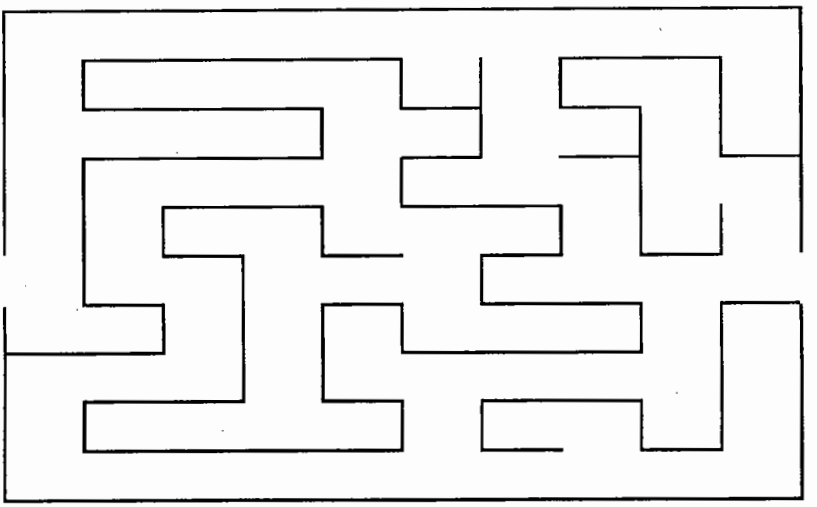
6

Case-Control Studies



Case-control studies constitute the major advance in epidemiologic methods of our time. In itself the case-control study has greatly improved the efficiency of research into the etiology of rare conditions. Since the case-control study differs from the cohort study only in that knowledge about the source population is garnered through sampling, elucidation of the case-control study has also shed light on many subtleties of the parent design. I will develop the notation here for the case of sampling from open cohorts whose experience is measured in person time. There is a precisely analogous line of argument for case-control sampling from closed cohorts. Since everything that follows is predicated on the notion of sampling from an underlying cohort, essentially all the terminology, definitions, and caveats that apply to the conduct of research in cohorts will apply to case-control investigations as well.

Case-Control Studies and the Cohorts that Underlie Them

Consider an open population with person time experience accumulating in two pools, which we will designate for convenience as "exposed" and "unexposed." Define the following quantities:

P_1, P_0 the accumulated person time in exposed and unexposed persons, respectively,

x_1, x_0 the corresponding observed numbers of cases,

IR_1, IR_0 the corresponding observed incidence rates of disease,

RR the ratio of the observed incidence rate in exposed to that in unexposed persons.

The following relations exist among these various quantities:

$$IR_1 = \frac{x_1}{P_1}$$

$$IR_0 = \frac{x_0}{P_0}$$

$$RR = \frac{IR_1}{IR_0}$$

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

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

If the quantities P_1, P_0, x_1, and x_0 are all known for a given population, then there is no obstacle to calculating the incidence rates IR_1 and IR_0 and therefore the rate ratio RR. However, the cost of actually identifying the exposure status of a large and changing population may be high. If the exposure status of members of the population is changing even as the members remain under observation, the technical obstacles to maintaining current data may be insurmountable, no matter what resources are available.

A way around this barrier lies in the last formulation given above for the rate ratio. Even if the population distribution of exposure is unknown, it will generally be possible to ascertain the exposure status of the x_1+x_0 individuals who develop disease. This might be accomplished by interview, for example, or by hand review of individual exposure records. Given x_1 and x_0, the remaining quantity to be identified in order to calculate the rate ratio is P_1/P_0. If it were possible to estimate the relative amounts of person time in the exposed and unexposed portions of the population, then estimation of the rate ratio could follow immediately, even in the absence of full exposure information on the underlying population.

If an estimate of P_1/P_0 were obtained by a sampling procedure, then the result would be subject to random error, but it would be valid in that the accuracy of the process producing it would be limited only by the number of individuals sampled. A study based on enumeration of the cases of disease and on random sampling of the person time giving rise to those cases provides a consistent[48] estimate of the rate ratio.

Sampling the Source Population

Consider mesothelioma, a rare malignancy for which all or nearly all cases in individuals under the age of 65 might be referred to a cancer center. Imagine that a study is being conducted in a region in which there is a single facility that will receive essentially all the cases from the area immediately surrounding it and a number of cases from more distant places as well. The population giving rise to cases of mesothelioma seen in the facility consists of all persons who would have been referred there, had they developed disease. While this is a logically valid definition, it is not one that leads immediately to any specific sampling policy. How can we know that a healthy person not from the immediately surrounding area would have been referred?

Instead of searching for ingenious methods of identifying the population at risk of referral from outside the immediate area, the investigator can simplify the task of sampling by incorporating a population specification into the case definition. If cases for the study are restricted to those who are treated at the facility and who are residents of districts surrounding the facility, for whom the probability of referral is nearly 100 percent, then the population at risk is immediately identified as the population resident in those districts during the period of case accrual.

48. Estimates which can be expected to come arbitrarily close to a value of interest as the sample size increases are said to be "consistent."

Sampling a geographically defined population in order to obtain a control series can be entirely straightforward. In many places, it is possible to obtain a complete list of residents. In this case the sampling procedure is as follows.

(1) Select a date at random from the case accrual period.

(2) Select a person at random from the population list.

(3) If the subject selected was actually resident within the predetermined area as of the random date chosen for that particular subject, then the subject is accepted as a control for the case-control study as of the randomly sampled day.

(4) Repeat the above three steps until the desired number of controls has been chosen.

(5) Any information collected from cases that invokes the date of onset of illness as a reference point (such as "A year prior to your diagnosis, did you ... ?") is collected from controls with reference to the random point in time that constituted part of the control definition.

Recall that P_0 and P_1 represent person time rather than persons, and that this distinction is necessary for calculation of incidence rates. The double control selection process (choosing a random time point from the case accrual period and a random subject from the population under study) samples person time rather than people. It can be shown that the probability of any given person day's becoming an index day in the control series is equal to that of any other person day experienced by the source population. By extension, the probability of any individual's being selected as a control can be demonstrated to be proportional to the duration of time that the individual spends in the source population at risk.

The investigator repeats the control selection procedure until he has obtained the desired number of controls. In so doing he establishes a probability for each person day at risk in the study person-time[49] that it will be chosen for the control series. If we represent this control selection probability as

f the probability of selecting a given person day from the person time at risk,

49. c.f. Figure 3.6.

then we can represent the number of controls obtained in a case-control study as an expected function of exposure. Define the numbers of exposed and unexposed controls as

y_1 the number of selected controls who are exposed on their designated index day,

y_0 the number of selected controls who are not exposed as of their designated index day.

The operational definition of f is

$$f = \frac{y_1 + y_0}{P_1 + P_0}$$

Since f is defined without respect to exposure status, it follows that the expected values of y_1 and y_0 are

$$E(y_1) = fP_1$$
$$E(y_0) = fP_0$$

from which it is possible to obtain an expression for the rate ratio as a function of case and control counts:

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

$$= \frac{x_1}{x_0} / \frac{fP_1}{fP_0}$$

$$= \frac{x_1}{x_0} / \frac{E(y_1)}{E(y_0)}$$

Substitution of the observed for the expected counts yields the *odds ratio* of the table, which is an estimate of the incidence rate ratio

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

The value x_1/x_0 is called the "exposure odds" for cases; y_1/y_0 is the exposure odds for controls. The term (like many concepts in statistics) is derived from gambling: the odds that a given case will

be exposed are "x_1 to x_0". The rate ratio estimate in a case-control study is the ratio of the exposure odds in cases to the exposure odds in controls, and is therefore often called the "odds ratio."

Example 6.1. *Cryptorchidism and testicular cancer.*[50]

Morrison conducted a study of testicular cancer, using medical records kept by the United States Army. From a pathology index, he identified 596 cases of testicular cancer among soldiers, and he noted the date of diagnosis of each. He then established a scheme for locating comparison records at random: for research purposes there had already been developed a list consisting of one tenth of one percent of military personnel who were active during the time of the study. Using the last digits of the soldiers' military identification numbers, Morrison obtained a subsample of this research list to serve as comparison subjects (controls). He reviewed the physical examinations carried out at the time of induction into the army of the testicular cancer patients (the cases) and 602 controls in order to establish the prevalence of treated or untreated undescended testis in each group. Since the exposure under study did not vary with time (and presumably did not affect duration of service in the army), Morrison did not need to make provision for a single date as of which to define the exposure status of controls. He obtained the data of Table 6.1.

Table 6.1 Cryptorchidism and testicular cancer in the U.S. Army

	Cryptorchid	Not cryptorchid
Testicular cancer	17	579
Controls	2	600

$$RR = \frac{(17)(600)}{(2)(579)} = 8.8$$

50. Morrison AS. Cryptorchidism, hernia, and cancer of the testis. J Natl Cancer Inst 1976;56:731-3

Pseudo-Sampling When the Source Population is not Identified

It may happen that there is no ready list of persons in the source population, nor even a list of some larger population from which the desired population members might be culled. In the preceding example, if there were two oncology centers serving the same region, then the referral of patients to one or the other center might be the result of a series of seemingly haphazard events related to the choice of physician, the presence of relatives near one center, and so forth. We cannot know exactly who those persons are who would have been referred, much less sample them. One frequently chosen solution is to select controls from those persons who were in fact referred to the oncology center that provides our cases. The date of disease that defines controls is taken as each control's index date. Controls so chosen must have diseases unrelated to the exposures under study.

Denote the incidence rate of a comparison disease by Q. Suppose that, within categories of all the ascertained characteristics of study subjects other than exposure, the comparison disease incidence rate is identical in persons with and without the exposures under study. Imagine furthermore that the referral patterns for the comparison disease and for the primary disease under study are identical. If the persons with the comparison disease are taken as controls, then the expected number of exposed and unexposed controls taken together is

$$E(y_1 + y_0) = Q(P_1 + P_0)$$

and since incidence is unrelated to exposure status,

$$E(y_1) = QP_1$$
$$E(y_0) = QP_0$$

From here the argument leading to the odds ratio as an estimate of the rate ratio proceeds exactly as before, except that Q, the incidence rate of the comparison disease, takes the place of f, the probability of selecting a given control day out of the person time at risk. Although there is no theoretical difficulty in utilizing diseased controls, there are practical problems in assuring the validity of the assumptions that underlie the practice. Is Q, the incidence rate of the comparison disease, unrelated to the exposures under study in the source population? A valid case-control analysis assumes not

that there is no causal association between exposure and the comparison disease, but that there is no variation *for any reason* in the comparison disease incidence rate across categories of exposure.

To some extent, the validity of the comparison series may be examined by choosing a number of disease entities for inclusion. If each truly portrays exposure in the source population, then each of the comparison entities ought to provide similar estimates of exposure prevalence, up to whatever level of accuracy is imposed by chance variation. Thus the various control diseases can be compared in terms of their exposure prevalence. If there are any outliers, these groups are removed on the presumption that they are biased with respect to exposure for some previously unsuspected reason. If no group's exposure experience is dissimilar to that of the others, then the hypothesis that each is a valid representation of exposure patterns in the base population is strengthened. All of the control groups are then collapsed together for the purposes of analysis.

The use of diseased controls brings up considerations peculiar to the medical facilities in which the cases and controls are ascertained. Are the probabilities of diagnosis and referral to the facility generating the cases identical for the comparison disease and the disease under study? If not, then the source population for controls differs from that for cases; the validity of the control series depends on a further assumption of homology between the source population for cases and that for controls. If the prevalence of exposure in the control source population is identical to that in the case source population, and if Q, the incidence of the control disease, is unrelated to exposure in the source population for controls, then the control series provides an unbiased estimate of exposure prevalence in the source population for cases, and the odds ratio estimate of the rate ratio is valid. The practice of comparing exposure prevalence in different control groups, noted above, does not provide strong evidence of the suitability of controls whose source population differs from that of the cases.

Clinical Aspects of Case Definition

As in cohort studies, the criteria for case definition in a case-control study depend on the amount of information routinely available on potential cases within the source population. While investigators may sometimes establish a dedicated surveillance network to detect all potential cases, it is much more common for epidemiologic research to depend on routinely collected data. In this latter circumstance, the investigator balances a desire for accuracy in case designation against the danger of obtaining a case series in which principal determinants of case status are the social or demographic determinants of diagnosis. An optimal case definition depends on criteria that can be applied to any potential case in the source population.

An epidemiologic study is at high risk of detecting spurious associations whenever case definition depends on a diagnostic procedure that is rarely performed, such as autopsy, or on the judgment of specialists. Most psychiatric disorders fall into the latter category, and Alzheimer's Disease would be an example of the former. By the same token, studies of generally benign conditions run a high risk of mistaking correlates of easy access to medical care for etiologic factors. Functional ovarian cysts and gallstones would be two examples: in each case there is a high prevalence of a mildly symptomatic disease whose diagnosis *may* occur in persons who consult physicians often for other reasons, but will not occur otherwise.

Note how very different are the roles of case ascertainment in an epidemiologic study and in a clinical therapeutic trial. In the clinical trial, where the goal is to evaluate the efficacy of therapy, the overriding concern is that subjects entered into the trial actually have the disease in question. Those excluded from the trial are of no interest. In an epidemiologic study the focus is on the relative frequency with which persons with different exposure statuses develop disease. Factors that lead to the exclusion of true cases are of concern, particularly if correlated with exposure. If case definition requires the use of diagnostic procedures that cannot be made uniformly available, then it may be that the population in question is not a suitable starting point for the conduct of a case-control study.

Almost as important as identifying cases correctly is the specification of a time at which the individuals in question undergo the transition from "healthy" to "ill." This is the case for two reasons. First, any comprehensible discussion of exposures that change with time or of exposure effects that change with time requires unambiguous definitions of those times at which the exposures (or their residua) are considered to be acting. Second, exposures measured after the onset of illness may provide a poor stand-in for earlier exposures, particularly when disease affects lifestyle or work patterns.

Seldom is it possible to identify the onset of a chronic disease, and the solutions that have been applied to the problem have a disconcerting, *ad hoc* quality. Occasionally researchers estimate presumed times of onset on the basis of (untestable) assumptions about the rate of progression from incipient to clinical stages of illness. For diseases that generally lead to a hospitalization, it is common simply to establish the date of first diagnosis or the date of first hospitalization leading to a diagnosis as an index date. Questions about exposures that might be affected by disease are backdated to an index date that precedes most durations of symptoms prior to onset, as determined clinically.

Vigorous pursuit of efforts at backdating disease onset frequently raises more problems than are resolved. An estimated date of onset far into the past may locate the disease event in a source population that cannot be properly sampled. An unobserved date of presumed onset that precedes case identification may raise the possibility of cases to be identified in the future that will be ascribed to the current study time: factors that lead to early detection can then masquerade as predictors of the disease. The best solution is to identify incidence-dates that are closely tied to events observable in all members of the source population. Then be clear that you are studying what has been defined, and separately speculate on or investigate the relation between the operational criterion and the imagined true event.

Alternatives to Simple Causal Interpretation

It often happens that there is some extra characteristic of individuals in the study population (beyond exposure and disease) that threatens to distort the apparent exposure-disease relation. This situation occurs when the extra factor is itself a predictor of disease risk and is not evenly distributed between the exposed and unexposed sectors of the population. An evaluation of the occurrence of disease according to exposure is then contaminated by the different expectations of disease occurrence in the exposed and unexposed groups, even in the absence of an effect of exposure itself. As in the analysis of cohort data, the admixture of an extraneous effect in the comparison of two exposure groups in a case-control study is called *confounding*. The presence of confounding is a characteristic of the population and disease under study, and it poses the same threat to validity in case-control studies that it does in cohort studies.

Just as in the analysis of cohort studies, the most common solution to the problem of confounding is to segregate study subjects into subgroups, or strata, within which there is little or no variation in the extent of the extra factor, the predictor of disease that threatens to confound the unstratified ("crude") analysis. After stratification, there is little or no residual potential for confounding within each stratum. By definition, the third factor is identical in exposed and unexposed persons within each stratum.

For any observational study, the search for confounding factors amounts to a search for alternative causal explanations for an observed exposure-disease relation. In those case-control studies in which it has been impossible to obtain for the control series a simple random sample of the population giving rise to the cases, it is imperative to ask whether or not the factors determining control selection are likely to have yielded a comparison group whose relevant exposure characteristics accurately reflect those of the source population for cases.

Example 6.2. *Ferruginous bodies and lung cancer.*[51]

> Warnock and Churg sought to evaluate the effects of low level asbestos exposure in the production of bronchogenic carcinoma of the lung through a case-control study. Their measure of low level exposure was based on an analysis of the concentration of a marker of cumulative asbestos exposure, ferruginous bodies (also called "asbestos bodies"), in lung tissue. The cases were 30 persons who had died with lung cancer; ferruginous body counts were carried out *post mortem*. Since Warnock and Churg could not directly measure ferruginous bodies in the lungs of persons in the population giving rise to the lung cancer cases, they performed similar measurements on 100 consecutive autopsies of persons over the age of 20 during a period that encompassed the accrual of lung cancer cases. The results of the analysis, shown in Table 6.2, indicate nearly a seven fold elevation of lung cancer mortality in persons with low-level exposure.

For the relation between ferruginous bodies and lung cancer, the strength of the association noted above would raise a warning signal to anyone familiar with the subsequent literature on the relation between asbestos and pulmonary disease. A relative mortality on the

51. Warnock ML, Churg AM. Association of asbestos and bronchogenic carcinoma in a population with low asbestos exposure. Cancer 1975;35:1236-42

Table 6.2 Lung cancer and ferruginous bodies in lung tissue

	Bodies per gram of wet tissue	
	≥50	<50
Lung cancer	8	22
Other	5	95

$$RR = \frac{(8)(95)}{(22)(5)} = 6.9$$

order of five has been observed for lung cancer in some (but by no means all) cohorts of workers heavily exposed to asbestos throughout their lifetimes. That a similar risk should obtain for the subset of the general population with low-to-moderate asbestos exposure seems improbable.

Several distortions are evident. The first is related to cigarette smoking. Tobacco is a potent inducer of bronchogenic carcinoma. Cigarette smoking in the United States during the period of study was related to social class, with laborers smoking more than white collar workers. As a result there existed an association between on-the-job asbestos exposure and smoking whose basis was entirely sociological. Connected to both exposure and risk of disease, tobacco use might then be expected to confound the apparent relation between the two. There may also be a selective distortion of the data (sometimes called "information bias"). Smokers suffer paralysis of the bronchial cilia responsible for clearing mucus, and are thus likely to retain for a longer time inhaled particulates, including asbestos fibers, which can give rise to ferruginous bodies. The measure of asbestos exposure itself, the concentration of ferruginous bodies in lung tissue, may therefore have been exaggerated by concomitant use of cigarettes. Whether exaggeration of the measure (ferruginous bodies) reflects an increase in the relevant carcinogenic exposure is a matter of speculation: the properties of asbestos that result in carcinogenicity are not well understood.

Age in this example carries a clear potential for confounding, in that it is related to both mortality from lung cancer (which rises as the fourth or fifth power of age) and to the accumulation of ferruginous bodies in lung tissue (as a more or less linear function of age for environmental asbestos pollution). A comparison of lung cancer cases with a true random sample from the general population would yield a very strong association between ferruginous body levels and disease, on the basis of age confounding alone. To an unknown extent this distortion has been reduced by the use of dead controls, who are more likely to have had the age distribution of lung cancer cases than did the population at large.

A comparison of lung cancer victims with other decedents depends on the assumption that exposure to low levels of asbestos does not appreciably affect mortality from causes other than lung cancer. On the basis of other studies, this seems to be a reasonable hypothesis. Warnock and Churg also presumed that differences in the source populations for cases and controls as well as the determinants of autopsy were unrelated to pulmonary ferruginous body concentrations. Since the exposure measurement was available only after autopsy, the decision to carry out the autopsy was almost certain not to have been related to the measure itself. However, there is a strong possibility that cases of different complexity (and therefore different likelihood of undergoing autopsy) derive from referral areas that are not coextensive. As ambient asbestos levels vary with locale, it is not at all obvious that the controls provide information on the precise population from which the cases were drawn. Since this latter source population has remained unspecified in the study design, there is no opportunity for further evaluation of the bias.