

5. Why is it incorrect to describe a rate ratio of 10 as indicating a high risk of disease among the exposed?
6. A newspaper article states that a disease has increased by 1200% in the past decade. What is the rate ratio that corresponds to this level of increase?
7. Another disease has increased by 20%. What is the rate ratio that corresponds to this increase?
8. From the data in Table 4-6, calculate the fraction of diarrhea cases among infants exposed to a low antibody level that is attributable to the low antibody level. Calculate the fraction of all diarrhea cases attributable to exposure to low antibody levels. What assumptions are needed to interpret the result as an attributable fraction?
9. What proportion of the 56 breast cancer cases in Table 4-7 is attributable to radiation exposure? What are the assumptions?
10. Suppose you worked for a health agency and had collected data on the incidence of lower back pain among people in different occupations. What measures of effect would you choose to look at, and why?
11. Suppose that the rate ratio measuring the relation between an exposure and a disease is 3 in two different countries. Would this situation imply that exposed people have the same risk in the two countries? Would it imply that the effect of the exposure is the same magnitude in the two countries? Why or why not?

REFERENCES

1. Gaylord Anderson, as cited in Cole P. The evolving case-control study. *J Chron Dis.* 1979;32:15-27.
2. Iskrant AP, Joliet PV. *Accidents and Homicides*. Cambridge, MA: Harvard University Press; 1968.
3. Snow J. *On the Mode of Communication of Cholera*. 2nd ed. London: John Churchill; 1860. (Facsimile of 1936 reprinted edition by Hafner, New York, 1965.)
4. Kurtzke JF, Hyllested K. Multiple sclerosis in the Faroe Islands: clinical and epidemiologic features. *Ann Neurol.* 1979;5:6-21.
5. Poser CM, Hibberd PL, Benedikz J, Gudmundsson G. *Neuroepidemiology*. 1988;7:168-180.
6. Cole TB, Chomba TL, Horan JM. Patterns of transmission of epidemic hysteria in a school. *Epidemiology*. 1990;1:212-218.
7. Glass RI, Svennerholm AM, Stoll BJ, et al. Protection against cholera in breast-fed children by antibiotics in breast milk. *N Engl J Med.* 1983;308:1389-1392.
8. Boice JD, Monson RR. Breast cancer in women after repeated fluoroscopic examinations of the chest. *J Natl Cancer Inst.* 1977;59:823-832.
9. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

Types of Epidemiologic Studies

Chapter 4 described measures of disease frequency, including risk, incidence rate, and prevalence; measures of effect, including risk and incidence rate differences and ratios; and attributable fractions. Epidemiologic studies may be viewed as measurement exercises undertaken to obtain estimates of these epidemiologic measures. The simplest studies aim only at estimating a single risk, incidence rate, or prevalence. More complicated studies aim at comparing measures of disease occurrence, with the goal of predicting such occurrence, learning about the causes of disease, or evaluating the impact of disease on a population. This chapter describes the two main types of epidemiologic study, the cohort study and the case-control study, along with several variants. More specialized study designs, such as two-stage designs and ecologic studies, are discussed in *Modern Epidemiology*.¹

COHORT STUDIES

In epidemiology, a cohort is defined most broadly as "any designated group of individuals who are followed or traced over a period of time."² A cohort study, which is the archetype for all epidemiologic studies, involves measuring the occurrence of disease within one or more cohorts. Typically, a cohort comprises persons with a common characteristic, such as an exposure or ethnic identity. For simplicity, we refer to two cohorts, *exposed* and *unexposed*, in our discussion. In this context, we use the term *exposed* in its most general sense; for example, an exposed cohort may have in common the presence of a specific gene. The purpose of following a cohort is to measure the occurrence of one or more specific diseases during the period of follow-up, usually with the aim of comparing the disease rates for two or more cohorts.

The concept of following a cohort to measure disease occurrence may appear straightforward, but there are many complications involving who is eligible to be followed, what should count as an instance of disease, how the incidence rates or risks are measured, and how exposure ought to be defined. Before exploring these

5. Why is it incorrect to describe a rate ratio of 10 as indicating a high risk of disease among the exposed?
6. A newspaper article states that a disease has increased by 1200% in the past decade. What is the rate ratio that corresponds to this level of increase?
7. Another disease has increased by 20%. What is the rate ratio that corresponds to this increase?
8. From the data in Table 4-6, calculate the fraction of diarrhea cases among infants exposed to a low antibody level that is attributable to the low antibody level. Calculate the fraction of all diarrhea cases attributable to exposure to low antibody levels. What assumptions are needed to interpret the result as an attributable fraction?
9. What proportion of the 56 breast cancer cases in Table 4-7 is attributable to radiation exposure? What are the assumptions?
10. Suppose you worked for a health agency and had collected data on the incidence of lower back pain among people in different occupations. What measures of effect would you choose to look at, and why?
11. Suppose that the rate ratio measuring the relation between an exposure and a disease is 3 in two different countries. Would this situation imply that exposed people have the same risk in the two countries? Would it imply that the effect of the exposure is the same magnitude in the two countries? Why or why not?

REFERENCES

1. Gaylord Anderson, as cited in Cole P. The evolving case-control study. *J Chron Dis.* 1979;32:15-27.
2. Iskrant AP, Joliet PV. *Accidents and Homicides*. Cambridge, MA: Harvard University Press; 1968.
3. Snow J. *On the Mode of Communication of Cholera*. 2nd ed. London: John Churchill; 1860. (Facsimile of 1936 reprinted edition by Hafner, New York, 1965.)
4. Kurtzke JF, Hyllested K. Multiple sclerosis in the Faroe Islands: clinical and epidemiologic features. *Ann Neurol.* 1979;5:6-21.
5. Poser CM, Hibberd PL, Benedikz J, Gudmundsson G. *Neuroepidemiology*. 1988;7:168-180.
6. Cole TB, Chorba TL, Horan JM. Patterns of transmission of epidemic hysteria in a school. *Epidemiology.* 1990;1:212-218.
7. Glass RI, Svennerholm AM, Stoll BJ, et al. Protection against cholera in breast-fed children by antibiotics in breast milk. *N Engl J Med.* 1983;308:1389-1392.
8. Boice JD, Monson RR. Breast cancer in women after repeated fluoroscopic examinations of the chest. *J Natl Cancer Inst.* 1977;59:823-832.
9. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

Types of Epidemiologic Studies

Chapter 4 described measures of disease frequency, including risk, incidence rate, and prevalence; measures of effect, including risk and incidence rate differences and ratios; and attributable fractions. Epidemiologic studies may be viewed as measurement exercises undertaken to obtain estimates of these epidemiologic measures. The simplest studies aim only at estimating a single risk, incidence rate, or prevalence. More complicated studies aim at comparing measures of disease occurrence, with the goal of predicting such occurrence, learning about the causes of disease, or evaluating the impact of disease on a population. This chapter describes the two main types of epidemiologic study, the cohort study and the case-control study, along with several variants. More specialized study designs, such as two-stage designs and ecologic studies, are discussed in *Modern Epidemiology*.¹

COHORT STUDIES

In epidemiology, a cohort is defined most broadly as "any designated group of individuals who are followed or traced over a period of time."² A cohort study, which is the archetype for all epidemiologic studies, involves measuring the occurrence of disease within one or more cohorts. Typically, a cohort comprises persons with a common characteristic, such as an exposure or ethnic identity. For simplicity, we refer to two cohorts, *exposed* and *unexposed*, in our discussion. In this context, we use the term *exposed* in its most general sense; for example, an exposed cohort may have in common the presence of a specific gene. The purpose of following a cohort is to measure the occurrence of one or more specific diseases during the period of follow-up, usually with the aim of comparing the disease rates for two or more cohorts.

The concept of following a cohort to measure disease occurrence may appear straightforward, but there are many complications involving who is eligible to be followed, what should count as an instance of disease, how the incidence rates or risks are measured, and how exposure ought to be defined. Before exploring these

issues, we consider an example of an elegantly designed epidemiologic cohort study.

John Snow's Natural Experiment

In Chapter 4 we looked at data compiled by John Snow regarding the cholera outbreak in London in 1854 (see Fig. 4-4). In London at that time, there were several water companies that piped drinking water to residents. Snow's so-called natural experiment consisted of comparing the cholera mortality rates for residents subscribing to two of the major water companies, the Southwark and Vauxhall Company, which piped impure Thames water contaminated with sewage, and the Lambeth Company, which in 1852 changed its collection from opposite the Hungerford Market upstream to Thames Ditton, obtaining a supply of water free of the sewage of London. Snow³ described it as follows:

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

The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentle folks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from impurity.

To turn this experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur...

From this natural experiment, Snow³ was able to estimate the frequency of cholera deaths, using households as the denominator, separately for people in each of the two cohorts:

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

Snow also obtained estimates of the size of the population served by the two water companies, enabling him to report the attack rate of fatal cholera among

Table 5-1 ATTACK RATE OF FATAL CHOLERA AMONG CUSTOMERS OF THE SOUTHWARK AND VAUXHALL COMPANY (EXPOSED COHORT) AND THE LAMBETH COMPANY (UNEXPOSED COHORT), LONDON, 1854

	Water Company	
	Southwark & Vauxhall	Lambeth
Cholera deaths	4,093	461
Population	266,516	173,748
Attack Rate	0.0154	0.0027

Data from Snow.³

residents of households served by them during the 1854 outbreak (Table 5-1). Residents whose water came from the Southwark and Vauxhall Company had an attack rate 5.8 times greater than that of residents whose water came from the Lambeth Company.

Snow saw that circumstance had created conditions that emulated an experiment, in which people who were otherwise alike in relevant aspects differed by their consumption of pure or impure water. In an actual experiment, the investigator assigns the study participants to the exposed and unexposed groups. In a natural experiment, as studies such as Snow's have come to be known, the investigator takes advantage of a setting that serves effectively as an experiment. It could be argued that the role of the investigator in a natural experiment requires more creativity and insight than in an actual experiment. In the natural experiment, the investigator has to see the opportunity for the research and define the study populations to capitalize on the setting. For example, Snow conducted his study within specific neighborhoods in London where the pipes from these two water companies were intermingled. In other districts, there was less intermingling of pipes from the various water companies that supplied water to dwellings. Comparing the attack rates across various districts of London would have been a less persuasive way to evaluate the effect of the water supply because many factors differed from one district to another. Within the area in which the pipes of the Southwark and Vauxhall and those of the Lambeth Companies were intermingled, however, Snow saw that there was little difference between those who consumed water from one company or the other, apart from the water supply itself. Part of his genius was identifying the precise setting in which to conduct the study.

Types of Experiments

Experiments are conceptually straightforward. Experiments are cohort studies, although not all cohort studies are experiments. In epidemiology, an experiment is a study in which the incidence rate or the risk of disease in two or more cohorts is compared after assigning the exposure to the people who constitute the cohorts. In an experiment, the reason for the exposure assignment is solely to suit the objectives of the study; if people receive their exposure assignment based

on considerations other than the study protocol, it is not a true experiment. The *protocol* is the set of rules by which the study is conducted.

Among the several varieties of epidemiologic experiment, the main types are clinical trials, field trials, and community intervention trials. The word *trial* is used as a synonym for an epidemiologic experiment. Epidemiologic experiments are most frequently conducted in a clinical setting, with the aim of evaluating which treatment for a disease is better. These studies are known as *clinical trials*. In clinical trials, all study subjects have been diagnosed with a specific disease, but that disease is not the disease event that is being studied. Rather, it is some consequence of that disease, such as death or spread of a cancer, that becomes the "disease" event studied in a clinical trial. The aim of a clinical trial is to evaluate the incidence rate or risk of disease complications in the cohorts assigned to the different treatment groups. The primary outcome of interest is often a stage in the natural history of the disease, such as recurrence of cancer, or deaths among patients with cardiovascular disease. Alternatively, the outcome may be an adverse effect, ranging from transitory malaise to extreme outcomes such as liver failure or sudden death. In most trials, treatments are assigned by *randomization*, using random number assignment. Randomization tends to produce comparability between the cohorts with respect to factors that may affect the outcome under study.

To take full advantage of random assignment, the groups that should be compared in the analysis of an experiment are the groups that are classified according to their random assignment. Suppose 100 patients are randomly assigned to receive a new treatment in a clinical trial, and 100 patients are randomly assigned to receive an old treatment. To benefit maximally from the random assignment, the investigator should compare these two groups of 100, regardless of what treatments actually were given. It is not unusual for a patient to be assigned a treatment and not to take it as instructed. The patient may reject treatment for a variety of reasons or change his or her mind after the treatment assignment is made. Patients assigned to receive an old treatment may find a way to get the new treatment. Even if treatments are disguised in the form of coded medications that are not readily identified, patients may not be treated as assigned, because they react poorly to an assigned medication or otherwise ignore their assigned treatment. Thus, the assigned treatment may differ from the actual treatment for a proportion of study participants. Nevertheless, a standard approach in analyzing data from a clinical trial is to follow the principle of *intent to treat*, which means that the treatment assignment, rather than the actual treatment, determines the classification of participants in the data analysis. The intent-to-treat approach to the data analysis avoids problems that can arise if there is a tendency for patients who are at especially high or low risk for the outcome to fail to adhere to their assigned treatment. For example, if high-risk patients were more likely than low-risk patients to switch from the old treatment to the new treatment, the new treatment would appear to be worse than it should if the patients were compared according to the actual treatment that they received.

An intent-to-treat analysis may misclassify a substantial proportion of study participants with respect to their actual exposure. This misclassification of actual exposure leads to an underestimate of the treatment effect (see Chapter 7). Although the intent-to-treat approach is often desirable because it preserves the advantages of random assignment and therefore can lead to better comparability

of the study groups, the fact that it will underestimate the treatment effect should be borne in mind. Underestimating the benefit of a new treatment may be considered a small problem, because adoption of the treatment will have even greater benefits than anticipated. If the randomized trial is intended to study adverse effects of treatment, however, underestimating the magnitude of those effects is a larger problem. In trials aimed at evaluating the safety (rather than efficacy) of a new treatment, the drawbacks of an intent-to-treat analysis may outweigh any advantages, and it may be preferable to analyze the data based on the actual exposure of participants, rather than the category determined by random assignment.

NATURAL EXPERIMENTS ARE NOT EXPERIMENTS

In John Snow's natural experiment, customers of the Southwark and Vauxhall and the Lambeth companies were not randomly assigned to their water supply, as they would be in an experiment. The *natural experiment* is not an actual experiment; it is a cohort study that simulates what would occur in an experiment. In Snow's description of the customers of the two water companies, he gives the impression that the comparability between them was nearly as good as might have been achieved by random assignment. Thus, we have an experiment created by "nature," or a natural experiment, which may be more accurately described as a cohort study designed by an ingenious epidemiologist.

The data in Table 5-2 come from a clinical trial of adult patients recently infected with human immunodeficiency virus (HIV) that was undertaken to determine whether early treatment with zidovudine was effective in improving the prognosis.⁴ Patients were randomly assigned to receive zidovudine or placebo and then followed for an average of 15 months. The data show that the risk of getting an opportunistic infection during the follow-up period was low among those who received early zidovudine treatment but considerably higher among those who received a dummy (placebo) treatment.

Clinical trials may be the most common type of epidemiologic experiment, but they are not the only type. Epidemiologists also conduct *field trials*, which differ

Table 5-2 RANDOMIZED TRIAL COMPARING THE RISK OF OPPORTUNISTIC INFECTION AMONG PATIENTS WITH RECENT HIV INFECTION WHO RECEIVED EITHER ZIDOVUDINE OR PLACEBO

	Treatment Group	
	Zidovudine	Placebo
Opportunistic infection	1	7
Total patients	39	38
Risk	0.026	0.184

Data from Kinloch-de Loes et al.⁴

from clinical trials mainly in that the study participants are not patients. In a field trial, the goal is to study the primary prevention of a disease, rather than treatment of an existing disease. For example, experiments of new vaccines to prevent infectious illness are field trials because the study participants have not yet been diagnosed with a particular disease. In a clinical trial, the study participants can be followed through regular clinic visits, whereas in a field trial, it may be necessary to contact participants for follow-up directly at home, work, or school. The largest formal human experiment ever conducted, the Salk vaccine trial of 1954, is a prominent example of a field trial.⁵ It was conducted to evaluate the efficacy of a new vaccine to prevent paralytic poliomyelitis, and it paved the way for the first widespread use of vaccination to prevent poliomyelitis.

Another type of experiment is a *community intervention trial*. In this type of study, the exposure is assigned to groups of people rather than singly. For example, the community fluoridation trials in the 1940s and 1950s that evaluated the effect of fluoride in a community water supply were community intervention trials. The data in Table 5-3 illustrate a community intervention trial that evaluated a program of home-based neonatal care and management of sepsis designed to prevent death among infants in rural India.⁶ This trial, as is often the case for community intervention trials, did not employ random assignment. Instead, the investigators selected 39 villages targeted for the new program and 47 villages in which the new program was not introduced. The program consisted of frequent home visits after each birth in the village by health workers who were trained for the study to deal with problems in neonatal care. This program resulted in reduced neonatal mortality for each of the 3 years of its implementation, with the reductions increasing with time. The data in Table 5-3 show the results for the third of the 3 years.

Population at Risk

Snow's study on cholera defined two cohorts on the basis of their water supply. One was the customers of the Southwark and Vauxhall Company, which drew polluted water from the lower Thames River, and the other was the customers of the Lambeth Company, which drew its water from the comparatively pure upper

Table 5-3 NEONATAL MORTALITY IN THE THIRD OF 3 YEARS AFTER INSTITUTING A COMMUNITY INTERVENTION TRIAL OF HOME-BASED NEONATAL CARE IN RURAL INDIAN VILLAGES

	Intervention Group	
	Home Care	Usual Care
Infant deaths	38	64
Number of births	979	940
Risk	0.039	0.068

Data from Bang et al.⁶

Thames. Any person in either of these cohorts could have contracted cholera. Snow measured the rate of cholera occurrence among the people in each cohort.

EXPERIMENTS ARE AN IMPERFECT GOLD STANDARD

Randomized trials are commonly described as the gold standard of epidemiologic studies, with all that implies. The random assignment does confer an important advantage, usually preventing or at least reducing confounding by measured and unmeasured risk factors. Nevertheless, randomized trials are far from perfect:

- The full benefits of random assignment depend on conducting an intent-to-treat analysis, which comes with its own bias, leading to an underestimate of the effect.
- In small trials, random assignment can lead to large imbalances between groups, thus failing to balance risk factors as hoped.
- For practical and ethical reasons, many research questions do not lend themselves to being studied in a randomized trial.
- Some trials evaluate treatments that are delivered more rigorously in the trial setting or differ in other ways from the real-world interventions that occur outside of trials.
- The expense of large trials may lead to substituting for the intended end point a more common intermediate end point, such as a change in a biomarker; this approach allows smaller studies, but the results may not correspond to the effect on the intended outcome.
- The small size of many trials leads to imprecise results that may not be replicable.

The gold standard does not necessarily provide certainty. If trials were perfect, trials of the same study question would always produce similar results, but that is seldom seen. Moreover, if the results of a trial and a non-experimental study differ, it is not guaranteed that the trial results are closer to the truth. Thorough consideration of the design and analysis of both studies is warranted to resolve discrepancies and is much more informative than simply assuming that results from randomized trials, being based on a supposed gold standard, are always correct.

To understand which people can belong to a cohort in an epidemiologic study, we must consider a basic requirement for cohort membership: Cohort members must meet the criteria for being at risk for disease. The members of a cohort to be followed are sometimes described as the *population at risk*. The term implies that all members of the cohort should be at risk for developing the specific diseases being measured. Determining who may be part of the population at risk may depend on which disease is being measured.

A standard requirement of any population at risk is that everyone be free of the disease being measured at the outset of follow-up. The reason is that a person

usually cannot develop anew a disease that he or she currently has. Someone with diabetes cannot develop diabetes, and someone with schizophrenia cannot develop schizophrenia. To be at risk for disease also implies that everyone in the population at risk must be alive at the start of follow-up; dead people are not at risk of getting any disease. Being alive and free of the disease are straightforward eligibility conditions, but other eligibility conditions may not be as simple. Suppose that the outcome is the development of measles. Should people who have received a measles vaccination be included in the population at risk? If they are completely immunized, they are not at risk for measles, but how can we know whether the vaccine has conferred complete immunity? In a cohort being studied for the occurrence of breast cancer, should men be considered part of the population at risk? Men do develop breast cancer, but it is rare compared with its occurrence in women. One solution is to distinguish male and female breast cancer as different diseases. In that case, if female breast cancer is being studied, men would be excluded from the population at risk.

Some diseases occur only once in a person, whereas others can recur. Death is the clearest example of a disease outcome that can occur only once for a given person. Other examples include diabetes, multiple sclerosis, chicken pox, and cleft palate. Disease can occur only once if it is incurable (eg, diabetes, multiple sclerosis), if recovery confers complete lifetime immunity (eg, chicken pox), or if there is a period of vulnerability that a person passes through only once (eg, cleft palate). If the disease can only occur once, anyone in a cohort who develops the disease is no longer at risk for it again and therefore exits from the population at risk as soon as the disease occurs. Also, any person who dies during the follow-up period, for whatever reason, is no longer part of the population at risk. The members of the population at risk at any given time must be people in whom the disease can still occur.

It may be possible, however, for someone with a disease to recover from the disease and then develop it again. For example, someone with a urinary tract infection can recover and then succumb to another urinary tract infection. In that case, the person is not part of the population at risk while he or she has the urinary tract infection but can become part of the population at risk again at the time of recovery. Being part of a population at risk is a dynamic process. People may enter and leave a population at risk depending on their health and other possible eligibility criteria (eg, geography).

Cohort Study of Vitamin A During Pregnancy: An Example

To study the relation between diet and other exposures of pregnant women and the development of birth defects in their offspring, Milunsky and colleagues⁷ interviewed more than 22,000 pregnant women early in their pregnancies. The original purpose was to study the potential effect of folate to prevent a class of birth defects known as neural tube defects. A later study, based on the same population of women, evaluated the role of dietary vitamin A in causing another class of birth defects that affect either the heart or the head, described as cranial neural crest defects.⁸ For the latter study, the women were divided into cohorts according to the amount of vitamin A in their diet from food or supplements.

DISEASE-FREE DOES NOT IMPLY HEALTHY

Although a population at risk should be free of disease at the outset of follow-up, it is incorrect to conclude that the population at risk is healthy. The requirement to be free of disease does not imply health; it merely implies that the people being followed do not have the specific disease being measured. The search for a population that is healthy in the sense of being free of all disease is fruitless. If *disease* is defined broadly, virtually every person has some disease or disorder at any given time. Acne, periodontal disease, back ailments, allergies, vision deficits, obesity, asthma, and respiratory infection are examples of the prevalent conditions that make it almost impossible to find even one person who is completely healthy. Being free of disease and therefore a member of a population at risk implies only that the person is free of the specific disease being followed, not all diseases.

The data in Table 5-4 summarize the results of this cohort study and show that the prevalence of these defects increased steadily and substantially with increasing intake of vitamin A supplements by pregnant women.

Table 5-4 gives results for four separate cohorts of the study population, each defined according to the level of supplemental intake of vitamin A that the women reported in the interview. The occurrence of cranial neural crest defects increased substantially for women who took supplements of vitamin A in doses greater than 8000 IU/day.

Closed and Open Cohorts

Epidemiologists follow two types of cohorts. A *closed cohort* is one with a fixed membership. After it is defined and follow-up begins, no one can be added to a closed cohort. The initial roster may dwindle, however, as people in the cohort die, are lost to follow-up, or develop the disease. Randomized experiments are examples of studies of closed cohorts (Fig. 5-1); the follow-up begins at randomization, a common starting point for everyone in the study. Another example of a

Table 5-4 PREVALENCE OF CRANIAL NEURAL-CREST DEFECTS AMONG THE OFFSPRING OF FOUR COHORTS OF PREGNANT WOMEN, CLASSIFIED ACCORDING TO THEIR INTAKE OF SUPPLEMENTAL VITAMIN A DURING EARLY PREGNANCY

	Level of Vitamin A Intake from Supplements (IU/Day)			
	0-5000	5001-8000	8001-10,000	≥10,001
Affected infants	51	54	9	7
Pregnancies	11,083	10,585	763	317
Prevalence	0.46%	0.51%	1.18%	2.21%

Data from Rothman et al.⁸

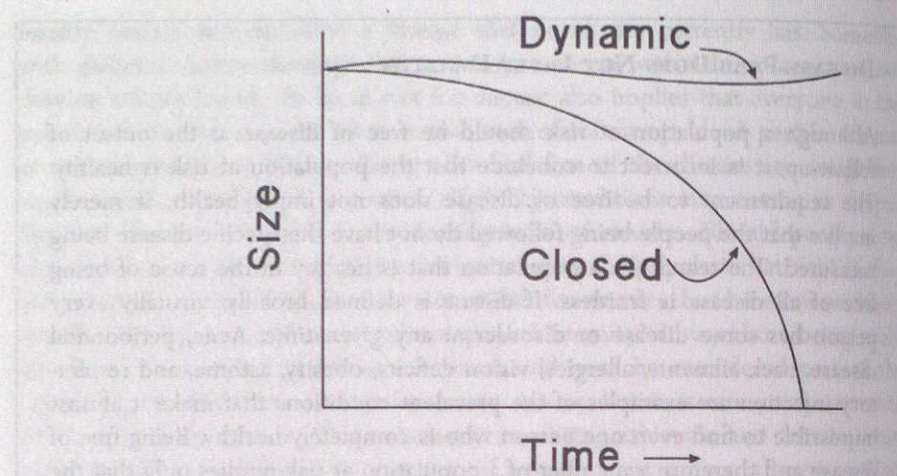


Figure 5-1 Size of hypothetical dynamic and closed cohorts over time.

closed cohort study is the landmark Framingham Heart Study, which was initiated in 1949 and is still ongoing.⁹

In contrast, an *open cohort*, also referred to as a *dynamic cohort* or a *dynamic population*, can take on new members as time passes (see Fig. 5-1). An example of an open cohort is the population of Connecticut, where one of the oldest cancer registries in the United States is found. The population studied in the Connecticut cancer registry may be considered a dynamic cohort that comprises the people of Connecticut. Cancer incidence rates over a period of time in Connecticut reflect the rate of cancer occurrence among a changing population as people move to or away from Connecticut. The population at risk at any given moment comprises current residents of Connecticut, but since residency may change, and in particular new residents may be added to the population, the population being described is dynamic, not closed. Another example of a dynamic population is the population of a school, with new students entering each year and others leaving. An extreme example of a dynamic cohort is the population of current U.S. presidents: whenever there is a new one sworn into office, the previous one leaves the cohort, and whenever one leaves, a new one takes over; the size of the population always remains constant at 1.

In contrast to a dynamic cohort, a closed cohort always becomes smaller with passing time. Ideally, investigators of a closed cohort attempt to track down cohort members if they leave the vicinity of the study. Members of a closed cohort constitute a group of people who remain members of the cohort even if they leave the area of the study. In a dynamic population that is defined geographically, people who leave the geographic boundaries of the study are leaving the cohort and will not be followed.

Counting Disease Events

In cohort studies, epidemiologists usually calculate incidence rates or risks by dividing the number of new *disease events* (ie, the number of disease onsets) by the appropriate denominator, based on the size of the population at risk. Usually

there are one or more categories of disease that are of special interest, and new cases of those diseases are counted. Occasionally, however, some disease onsets are excluded, even if they represent the disease under study.

One reason to exclude a disease event might be that it is not the first occurrence of the disease in that person. For example, suppose a woman develops breast cancer in one breast and later develops breast cancer in the other breast. In many studies, the second onset of breast cancer would not be counted as a new case, despite all biologic indications that it represents a separate cancer rather than spread of the first cancer. Similarly, in many studies of myocardial infarction, only the first myocardial infarction is counted as a disease event, and subsequent heart attacks are excluded. Why should investigators make this distinction between the first occurrence of a disease and subsequent occurrences? First, it may be difficult to distinguish between a new case of disease and a recurrence or exacerbation of an earlier case. Second, recurrent disease may have a different set of causes than the primary occurrence. If the investigator limits his or her interest to the first occurrence of a disease, all subsequent occurrences will be excluded, but there will also have to be a corresponding adjustment in the population at risk. If only the first occurrence of disease is of interest, any person who develops the disease is removed from the at-risk population at the time the disease develops. This procedure is consistent with the requirement that members of the population at risk must be eligible to develop the disease. If only the first occurrence is counted, people who develop the disease terminate their eligibility to get disease at the point at which they develop disease.

If the epidemiologist is interested in measuring the total number of disease events, regardless of whether they are first or later occurrences, then a person who is in the cohort would remain as part of the population at risk even after getting the disease. In such an analysis, the first disease event would be counted just the same as a subsequent event, and there would be no way to distinguish the occurrence of first versus later events. The distinction can be made, however. One way is to calculate separate occurrence measures for first and subsequent events. For example, it is possible to calculate the incidence rate of first events, second events, third events, and so forth. The population at risk for second events would be those who have had a first event. On having a first event, a person would leave the population at risk for a first event and enter the population at risk for a second event.

Another reason not to count a disease event is that there was insufficient time for the disease to be related to an exposure. This issue is addressed in the "Exposure and Induction Time" section.

Measuring Incidence Rates or Risks

From a closed cohort, we can estimate a risk or an incidence rate to measure disease occurrence. Calculation of a risk is complicated by the problem of competing risks (see Chapter 4). Because of competing risks, the population at risk will not remain constant in size over time, which means that some people will be removed from the population at risk before they have experienced the entire period of follow-up. Despite this problem, there are many cohort studies in which risks are estimated directly. Usually, the period of follow-up is short enough or

the competing risks are small enough in relation to the disease under study that there is relatively little distortion in the risk estimates. In these studies, the risk in each cohort is calculated by dividing the number of new disease events by the total number of people who are being followed in the closed cohort. This approach was used to calculate the risk for cholera in Snow's analysis depicted in Table 5-1. Essentially the same approach was used in the study of vitamin A and birth defects described earlier, although the measure reported is the prevalence, rather than the risk, of birth defects.

It is problematic to measure risk directly in a dynamic cohort, in which new people are added to the cohort during the follow-up period. To get around this problem, the investigator can take into account the amount of time that each person spends in the population at risk and calculate an incidence rate by dividing the number of new disease events by the amount of person-time experienced by the population at risk. The same approach can be applied to a closed cohort, addressing the problem of competing risks.

In the calculation of an incidence rate, the ideal situation is to have precise information on the amount of time that each person has been in the population at risk. Often, this time is calculated for each person in terms of days at risk, although the final results may be expressed in terms of years after converting the time units.

Cohort Study of X-Ray Fluoroscopy and Breast Cancer: An Example

The data in Table 4-6 (see Chapter 4) are taken from a cohort study of radiation exposure and breast cancer. As part of their treatment for tuberculosis, many of the women received substantial doses of x-rays for fluoroscopic monitoring of their lungs. Because the women were followed for highly variable lengths of time, it would not have been reasonable to calculate directly the risk of breast cancer; to do so requires a fixed length of follow-up or at least a minimum follow-up time for all the women in the cohort. (They could have calculated the risk of breast cancer for segments of the follow-up time using the life-table method described in Chapter 4.) Instead, the investigators measured the incidence rate of breast cancer among these women with x-ray exposure. They compared this rate with the rate of breast cancer among women treated during the same period for tuberculosis but not with x-rays. The data in Table 4-6 show that the women who received x-ray exposure had nearly twice the incidence rate of breast cancer as the women who did not receive x-ray exposure.

Exposure and Induction Time

After World War II, the United States and Japan jointly undertook a cohort study of the populations of Hiroshima and Nagasaki who survived the atomic bomb blasts. These populations have been followed for decades, initially under the aegis of the Atomic Bomb Casualty Commission and later under its successor, the Radiation Effects Research Foundation. A category of outcome that has been

of primary interest to the researchers has been cancer occurrence. Leukemia is one of the types of cancer that or substantially increased in incidence by ionizing radiation. Consider the survivors of the bombs to constitute several closed cohorts, each corresponding to a different dose category of ionizing radiation. The main factors that determined the dose of exposure were the distance from the epicenter of the blast and the shielding provided by the immediate environment, such as buildings, at the time of the blast.

Suppose that we wish to measure the incidence rate of leukemia among atomic bomb survivors who received a high dose of ionizing radiation and compare this rate with the rate experienced by those who received little or no radiation exposure. The cohorts are defined as of the time of the blasts, and their subsequent experience is tracked as part of the cohort study. We might consider that those who received a high dose of ionizing radiation immediately entered the population at risk for leukemia. The difficulty with beginning the follow-up immediately after the exposure is that it does not allow a sufficient induction time for leukemia to develop as a result of the radiation exposure. For example, an exposed person who was diagnosed with leukemia 2 weeks after exposure is unlikely to have developed his or her leukemia as a consequence of the radiation exposure. After the exposure, disease does not occur until the induction period has passed (see Chapter 3). The induction period corresponds to the time that it takes for the causal mechanism to be completed by the action of the complementary component causes that act after radiation exposure. Suppose that the average time it takes before causal mechanisms that involve radiation are completed and leukemia occurs is 5 years and that few causal mechanisms if any are completed until 3 years have passed. After disease occurs, there is an additional interval, the latent period, during which disease exists but has not yet been diagnosed. It is important to consider the induction period and the latent period in the calculation of incidence rates. To measure the effect of radiation exposure most clearly, the investigator should define the time period at risk for leukemia among exposed people in a way that allows for the induction time and perhaps for the latent period. It would make more sense to allow exposed people to enter the population at risk for leukemia only after a delay of at least 3 years, if we assume that any case occurring before that time could not plausibly be related to exposure.

Typically, the investigator cannot be sure what the induction time is for a given exposure and disease. In that case, it may be necessary to hypothesize various induction times and reanalyze the data under each separate hypothesis. Alternatively, there are statistical methods that estimate the most appropriate induction time.¹⁰

Among exposed people, what happens to the person-time that is not related to exposure under the hypothesis of a specific induction time? Consider the previous example of studying the effect of radiation exposure from the atomic bomb blasts on the development of leukemia. If we hypothesize that no leukemia can occur as a result of the radiation until at least 3 years have elapsed since the blast, what happens to the first 3 years of follow-up for someone who was exposed? How should we treat the experience of exposed people before they are exposed? Although the induction time comes after exposure, it is a period during which the exposure is presumed not to have any effect and is therefore like the time that comes before exposure. There are two reasonable options for dealing with this

time: ignore it or combine it with the follow-up time of people who were never exposed.

The hypothetical data in Figure 5-2 can be used to calculate incidence rates for exposed and unexposed cohorts in a cohort study. Figure 5-2 depicts the follow-up time for 10 people, 5 exposed and 5 unexposed, who were followed for up to 20 years after a point exposure. There are three ways in which follow-up can end: the person can be followed until the end of the study follow-up period of 20 years, the person can be lost to follow-up, or the person can get the disease. Those who are followed for the full 20 years are said to be withdrawn at the end of follow-up. We can calculate the incidence rate among exposed people during the 20 years after exposure. Figure 5-2 shows that the follow-up times for the first 5 people are 12, 20, 15, 2, and 10 years, which sum to 59 years. In this experience of 59 years, three disease events have occurred, for an incidence rate of 3 events per 59 years, or $3/59 \text{ yr}^{-1}$. We also can express this rate as 5.1 cases per 100 person-years, or $5.1/100 \text{ yr}^{-1}$. For the unexposed group, the follow-up times were 20, 18, 20, 11, and 20 years, for a total of 89 years, and there was only one disease event, for a rate of $1/89 \text{ yr}^{-1}$, or $1.1/100 \text{ yr}^{-1}$.

The rate for the exposed group, however, does not take into account the 3-year induction period for exposure to have an effect. To take that into account, we must ignore the first 3 years of follow-up for the exposed group. The follow-up times for the exposed cohort that comes after the induction period for exposure are 9, 17, 12, 0, and 7 years, for a total of 45 years, with only two disease events

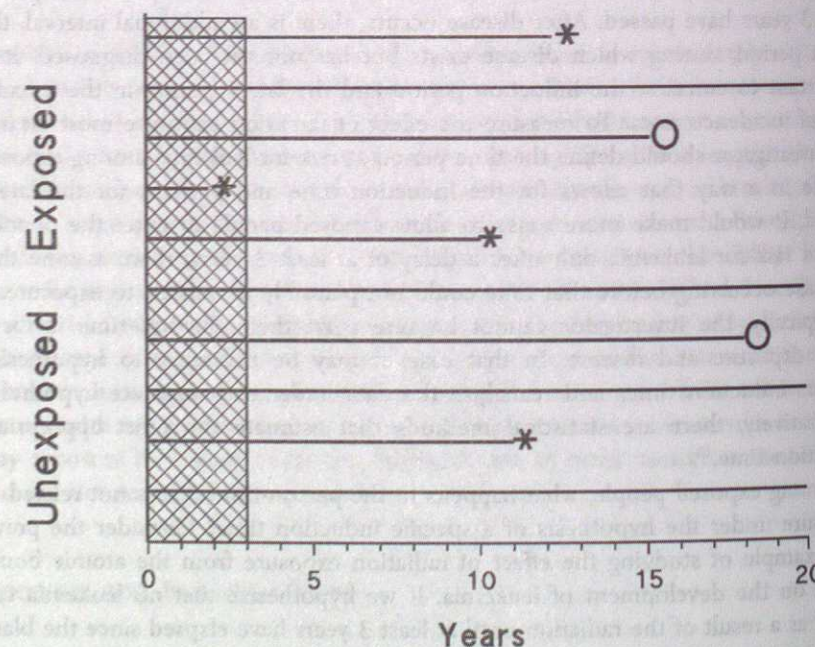


Figure 5-2 Follow-up data for 10 people in a hypothetical cohort study that followed 5 exposed people (top five lines) and 5 unexposed people (bottom five lines). The exposure was a point exposure that is hypothesized to have a minimum 3-year induction time (cross-hatched area) before any case of disease could result from it.

occurring during this follow-up experience. The rate after taking into account the 3-year induction period would be $2/45 \text{ yr}^{-1}$, or $4.4/100 \text{ yr}^{-1}$. There is no reason to exclude the first 3 years of follow-up for the unexposed group, because there is no induction period among those who are not exposed. An investigator may also consider including the first 3 years of follow-up for each exposed person as unexposed experience, because under the study hypothesis, this experience is not related to exposure. With that approach, the denominator of the rate for unexposed would include 14 additional years of follow-up and one additional event, giving a rate of $2/103 \text{ yr}^{-1}$, or $1.9/100 \text{ yr}^{-1}$.

The assumption that the induction time is 3 years is only a hypothesis, and it may be wrong. Other possible induction times can be considered as alternatives, leading to different results. Many epidemiologists ignore the issue of induction time and do not exclude any period of time following exposure. That practice is equivalent to assuming that the induction time is zero, which may be a reasonable assumption, but it may be unreasonable for many study questions. To the extent that the induction time hypothesis is incorrect, there will be nondifferential misclassification of exposure, which introduces a bias (see Chapter 7).

Eligibility Criteria, Exposure Classification, and Time Loops

In a prospective cohort study, the investigator selects subjects who meet the study eligibility criteria and then assigns them to exposure categories as they meet the conditions that define those categories. For example, in a prospective cohort study of smoking, subjects who meet age and other entry criteria may be invited into the cohort and then classified in appropriate smoking categories as they meet the definitions for those categories. A person classified as a nonsmoker at the start of the follow-up may be reclassified as a smoker if he takes up smoking during the follow-up, or a smoker who gives up smoking may be reclassified as an ex-smoker if he gives it up. In retrospective cohort studies, it is important to ensure that decisions about eligibility of participants and any exposure categorization are based on information that is known at the time to which these decisions or assignments pertain, rather than later. The investigator should only use information that would have been known at that time if the investigator had been conducting a prospective cohort study. If this rule is not observed, the result may be the formation of a "time loop," in which a decision is made to include or exclude or classify a study subject at a point in time before the information is known that the decision is based on. For example, suppose the intent is to exclude ex-smokers from the study. A smoker who gives up smoking during the follow-up and becomes an ex-smoker could not be prevented from enrolling in a prospective cohort study if the discontinuation of smoking is in the future at the time the study begins, because the information is not yet known. If the smoker becomes an ex-smoker during follow-up, it would create problems to exclude retroactively his already accumulated experience from the study. An investigator can, however, exclude or censor the person's future experience starting when he becomes an ex-smoker.

It is permissible to change the classification of a study participant as circumstances change during follow-up, but those changes should influence only the follow-up time that comes after the change. An unexposed person can become

WHICH MEASURES TO REPORT FROM COHORT STUDIES?

In a cohort study, the epidemiologist often has data that allow the calculation of a risk or a rate of disease. The choice depends on whether the denominators available are the number of people in the cohort, which gives risks, or the amount of person-time, which gives rates. To measure risks, everyone in the cohort should be followed for at least the length of the risk period. Risks are often reported in experimental studies, which usually aim for a uniform length of follow-up. If the follow-up time varies considerably from person to person, it may be preferable to use person-time as the denominator measure and report rates. In some cohort studies, the actual risks or rates in each cohort are not reported; instead, a risk or rate ratio is reported for one or more levels of the study exposure. Reporting only risk or rate ratios is a disservice to readers, who deserve to know the underlying risks or rates if these are obtainable. Ideally, the investigator should report the risk or rate for each level of exposure, as well as the numerators and denominators from which these risks or rates are calculated. The risk or rate differences are still of interest, although they are secondary to the actual risks or rates, because they can be derived from the risks or rates.

One reason that some studies report only ratio measures is that the investigators may have used a statistical model to analyze their data that only produces ratio measures. Nevertheless, it is not difficult to use stratification or other analytic methods to obtain the risks or rates themselves (see Chapter 10). Some cohort studies report odds ratios rather than risk ratios. Usually, odds ratios are reported because the statistical model used is a logistic model, which estimates odds ratios (see Chapter 12). Odds ratios are a fundamental measure in case-control studies, where they are used to estimate risk or rate ratios. In cohort studies, the risks or rates are obtainable directly, and there is little reason to consider an odds ratio. Although odds ratios are often reported in experiments, they should not be used, because in experiments, the outcome is typically frequent enough that the odds ratio is a poor estimate of the risk ratio, which could be obtained directly. Odds ratios are appropriate when analyzing case-control studies, but odds ratios usually have little reason to appear in cohort studies and should not be reported. If they are reported, it is better for readers to ignore them and look for information on the actual risks or rates.

exposed during the follow-up period of a cohort study. That information should not be used to change that person's categorization at the start of follow-up, when the person was unexposed. It can be used, however, to change the exposure category in which the person's follow-up time is tallied after the time he or she became exposed.

One example of the effect of a time loop is the creation of *immortal person-time*. Suppose we are conducting a cohort study of mortality among workers in a factory who are exposed to mercury vapor on the job. A common feature of many exposure measures for occupational (and other) exposures is that the

measure is based on the amount of time exposed. For example, the number of years of employment for workers exposed to mercury vapor on the job is a crude index of cumulative exposure, especially if the exposure in the workplace has been relatively stable over time. It may seem reasonable to compare the mortality among workers who were employed for only a few years with the mortality among workers employed for longer periods. Consider classifying workers into the categories of 0 to 9 years, 10 to 19 years, and 20+ years of employment, which we hope will separate workers with different levels of exposure to mercury vapor. A worker who ends up in the category of 20+ years of employment must pass through the other two categories first. How do we tabulate the follow-up time for a worker at this factory, starting at the beginning of employment? For the first 10 years of employment, the follow-up time must be tallied in the category of 0-9 years and, for the next 10 years, in the second category. It is only after 20 years of employment that a worker can begin to contribute follow-up time to the third category of employment. If a worker with 40 years of employment had been inappropriately classified in the 20+ category with respect to all 40 years, the first 20 of those 40 years would constitute immortal person-time. Those long-term workers were not actually immortal during their first 20 years on the job, but if any of them had died before reaching the 20th anniversary of employment, he or she could not have reached the category of 20+ to contribute any time at all. Everyone in the category of 20+ would have had 20 years during which they could not have died, because those who did were not classified in this category. This mistake would lead to a severe underestimate of mortality among the longest employed workers and an overestimate of mortality for those employed for shorter periods. This kind of problem can be avoided by avoiding any time loops that come from using future information to classify person-time before that information could have been known.

Retrospective Cohort Studies

A prospective cohort study is one in which the exposure information is recorded at the beginning of the follow-up (with possible updates if exposure status changes), and the period of time at risk for disease runs concurrently with the conduct of the study. This is always the case with experiments and with many nonexperimental cohort studies. Nevertheless, a cohort study is not always prospective; cohort studies can also be retrospective. In a *retrospective cohort study* (also known as a *historical cohort study*), the cohorts are identified from recorded information, and the time during which they are at risk for disease occurred before the beginning of the study.

An outstanding example of a retrospective cohort study was conducted by Morrison et al.¹¹ They studied young women who were born in Florence in the 15th and 16th centuries and who were enrolled in a dowry fund soon after they were born. The dowry fund was an insurance plan that would pay the family a sizable return if an enrolled woman married. If the woman died or joined a convent first, the fund did not have to pay a dowry. The fund records contain the date of birth, date of investment, and date of dowry payment or death of 19,000 girls and women. More than 500 years after the first women were enrolled in the dowry fund, epidemiologists were able to use the fund records to chart waves

of epidemic deaths from the plague and show how successive plague epidemics became milder over a period of 100 years. This retrospective cohort study, conducted centuries after the data were recorded, illustrates well that a cohort study need not be prospective.

Because a retrospective cohort study must rely on existing records, important information may be missing or otherwise unavailable. Nevertheless, when a retrospective cohort study is feasible, it offers the advantage of providing information that is usually much less costly than that from a prospective cohort study, and it may produce results much sooner because there is no need to wait for the disease to occur.

Tracing of Subjects

Cohort studies that span many years present a challenge with respect to maintaining contact with the cohort to ascertain disease events. Whether the study is retrospective or prospective, it is often difficult to locate people or their records many years after they have been enrolled in study cohorts. In prospective cohort studies, the investigator may contact study participants periodically to maintain current information on their location. Tracing subjects in cohort studies is a major component of their expense. If a large proportion of participants are lost to follow-up, the validity of the study may be threatened. Studies that trace less than about 60% of subjects usually are regarded with skepticism, but even follow-up of 70%, 80%, or more can be too low if the subjects lost to follow-up are lost for reasons related to both the exposure and the disease. Increasing access to the Internet may provide more efficient ways to enroll and trace participants in cohort studies.¹² We later consider the relative importance of successful tracing of subjects versus successful recruitment of subjects for cohort studies.

Special Exposure and General Population Cohorts

Cohort studies permit the epidemiologist to study many different disease end points at the same time. A mortality follow-up can be accomplished just as easily for all causes of death as for any specific cause. Health surveillance for one disease end point can sometimes be expanded to include many end points without much additional work. A cohort study can provide a comprehensive picture of the health effect of a given exposure. Cohort studies that focus on people who share a particular exposure are called *special-exposure cohort studies*. Examples of special-exposure cohorts include occupational cohorts exposed to substances in the workplace; soldiers exposed to Agent Orange in Vietnam; residents of the Love Canal area of Niagara, New York, exposed to chemical wastes; Seventh Day Adventists adhering to vegetarian diets; and atomic bomb victims exposed to ionizing radiation. Each of these exposures is uncommon; therefore, it is usually more efficient to study them by identifying a specific cohort of people who have sustained that exposure and comparing their disease experience with that of a cohort of people who lack the exposure.

In contrast, common exposures are sometimes studied through cohort studies that survey a segment of the population that is identified initially without regard to their exposure status. These *general-population cohorts* typically focus on exposures that a substantial proportion of people have experienced. Otherwise, there would be too few people in the study who are exposed to the factors of interest. After a general-population cohort is assembled, the cohort members can be classified according to smoking, alcoholic beverage consumption, diet, drug use, medical history, and many other factors of potential interest. The study described earlier of vitamin A intake in pregnant women and birth defects among their offspring⁸ is an example of a general-population cohort study. No women in that study were selected for the study because they had vitamin A exposure. Their exposure to vitamin A was determined after they were selected for the study during the interview. Although the study was a general-population cohort study, a high level of vitamin A intake during pregnancy was not a common exposure. Table 5-4 shows that only 317 of the total of 22,058 women, or 1.4%, were in the highest category of vitamin A intake. Fortunately, the overall study population was large enough that the vitamin A analysis was feasible; it would have been difficult to identify or recruit a special-exposure cohort of women who had a high intake of vitamin A during pregnancy.

In both special-exposure and general-population cohort studies, the investigator must classify study participants into the exposure categories that form the cohorts. This classification is easier for some exposures than for others. When the female offspring of women who took diethylstilbestrol (DES) were assembled for a special population cohort study, defining their exposure was comparatively clear-cut, based on whether their mothers took DES while they were pregnant.¹³ For other exposures, such as secondhand smoke or dietary intake of saturated fat, almost everyone is exposed to some extent, and the investigator must group people together according to their level of intake to form cohorts.

CASE-CONTROL STUDIES

The main drawback of conducting a cohort study is the necessity in many situations to obtain information on exposure and other variables from large populations to measure the risk or rate of disease. In many studies, however, only a tiny minority of those who are at risk for disease actually develop the disease. The case-control study aims at achieving the same goals as a cohort study, but more efficiently, using sampling. Properly carried out, case-control studies provide information that mirrors what could be learned from a cohort study, usually at considerably less cost and time.

Case-control studies are best understood by considering as the starting point a *source population*, which represents a hypothetical study population in which a cohort study might have been conducted. The source population is the population that gives rise to the cases included in the study. If a cohort study were undertaken, we would define the exposed and unexposed cohorts (or several cohorts), and from these populations obtain denominators for the incidence rates or risks that would be calculated for each cohort. We would then identify the number of cases occurring in each cohort and calculate the risk or incidence rate for each.

In a case-control study, the same cases are identified and classified according to whether they belong to the exposed or unexposed cohort. Instead of obtaining the denominators for the rates or risks, however, a control group is sampled from the entire source population that gives rise to the cases. Individuals in the control group are then classified into exposed and unexposed categories. The purpose of the control group is to determine the relative size of the exposed and unexposed components of the source population. Because the control group is used to estimate the distribution of exposure in the source population, the cardinal requirement of control selection is that the controls be sampled independently of exposure status.

Figure 5-3 shows the relation between a case-control study and the cohort study that it replaces. In the illustration, 25% of the 288 people in the source population are exposed. Suppose that the cases, illustrated at the right, arise during 1 year of follow-up. For simplicity, assume that the cases all occur at the end of the year; ordinarily they are spread out through time, which would necessitate discontinuing the person-time contribution to follow-up for each case at the time of the event. The rate of disease among exposed people is 8 cases occurring in 72 person-years, for a rate of 0.111 cases per person-year. Among the 216 unexposed people, 8 additional cases arise during the 1 year of follow-up, for a rate of 0.037 cases per person-year. In this hypothetical example, the incidence rate among the exposed cohort is three times the rate among the unexposed cohort. Now consider what would happen if a case-control study were conducted. The rectangle drawn around a portion of the source population represents a sample that could represent the control group. This sample must be taken independently of the exposure. Among the 48 people in the control group, 12 are exposed. If the sample is taken independently of the exposure, the same proportion of controls will be exposed as the proportion of people (or person-time) exposed in the original source population, apart from sampling error. The same cases that were included in the cohort study are also included in the case-control study as the case group. Later, we will see how the data in the case-control study can be used to estimate the ratio of the incidence rates in the source population, giving the same result for the incidence rate ratio that the cohort study provides.

Nested Case-Control Studies

It is helpful to think of every case-control study as being nested, or conducted, within cohorts of exposed and unexposed people, as illustrated in Fig. 5-3. Epidemiologists sometimes refer to specific case-control studies as *nested* case-control studies when the population within which the study is conducted is a well-defined cohort, but almost any case-control study can be thought of as nested within some source population. In many instances, this population may be identifiable, such as all residents of Rio de Janeiro during the year 2001; in other instances, the members of the source population may be hard to identify.

In occupational epidemiology, a commonly used approach is to conduct a case-control study nested within an occupational cohort that has already been enumerated. The reason for conducting a case-control study even when a cohort can be enumerated is usually that more information is needed than is readily

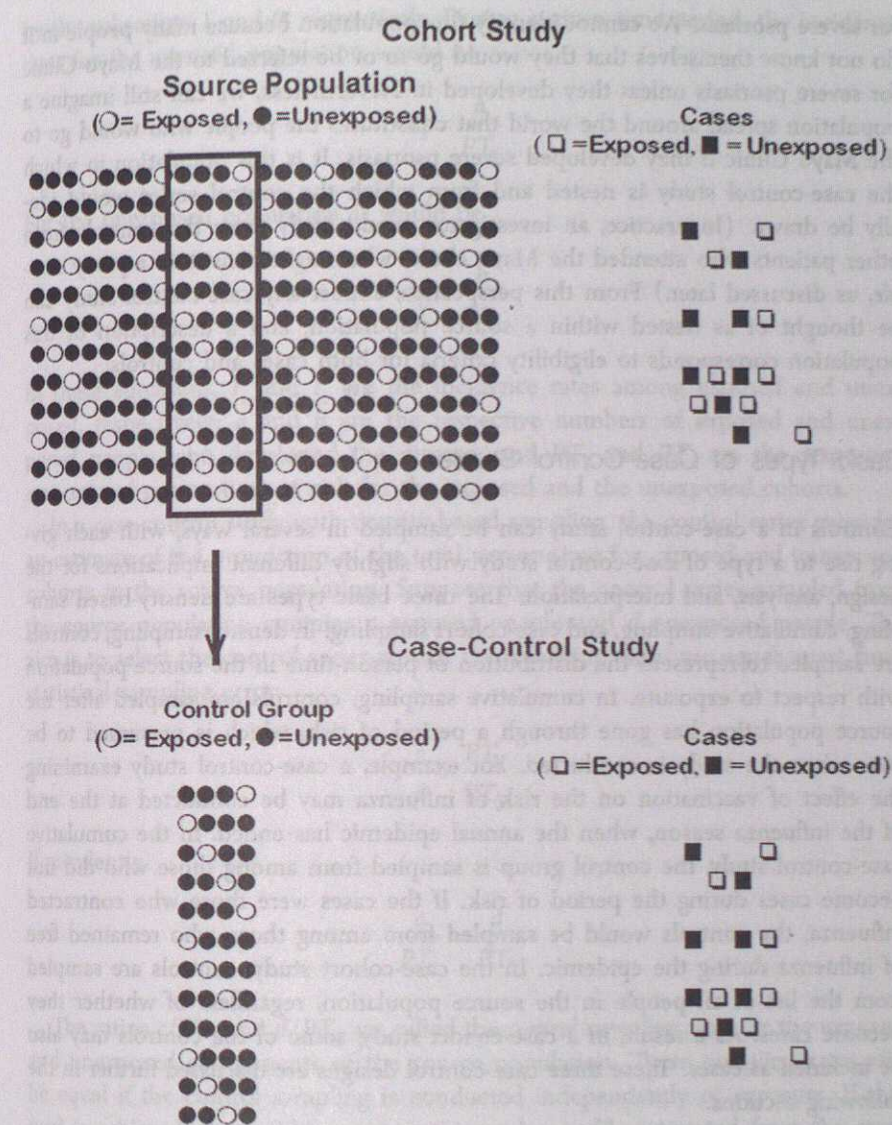


Figure 5-3 Schematic of a cohort study and a nested case-control study within the cohort shows how the control group is sampled from the source population.

available from records and that it would be too expensive to seek this information for everyone in the cohort. A nested case-control study is then more efficient. In these studies, the source population is easy to identify. It is the occupational cohort. A control group can be selected by sampling randomly from this source population.

As an example of a case-control study in which the source population is hard to identify, consider one in which the cases are patients treated for severe psoriasis at the Mayo Clinic. These patients come to the Mayo Clinic from all corners of the world. What is the specific source population that gives rise to these cases? To answer this question, we need to know exactly who goes to the Mayo Clinic

for severe psoriasis. We cannot identify this population because many people in it do not know themselves that they would go to or be referred to the Mayo Clinic for severe psoriasis unless they developed it. Nevertheless, we can still imagine a population spread around the world that constitutes the people who would go to the Mayo Clinic if they developed severe psoriasis. It is this population in which the case-control study is nested and from which the control series would ideally be drawn. (In practice, an investigator would likely draw the controls from other patients who attended the Mayo clinic, who might constitute a *proxy sample*, as discussed later.) From this perspective, almost any case-control study can be thought of as nested within a source population, and a description of this population corresponds to eligibility criteria for both cases and controls.

Basic Types of Case-Control Studies

Controls in a case-control study can be sampled in several ways, with each giving rise to a type of case-control study with slightly different implications for the design, analysis, and interpretation. The three basic types are density-based sampling, cumulative sampling, and case-cohort sampling. In density sampling, controls are sampled to represent the distribution of person-time in the source population with respect to exposure. In cumulative sampling, controls are sampled after the source population has gone through a period of risk, which is presumed to be over when the study is conducted. For example, a case-control study examining the effect of vaccination on the risk of influenza may be conducted at the end of the influenza season, when the annual epidemic has ended. In the cumulative case-control study, the control group is sampled from among those who did not become cases during the period of risk. If the cases were those who contracted influenza, the controls would be sampled from among those who remained free of influenza during the epidemic. In the case-cohort study, controls are sampled from the list of all people in the source population, regardless of whether they become cases. As a result, in a case-cohort study, some of the controls may also be included as cases. These three case-control designs are discussed further in the following sections.

Density Case-Control Studies

The phrase *density-based sampling* comes from the term *incidence density*, which is sometimes used as a synonym for incidence rate. The aim of this type of control sampling is to have the distribution of controls mirror the distribution of person-time in the source population with respect to exposure. If 20% of the person-time in the source population is classified as exposed person-time, the aim of a density case-control study will be to sample controls in such a way that 20% of them are exposed. In an actual study, we ordinarily do not know the exposure distribution in the source population, and we rely on our sampling methods to reveal it through our control series.

Suppose that we have a dichotomous exposure. We can consider the source population to have two subcohorts, exposed and unexposed, which we denote

by the subscripts 1 and 0, respectively. During a given time period, the incidence rates for the exposed population would be

$$I_1 = \frac{a}{PT_1}$$

For the unexposed population, it would be

$$I_0 = \frac{b}{PT_0}$$

In these equations, I_1 and I_0 are the incidence rates among exposed and unexposed, respectively; a and b are the respective numbers of exposed and unexposed people who developed the disease; and PT_1 and PT_0 are the respective amounts of person-time at risk for the exposed and the unexposed cohorts.

In a case-control study with density-based sampling, the control series provides an estimate of the proportion of the total person-time for exposed and unexposed cohorts in the source population. Suppose that the control series sampled from the source population contains c exposed people and d unexposed people. The aim is to select the control series so that the following ratios are equal, apart from statistical sampling error:

$$\frac{c}{d} = \frac{PT_1}{PT_0}$$

Equivalently,

$$\frac{c}{PT_1} = \frac{d}{PT_0}$$

The ratios c/PT_1 and d/PT_0 are called the *control sampling rates* for the exposed and unexposed components of the source population. These sampling rates will be equal if the control sampling is conducted independently of exposure. If this goal is achieved, the incidence rate ratio can be readily estimated from the case-control data as follows:

$$\frac{I_1}{I_0} = \frac{a/PT_1}{b/PT_0} = \frac{a}{b} \times \frac{PT_0}{PT_1} = \frac{a}{b} \times \frac{d}{c} \quad \text{because} \quad \frac{d}{c} = \frac{PT_0}{PT_1}$$

The quantity ad/bc , which in a case-control study provides an estimate of the incidence rate ratio, is called the *cross-product ratio* or, more commonly, the *odds ratio*. Using the odds ratio in a case-control study with density-based sampling, an investigator can obtain a valid estimate of the incidence rate ratio in a population without having to obtain individual information on every person in the population.

What disadvantage is there in using a sample of the denominators rather than measuring the person-time experience for the entire source population? Sampling

of the source population can lead to an inaccurate measure of the exposure distribution, giving rise to an incorrect estimate. A case-control study offers less statistical precision in estimating the incidence rate ratio than a cohort study of the same population. A loss in precision is to be expected whenever sampling is involved. This loss can be kept small if the number of controls selected per case is large. The loss is offset by the cost savings of not having to obtain information on everyone in the source population. The cost savings may allow the epidemiologist to enlarge the source population and therefore obtain more cases, resulting in a better overall estimate of the incidence rate ratio statistically and otherwise than would be possible using the same expenditures to conduct a cohort study.

DEFINING THE SOURCE POPULATION

The earlier discussion presumes that all people who develop the disease of interest in the source population are included as cases in the case-control study. The definition of the source population corresponds to the eligibility criteria for cases to enter the study. In theory, it is not necessary to include all cases occurring within an identifiable population, such as within a geographic boundary. The cases identified in a single clinic or treated by a single medical practitioner can be used for case-control studies. The corresponding source population for the cases treated in a clinic is all people who would attend that clinic and be recorded with the diagnosis of interest if they had the disease in question. It is important to specify "if they had the disease in question" because clinics serve different populations for different diseases, depending on referral patterns and the reputation of the clinic in specific specialty areas. Unfortunately, without a precisely identified source population, it may be difficult or impossible to select controls in an unbiased fashion.

CONTROL SELECTION

In density case-control studies, the control series is sampled to represent the person-time distribution of exposure in the source population. If the sampling is conducted independently of the exposure, the case-control study can provide a valid estimate of the incidence rate ratio. Each control sampled represents a certain amount of person-time experience. The probability of any given person in the source population being selected as a control should be proportional to his or her person-time contribution to the denominators of the incidence rates in the source population. For example, a person who is at risk of becoming a study case for 5 years should have a five times higher probability of being selected as a control than a person who is at risk for only 1 year.

For each person contributing time to the source population experience, the time that he or she is eligible to be selected as a control is the same time during which he or she is also eligible to become a case if the disease occurs. A person who has already developed the disease or has died is no longer eligible to be selected as a control. This rule corresponds to the treatment of subjects in cohort studies: Every case that is tallied in the numerator of a cohort study contributes to the denominator of the rate until the time that the person becomes a case, when the contribution to the denominator ceases.

One way to implement control sampling according to these guidelines is to choose controls from the unique set of people in the source population who are at risk of becoming a case at the precise time that each case is diagnosed. This set, which changes from one case to the next as people enter and leave the source population, is sometimes referred to as the *risk set* for the case. Risk-set sampling allows the investigator to sample controls so that each control is selected in proportion to his or her time contribution to the person-time at risk.

A conceptually important feature of the selection of controls with density-based sampling is their continuous eligibility to become cases if they develop the disease. Suppose that the study period spans 3 years and that a given person free of disease in year 1 is selected as a control. The same person may develop the disease in year 3, becoming a case. How is such a person treated in the analysis? If the disease is uncommon, it will matter little, because a study is unlikely to have many subjects eligible to be both a case and a control, but the question is nevertheless of some theoretical interest. Because the person in question did develop disease during the study period, many investigators would be tempted to count the person as a case, not as a control. Recall, however, that if a cohort study were being conducted, each person who developed disease would contribute not only to the numerator of the disease rate, but also to the person-time experience counted in the denominator, until the time of disease onset. The control group in density case-control studies is intended to provide estimates of the relative size of the denominators of the incidence rates for the compared groups. Therefore, each case should have been eligible to be a control before the time of disease onset; each control should be eligible to become a case as of the time of selection as a control. A person selected as a control who later develops the disease and is selected as a case should be included in the study both as a control and as a case.

As an extension of the previous point, with density-based sampling, a person selected as a control should remain eligible to be selected again as a control as long as he or she remains at risk for disease in the study population. Although unlikely in typical studies, the same person may appear in the control group two or more times. Note, however, that including the same person at different times does not necessarily lead to exposure (or confounder) information being repeated, because this information may change with time. For example, in a case-control study of viral hepatitis, the investigator may ask about raw shellfish ingested within the previous 6 weeks. Whether a person has consumed raw shellfish during the previous 6 weeks will change with time, and a person included more than once, first as a control and then later as either a control or as a case, may have different exposure information at the different points in time. The same can be true for confounding variables, which may also change with time.

ILLUSTRATION OF DENSITY-BASED CASE-CONTROL DATA

Consider the data for the cohort study in Table 4-7 (see Chapter 4). These data are shown again in Table 5-5 along with a hypothetical control series of 500 women that might have been selected from the two cohorts.

The ratio of the rates for the exposed and unexposed cohorts is $14.6/7.9 = 1.86$. Suppose that instead of conducting a cohort study, the investigators conducted a density case-control study by identifying all 56 breast cancer cases that

Table 5-5 HYPOTHETICAL CASE-CONTROL DATA FROM A COHORT STUDY OF BREAST CANCER AMONG WOMEN TREATED FOR TUBERCULOSIS WITH X-RAY FLUOROSCOPIES AND FULL COHORT DATA FOR COMPARISON

	Radiation Exposure		Total
	Yes	No	
Breast cancer cases	41	15	56
(Person-years)	(28,010)	(19,017)	(47,027)
Control series (people)	298	202	500
Rate (cases/10,000 person-years)	14.6	7.9	11.9

occurred in the two cohorts and a control series of 500 women. The control series should be sampled from the person-time of the source population so that the exposure distribution of the controls sampled mirrors the exposure distribution of the person-time in the source population. Of the 47,027 person-years of experience in the combined exposed and unexposed cohorts, 28,010 (59.6%) are person-years of experience that relate to radiation exposure. If the controls are sampled properly, we would expect that more or less 59.6% of them would be exposed and the remainder unexposed. If we happened to get just the proportion that we would expect to get on the average, we would have 298 exposed controls and 202 unexposed controls, as indicated in Table 5-5.

Table 5-5 shows the case and control series along with the full cohort data for comparison. In an actual case-control study, the data would look like those in Table 5-6. Because there are two rows of data and two columns with four cell frequencies in the table (not counting the totals on the right), this type of table is often referred to as a 2×2 table.

From these data, we can calculate the odds ratio to get an estimate of the incidence rate ratio.

$$\text{Odds ratio} = \frac{41 \times 202}{15 \times 298} = 1.85 = \text{Incidence rate ratio}$$

This result differs from the incidence rate ratio from the full cohort data by only a slight rounding error. Ordinarily, we would expect to see additional error because the control group is a sample from the source population, and there may be some difference between the exposure distribution in the control series and the exposure distribution in the source population. In Chapter 9, we see how to take this sampling error into account.

Table 5-6 CASE-CONTROL DATA ALONE FROM TABLE 5-5

	Radiation Exposure		Total
	Yes	No	
Breast cancer cases	41	15	56
Controls	298	202	500

Cumulative Case-Control Studies

Density case-control studies correspond to cohort studies that measure person-time and estimate rates. The effect estimates from density case-control studies are estimates of rate ratios, with each control representing a certain amount of person-time. In cumulative case-control studies or in case-cohort studies, each control represents a certain number of people. These studies correspond to cohort studies that follow a closed population and measure risks, rather than rates. The effect estimate obtained from cumulative case-control studies or from case-cohort studies is a risk ratio rather than a rate ratio.

When sampling controls from a closed source population, an investigator may choose to sample the controls from the entire source population at the start of follow-up or at the end of the follow-up from the noncases that remain after the cases have been identified. If the control sample is drawn from the entire source population at the start of the follow-up, the design is called a *case-cohort study*, which is described later. If the control sample is drawn from the noncases at the end of the follow-up, the design is called a *cumulative case-control study*. Cumulative case-control studies are often conducted at the end of an epidemic period or a specific but time-limited risk period. For example, an investigator may be interested in the effect of specific drug exposures during early pregnancy on the occurrence of birth defects. To conduct a case-control study that addresses this issue, the investigator may identify cases who are born with birth defects. Typically, the control series is sampled from babies born without birth defects. At birth, the period of risk for birth defects is over, so the case and control sampling occurs after the risk period has ended. There is an intuitive appeal to choosing controls from among those babies who did not develop a birth defect, but such babies do not represent the experience of the entire source population. Some babies who were at risk for birth defects may not have survived to be born alive, but even if all had, selecting controls from among those born without birth defects omits the experience of the cases from the population at risk. Because the experience of cases is part of the overall experience of the source population, omitting them from the control series can result in a bias that will overestimate the risk ratio.

In a cumulative case-control study, the risk ratio is estimated from the same measure used in density case-control studies, the odds ratio.

$$\text{Odds ratio} = \frac{ad}{bc}$$

In this equation, a and b are the number of exposed and unexposed cases, respectively, and c and d are the number of exposed and unexposed controls. Because of the sampling approach in a cumulative case-control study, this odds ratio is an estimate of the risk ratio, rather than the rate ratio obtained from density case-control studies. If the disease is rare, the experience of cases will be a small part of the overall experience of the source population, and the odds ratio obtained from cumulative control sampling will be very close to the risk ratio. If the risk for disease is high enough, however, the cumulative case-control study can seriously overestimate the risk ratio.

We can illustrate this phenomenon with a hypothetical example. In Table 5-7, the top section shows hypothetical data from a cohort study of a closed population of 200 people, one half of whom are exposed. The risk of disease is 40% among exposed and 10% among unexposed, for a risk ratio of 4.0. The next section in the table shows the result if a case-control study had been conducted, with all cases included and 50 controls, using cumulative sampling. At the end of follow-up, there were a total of 50 cases, leaving 150 people who did not get the disease. Suppose that 50 controls were sampled from these 150 noncases. The exposure distribution of these controls is shown in the next section of Table 5-7. Although one half of the closed source population was exposed, only 40% of the noncases at the end of follow-up are exposed (there are 60 noncases among those exposed and 90 among those unexposed). The reason for this discrepancy is that exposure is associated with disease, and the disease is common, leaving fewer exposed noncases than unexposed noncases. The estimate of the risk ratio measuring the effect of exposure is obtained from the odds ratio. For the cumulative sampling, the odds ratio = $(40 \times 30)/(10 \times 20) = 6.0$, considerably greater than the correct value for the risk ratio of 4.0. This departure is a bias that results from the sampling method, combined with high risk of disease. If the risks were 4% and 1% among exposed and unexposed, the odds ratio from the same sampling approach would be about 4.1 instead of 6, much closer to the correct value of 4.0 for the risk ratio. As the risk for disease becomes very small, sampling from the noncases at the end of follow-up becomes almost identical to sampling from the entire cohort.

Cumulative sampling results in valid estimates of the risk ratio if the risk of disease is sufficiently low. This condition is described as the *rare disease assumption*. In the past, it has been mistakenly thought to be a necessary assumption to get a valid result for any case-control study, but the rare disease assumption is needed only for cumulative case-control studies. Density case-control studies, for example, provide unbiased estimates of the rate ratio even when the disease is common. When the rare disease assumption is met, it is worth keeping in mind that the risk ratio will approximate the rate ratio, which is to say that in a cumulative case-control study when risks are small, the odds ratio, the risk ratio, and the rate ratio will all be close to the same value.

Case-Cohort Studies

The third basic approach to control sampling is the *case-cohort study*. In this type of study, each control represents a certain number of people in the source

Table 5-7 CUMULATIVE SAMPLING VERSUS CASE-COHORT SAMPLING
IN A CASE-CONTROL STUDY

	Exposed	Unexposed	Risk or Odds Ratio
Cases	40	10	
Cohort denominator	100	100	4.0
Controls (cumulative)	20	30	6.0
Controls (case-cohort)	25	25	4.0

population, just as in the cumulative case-control study. As in the cumulative case-control study, the case-cohort study ordinarily provides estimates of risk ratio, rather than rate ratio. In the case-cohort study, however, the controls are sampled from the entire source population rather than from the noncases. Every person in the source population has the same chance of being included in the study as a control, regardless of whether that person becomes a case. The case-cohort design may also be used even if subjects are followed for various amounts of time. With this type of sampling, each control participant represents a fraction of the total number of people in the source population, rather than a fraction of the total person-time. The risk ratio is, as with cumulative case-control studies, estimated from the odds ratio. Because the controls are sampled from the entire source population, however, there is no need for the rare disease assumption in case-cohort studies. As seen in the lower section of Table 5-7, the exposure distribution among controls in a case-cohort study will, on average, reflect the exposure distribution among all persons followed in the source population, even if disease is common.

If the proportion of subjects that is sampled and becomes part of the control series is known, it is possible to estimate the actual size of the cohorts being followed and to calculate separate risks for the exposed and the unexposed cohorts. Usually, this sampling proportion is not known, in which case the actual risks cannot be calculated. As long as the controls are sampled independently of the exposure, however, the odds ratio will still be a valid estimate of the risk ratio, just as the odds ratio in a density case-control study is an estimate of the incidence rate ratio. No rare disease assumption is needed, because the controls are sampled from the entire source population.

One advantage of a case-cohort study over a density case-control study is convenience. Sufficient data to allow risk-set sampling (discussed earlier) may not be available, for example. Moreover, the investigators may intend to study several diseases. In risk-set sampling, each control must be sampled from the risk set (ie, the set of people in the source population who are at risk for disease at that time) for each case. The definition of the risk set changes for each case, because the identity of the risk set is related to the timing of the case. If several diseases are to be studied, each disease will require its own control group to maintain risk-set sampling. Control sampling for a case-cohort study requires only a single sample of people from the roster of people who constitute the cohort. The same control group can be used to compare with various case series, just as the same denominators for calculating risks can be used to calculate the risk for various diseases in the cohort. This approach can therefore be considerably more convenient than density sampling.

In a case-cohort study, a person who is selected as a control may also be a case in the study (the same possibility exists in a density case-control study). This possibility may seem bothersome; some epidemiologists take pains to avoid the possibility that a control subject may have even an undetected stage of the disease under study. Nevertheless, there is no theoretical difficulty with a control participant also being a case. The control series in a case-cohort study is a sample of the entire list of people who are in the exposed and unexposed cohorts. If we did not sample at all but included the entire list, we would have a cohort study from which we could directly calculate risks for exposed and unexposed

groups. In a cohort study risk calculation, every person in the numerator (ie, every case) is also included in the denominator (ie, is a member of the source population). This situation is analogous to the possibility that a person who is sampled as a control subject in a case-cohort study might be someone who has been included as a case. It may be helpful to consider the timing of control versus case selection. If the control series is seen as a sample of the exposed and unexposed cohorts at the start of their follow-up, the control sampling represents people who were free of disease, because everyone at the start of follow-up in a cohort study is free of disease. It is only later that disease develops in some of these people, who then become cases. These parallels in thinking between case-control and cohort studies help to clarify the principles of control selection and illustrate the importance of viewing case-control studies as cohort studies with sampled denominators. If the same subject is included with the same data as both case and control in a case-cohort study, some refined formulas may be used to analyze the data, taking account of the status of subjects who serve a dual role as case and control. *Modern Epidemiology* offers a more detailed discussion of case-cohort studies.¹

ILLUSTRATION OF CASE-COHORT DATA

Consider the data in Table 5-1 describing John Snow's natural experiment. Imagine that he had conducted a case-cohort study instead, with a sample of 10,000 controls selected from the source population of the London neighborhoods that he was investigating. If the control series had the same distribution by water company that the entire population in Table 5-1 had, the data might resemble the 2 × 2 table shown in Table 5-8. We would obtain the odds ratio from these hypothetical case-cohort data as follows:

$$\text{Odds ratio} = \frac{4093 \times 3946}{461 \times 6054} = 5.79 = \text{Risk ratio}$$

This result is essentially the same value that Snow obtained from his natural experiment cohort study. In this hypothetical case-cohort study, 10,000 controls were included, instead of the 440,000 people Snow included in the full cohort-study comparison. If Snow had to determine the exposure status of every person in the population, it would have been much easier to conduct the case-cohort study and sample from the source population. As it happened, Snow derived his exposure distribution by estimating the population of water company customers from business records, making it unnecessary to obtain information on each person. He did have to ascertain the exposure status of each case, however.

Table 5-8 HYPOTHETICAL CASE-COHORT DATA FOR JOHN SNOW'S NATURAL EXPERIMENT

	Water Company	
	Southwark & Vauxhall	Lambeth
Cholera deaths	4,093	461
Controls	6,054	3,946

Sources for Control Series

The ideal method of control selection in a case-control study is to sample controls directly from the source population of cases. If the cases represent all cases or a representative sample of cases within a geographic area, the controls should be sampled from the entire at-risk population of that geographic area. This is a *population-based study*, which means it is based on a geographically defined population. The at-risk subset of the population, which is the source population for cases, is those who met the study inclusion criteria for age, sex, and other factors. This subset also excludes current cases or any other people who were not able to become study cases, such as women with a hysterectomy in a study of endometrial cancer. Control sampling in a population-based study is facilitated if a population registry is available, from which potential controls may be identified, perhaps through random sampling.

Random sampling of controls does not necessarily mean that every person should have an equal probability of being selected to be a control. With density sampling, a person's control selection probability is proportional to the person's time at risk. For example, in a case-control study nested within an occupational cohort, workers on an employee roster have been followed for various lengths of time. Random sampling for a density-based case-control study should reflect the variation in time followed. Random sampling for a case-cohort study in the same setting, however, involves every person on the employee roster having an equal probability of being sampled as a control.

If no registry or roster of the source population is available, other approaches must be found to sample controls from the source population. One approach that has often been used is random-digit dialing. This method is based on the assumption that randomly calling telephone numbers simulates a random sample of the source population. Random-digit dialing offers the advantage of approaching all households in a designated area, even those with unlisted telephone numbers, through a telephone call. The method poses a few challenges, however.

First, the method assumes that every case can be reached by telephone, because the source population being sampled is that part of the total population that is reachable by telephone. If some cases have no telephone, in principle, they should be excluded from a study employing random-digit dialing. The second issue is that random dialing gives every telephone an equal probability of being called, but that is not equivalent to giving every person an equal probability of being called. Households vary in the number of people who reside in them and in the amount of time someone is at home. Third, making contact with a household may require many calls at various times of day and various days of the week. Fourth, it may be challenging to distinguish business from residential telephone numbers, a distinction that affects calculating the proportion of nonresponders. Fifth, the increase in telemarketing in many areas and the availability of caller identification has further compromised response rates to cold calling. Obtaining a control subject meeting specific eligibility characteristics can require dozens of telephone calls. Other problems include answering machines and households with multiple telephone numbers, a rapidly increasing phenomenon. Because telephony

is rapidly changing in rich and poor countries, the use of random-digit dialing is becoming more complicated, and the possible biases introduced by identifying controls using this method must be carefully considered in each study in which the method is contemplated.

Another method to identify population-based controls when the source population cannot easily be enumerated is sampling residences in some systematic fashion. If a geographic roster of residences is not available, some scheme must be devised to sample residences without enumerating them all. Often, matching is employed as a convenience. After a case is identified, one or more controls who reside in the same neighborhood as that case are identified and recruited into the study. With this type of design, neighborhood must be treated as a matching factor (see Chapter 7).

If the case-control study is not population based, it may be based on a referral population in a hospital or clinic. In these studies, the source population represents a group of people who would be treated in a given clinic or hospital if they developed the disease in question. This population may be hard to identify, because it does not correspond to the residents of a specific geographic area. Any clinic-based study can be restricted to a given geographic area, but the hospitals or clinics that provide the cases for the study often treat only a small proportion of those in the geographic area, making the actual source population unidentifiable. A case-control study is still possible, but the investigator must take into account referral patterns to the hospital or clinic in the sampling of controls. Typically, he or she would draw a control series from patients treated at the same hospitals or clinics as the cases. The source population does not correspond to the population of the geographic area, but only to those who would attend the hospital or clinic if they contracted the disease under study. Other patients treated at the same hospitals or clinics as the cases will constitute a sample, albeit not a random sample, of this source population.

The major problem with any nonrandom sampling of controls is the possibility that they are not selected independently of exposure in the source population. Patients hospitalized with other diseases at the same hospitals, for example, may not have the same exposure distribution as the entire source population, because exposed people are more or less likely than nonexposed people to be hospitalized for the control diseases if they develop them or because the exposure may cause or prevent these control diseases in the first place. Suppose the study aims to evaluate the relation between tobacco smoking and leukemia. If controls are people hospitalized with other conditions, many of them will have been hospitalized for conditions that are caused by smoking. A variety of other cancers, cardiovascular diseases, and respiratory diseases are related to smoking. Thus, a series of people hospitalized for diseases other than leukemia may include more smokers than the source population from which they came. One approach to this problem is to exclude any diagnosis from the control series that is likely to be related to the exposure. For example, in the imagined study of smoking and leukemia, it would be reasonable to exclude from the control series anyone who was hospitalized with a disease thought to be related to smoking. This approach may lead to the exclusion of many diagnostic categories, but even a few remaining diagnostic categories should suffice to find enough control subjects.

It is risky, however, to reduce the control eligibility criteria in a hospital-based case-control study to a single diagnosis. Using a variety of diagnoses has the advantage of diluting any bias that may result from including as the control series only a specific diagnostic group that turns out to be related to the exposure. For the diagnostic categories that constitute exclusion criteria for controls, the exclusion should be based only on the cause of the hospitalization used to identify the study subject, rather than on any previous hospitalization. In the example of a hospital-based case-control study of tobacco smoking and leukemia, a person who was hospitalized because of a traumatic injury and is therefore eligible to be a control should not be excluded if he or she had previously been hospitalized for cardiovascular disease. The reason is that the source population includes people who have had cardiovascular disease, and they must also be included in the control series. In considering whether to exclude potential controls, the investigator must distinguish between the current hospitalization and past hospitalizations.

In some situations, it is impractical or impossible to identify the actual source population for cases. It may still be possible to conduct a valid study, however, by the use of *proxy sampling*. Theoretically, if a control series can be identified that has the same exposure distribution as does the source population for cases, that control series should give the same results as one that draws controls directly from the source population. A control series comprising people who are not in the source population but who serve as valid proxies for those who are in the source population is a reasonable study design.

Consider a case-control study examining the relation between ABO blood type and female breast cancer. Could such a study have a control series comprising the brothers of the (female) cases? The brothers of the cases are not part of the source population. Nevertheless, the distribution of ABO blood type among the brothers should be identical to the distribution of ABO blood type among the source population of women who might have been included as cases, because ABO blood type is not related to sex. Clinic-based studies that use as controls clinic patients with disease diagnoses different from that of the cases may also involve proxy sampling if those control patients would not have come to the same clinic if they had been diagnosed with the disease that the cases have. In studies in which cases who have died are compared with a control series comprising dead people, a comparison sometimes justified by the interest in getting comparable information for cases and controls, the controls cannot be part of the source population. Death precludes the occurrence of any further disease, and dead people therefore are not at risk to become cases. Nevertheless, if a series of dead controls can provide the same exposure distribution as exists in the source population, it may be a reasonable control series to use.

Prospective and Retrospective Case-Control Studies

Case-control studies, like cohort studies, can be prospective or retrospective. In a retrospective case-control study, cases have already occurred when the study begins; there is no waiting for new cases to occur. In a prospective case-control

IS REPRESENTATIVENESS IMPORTANT?

Some textbooks claim that cases should be representative of all persons with the disease and that controls should be representative of the entire non-diseased population. Such advice can be misleading. Cases can be defined in any way that the investigator wishes and need not be representative of all cases. Older cases, female cases, severe cases, or any clinical subset of cases can be studied. These groups are not representative of all cases but are allowable as case definitions. Any type of case that can be used as the disease event in a cohort study also can be used to define the case series in a case-control study.

The case definition implicitly defines the source population for cases, from which the controls should be drawn. It is this source population for the cases that the controls should represent, not the entire nondiseased population.

study, the investigator must wait, just as in a prospective cohort study, for new cases to occur.

COHORT/CASE-CONTROL STUDIES VERSUS PROSPECTIVE/RETROSPECTIVE STUDIES

Early descriptions of cohort studies often referred to them as prospective studies and to case-control studies as retrospective studies. We now reserve the terms *prospective* and *retrospective* to refer to the timing of the information and events of the study, and we use the term *case-control* to describe studies in which the source population is sampled rather than ascertained in its entirety, as in a cohort study. The early descriptions carried the implication that retrospective studies were less valid than prospective studies, an idea that lingers. It is still commonly thought that case-control studies are less valid than cohort studies. The truth is that validity issues can affect both case-control studies and cohort studies (including randomized trials), whether they are prospective or retrospective. Nevertheless, there is no reason to discount a study simply because it is a case-control study or a retrospective study. Case-control studies represent a high achievement of modern epidemiology, and if conducted well, they can reach the highest standards of validity.

Case-Crossover Studies

Many variants of case-control studies have been described in recent years. One that is compelling in its simplicity and its elegance is the case-crossover study, which is a case-control analog of the crossover study. A *crossover study* is a self-matched cohort study, usually an experimental study, in which two or more interventions

are compared, with each study participant receiving each of the interventions at different times. If the crossover study is an experiment, each subject receives the interventions in a randomly assigned sequence, with some time interval between them so that the outcome can be measured after each intervention. A crossover study requires the effect period related to the intervention to be short enough so that it does not persist into the time period during which the next treatment is administered.

The *case-crossover study*, first proposed by Maclure,¹⁴ may be considered a case-control version of the crossover study. Unlike an ordinary case-control study, however, all the subjects in a case-crossover study are cases. The control series, rather than being a different set of people, is represented by information on the exposure distribution drawn from the cases themselves, outside of the time window during which the exposure is hypothesized to cause the disease. Usually, this information is drawn from the experience of cases before they develop disease to address the concern that after getting the disease, a person may modify the exposure, as in someone who cuts back on caffeine after having a myocardial infarction. In some situations, however, when the disease cannot or does not influence subsequent exposure, the experience of cases before and after their disease event may be used. For example, if studying the effect of transient air pollution on asthma attacks, air pollution levels after a case's asthma attack may be used to describe the frequency of air pollution episodes, because the asthma attack cannot affect ambient air pollution levels.

The case-crossover study design can be implemented successfully only for an appropriate study hypothesis. As in a crossover study, in a case-crossover study, the effect of the exposure must be brief, and the disease event ideally will have an abrupt onset. The study hypothesis defines a time window during which the exposure may cause the disease event. The window should be brief in relation to the time between typical successive exposure intervals, so that the effect of exposure will have sufficient time to fade before the next episode of exposure according to the study hypothesis. Each case is classified as exposed or unexposed, depending on whether there was any exposure during the hypothesized time window just before the disease event. Maclure used the example of studying whether sexual intercourse causes myocardial infarction. The period of increased risk after sexual intercourse was hypothesized to be 1 hour. The cases would be a series of people who have had a myocardial infarction. Each case would then be classified as exposed if he or she had sexual intercourse within the hour preceding the myocardial infarction. Otherwise, the case would be classified as unexposed.

This process appears to differ little from what may be done for any case-control study. The key difference is that there is no separate control series; instead, the control information is obtained from the cases themselves. In the example of sexual intercourse and myocardial infarction, the average frequency of sexual intercourse would be ascertained for each case during a period (eg, 1 year) before the myocardial infarction occurred. Under the study hypothesis, after each instance of sexual intercourse, the risk of myocardial infarction during the following hour is elevated, and that hour is considered exposed person-time. All other time would be considered unexposed. If a person had sexual intercourse once per week, 1 hour per week would be considered exposed and the remaining

167 hours would be considered unexposed. Such a calculation can be performed for each case, and from the distribution of these hours within the experience of each case, the incidence rate ratio of myocardial infarction after sexual intercourse in relation to the incidence rate at other times can be estimated. The analysis uses analytic methods based on matching, with each case being self-matched to his or her own experience outside the case time window. Thus, all the information for the study is obtainable from a series of cases.

Only certain types of study questions can be studied with a case-crossover design. The exposure must be something that varies from time to time for a person. The effect of blood type cannot be examined in a case-crossover study, because it does not change. An investigator can study whether coffee drinking triggers an asthma attack within a short time, however, because coffee is consumed intermittently. It is convenient to think of the case-crossover study as evaluating exposures that trigger a short-term effect. The disease also must have an abrupt onset. The causes of multiple sclerosis cannot be considered in a case-crossover study, but whether an automobile driver who is talking on a telephone is at higher risk of having a collision can. The effect of the exposure must be brief. If the exposure had a long effect, it would not be possible to relate the disease to a particular episode of exposure.

CROSS-SECTIONAL VERSUS LONGITUDINAL STUDIES

All of the study types previously described in this chapter can be described as *longitudinal* studies. In epidemiology, a study is considered to be longitudinal if the information obtained pertains to more than one point in time. Implicit in a longitudinal study is the universal premise that the causal action of an exposure comes before the subsequent development of disease as a consequence of that exposure. This concept is integral to the thinking involved in following cohorts over time or in sampling from the person-time at risk based on earlier exposure status. All cohort studies and most case-control studies rely on data in which exposure information refers to an earlier time than that of disease occurrence, making the study longitudinal.

Occasionally, epidemiologists conduct cross-sectional studies, in which all of the information refers to the same point in time. These studies are basically snapshots of the population status with respect to disease or exposure variables, or both, at a specific point in time. A population survey, such as the decennial census in the United States, is a cross-sectional study that attempts to enumerate the population and to assess the prevalence of various characteristics. Surveys are conducted frequently to sample opinions, but they may also be used to measure disease prevalence or even to assess the relation between disease prevalence and possible exposures.

A cross-sectional study cannot measure disease incidence, because risk or rate calculations require information across a time period. Nevertheless, cross-sectional studies can assess disease prevalence. It is possible to use cross-sectional data to conduct a case-control study if the study includes prevalent cases and uses concurrent information about exposure. A case-control study that is based on prevalent cases, rather than new cases, does not necessarily provide information about the causes of disease. Because the cases in such a study are those who

have the disease at a given point in time, the study is more heavily weighted with cases of long duration than any series of incident cases would be. A person who died soon after getting disease, for example, would count as an incident case but likely would not be included as a case in a prevalence survey, because the disease duration is so brief.

Sometimes, cross-sectional information is used because it is considered a good proxy for longitudinal data. For example, an investigator may wish to know how much supplemental vitamin E a person consumed 10 years in the past. Because no written record of this exposure is likely to exist, the basic choices are to ask people to recall how much supplemental vitamin E they consumed in the past or to find out how much they consume now. Recall of past use is likely to be hazy, whereas current consumption can be determined accurately. In some situations, accurate current information may be a better proxy for the actual consumption 10 years earlier than the hazy recollections of that past consumption. Current consumption may be cross-sectional, but it would be used as a proxy for exposure in the past. Another example is blood type. Because it remains constant, cross-sectional information on blood type is a perfect proxy for past information about blood type. In this way, cross-sectional studies can sometimes be almost as informative as longitudinal studies with respect to causal hypotheses.

RESPONSE RATES

In a cohort study, if a substantial proportion of subjects cannot be traced to determine the disease outcome, the study validity can be compromised. In a case-control study, if exposure data is missing on a sizable proportion of subjects, it can likewise be a source of concern. The concern stems from the possibility of bias from selectively missing data, which is a form of selection bias (see Chapter 7). The more missing data on outcome there is in a cohort study, or analogously the more missing exposure data there is in a case-control study, the greater the potential for selection bias. For that reason, critics are often skeptical about cohort studies with a high proportion of subjects with unknown disease outcome or about case-control studies with a high proportion of subjects lacking exposure information. These two proportions are sometimes referred to as *response rates*, with the disease outcome corresponding to the response in a cohort study and the exposure information corresponding to the response in a case-control study.

There is no absolute threshold for what a response rate ought to be, but as discussed for tracing subjects in a cohort study, if the response rate or proportion traced is less than 70% to 75%, it may engender some skepticism about the study. Nevertheless, in some settings, a low response rate need not be an important validity concern. Suppose that in a case-control study of risk of acquired immunodeficiency syndrome (AIDS) after transfusion, the exposure information, a history of transfusion, was ascertained from medical records but that only one half of the desired medical records were obtainable for review. Even so, if there is no association between a history of transfusion and the availability of the records for review, an unbiased estimate of the effect of transfusion should be obtainable

from the records that are available. Selection bias is a concern when the interrelation between the study variables in the missing data is different from the corresponding relation in the available data.

In many cohort studies, subjects are recruited as volunteers from a larger population, perhaps from the general population. Recruitment of volunteers is a well-known feature of experimental studies, but it is also common in nonexperimental cohort studies. The recruitment proportion should not be viewed as a response rate. Even if recruitment is difficult, there may be little reason to be concerned with the validity of the study results. In a randomized experiment, the internal validity of the study flows from the random assignment, which is implemented among those who volunteer to be studied, even if they represent a small proportion of a larger population. In other cohort studies, the participants who are recruited should all be free of disease at the start of follow-up. Regardless of the success or lack of success in recruiting volunteers, there will not be any selection bias from volunteering to be studied unless volunteering is related to both exposure and disease risk. Because disease has not yet occurred in cohort participants, the presence of disease cannot influence volunteering, apart from prodromal effects. The internal comparisons of a cohort study based on a select group of volunteers should ordinarily be free of selection bias stemming from volunteering even if recruitment success is poor. The external validity, or generalizability, of the study may be affected by a low recruitment rate, but only if the study participants represent a subset of the population in which the relation between exposure and disease is different from the relation for those who did not volunteer. These considerations have implications for the effort expended to recruit study participants. In some cohort studies, it may be most sensible to have a gentle approach to recruit volunteers as opposed to a hard sell that recruits more reluctant volunteers. If reluctant participants are persuaded to volunteer for the study but later drop out, their missing follow-up data may have a more profound effect on the study's validity than their nonparticipation in the study in the first place. Because dropouts are usually worse for a study than refusals to participate from the beginning, a better strategy is to concentrate efforts more on follow-up than on recruitment. On the other hand, in case-control studies in which study participants know their exposure status, getting high levels of participation is important, because agreement to be studied may depend on exposure. If study participants do not know their exposure status, because, for example, it is obtainable only from a laboratory test that most people would not have had, low recruitment into a case-control study is less of a concern.

COMPARISON OF COHORT AND CASE-CONTROL STUDIES

It may be helpful to summarize some of the key characteristics of cohort and case-control studies. The primary difference is that a cohort study involves complete enumeration of the denominator (ie, people or person-time) of the disease measure, whereas case-control studies sample from the denominator. As a result, case-control studies usually provide estimates only of ratio measures of effect, whereas cohort studies provide estimates of disease rates and risks for each cohort, which

Table 5-9 COMPARISON OF THE CHARACTERISTICS OF COHORT AND CASE-CONTROL STUDIES

Cohort Study	Case-Control Study
Complete source population denominator experience tallied	Sampling from source population
Can calculate incidence rates or risks, and their differences and ratios	Can calculate only the ratio of incidence rates or risks (unless the control sampling fraction is known)
Usually very expensive	Usually less expensive
Convenient for studying many diseases	Convenient for studying many exposures
Can be prospective or retrospective	Can be prospective or retrospective

can then be compared by taking differences or ratios. Case-control studies can be thought of as modified cohort studies, with sampling of the source population being the essential modification.

Consistent with this theme is the idea that many issues that apply to cohort studies apply to case-control studies in the same way. For example, if a person gets a disease, he or she no longer contributes time at risk to the denominator of a rate in a cohort study (assuming that only the first occurrence of disease in a person is being studied). Analogously, in a density case-control study, a person who gets disease is from that point in time forward no longer eligible to be sampled as a control. Another example of the parallels between cohort studies and case-control studies is the classification of studies into prospective and retrospective studies.

Case-control studies are usually more efficient than cohort studies because the cost of the information that they provide is often much lower. With a cohort study, it is often convenient to study many different disease outcomes in relation to a given exposure. With a case-control study, it is often convenient to study many different exposures in relation to a single disease. This contrast, however, is not absolute. In many cohort studies, a variety of exposures can be studied in relation to the diseases of interest. Likewise, in many case-control studies, the case series can be expanded to include more than one disease category, which in effect leads to several parallel case-control studies conducted within the same source population. These characteristics are summarized in Table 5-9.

QUESTIONS

1. During the second half of the 20th century, there was a sharp increase in hysterectomy in the United States. Concurrent with that trend, there was an epidemic of endometrial cancer that has been attributed to widespread use of replacement estrogens among menopausal women. The epidemic of endometrial cancer was not immediately evident, however, in data on endometrial cancer rates compiled from cancer registries. Devise a hypothesis based on considerations of the population at risk for endometrial cancer that can explain why the epidemic went unnoticed.

2. What is the purpose of randomization in an experiment? How is the same goal achieved in nonexperimental studies?

3. When cancer incidence rates are calculated for the population covered by a cancer registry, the usual approach is to take the number of new cases of cancer and divide by the person-time contributed by the population covered by the registry. Person-time is calculated as the size of the population from census data multiplied by the time period. This calculation leads to an underestimate of the incidence rate in the population at risk. Explain.

4. In the calculations of rates for the data in Figure 5-2, the rate in the exposed group declined after taking the induction period into account. If exposure does cause disease, would you expect that the rate in exposed people would increase, decrease, or stay the same after taking into account an appropriate induction period?

5. If a person already has disease, can that person be selected as a control in a case-control study of that disease?

6. If a person has already been selected as a control in a case-control study and then later during the study period develops the disease that is being studied, should the person be kept in the study as (1) a case, (2) a control, (3) both, or (4) neither?

7. In case-cohort sampling, a single control group can be compared with various case groups in a set of case-control comparisons, because the control sampling depends on the identity of the cohorts and has nothing to do with the cases. Analogously, the denominators of risk for a set of several different diseases occurring in the cohort will be the same. Risk-set sampling, in contrast, requires that the investigator identify the risk set for each case; the sample of controls will be different for each disease studied. If the analogy holds, this observation implies that the denominators of incidence rates will differ when calculating the rates for different diseases in the same cohort. Is that true? If not, why not? If so, why should the denominators for the risks not change no matter what disease is studied, whereas the denominators for the rates change from studying one disease to another?

8. Explain why it would not be possible to study the effect of cigarette smoking on lung cancer in a case-crossover study but why it would be possible to study the effect of cigarette smoking on sudden death from arrhythmia using that design.

9. Cumulative case-control studies are conducted by sampling the controls from people who remain free of disease after the period of risk for disease (eg, an epidemic period) has ended. With this sampling strategy, demonstrate why the odds ratio will tend to be an overestimate of the risk ratio.

10. Often, the time at which disease is diagnosed is used as the time of disease onset. Many diseases, such as cancer, rheumatoid arthritis, and schizophrenia, may be present in an undiagnosed form for a considerable time before the diagnosis. Suppose you are conducting a case-control study of a cancer. If it were possible, would it be preferable to exclude people with undetected cancer from the control series?

REFERENCES

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
2. Porta M. *A Dictionary of Epidemiology*. 5th ed. New York, NY: Oxford University Press; 2008.
3. Snow J. *On the Mode of Communication of Cholera*. 2nd ed. London, England: John Churchill; 1860. (Facsimile of 1936 reprinted edition by Hafner, New York, 1965.)
4. Kinloch-de Loes S, Hirschel BJ, Hoen B, et al. Controlled trial of zidovudine in primary human immunodeficiency virus infection. *N Engl J Med*. 1995;333:408-413.
5. Francis TF, Korn RF, Voight RB, et al. An evaluation of the 1954 poliomyelitis vaccine trials. *Am J Public Health*. 1955;45(suppl):1-63.
6. Bang AT, Bang RA, Baitule SB, Reddy MH, Deshmukh MD. Effect of home-based neonatal care and management of sepsis on neonatal mortality: field trial in rural India. *Lancet*. 1999;354:1955-1961.
7. Milunsky A, Jick H, Jick SS, et al. Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects. *JAMA*. 1989;262:2847-2852.
8. Rothman KJ, Moore LL, Singer MR, Nguyen US, Mannino S, Milunsky A. Teratogenicity of high vitamin A intake. *N Engl J Med*. 1995;333:1369-1373.
9. Dawber TR, Moore FE, Mann GV. Coronary heart disease in the Framingham study. *Am J Public Health*. 1957;47:4-24.
10. Richardson DB, MacLehose RF, Langholz B, Cole SR. Hierarchical latency models for dose-time-response associations. *Am J Epidemiol*. 2011;173:695-702.
11. Morrison AS, Kirshner J, Molho A. Epidemics in Renaissance Florence. *Am J Public Health*. 1985;75:528-535.
12. Huybrechts KF, Mikkelsen EM, Christensen T, et al. A successful implementation of e-epidemiology: evidence from the Danish pregnancy planning study "Snart-Gravid." *Eur J Epidemiol*. 2010;25:297-304.
13. Labarthe D, Adam E, Noller KL, et al. Design and preliminary observations of National Cooperative Diethylstilbestrol Adenosis (DESAD) Project. *Obstet Gynecol*. 1978;4:453-458.
14. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133:144-153.