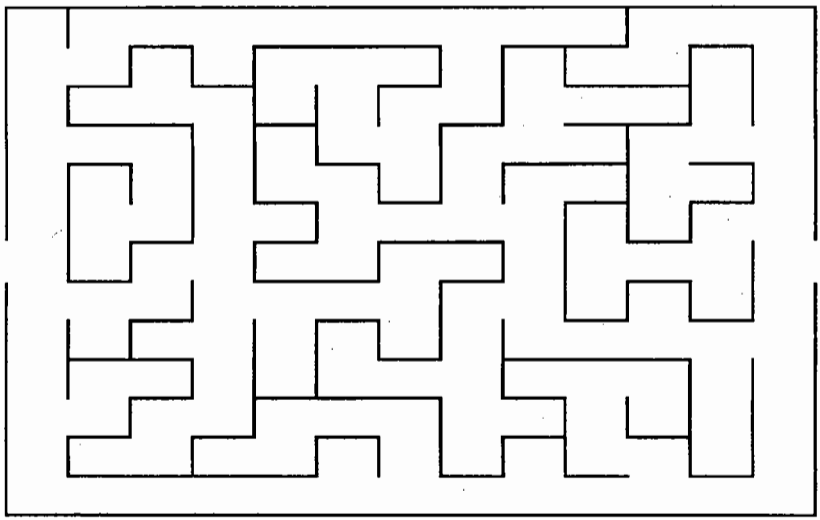
3

Study Types



Investigators who want to learn about disease prevalence or incidence content themselves with observations made on small portions of the populations that they would like to characterize. The process of choosing study subjects from the populations of interest is known as *sampling*. The strategies used to carry out sampling define three classes of epidemiologic study: the *survey*, the *cohort study*, and the *case-control study*. Each is valid, has its own logic, has its pitfalls. The purpose of the present chapter is to introduce the three study types, with an emphasis on their interrelations. Chapters 4 and 5 will explore cohort and case-control studies in greater detail.

Surveys

Surveys describe prevalence. The sampling that yields survey data has the goal of obtaining a study population that is a replica in miniature of a *source population*[23] (Figure 3.1).

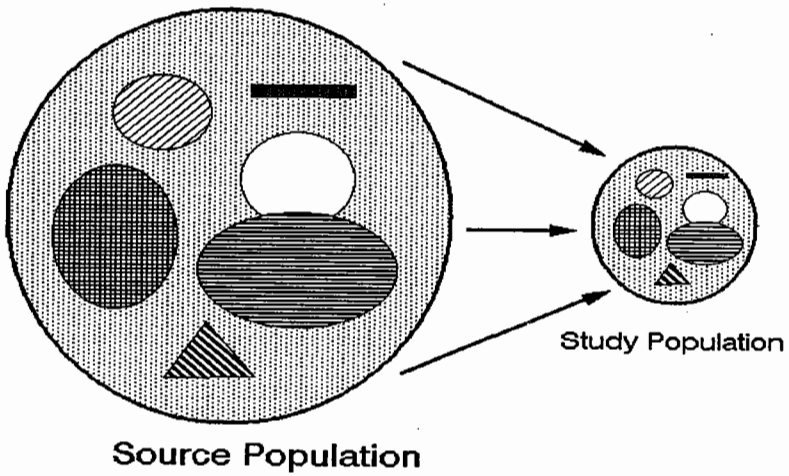


Figure 3.1 A survey

Source population. *The individuals about whose experience or condition a study yields data are the source population. A source population is defined by the identity of the individuals whom it comprises and by the time periods during which each individual is considered to be a member.*

Study population. *The study population is the group of individuals that an investigator observes.*

Sampling. *The process of selecting a study population from a source population, with the goal of learning about characteristics of the source population, is known as sampling.*

23. The source population is also sometimes called the "base population," a term that connotes the role of the source population as the base upon which the structure of a study is erected.

Surveys generally address themselves to the status of individuals, assessed at a single point in time. In surveys, measurements that might be too expensive to make on the entire source population can be made in a study population, and knowledge of the sampling scheme allows you to generalize the particular data obtained to the larger group.

Table 3.1 Prevalence of blood pressure readings among white males aged 35-44 years in the United States in 1970 (percent)

	Diastolic blood pressure (mm Hg)					
Systolic blood pressure	Under 70	70-	80-	90-	100-	Over 109
Under 110	2.3	4.7	1.7	-	-	-
110-119	1.5	6.9	9.1	0.6	-	-
120-129	0.4	6.1	14.9	5.2	0.1	-
130-139	0.4	3.8	10.5	7.0	2.6	0.3
140-149	0.2	0.1	3.4	4.4	2.9	0.9
150-159	-	-	0.4	2.5	2.4	0.8
160-169	-	-	0.2	0.6	0.1	1.4
170 and over	-	0.5	-	-	0.2	0.8

Table 3.1 is drawn from a National Center for Health Statistics survey. While it purports to describe all white males aged 35-44 years in the United States in 1970, the measurements pertain only to some thousand men, chosen as the study population from a source population that included all white males in the United States. The entries in the table are the percentages of the full sample found in each category of systolic and diastolic blood pressure.

Prevalence data are key to health planning, and they provide the underpinnings for the standards on which most diagnostic practice rests. However useful, prevalence tends to be dissatisfying to someone who is looking into the origins of disease, because it offers little insight into the direction of causal relations. The result is seldom

persuasive, in part because of concern about differential mortality rates.[24] Figure 3.2[25] presents an example of cross-sectional data that tantalize because they fall just short.

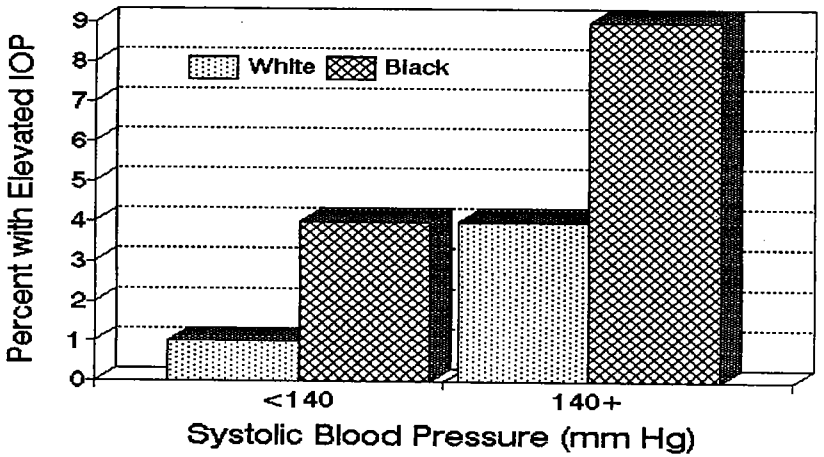


Figure 3.2 Prevalence of elevated intraocular pressure (≥22 mm Hg) according to race and systolic hypertension (females)

Elevated intraocular pressure is more common in hypertensives and in blacks. Does systolic hypertension lead to (or at least predate) ocular hypertension? Are there common factors giving rise to both? Do white and black women differ in their reasons for developing an elevated IOP? The data are silent: they give an end result, not the natural history.

Closed Cohort Studies

It is the search for cause and effect that leads to the introduction of elapsed time into epidemiologic studies. The most direct form of this introduction is in the cohort study, whose most classical form is the *closed cohort*, the nonrandomized cousin of a clinical trial.

24. See the discussion of survivor cohorts in Chapter 4.

25. Klein BE, Klein R. Intraocular pressure and cardiovascular risk variables. Arch Ophthalmol 1981;99:837-9

Cohort. *Any group of individuals whose disease or mortality is measured over time is a cohort.*

Closed cohort. *A closed cohort consists of individuals who are followed from a defined starting point to a defined end point. The membership of the group does not change, apart from mortality, from the beginning of observation to the end.*[26]

The term "cohort" was originally a Roman military term: a cohort was one-tenth of a legion. Typically the members of a cohort would be recruited from young men of a single age from one locale. The cohort would then undergo attrition, never being replenished, and would be disbanded when the term of enlistment was over. The word "cohort" has come to be used in epidemiology to designate individuals whose experience we observe in order to learn about the occurrence of disease.

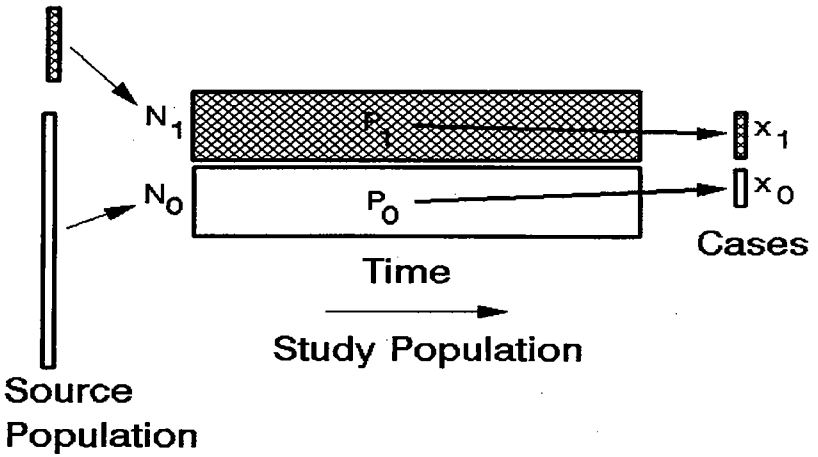


Figure 3.3 A closed cohort study

Typically, closed cohort studies begin with study groups of approximately equal size, one of which has undergone some exposure or experience that is thought to influence the risk of acquiring disease. The study groups "1" and "0" in Figure 3.3 are followed through time,

26. Closed cohorts are also referred to as "fixed" cohorts.

and eventually cases of disease (x_1 and x_0) begin to appear. Notice from Figure 3.3 that the sizes of the exposed and unexposed study groups (N_1 and N_0 respectively) are not necessarily at all representative of the relative numbers of exposed and unexposed people in the source population. Nor is the person time of experience (P_1 and P_0) in the two populations "representative" of any particular distribution of person time elsewhere. Through subject selection there is an intentional distortion of the exposure distribution in the study population relative to the source population. The distortion serves to increase the efficiency of the study by reducing the imbalance in expenditures for the observation of exposed and unexposed subjects.

Table 3.3 Available comparisons in closed cohort studies

	Difference	Ratio
Cumulative incidence	$\frac{x_1}{N_1} - \frac{x_0}{N_0}$	$\frac{x_1}{N_1} / \frac{x_0}{N_0}$
Incidence rate	$\frac{x_1}{P_1} - \frac{x_0}{P_0}$	$\frac{x_1}{P_1} / \frac{x_0}{P_0}$

At the end of follow-up, the occurrence of disease can be measured and compared as shown in Table 3.3. Cumulative incidences are the proportions of persons in the original cohorts who become diseased. Incidence rates are the numbers of cases per unit of person time (e.g. person years) at risk. Cumulative incidences are the closed cohort measures that are easiest to interpret: they answer a question that is implicit in the structure of the study: what proportion becomes ill? Most of us see our own futures as tiny closed cohorts, with risk interpreted in a probabilistic sense.

In order for any statement of risk to be interpretable, it must be presented with a specification of the elapsed time over which the cumulative incidence is manifest. The "risk of relapse" or the "risk of myocardial infarction" are meaningless by themselves: they need to be fleshed out to the "risk of relapse within the first year following successful induction of a disease remission" or the "risk of having an MI between the ages of 50 and 59."

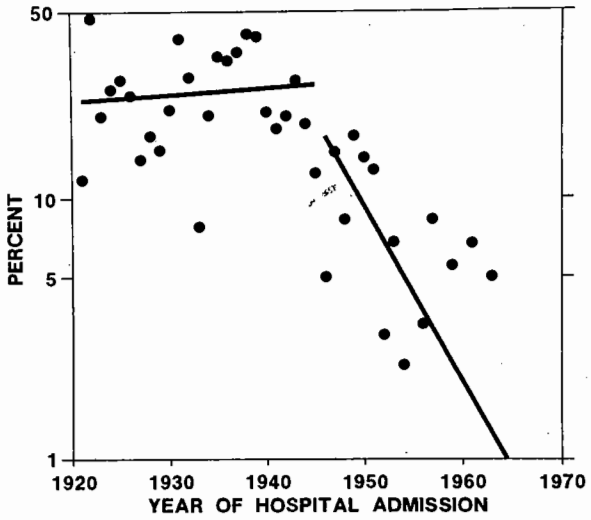


Figure 3.4 Five-year case-fatality rates for patients less than 15 years of age with rheumatic carditis

Example 3.1. *Case-fatality for rheumatic carditis.*[27]

Patients with rheumatic heart disease who were admitted to the House of the Good Samaritan in Boston at age 15 years or younger were followed in order to determine the course of their disease. Two outcomes were assessed at the end of five years' follow-up: death and clinical resolution of all murmurs. There were 2090 patients, of whom 339 died in the five years following first admission. Figure 3.4 displays the five year *case fatality rates* (i.e. cumulative incidences of death over five years) for each closed cohort defined by hospital admission in each calendar year from 1921 through 1970. The black dots represent the observed cohort-specific case fatality rates,[28] and the dark lines are

27. Massell BF, Chute CG, Walker AM, Kurland GS. Penicillin and the marked decrease in morbidity and mortality from rheumatic fever in the United States. N Engl J Med 1988;318:280-6

28. There were no deaths after 1963, so that the location of the dots for later years is outside the range graphed in Figure 3.3.

estimates of the linear components of the secular trends. There is an evident inflection in the trend around the years 1945 and 1946. Figure 3.5 shows the proportion of loss of all cardiac murmurs (i.e. clinical recovery) over the same periods in the same cohorts. There were 343 such children. Again, there is a discontinuity just after the end of the second World War. The cumulative incidence difference, comparing 1945 with 1946, was estimated at 13 percent. In all likelihood, the source of the abrupt changes in Figures 3.4 and 3.5 was the widespread civilian availability of penicillin. Trends in national mortality rates from rheumatic fever showed the same discontinuity, with an acceleration in the rate of decline after 1946.

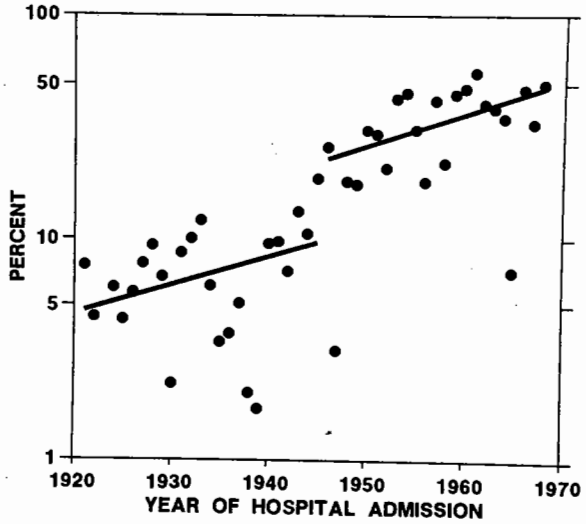


Figure 3.5 Proportion of rheumatic carditis patients less than 15 years of age in whom all evidence of murmurs disappeared within five years

The foregoing example describes cumulative incidences. As indicated in Chapter 1, for closed cohort studies, (and for clinical trials), the incidence rate of disease can be calculated simply as the proportion becoming diseased, divided by the average time of disease-free follow-up. For closed cohort studies, rate calculations

may appear to provide an unnecessary flourish, but there is an important class of cohort studies in which incidence rates are the only measure that is directly accessible: these are *open cohort studies*.

Open Cohort Studies

Consider a study of smoking in men aged 50-59. For five calendar years, we recruit both smokers and nonsmokers into the study. We follow them during the same period: smokers and nonsmokers in the appropriate age range are observed for the occurrence of myocardial infarction for as long as they remain eligible. This study situation differs in crucial respects from the closed cohort design:

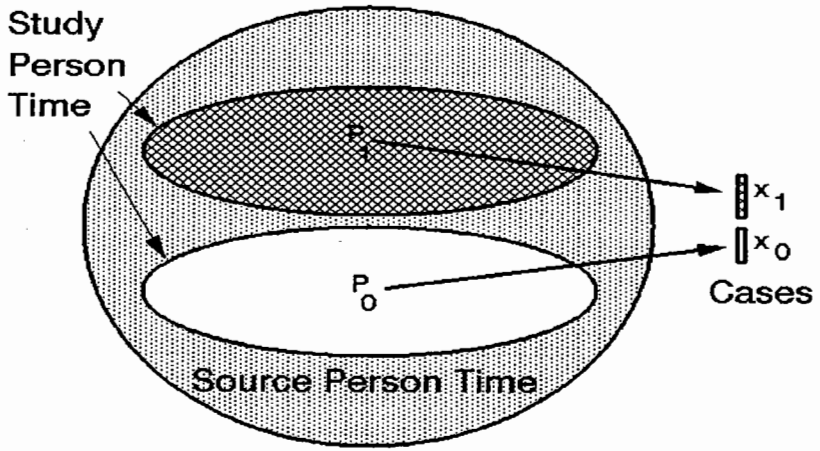


Figure 3.6 An open cohort study

(1) "Exposure" does not represent an event, fixed in time, but rather a status, which continues over time, of being a smoker or

a nonsmoker.[29]

(2) There is no fixed ending time for observation.

(3) Subjects can enter the study at any time, and they may leave it (by dying or leaving town or through termination of the study) at any time.

Open cohort. *An open cohort is a cohort whose studied composition may change with the passage of time.*[30]

Even in a closed cohort, the composition of the population may change with time prior to the end of follow-up, due to the onset of disease or to change in the defining exposure. The difference between an open and a closed cohort is that in an open cohort the evolving population composition (describable in terms of persons and exposures) is directly monitored and becomes the studied experience. Allowance for change in the observed population opens up the possibility of cohort studies that can be adapted to populations that do not fit the static presumptions of closed cohort analysis.

In an open cohort, cases arise out of a pool of human experience whose size is measured in units of person time. The x_1 and x_0 of Figure 3.6 represent the cases that arise out of the P_1 exposed and P_0 unexposed person years contributed by exposed and unexposed persons to the study person-time. The study person-time in turn may represent just a portion of the source person-time, that is, the person-time experience of the source population.

Although open cohort studies permit more flexible designs, they do not offer all the comparisons that can be derived from closed cohorts. The three conditions outlined above render cumulative incidence differences or ratios unavailable as measures of disease frequency: you cannot talk about "MIs per 100 men" when a man may represent a quantity of observation that ranges anywhere from days to years. You can, however, still talk about incidence rates, as

29. It would be possible to perform a closed cohort study of smokers and nonsmokers. The chief advantage of closed cohort designs is that they permit a direct and natural estimation of risk (c.f. Table 3.3). For the closed cohort design to offer an advantage in this example, there should be the achievement of some state, such as "smoker" or "smoker for 20 years" as the definition of t_0. The risk over a specified interval following such a well defined t_0 could be a generally useful figure; the risk over the same interval following an otherwise unrestricted study entry date is unlikely to be of general interest.

30. The term "dynamic cohort" is sometimes used to describe open cohorts.

Table 3.4 Available comparisons in open cohort studies

	Difference	Ratio
Cumulative incidence	*	*
Incidence rate	$\frac{x_1}{P_1} - \frac{x_0}{P_0}$	$\frac{x_1}{P_1} / \frac{x_0}{P_0}$

* Not defined for open cohort studies

"MIs per man year (or man month or 100 man years) of observation." The available comparisons between groups are reduced to those in Table 3.4.

Example 3.2. *Vasectomy and myocardial infarction.*[31]

Table 3.5 presents the results of an open cohort study of the incidence of myocardial infarction in vasectomized men. Men in an HMO who had undergone vasectomy between 1963 and 1978 were observed while they were members of the HMO for the occurrence of myocardial infarction. Their experience was compared to that of a group of HMO members without vasectomy. Men with and without vasectomies could move between age categories so that each could contribute person years of experience to several age categories.

This open cohort study is based on a well-defined list of persons, and could readily incorporate a clear t_0, the date of vasectomy. What makes this an open cohort is the opportunity for men to enter or leave follow-up at any moment. The opportunity to change some aspect of their "exposure" (in this case age) with every moment would also be sufficient to define this as an open cohort. The incidence rate analysis is characteristic of an open cohort study, but it does not define it, since closed cohort experience can also be subjected to incidence rate analysis.

31. Walker AM, Jick H, Hunter JR, McEvoy J. Vasectomy and nonfatal myocardial infarction: Continued observation indicates no elevation of risk. J Urol 1983;130:936-8

Table 3.5 First time myocardial infarction rates in vasectomized and nonvasectomized men

Age	Cases	Person Years	Rate per 1,000
Vasectomized			
35-44	14	16,806	0.8
45-54	24	8,133	3.0
55-64	7	1,700	4.1
Total	45	26,639	1.7
Not Vasectomized			
35-44	56	83,057	0.7
45-54	110	40,971	2.7
55-64	49	8,570	5.7
Total	215	132,598	1.6

Cohort studies provide an opportunity for direct observation of time relations and are well-adapted to the study of relatively common disease outcomes. However, observation for even moderately rare diseases that do not occur within a short time requires large numbers of subjects (numbering in the thousands or ten of thousands) observed over many years. When the occurrence of disease is not a common event, the expense of maintaining information on a large enough cohort of persons at risk may be prohibitive. Investigators deal with this common situation through a study design whose cardinal feature is that data are not collected on all persons, but rather on samples of the diseased and nondiseased populations.

Case-Control Studies

Figure 3.7 presents a schematic picture of the relation of a case-control study to an underlying open cohort. Illustrated is the common situation in which the sampling fraction for cases is 100 percent. As before, there is experience among persons exposed and among those unexposed. As before, cases arise out of the pool of experience: x_1 exposed and x_0 unexposed. If P_1 and P_0 were known, the ratio of rates in exposed versus unexposed person time could be

calculated as $(x_1/P_1)/(x_0/P_0)$, which equals algebraically $(x_1/x_0)/(P_1/P_0)$. (x_1/x_0) is called the exposure odds among the cases; (P_1/P_0) is the exposure odds in the source population.

Exposure odds. *The number of exposed persons divided by the number of unexposed persons in a group yields the exposure odds. The exposure odds in a pool of person time are obtained by dividing the amount of exposed person-time by the amount of unexposed person-time.*

In both open cohort studies and case-control studies, the exposure odds in cases is observed directly. The difference between an open cohort study and the corresponding case-control study is that the exposure odds in the source population (P_1/P_0) are not observed in the latter design. Instead, in the case-control study, the exposure odds in the source population are estimated from a sample of person days. Like a marine biologist counting different species of algae, we dip a test tube into the pool of experience of an open cohort to learn its composition. In so doing, we return to the cross-sectional survey discussed at the beginning of the chapter. The goal is to sample a small proportion of the population giving rise to the cases in such a way as to estimate the exposure odds in that source population.

Note that since we are sampling person time, the sampled population is a population of person days, defined operationally as specified days in the lives of specified individuals. Both the person and the date are chosen by some suitably random procedure. The sampled group of person days is called the control series. The persons whose days are sampled are called the *controls*.

Controls. *The controls in a case-control study are a group of persons whose exposure status collectively provides information about the distribution of exposure in the persons or person time giving rise to the cases.*

Table 3.6 gives the layout of data from a case-control study. The ratio of exposed to unexposed controls (y_1/y_0) is an estimate of (P_1/P_0). The incidence rate ratio from a case-control study is derived then as

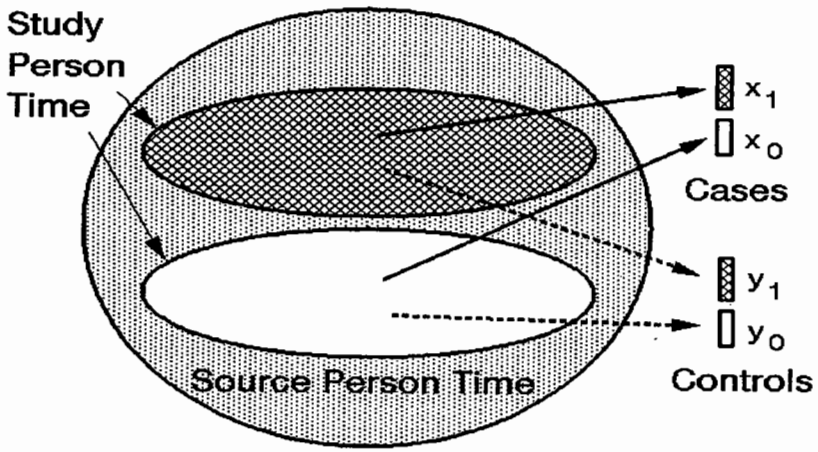


Figure 3.7 A case-control study that arises from an open cohort

$$RR = \frac{IR_1}{IR_0} = \frac{x_1/P_1}{x_0/P_0}$$

$$\doteq \frac{x_1 y_0}{x_0 y_1}$$

The symbol "≐" means "approximately equal to." The approximation involved has to do with the additional, statistical error that results from taking a sample of all person days to represent the total. The quantity in the last line is the ratio of exposure odds in cases to the exposure odds in controls, and is referred to as the *odds ratio* (*OR*). The odds ratio in a case-control study is an estimate of the rate ratio in the source population giving rise to cases and controls.

Case-control studies are also conducted in situations for which the most natural corresponding cohort study is the closed cohort. In perinatal epidemiology, an infant's health status might typically be assessed at 30 days of life. If cases were drawn from those who had died, then controls would be chosen at random from those who have survived. When the controls are drawn from the healthy survivors of a closed cohort and when the cases are a small proportion of the

Table 3.6 Data layout for a case-control study

	Exposed	Unexposed
Cases	x_1	x_0
Controls	y_1	y_0

$$RR \doteq OR = \frac{x_1 y_0}{x_0 y_1}$$

total study subjects, the ratio y_1/y_0 provides an estimate of the ratio of exposed to unexposed persons in the source population (N_1/N_0), so that the cumulative incidence ratio can be estimated as

$$CIR = \frac{CI_1}{CI_0} = \frac{x_1/N_1}{x_0/N_0}$$

$$\doteq \frac{x_1 y_0}{x_0 y_1}$$

The approximation indicated on the last line refers both to the statistical uncertainty and to the approximation taken from the use of exposure odds in the controls as an estimate of the exposure odds in the source population. The numbers of exposed and unexposed persons in the control series are used to estimate the ratio of exposed to unexposed persons in the source population.

Notice that the moment as of which the exposure status of the controls is assessed depends on whether they are drawn from an underlying cohort that is open or one that is closed. In the latter, the exposure status of interest is that which characterizes each control as of t_0, the initial time defining cohort eligibility. When the controls are drawn from open cohorts, the exposure status of each control is relevant only on the day(s) sampled when the controls are drawn from an open cohort.

The relative measures of disease occurrence derivable from a case-control study are displayed in Table 3.7.

which early middle-aged adults were identified in the early 1950s, and have been observed ever since. Most important case-control studies, beginning with the earliest investigations of smoking and lung cancer, have been retrospective. Nonetheless these terms, which describe the relation in time between the researcher and the object of study are quite independent of the sampling design.

Prospective. *A prospective study is one in which the disease events under study occur after the protocol for data collection has been implemented.*

Retrospective. *A retrospective study is one in which the protocol is implemented after the disease events have occurred.*

Cohort studies are frequently retrospective. The groups to be compared are defined by exposures that occurred in the distant past, and data on subsequent health events is drawn from vital statistics or medical records extending up to the date of data collection. This is typical for occupational cohort studies, and is becoming more common in other areas as long-term medical records become available for special population groups. The vasectomy study of Table 3.5 was a retrospective cohort study carried out inside a health maintenance organization.[34] More important than the time orientation of a study is the quality of the data that it yields. The greatest potential advantage of a prospective study is that the investigator can arrange the administration and data collection so that the necessary information flows into the study in a usable form. A prospective study in which the data flow is incomplete or poorly monitored has no advantage over a retrospective study carried out in an "information rich" environment.

34. Often studies are designed with both retrospective and prospective collection of data. A useful neologism to designate such designs is "ambispective."