and well-trained interviewer. Confounding by other risk factors was assessed and found not to be present or to have been controlled for. Data entry and calculations were verified. But before you conclude that the association is causal, you should consider the strength of the association, its biological plausibility, consistency with results from other studies, temporal sequence, and dose-response relationship, if any.

Strength of the association

In general, the stronger the association, the more likely one is to believe it is real. Thus we are generally more willing to believe that a relative risk of 9.0 may be causal than a relative risk of 1.5. This is not to say that a relative risk of 1.5 cannot reflect a causal relationship; it can. It is just that a subtle selection bias, information bias, or confounding could easily account for a relative risk of 1.5. The bias would have to be quite dramatic to account for a relative risk of 9.0!

Biological plausibility

Does the association make sense? Is it consistent with what is known of the pathophysiology, the known vehicles, the natural history of disease, animal models, or other relevant biological factors? For an implicated food vehicle in an infectious disease outbreak, can the agent be identified in the food, or will the agent survive (or even thrive) in the food? While some outbreaks are caused by new or previously unrecognized vehicles or risk factors, most are caused by those that we already know.

Consistency with other studies

Are the results consistent with those from other studies? A finding is more plausible if it can be replicated by different investigators, using different methods in different populations.

Exposure precedes disease

This criterion seems obvious, but in a retrospective study it may be difficult to document that exposure precedes disease. Suppose, for example, that persons with a particular type of leukemia are more likely to have antibodies to a particular virus, It might be tempting to conclude that the virus causes the leukemia, but from the serologic evidence at hand you could not be certain that exposure to the virus preceded the onset of leukemic changes.

Dose-response effect

Evidence of a dose-response effect adds weight to the evidence for causation. A dose-response effect is not a *necessary* feature for a relationship to be causal; some causal relationships may exhibit a threshold effect, for example. In addition, a dose-response effect does not rule out the possibility of confounding. Nevertheless, it is usually thought to add credibility to the association.

In many field investigations, a likely culprit may not meet all the criteria listed above. Perhaps the response rate was less than ideal, or the etiologic agent could not be isolated from the implicated food, or the dose-response analysis was inconclusive. Nevertheless, if the public's health is at risk, failure to meet every criterion should not be used as an excuse for inaction. As stated by George Comstock, "The art of epidemiologic reasoning is to draw sensible conclusions from imperfect data."[11] After all, field epidemiology is a tool for public health action to promote and protect the public's health based on science (sound epidemiologic methods), causal reasoning, and a healthy dose of practical common sense.

> "All scientific work is incomplete—whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action it appears to demand at a given time."[12]
>
> Sir Austin Bradford Hill

APPENDIX 8-1. FISHER EXACT TEST

The probability that the value in cell "a" is equal to the observed value, under the null hypothesis, is

$$\Pr(a) = \frac{(v_1)!\,(v_0)!\,(h_1)!\,(h_0)!}{t!a!b!c!d!}$$

where k! ("k factorial") = 1 x 2 x...x k,
(e.g.,5! =1x2x3x4x5= 120)

The easiest way to compute a two-tailed Fisher exact test is compute a one-tailed test and multiply by 2. Computing the one-tailed test is the hard part!

To compute the one-tailed Fisher exact test, first calculate the exact probability that cell a equals the observed value, using the formula shown above. Next, keeping all of the row and column totals the same, add or subtract 1 to the observed value in cell a to get a value even more extreme than the value observed. Modify the values in the other cells as necessary (add or subtract 1 to get the right row and column totals), and use the formula shown above to compute this new value's exact probability. Continue adding or subtracting 1 and computing probabilities until no more extreme values are possible without changing the marginal totals. Finally, sum these individual probabilities to get the one-tailed *P* value. For a two-tailed *P* value, add any smaller probabilities from the other tail.

Example:

	ILL	WELL	TOTAL	
EXPOSED	4	17	21	Odds ratio = undefined (cannot divide by zero)
UNEXPOSED	0	19	19	
TOTAL	4	36	40	

Based on the margins of this two-by-two table, cell a can take on values from 0 to 4, but none more extreme than 4. The probabilities for each value from 0 to 4 are

a	*b*	*c*	*d*	*Probability*	
4	17	0	19	4!36!21!19! / 40!4!17!0!19!	= 0.07
3	18	1	18	4!36!21!19! / 40!3!18!1!18!	= 0.28
2	19	2	17	4!36!21!19! / 40!2!19!2!17!	= 0.39
1	20	3	16	4!36!21!19! / 40!1!20!3!16!	= 0.22
0	21	4	15	4!36!21!19! / 40!0!21!4!15!	= 0.04.

Since there are no possible values of cell a more extreme than 4, the one-tailed *P* value is simply 0.07. The two-tailed *P* value is 0.07 + 0.04 = 0.11. Given a cutoff of 0.05, we could not reject the null hypothesis.

APPENDIX 8-2.

Chi-Square Table

DEGREE OF FREEDOM	PROBABILITY						
	0.50	0.20	0.10	0.05	0.02	0.01	0.001
1	0.455	1.642	2.706	3.841	5.412	6.635	10.827
2	1.386	3.219	4.605	5.991	7.824	9.210	13.815
3	2.366	4.642	6.251	7.815	9.837	11.345	16.268
4	3.357	5.989	7.779	9.488	11.668	13.277	18.465
5	4.351	7.289	9.236	11.070	13.388	15.086	20.517
10	9.342	13.442	15.987	18.307	21.161	23.209	29.588
15	14.339	19.311	22.307	24.996	28.259	30.578	37.697
20	19.337	25.038	28.412	31.410	35.020	37.566	43.315
25	24.337	30.675	34.382	37.652	41.566	44.314	52.620
30	29.336	36.250	40.256	43.773	47.962	50.892	59.703

Note: The Pearson chi-square test and the Yates corrected chi-square test from a two-by-two table have one degree of freedom. The Mantel-Haenszel chi-square also has one degree of freedom, whether from a single two-by-two table or from stratified analysis.

REFERENCES

1. Shem, S. (1978). *The house of God*. Richard Marek Publishers, New York.
2. Dean A.G. Dean J.A., Coulomvier D., et al. (1994). Epi Info, Version 6: A word processing, database, and statistics program for epidemiology or microcomputers. Centers for Disease and Prevention, Atlanta, Georgia.
3. Luby, S. P., Jones, J. L., Horan, J. M. (1993). A large salmonellosis outbreak catered by a frequently penalized restaurant. *Epidemiology and Infection*, 110, 31-39.
4. Berkelman, R. L., Martin, D., Graham, D. R., et al. (1982). Streptococcal wound infections caused by a vaginal carrier. *Journal of the American Medical Association*, 247, 2680-82.
5. Landrigan, P. J. (1972). Epidemic measles in a divided city. *Journal of the American Medical Association*, 221, 567-70.
6. Kleinbaum, D. G., Kupper, L. L., Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*. Lifetime Learning Publications, Belmont, California.
7. Schlesselman, J. J. (1982). *Case control studies: Design, conduct, analysis*. Oxford University Press, New York.
8. Shands, K. N., Schmid, G. P., Dan, B. B., et al. (1980). Toxic-shock syndrome in menstruating women: Association with tampon use and *Staphylococcus aureus* and clinical features in 52 cases. *New England Journal of Medicine*, 303, 1436-42.
9. Dicker, R. C. (1986). Kawasaki syndrome. *Washington Morbidity Report*, (Oct); 1-4.
10. Robins, J., Greenland, S., Breslow, N. E. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology*, 124, 719-23.
11. Comstock, G. W. (1990). Vaccine evaluation by case-control or prospective studies. *American Journal of Epidemiology*, 131, 205-207.
12. Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295-300.