Chapter 9
Case–control studies

A case–control study involves the identification of individuals with ('cases') and without ('controls') a particular disease or condition. The prevalence (or level) of exposure to a factor is then measured in each group. If the prevalence of exposure among cases and controls is different, it is possible to infer that the exposure may be associated with an increased or decreased occurrence of the outcome of interest (see Section 9.5).

Example 9.1. The relationship between use of conjugated estrogens and the risk of endometrial cancer was examined among 188 white women aged 40–80 years with newly diagnosed endometrial cancer and 428 controls of similar age hospitalized for non-malignant conditions requiring surgery at the Boston Hospital for Women Parkway Division, Massachusetts, between January 1970 and June 1975. The data on drug use and reproductive variables were extracted from hospital charts and from the medical records of each woman's private physician. Thirty-nine per cent of the cases and 20% of the controls had used conjugated estrogens in the past (Buring et al., 1986).

In Example 9.1, women with endometrial cancer ('cases') or without ('controls') were identified and information on their past use of conjugated estrogens ('exposure') was extracted from hospital and other medical records. The prevalence of use of conjugated estrogens was much higher among the cases (39%) than among the controls (20%), suggesting that the use of this drug was associated with an increase in the incidence of endometrial cancer.

The major difference between cohort and case–control methods is in the selection of the study subjects. In a cohort study, we start by selecting subjects who are initially free of disease and classify them according to their exposure to putative risk factors (see Chapter 8), whereas in a case–control study, we identify subjects on the basis of presence or absence of the disease (or any other outcome) under study and determine past exposure to putative risk factors.

Case–control studies are particularly suitable for the study of relatively rare diseases with long induction period, such as cancer. This is because a case–control study starts with subjects who have already developed the condition of interest, so that there is no need to wait for time to elapse

between exposure and the occurrence of disease, as in prospective cohort studies. Historical cohort studies allow similar savings in time, but can be conducted only in the rare situations when past records with data on relevant exposures have been kept or when banks of biological specimens have been properly stored and appropriate laboratory assays are available for measurement of the exposures of interest.

9.1 Study hypothesis

As with any other type of study, the specific hypothesis under investigation must be clearly stated before a case–control study is designed in detail. Failure to do this can lead to poor design and problems in interpretation of results. Case–control studies allow the evaluation of a wide range of exposures that might relate to a specific disease (as well as possible interactions between them). Example 9.2 clearly illustrates this feature.

> *Example 9.2.* *A population-based case–control study was carried out in Spain and Colombia to assess the relationship between cervical cancer and exposure to human papillomavirus (HPV), selected aspects of sexual and reproductive behaviour, use of oral contraceptives, screening practices, smoking, and possible interactions between them. The study included 436 incident cases of histologically confirmed invasive squamous-cell carcinoma of the cervix and 387 controls of similar age randomly selected from the general population that generated the cases (Muñoz et al., 1992a).*

Case–control studies often constitute one of the first approaches to study the etiology of a disease or condition, as in Example 9.3. This is partly because of their ability to look at a wide range of exposures and partly because they can be conducted relatively cheaply and quickly.

> *Example 9.3.* *Because of their rarity, very little is known about the etiology of malignant germ-cell tumours in children. To explore risk factors for these malignancies and generate etiological hypotheses, a population-based case–control study of 105 children with malignant germ-cell tumours and 639 controls was conducted (Shu et al., 1995).*

The results from these *exploratory* case–control studies may suggest specific hypotheses which can then be tested in specifically designed studies.

9.2 Definition and selection of cases

9.2.1 Case definition

Precise criteria for the definition of a case are essential. It is usually advisable to require objective evidence that the cases really suffer from

the disease or condition of interest, even if, as a result, some true cases have to be eliminated. For instance, a histologically confirmed diagnosis should be required for most cancers. By accepting less well documented cases, the investigator runs the risk of diluting the case group with some non-cases and lessening the chances of finding real exposure differences between cases and controls.

It is sometimes impossible to eliminate all cases whose diagnosis is not properly documented, particularly if the pool of available cases is relatively small. In these circumstances, it may be possible to classify the cases according to diagnostic certainty. Such classification allows assessment of the extent to which the results are likely to be affected by disease misclassification (see Chapter 13). Suppose, for instance, that cases in a particular case–control study are classified as 'definite', 'probable' or 'possible'. If there is disease misclassification, a gradual decline in relative risk from the 'definite' to the 'possible' category should become apparent in the analysis, since the probability that non-cases may have been misdiagnosed as cases increases from the 'definite' to the 'possible' category.

The case definition should be established in such a way that there is no ambiguity about types of cases and stages of disease to be included in, or excluded from, the study. The choice of cases should be guided more by concern for validity than for generalizability. For example, in a study of breast cancer, we may learn more by limiting the cases (and the controls) to either pre- or post-menopausal women than by including women of all ages (unless the number of cases in each group is large enough to allow separate analyses), since the risk factors for pre- and post-menopausal breast cancers may be different. By ensuring that the cases are a relatively homogeneous group, we maximize the chances of detecting important etiological relationships. The ability to generalize results to an entire population is usually less important than establishing an etiological relationship, even if only for a small subgroup of the population.

Cases should also be restricted to those who have some reasonable possibility of having had their disease induced by the exposure under investigation.

> *Example 9.4. A multinational, hospital-based case–control study was conducted to evaluate the relationship of combined oral contraceptive use to the risk of developing five different site-specific cancers. The study was conducted in 10 participating centres in eight countries (Chile, China, Colombia, Israel, Kenya, Nigeria, Philippines and Thailand) from October 1979 to September 1986. Women with newly diagnosed cancers of the breast, corpus uteri, cervix uteri, ovary and liver were eligible if born after 1924 or 1929 (depending on when oral contraceptives became locally available) and had been living in the area served by the participating hospital for at least one year (WHO Collaborative Study of Neoplasia and Steroid Contraceptives, 1989).*

In Example 9.4, cases were restricted to women born since the 1920s because only women born since then could have been exposed to the factor of interest (oral contraceptives).

Although most case–control studies include only one case group, it is possible to study simultaneously two or more cancers whose risk factors are thought to share the same, or related, risk factors. Example 9.4 illustrates this point. Such multiple-disease case–control studies may be regarded as a series of case–control studies. This approach provides two main advantages. First, it provides the possibility of studying more than one cancer for relatively little extra cost. Second, the control groups may be combined to give each case–control comparison increased statistical power, that is, the ability of the study to detect a true effect, if one really exists, is enhanced because of the larger number of controls per case (see Chapter 15).

If the disease or condition of interest is very rare, the study may have to be carried out in various participating centres, possibly located in various countries. The study cited in Example 9.4 was conducted in 10 centres in eight countries. Despite this, only 122 newly diagnosed liver cancers were accrued during the seven-year study period. Some studies deliberately include participating centres from low- and high-incidence areas to assess whether the risk factors are similar. For instance, the cervical cancer study mentioned in Example 9.2 was conducted in Colombia and Spain, countries with an eight-fold difference in cervical cancer incidence (Muñoz *et al.*, 1992a).

The eligibility criteria should include not only a clear case definition but also any other inclusion criteria (Example 9.5). Persons who are too ill to cooperate or for whom the study procedures may cause considerable physical or psychological distress should be excluded. It is also usual to exclude elderly people in cancer case–control studies because their diagnosis is likely to be less valid and because of their difficulty in recalling past exposures.

> *Example 9.5. In the cervical cancer case–control study mentioned in Example 9.2, eligible cases were incident, histologically confirmed, invasive squamous-cell carcinomas of the cervix identified among patients resident in the study areas for at least six months. Patients were excluded if their physical and/or mental condition was such that interview and/or collection of specimens was inadvisable or if they were older than 70 years (Muñoz et al., 1992a).*

Usually, the inclusion of all patients who meet the eligibility criteria is not possible for a variety of reasons. Subjects may move out of the area, die or simply refuse to cooperate. The investigator should report how many cases met the initial criteria for inclusion, the reasons for any exclusion, and the number omitted for each reason (as in Example 9.6). This information allows us to assess the extent to which the results from the study may have been affected by selection bias (see Chapter 13).

Example 9.6. A large multi-centre case–control study was conducted in high- and low-risk areas of Italy to evaluate the role of dietary factors in the etiology of gastric cancer and their contribution to the marked geographic variation in mortality from this cancer within the country. All patients with new histologically confirmed gastric cancer diagnosed between June 1985 and December 1987, resident in the study areas, and aged 75 years or less were eligible as cases. A total of 1129 eligible cases were identified in surgery and gastroenterology departments and outpatient gastroscopic services of private and public hospitals. Approximately 83% of these cases were successfully interviewed using a structured questionnaire (Buiatti et al., 1989a,b). Table 9.1 shows the numbers of eligible patients in each recruitment centre, how many were recruited and the reasons for exclusion.

Table 9.1.
Recruitment levels among gastric cancer cases and reasons for non-participation by recruitment centre[a].

Recruitment centre	Eligible cases	Recruited No. (%)	Excluded due to		
			Refusal No. (%)	Poor health No. (%)	Deceased[b] No. (%)
Cagliari	104 (100)	82 (78.9)	3 (2.9)	4 (3.8)	15 (14.4)
Cremona	71 (100)	66 (93.0)	0 (0.0)	4 (5.6)	1 (1.4)
Florence	435 (100)	382 (87.8)	9 (2.1)	28 (6.4)	16 (3.7)
Forli	255 (100)	232 (91.0)	8 (3.1)	14 (5.5)	1 (0.4)
Genoa	155 (100)	122 (78.7)	3 (1.9)	24 (15.5)	6 (3.9)
Imola	76 (100)	47 (61.8)	9 (11.8)	8 (10.5)	12 (15.9)
Siena	133 (100)	85 (63.9)	18 (13.5)	29 (21.8)	1 (0.8)
Total	**229 (100)**	**1016 (82.7)**	**50 (4.1)**	**111 (9.0)**	**52 (4.2)**

[a] Data from Buiatti *et al.* (1989a).
[b] Deceased after being identified as potential cases.

Information on the entire eligible case series should be sought, whenever possible, regarding characteristics such as age, gender, education, socioeconomic status, so that selection factors for the non-participating subjects may be evaluated. This information may be available from routine data sources such as hospital records and cancer registries (as in Example 9.7).

9.2.2 Incident versus prevalent cases

An important issue to consider at the design stage of a case–control study is whether to include *prevalent* or only *incident* cases. Incident cases are all *new* cases occurring in a population within a fixed period of time. Prevalent cases are all *existing* (new and old) cases who are present in a population at a particular point in time (or within a very short period) (see Section 4.2). The main disadvantage of using a prevalent case series is that patients with a long course of disease tend to be over-represented since, by

Example 9.7. In the stomach cancer case–control study discussed in Example 9.6, the number and characteristics of the cases recruited in each participating centre were compared with the information collected by local cancer registries or pathology departments. Table 9.2 shows that the cases recruited to the study in Florence were slightly younger and more often females than the cases notified to the local cancer registry. This was because cases without histological confirmation were excluded from the study, and these were generally men in older age-groups (Buiatti et al., 1989a, b).

Table 9.2.
Age and sex distribution (%) of gastric cancer cases recruited by the Florence centre during 1985–87 and of gastric cancer cases notified to the local cancer registry in 1985.[a]

Age (years)	Males (M)				Females (F)				M:F ratio
	<45	45–54	55–64	65–74	<45	45–54	55–64	65–74	
Cases recruited, 1985-87	4.4	13.3	33.2	49.1	3.8	12.1	18.2	65.9	1.7
Cases notified to the registry, 1985	2.6	10.7	33.2	53.6	3.1	13.5	20.8	62.5	2.0

[a] Data from Buiatti *et al.* (1989a,b). (The numbers of cases on which these percentages are based were not given in these papers).

definition, all those with a short duration leave the pool of prevalent cases because of either recovery or death. Unless we can justify the assumption that the exposure being studied is not associated with recovery or survival, every effort should be made to limit recruitment to incident cases. By using only newly diagnosed (incident) cases and selecting controls to be representative of subjects from the population from which the cases arise, the case–control study aims to identify factors responsible for disease development, much like a cohort study. Moreover, prevalent cases may not be representative of all cases if some affected patients are institutionalized elsewhere or move to another city where there are special facilities for treatment.

There are other advantages to the use of incident cases. Recall of past events in personal histories tends to be more accurate in newly diagnosed cases than in prevalent cases. Besides, incident cases are less likely to have changed their habits (or 'exposures') as a result of the disease.

If constraints on time or resources make the use of prevalent cases inevitable, we should choose those that were diagnosed as close as possible to the time of initiation of the study. A check on the characteristics of the prevalent cases may be possible by comparing the frequency (or level) of exposure among subjects with different times of diagnosis. If, among the cases, the frequency of exposure to a factor suspected of being associated with the disease changes with time since diagnosis, we should suspect *survival bias*. For instance, if those cases who were exposed to the factor under study have poorer survival than those unexposed, they will become under-represented in a prevalent case series as time since diagnosis increases. As a result, the prevalence of exposure among the surviving cases will decrease.

Prevalent cases may have to be used for conditions for which it is difficult to establish a specific date of onset. For instance, case–control studies to examine risk factors for *Helicobacter pylori* infection have to be based on prevalent cases, because it is difficult to establish the date of onset of this condition.

9.2.3 Source of cases

Which cases are to be recruited into a study needs to be carefully considered. The study may be '*hospital-based*' and the cases taken from all patients fulfilling the eligibility criteria and attending a certain hospital or a group of hospitals. In Example 9.1, the cases were white women, aged 40–80 years, who were admitted to a certain hospital in Boston from January 1970 to June 1975 with a first diagnosis of endometrial cancer.

Alternatively, the study may be '*population-based*' and cases taken from a defined population over a fixed period of time. This is illustrated in Example 9.8.

> **Example 9.8.** *In the cervical cancer case–control study mentioned in Example 9.2, an active case-finding network was organized with periodic visits to all hospitals, clinics and pathology departments in the public and private sector in each study area to identify and interview the cases before any treatment was applied. All cervical intraepithelial neoplasia (CIN) III cases diagnosed during the study period were also identified and the histological slides were reviewed by a panel of pathologists to ensure completeness of recruitment of the invasive cancer cases (Muñoz et al., 1992a).*

In population-based case–control studies, it is essential to ensure completeness of case-finding. Issues that need to be considered are completeness of patient referral to health centres (which is likely to be a minor problem in cancer studies in countries where medical care is generally available but a much greater one elsewhere), difficulty in tracing the subjects, and refusal to participate.

Population-based cancer registries may be used to recruit all incident cases from their catchment population, but their value as a source of cases may be limited if there is a substantial time lag between diagnosis and registration. Moreover, cases with poor survival may have died in the meantime and others may have moved out of the catchment area as a result of their disease. Thus, by the time cases are registered, it may not be possible to regard them as incident.

9.3 Definition and selection of controls

9.3.1 Definition of controls

Controls must fulfil all the eligibility criteria defined for the cases apart from those relating to diagnosis of the disease. For example, if the cases are

women with breast cancer aged 45 years and over, the controls must be selected from women in the same age group without the disease.

If the disease being studied is uncommon in the group serving as a source of controls, little, if any, diagnostic effort or documentation is needed to rule out the disease in the selected controls. A simple interview question will often suffice. However, if the disease is common, a greater effort to minimize misclassification, such as a review of the individuals' medical records, is desirable (as in Example 9.9).

> **Example 9.9.** *In the cervical cancer case–control study mentioned in Example 9.2, controls were eligible if they were 70 years of age or younger, had not received previous treatment for cervical cancer or had not been hysterectomized, and if the cytological smear taken at the time of recruitment was normal or had only inflammatory changes (Pap classes I and II) (Muñoz* et al., *1992a).*

9.3.2   Source of controls

In case–control studies, controls should represent the population from which the cases are drawn, i.e., they should provide an estimate of the exposure prevalence in the population from which the cases arise. If not, the results of the study are likely to be distorted because of *selection bias*.

In a *nested case–control study*, it is relatively straightforward to ensure that the cases and controls are drawn from the same study population, since both will arise from a clearly defined population—the cohort (see Section 8.8). In general, all the cases arising as the cohort is followed prospectively become the 'cases' in the case–control study, while a sample of unaffected members of the cohort become the 'controls'.

In Example 9.10, both the cases and the controls were drawn from the same population—the cohort of 5908 Japanese American men living in Hawaii.

Conceptually, we can assume that all case–control studies are 'nested' within a particular population. In a *population-based case–control*

> **Example 9.10.** *The relationship between* Helicobacter pylori *infection and gastric carcinoma was examined in a cohort of Japanese American men living in Hawaii. A total of 5908 men were enrolled from 1967 to 1970. At that time each man provided a blood sample. By 1989, a total of 109 new cases of pathologically confirmed gastric carcinoma had been identified among the cohort members. The stored serum samples from all the patients with gastric carcinoma ('cases') and from a selection of subjects who did not develop gastric cancer ('controls') were then tested for the presence of serum IgG antibody to* Helicobacter pylori *(Nomura* et al., *1991).*

study, a study population can be defined from which all incident cases are obtained; controls should be randomly selected from the disease-free members of the same population. Consider, for example, all the newly diagnosed cases of childhood cancer in the catchment area of a regional cancer registry. Controls for these cases would appropriately be drawn from the population of the same area in the same sex- and age-groups. Even when the cases are identified exclusively from hospitals, it still may be reasonable to assume that they represent all the cases in the catchment area if the disease is serious enough that all cases end up in hospital (which is likely to be true for most cancer cases in countries with good health care).

It is generally expensive and time-consuming to draw controls from a random sample of the catchment population. A list of all eligible subjects or households must be available for sampling, or has to be created (as in Example 9.11). (Methods to select a random sample from the study population are discussed in Chapter 10.) Besides, healthy people may be disinclined to participate, which may introduce selection bias due to non-response.

> **Example 9.11.** *In the cervical case–control study mentioned above, controls were randomly selected from the general population that generated the cases. In Colombia, up-to-date aerial pictures of the city were used as the sampling frame. From these pictures, houses were selected at random and door-to-door searching following pre-determined routines was employed to identify suitable controls. In Spain, the provincial census of 1981, the latest available, was used as the sampling frame (Muñoz et al., 1992a).*

Controls may also be selected from close associates of the case, such as *friends and relatives* who are from the same catchment population as the cases. Although a relatively small effort is required to identify these controls and obtain their cooperation, there is a danger that they will be too similar (overmatched) to cases in terms of exposures and other characteristics (see Section 9.3.4). *Neighbourhood controls* can also be used, but people living in the same neighbourhood are likely to be similar in many respects, so such controls may also be overmatched. Moreover, if the interviewer has to visit each neighbourhood to contact these controls, the cost of the study may become extremely high.

When using hospital-based cases, it may not be possible to define the population from which the cases arose, either because the exact catchment area of the hospital cannot be defined or because not all the cases in the area are referred to the hospital, and those referred may be selected according to particular criteria (e.g., the more serious). In these circumstances, *hospital-based controls* may be used because the study population can then be defined as potential 'hospital users'.

Hospitalized controls have several advantages. There are many selective factors that bring people to hospitals (e.g., financial standing, area of residence, ethnicity, religious affiliation) and by selecting controls from the same pool of patients that gave rise to the cases, we reduce the effect of these factors. These controls are generally easily identified and they tend to be cooperative. In addition, since they have also experienced illness and hospitalization, they may resemble the cases with respect to their tendency to give complete and accurate information, thus reducing potential differences between cases and controls in the quality of their recall of past exposures.

Choosing suitable hospital controls is often difficult and great care must be taken to avoid selection bias. A major disadvantage of a control group selected from diseased individuals is that some of their illnesses may share risk factors with the disease under study, that is, they may have a higher, or lower, exposure prevalence compared with the population from which the cases arise. For instance, in a study investigating the role of alcohol and breast cancer, the use of controls from the accident and emergency department of the same hospital would introduce bias because this group is known to have a higher alcohol consumption than the general population from which the cases arise. One way of minimizing this bias is to select controls with different conditions so that biases introduced by specific diseases will tend to cancel each other out.

Choice of a suitable control group is the most difficult part of designing a case–control study. Some studies use more than one type of control group. The conclusions from a study are strengthened if similar results are obtained with each of the control groups.

After the source and number of control groups for a study have been determined, it is necessary to decide how many controls per case should be selected. This issue is considered in detail in Chapter 15, but when the number of available cases and controls is large and the cost of obtaining information from both groups is comparable, the optimal control-to-case ratio is 1:1. When the number of cases available for the study is small, or when the cost of obtaining information is greater for cases than controls, the control-to-case ratio can be altered to ensure that the study will be able to detect an effect, if one really exists (i.e., that the study has the necessary statistical power). The greater the number of controls per case, the greater the power of the study (for a given number of cases). However, there is generally little justification to increase this ratio beyond 4:1, because the gain in statistical power with each additional control beyond this point is of limited magnitude. Sample size issues (and the concept of 'statistical power' of a study) are discussed in Chapter 15.

As for cases, it is important to collect information on reasons for non-participation of controls and, whenever possible, to obtain additional information on their sociodemographic characteristics (e.g., sex, age, socioeconomic status) (as in Example 9.12).

Example 9.12. In the Italian gastric cancer case–control study mentioned in Examples 9.6 and 9.7, controls were randomly selected from population lists within five-year age and sex strata. A total of 1423 population-based controls were sampled, of whom 1159 (81%) were successfully interviewed using the same structured questionnaire as for the cases (Buiatti et al., 1989a,b). Table 9.3 shows the numbers of controls that were sampled, how many were recruited and the reasons for non-participation by recruitment centre.

| Recruitment centre | Sampled No. (%) | Recruited No. (%) | Excluded due to | |
			Refusal No. (%)	Poor health No. (%)
Cagliari	118 (100)	108 (91.5)	8 (6.8)	2 (1.7)
Cremona	61 (100)	51 (83.6)	5 (8.2)	5 (8.2)
Florence	547 (100)	440 (80.4)	74 (13.6)	33 (6.0)
Forli	291 (100)	259 (89.0)	20 (6.9)	12 (4.1)
Genoa	205 (100)	137 (66.8)	17 (8.3)	51 (24.9)
Imola	74 (100)	61 (82.4)	10 (13.5)	3 (4.1)
Siena	127 (100)	103 (81.1)	6 (4.7)	18 (14.2)
Total	**1423 (100)**	**1159 (81.4)**	**140 (9.9)**	**124 (8.7)**

[a] Data from Buiatti *et al.* (1989a).

Table 9.3.
Recruitment levels among controls and reasons for non-participation by recruitment centre.[a]

9.3.3. Sampling schemes for controls

If the source of incident cases is a closed cohort with a fixed follow-up period (as is the case in nested case–control studies), controls may be selected in three different ways, as illustrated in Example 9.13.

In this example, the first option for the investigators is to sample controls from the population still at risk by the end of the study, that is from those subjects who were still disease-free by the end of the follow-up period (A in Figure 9.1). In this design, each woman can be either a case or a control, but not both.
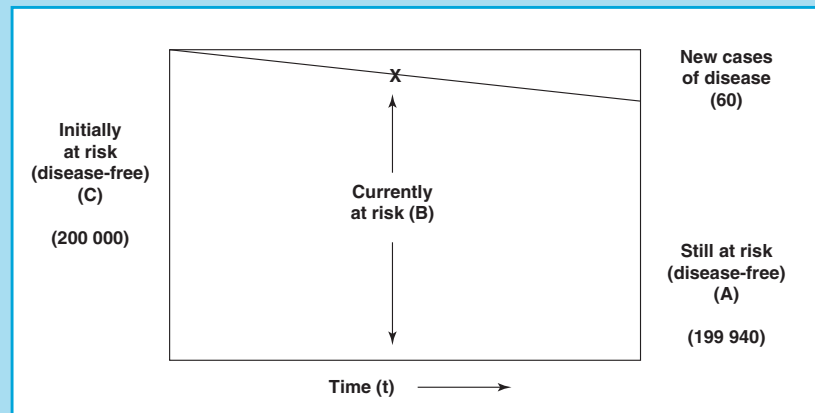
The second option is to sample controls from those who are still at risk at the time each case is diagnosed, that is, controls are *time-matched* to the cases (see Section 9.3.4) (B in Figure 9.1). In this sampling design, a subject originally selected as a control can become a case at a later stage. The opposite cannot happen, since once a woman has acquired endometrial cancer she is no longer at risk, and therefore not eligible for selection as a control. Subjects selected as controls who then become cases should be retained as, respectively, controls and cases in the appropriate sets.

Thirdly, controls can be a sample from those who were at risk at the start of the study (C in Figure 9.1). Studies of this type are called *'case–cohort studies'*. Since the control group reflects the total population and not just those who did not get the disease, a woman ascertained as a case may also be selected as a control, and vice versa. Such women should

Example 9.13. Suppose that a cohort of 200 000 healthy women was followed up for ten years to assess the relationship between lifestyle variables and the risk of developing various types of cancer and that a total of 60 women were diagnosed with endometrial cancer during the follow-up period. Suppose also that the investigators decided to conduct a case–control study nested within this cohort to assess whether oral contraceptive use protected against endometrial cancer. The investigators could sample the controls in three different ways, as indicated in Figure 9.1: (1) from those who were still disease-free by the end of the follow-up period (situation A); (2) from those who were still at risk at the time each case was diagnosed (situation B); or (3) from those who were at risk at the start of the study (situation C).

Figure 9.1.
Sampling schemes for controls when the source of incident cases is a closed cohort with a fixed follow-up (x=incident case).

be included in the study as both cases and controls.

If the source of incident cases is a dynamic population, as in most hospital and population-based case–control studies, it may be difficult to establish the exact population at risk at the start and end of the follow-up period, so the preferred method is to choose for each incident case one or more controls from those subjects who are members of the same population and still at risk of developing the disease at the time of the diagnosis of the case.
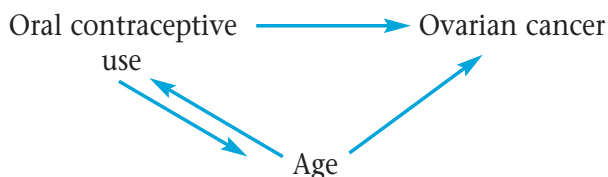
As we shall see later in this chapter (Section 9.5), the specific relative measure of effect (rate ratio, risk ratio or odds (of disease) ratio) that can be estimated from a case–control study depends on the type of sampling design used in the selection of the controls.

9.3.4. Matching

Individual matching refers to the procedure whereby one or more controls are selected for each case on the basis of similarity with respect to certain characteristics other than the exposure under investigation. Since cases and controls are similar on the matching variables, their difference with respect to disease status may be attributable to differences in some other factors. It is, however, important that *matching is restricted to con-*

founding factors and is not performed for the exposure under investigation. The characteristics generally chosen for matching are those that are known to be strong *confounders*. Common matching variables are age, sex, and ethnicity but others might be place of residence, or socioeconomic status.

Let us suppose that we are interested in examining the relationship between current use of oral contraceptives and ovarian cancer. In this example, it is appropriate to match on age, since age is associated with the exposure of interest (current oral contraceptive use) and is an independent risk factor for ovarian cancer. In other words, age is a confounding factor. Failure to match, or otherwise control, for age would result in a biased assessment of the effect of oral contraceptive use.

Oral contraceptive use $\longrightarrow$ Ovarian cancer

Age

When controls are chosen so as to be similar to the cases for a characteristic and when this similarity tends to mask the disease's association with the exposure of interest, cases and controls are said to be *overmatched*. This can happen when controls are matched to cases for a characteristic that is part of the pathway through which the possible cause of interest leads to disease. Imagine a case–control study conducted in West Africa to investigate the role of hepatitis B virus in the etiology of liver cancer in which controls were matched to cases on the basis of previous history of liver disease.

Hepatitis B virus $\longrightarrow$ Chronic liver disease $\longrightarrow$ Liver cancer

If chronic liver disease is on the pathway between hepatitis B infection and liver cancer, matching on that condition would result in an underestimation of the effect of the virus on the occurrence of liver cancer, since controls would have been made similar to the cases in relation to this variable.

Another form of overmatching relates to matching for a variable which is correlated to the exposure of interest but is not an independent risk factor for the disease under study (and so cannot be a confounding factor) and is not on its causal pathway. For instance, suppose we wish to examine the relationship between smoking and lung cancer in a population where smoking levels are positively correlated with alcohol intake, that is, the more someone drinks the more he/she is likely to smoke. In this example, matching on alcohol intake would result in overmatching because controls would be made similar to the cases not only in relation to their alcohol intake but also in relation to their smoking habits, which is the exposure of interest in this study.

Hence, caution should be exercised in determining the number of variables selected for matching, even when there are no practical restrictions. If the role of a variable is in doubt, the preferable strategy is not to match but to adjust for it in the statistical analysis (see Chapters 13–14).

In most case–control studies, there are a small number of cases and a large number of potential controls to select (or sample) from. In practice, each case is classified by characteristics that are not of direct interest, and a search is made for one or more controls with the same set of characteristics.

Example 9.14. Adenocarcinoma of the vagina in young women was recorded rarely until the report of several cases treated at the Vincent Memorial Hospital (in Boston, MA, USA) between 1966 and 1969. The unusual diagnosis of this tumour in eight young patients led to the conduct of a case–control study to search for possible etiological factors. For each of the eight cases with vaginal carcinoma, four matched female controls born within five days and on the same type of hospital service (ward or private) as the case were selected from the birth records of the hospital in which the case was born. All the mothers were interviewed personally by a trained interviewer using a standard questionnaire (Herbst et al., 1971).

In Example 9.14, cases and controls were individually matched on date of birth (and, hence, age), hospital of birth and type of hospital service (ward versus private). If the factors are not too numerous and there is a large reservoir of persons from which the controls can be chosen, case–control individual matching may be readily carried out. However, if several characteristics or levels are considered and there are not many more potential controls than cases, matching can be difficult and it is likely that for some cases, no control will be found. Moreover, when cases and controls are matched on any selected characteristic, *the influence of that characteristic on the disease can no longer be studied*. The number of characteristics for which matching is desirable and practical is actually rather small. It is usually sensible to match cases and controls only for characteristics such as age, sex and ethnicity whose association with the disease under study is well known.

As an alternative to individual matching, we may *frequency match* (or *group match*). This involves selecting controls so that a similar proportion to the cases fall into the various categories defined by the matching variable. For instance, if 25% of the cases are males aged 65–75 years, 25% of the controls would be taken to have similar characteristics. Frequency matching within rather broad categories is sufficient in most studies.

9.4　Measuring exposures

Data on the relevant exposures can be obtained by personal, postal or telephone interview, by examining medical, occupational or other records, or by taking biological samples. Whatever method is chosen, it is fundamental to ensure that the information gathered is unbiased, i.e., it is not influenced by the fact that an individual is a case or a control. Ideally, the investigator or interviewer should be 'blind' to the hypothesis under study and to the case/control status of the study subjects. In practice, this may be difficult to accomplish, but all possible efforts should be made to ensure unbiased collection of data to minimize *observer bias*. Particular effort is required in multicentric studies to ensure standardization of data collection techniques across the different participating centres.

Bias can also occur when the validity of the exposure information supplied by the subjects differs for cases and controls (*responder bias*). Subjects with a serious disease are likely to have been thinking hard about possible causes of their condition and so cases may be inclined to give answers that fit with what they believe (or think is acceptable to say) is the cause of their illness. This type of responder bias is called *recall bias*. Responder bias can be minimized by keeping the study members unaware of the hypotheses under study and, where possible, ensuring that both cases and controls have similar incentives to remember past events. These issues are further discussed in Chapter 13.

9.5　Analysis

The analysis of data from case–control studies depends on their design. Individual-matched studies require a different type of analysis from ummatched (or frequency-matched) studies.

9.5.1 Unmatched (and frequency-matched) studies

The first step in the analysis of an unmatched case–control study is to construct a table showing the frequency of the variables of interest separately for cases and controls. The frequency of some of these variables in the controls may help to judge whether they are likely to represent the population from which the cases arise. For instance, in Example 9.15, the distribution of schooling, parity, smoking, etc. in the control group of this population-based study may be compared with governmental statistics or results from surveys conducted in the same areas.

In Example 9.15, the distributions of some of the variables known to be risk factors for cervical cancer are consistent with those found in other studies in that cases were more likely to have a lower educational level, higher parity and a greater number of sexual partners than controls. They were also more likely to have ever used oral contraceptives or smoked.

Table 9.4.
Distribution of selected socio-demographic and reproductive variables among cervical cancer cases and controls in each of the participating centres[a].

Example 9.15. In the cervical cancer case–control study described in Example 9.2, the distribution of variables known to be risk factors for cervical cancer was examined among cases and controls. Table 9.4 shows the results for some of these variables.

	Spain		Colombia	
	Cases No. (%)	Controls No. (%)	Cases No. (%)	Controls No. (%)
Age (years)				
<30	7 (2.8)	7 (2.9)	15 (8.1)	10 (6.7)
30–39	41 (16.4)	39 (16.4)	48 (25.8)	30 (20.1)
40–44	30 (12.0)	27 (11.3)	27 (14.5)	24 (16.1)
45–54	61 (24.4)	58 (24.4)	50 (26.9)	40 (26.9)
55+	111 (44.4)	107 (45.0)	46 (24.7)	45 (30.2)
Schooling				
Ever	162 (64.8)	179 (75.2)	155 (83.3)	140 (94.0)
Never	88 (35.2)	59 (24.8)	31 (16.7)	9 (6.0)
Parity				
0–2	102 (40.8)	136 (57.1)	41 (22.0)	45 (30.2)
3–5	119 (47.6)	86 (36.1)	62 (33.3)	53 (35.6)
6+	29 (11.6)	16 (6.7)	83 (44.6)	51 (34.2)
Number of sexual partners				
0–1	189 (75.6)	218 (91.6)	76 (40.9)	87 (58.4)
2–5	48 (19.2)	16 (6.7)	77 (41.4)	58 (38.9)
6+	13 (5.2)	4 (1.7)	33 (17.7)	4 (2.7)
Oral contraceptives				
Never	182 (72.8)	175 (73.5)	109 (58.6)	95 (63.8)
Ever	64 (25.6)	53 (22.3)	77 (41.4)	53 (35.6)
Unknown	4 (1.6)	10 (4.2)	0 (0.0)	1 (0.6)
Smoking				
Never	184 (73.6)	198 (83.2)	92 (49.5)	90 (60.4)
Ever	66 (26.4)	40 (16.8)	94 (50.5)	59 (39.6)

[a]   Data from Bosch *et al.* (1992)

In an unmatched study, the numbers of cases and controls found to have been exposed and not exposed to the factor under investigation can be arranged in a 2 × 2 table as shown in Table 9.5:

Table 9.5.
Layout of a 2 × 2 table showing data from an unmatched case–control study

	Exposed	Unexposed	Total
Cases	a	b	n_1
Controls	c	d	n_0
Total	m_1	m_0	**N**

In Section 4.2.2, we presented the three measures of incidence (risk, odds of disease and rate) that can be estimated from a cohort study. These three measures use the same numerator—the number of new cases that occurred during the follow-up period—but different denominators. Risk takes as the denominator people who were at risk at the start of the fol-

low-up; odds of disease takes those who were still disease-free by the end of the follow-up; and the rate uses the total person-time at risk, which takes into account the exact time when the cases occurred. Comparison of these measures of incidence in those exposed relative to those unexposed yields three different measures of relative effect: the risk ratio, the odds (of disease) ratio and the rate ratio, respectively (see Section 5.2.1).

In case–control studies, it is not possible to directly estimate disease incidence in those exposed and those unexposed, since people are selected on the basis of having or not having the condition of interest, not on the basis of their exposure status[a]. It is, however, possible to calculate the odds of exposure in the cases and in the controls:

Odds of exposure in the cases = a/b

Odds of exposure in the controls = c/d

The *odds (of exposure) ratio* can then be calculated as

$$\text{Odds (of exposure) ratio} = \frac{\text{Odds of exposure in the cases}}{\text{Odds of exposure in the controls}} = \frac{a/b}{c/d}$$

Example 9.16. *In the case–control study illustrated in the previous example, the risk of cervical cancer was examined in relation to education (schooling). Data from Spain and Colombia were pooled in this analysis (Table 9.6) (Bosch* et al., *1992).*

	Schooling		Total
	Never ('exposed')	**Ever ('unexposed')**	
Cervical cancer cases	119 (a)	317 (b)	436 (n_1)
Controls	68 (c)	319 (d)	387 (n_0)
Total	**187 (m_1)**	**636 (m_0)**	**823 (N)**

[a] Data from Bosch *et al.* (1992)

Odds ratio = (119 / 317) / (68 / 319) = 1.76

95% confidence interval = 1.24–2.46

$\chi^2 = 11.04$, 1 d.f.; $P = 0.0009$

(Confidence intervals and test statistics for the odds ratio were calculated as shown in Appendix 6.1.)

Table 9.6.
Schooling among cervical cancer cases and controls.[a]

It can be shown algebraically that the odds (of exposure) ratio obtained from a case–control study provides an unbiased estimate of one of the three relative measures of effect that can be obtained from

[a] Indirect calculations are possible in population-based case–control studies (see Appendix A16.1)

a cohort study, depending on the sampling scheme used to select the controls (see Section 9.3.3). If controls are selected from all those who are initially at risk, the case–control study will directly estimate the *risk ratio*. If controls are sampled from those who are still disease-free by the end of the follow-up, the study will estimate the *odds (of disease) ratio*. If controls are selected from those still at risk at the time each case is ascertained, the study will provided an unbiased estimate of the *rate ratio*. In this last instance, the analysis should respect the fact that cases and controls are *matched* with respect to time. (An unmatched analysis will also yield an unbiased estimate of the rate ratio if the rates of acquiring disease remain constant over time among both the exposed and unexposed populations and the total numbers at risk remain relatively constant in both populations.)

As we saw in Section 5.2.1, when the disease is rare, as with cancer, cases constitute a negligible fraction of the population. The number of people at risk in a cohort study remains practically constant over time and therefore the three measures of effect yield similar results. Consequently, the three sampling schemes used to select controls in a case–control study will also provide similar results. If the disease is common, however, different control sampling schemes will yield different results and the choice of the most appropriate one will depend on the specific problem being addressed (Smith *et al.*, 1984; Rodrigues & Kirkwood, 1990).

In strict terms, the odds ratio obtained from a case–control study tells us how many more (or less, if the exposure is associated with a reduced risk) times likely the cases are to have been exposed to the factor under study compared with the controls.

In Example 9.16, cervical cancer cases were 76% more likely to have never attended school than controls. Since the odds ratio obtained from a case–control study provides an estimate of one of the three relative measures of effect that can be calculated from a cohort study, we can also interpret it as an indication of the likelihood of developing the disease in the exposed individuals relative to those unexposed. In our example, the odds ratio indicates that women who never attended school were 76% more likely to develop cervical cancer than those who attended.

As in other types of study, inferences about the association between a disease and a factor are considerably strengthened if there is evidence of a gradient between the level (or intensity) of exposure and risk of the disease in question. Odds ratios can be computed separately for each level of the exposure. The general approach is to treat the data as a series of 2×2 tables, comparing controls and cases at different levels of exposure, and then calculating the odds ratio at each level.

In Example 9.17, there is a trend of increasing risk of cervical cancer with increasing number of sexual partners.

Example 9.17. In the cervical cancer case–control study conducted in Colombia and Spain and described in Examples 9.2 and 9.15, the risk of developing cervical cancer was examined in relation to the lifetime number of sexual partners (Bosch et al., 1992). Data from Spain and Colombia were pooled together in this analysis (Table 9.7).

Number of sexual partners	Cervical cancer cases	Controls	Odds ratio (95% confidence interval)
0–1[b]	265	305	1.0[c]
2–5	125	74	1.94 (1.39–2.70)
6+	46	8	6.62 (3.07–14.27)

[a] Data from Bosch *et al.* (1992)
[b] Taken as the baseline (reference) category.
[c] χ^2 test for trend = 39.48; 1 d.f.; *P*<0.00001
(Confidence intervals and χ^2 test for trend in odds ratio calculated as shown in Appendix 6.1.)

The odds ratios for each category of exposure were calculated in the following way:

	Number of sexual partners	
	2–5 ('exposed')	0–1 ('unexposed')
Cervical cancer cases	125	265
Controls	74	305

Odds ratio = (125 / 265) / (74 / 305) = 1.94

	Number of sexual partners	
	6+ ('exposed')	0–1 ('unexposed')
Cervical cancer cases	46	265
Controls	8	305

Odds ratio = (46 / 265) / (8 / 305) = 6.62

Special statistical techniques can be used to adjust for potential confounding factors in the analysis. These are discussed in Chapters 13 and 14. One of these techniques was used to examine the association between schooling and the risk of developing cervical cancer found in Example 9.16. The *crude odds ratio* was 1.8 (Table 9.6). After taking into account differences in age, participating centre, human papillomavirus status, number of sexual partners, education, age at first birth, and history of previous screening between cases and controls, the resulting *adjusted odds ratio* was 2.5 (95% confidence interval = 1.6–3.9). Thus, the association between never having attended school and cervical cancer observed in the crude analysis could not be explained by differences in the distribution between

cases and controls of any of these factors (in fact, the adjusted odds ratio was higher than the crude odds ratio) (see Section 13.2 and Chapter 14 for further discussion of these issues).

9.5.2 Individual-matched studies

Individual-matched studies require a special type of analysis, in which the 2×2 table takes a different form. Let us consider the simplest situation where there is only one control per case. The status of the cases with regard to the presence or absence of the exposure of interest is cross-tabulated against the exposure status of their respective controls (Table 9.8).

Table 9.8.
Layout of a 2×2 table with data from an individual-matched case–control study (control-to-case ratio = 1:1).

		Controls		Total
		Exposed	**Unexposed**	
Cases	Exposed	*r*	*s*	*a*
	Unexposed	*t*	*u*	*b*
	Total	***c***	***d***	***N/2***

In this table, *r*, *s*, *t*, *u* represent the number of pairs in which

r = case exposed and control exposed (+ +)
s = case exposed but control not exposed (+ –)
t = case not exposed and control exposed (– +)
u = case not exposed and control not exposed (– –)

The marginal totals (*a*, *b*, *c*, *d*) of this table correspond to the entries in the cells of the table for the unmatched studies. The total for the entire table is *N/2* pairs, where *N* represents the total number of paired individuals.

The *matched odds ratio* can be calculated as

Odds ratio = *s/t* (provided *t* is not equal to 0)

This odds ratio calculation considers only the discordant pairs. It can be explained intuitively: pairs where both case and control were exposed or where both were unexposed give no information about the relationship of the exposure to disease (Example 9.18).

The analysis is more complex than shown here if there is more than one control per case (see Breslow & Day (1980), chapter 5).

9.6   Interpretation of results

Case–control studies are well suited to study diseases of long induction, because no lengthy follow-up is involved. They are also suitable for studying rare diseases, since a prospective cohort study would require the recruitment of a very large number of individuals and a long follow-up period to ensure the accrual of a sufficient number of cases.

The interpretation of case–control studies is, however, less straightforward than that of cohort studies and the investigator must always consider

Example 9.18. A case–control study was carried out in Canada to assess whether artificial sweeteners, particularly saccharin, increased the risk of bladder cancer. Newly diagnosed cases of bladder cancer that occurred among residents in the provinces of British Columbia, Nova Scotia and Newfoundland between April 1974 and June 1976 were identified through provincial cancer registries and cooperative pathologists and urologists. A total of 821 eligible cases were ascertained, and 632 of these were personally interviewed in their homes using a structured questionnaire. Reasons for failure to interview included death (56), refusal (65), too ill to be interviewed (25), and refusal of permission by the attending physician (34). Most interviews were done within three months of diagnosis, and all within six months. For each case, an individual matched on sex, age (within 5 years), and neighbourhood residence was interviewed (Howe et al., 1977). The main results are shown in Table 9.9.

Table 9.9.
Frequency of exposure to artificial sweeteners among 632 bladder cancer cases and their individual-matched controls.[a]

		Controls		Total
		Exposed	**Unexposed**	
Cases	Exposed	468 (*r*)	87 (*s*)	555 (*a*)
	Unexposed	73 (*t*)	4 (*u*)	77 (*b*)
	Total	**541 (*c*)**	**91 (*d*)**	**632 (*N/2*)**

[a] Data from Howe *et al.* (1977)

Matched odds ratio = 87/73 = 1.19

95% confidence interval for the matched odds ratio = 0.86–1.65

McNemar's χ^2 = 1.23; P = 0.27.

(The calculation of confidence intervals and significance tests for matched case–control studies is explained in Breslow & Day (1980)).

whether the result could have arisen as a result of selection bias in the choice of cases and controls, information bias in the gathering of exposure data, or failure to take proper account of confounding factors.

The most serious potential problem in case–control studies is that the procedures used to select cases and controls may produce groups that are not truly comparable.

In Example 9.19, selection bias could have affected the results of this study since only 62% (314/510) of all eligible patients were included in the final analysis. Low participation levels can introduce bias if cases who used oral contraceptives were more or less likely to participate in the study. If, for instance, users of oral contraceptives were more likely to have a less aggressive form of breast cancer than non-users and, hence, a better survival, this would lead to over-estimation of the effect of oral contraceptives since a high proportion of the deaths would have occurred among non-users.

Selection of an appropriate control group is one of the most difficult problems in case–control studies. Controls must come from the same defined population as the cases. The use of hospital-based controls works only if patients with different diseases came from the same general population (i.e.,

> **Example 9.19.** *The relation between use of oral contraceptives by young women and their risk of breast cancer was investigated in a population-based case–control study conducted in Los Angeles County. The cases were patients with histologically confirmed breast cancer, first diagnosed between July 1972 and May 1982, diagnosed before age 37 years, and without a prior history of malignancy. A total of 510 eligible cases were identified through the local population-based cancer registry, of whom 458 were still alive at the time of the first contact through their doctors. Physicians gave permission to contact 393 (86%) of these patients. Of these, 26 could not be located and 37 refused to be interviewed. Thus, completed questionnaires were obtained from 330 patients. Sixteen of these patients were later excluded because no suitable individually matched control was found (Pike et al., 1983).*

if the referral patterns are the same for the disease under investigation and the control diseases) and if the control diseases are themselves unrelated to the exposure. In many situations, it is difficult to be sure that these conditions are satisfied. The use of population-based controls avoids these problems, but selection bias may still be introduced if the levels of non-response are high either because some eligible controls cannot be traced or because they refuse to participate (as in Example 9.20). In this instance, the control series may not be representative of the population from which the cases arise.

> **Example 9.20.** *The possible association between oral contraceptive use and the risk of breast cancer at young ages (under 45 years) was investigated in a population-based case–control study conducted in Sweden and Norway. In Norway, where notification of all cancer diagnoses is mandatory, cases were identified from population-based cancer registries. A total of 114 eligible women were identified of whom 105 (92%) participated. For each case who agreed to participate, two controls were chosen from an up-to-date national population register. Potential controls were mailed a request to participate. If an answer was not received within four weeks or if the control refused to participate, a new control was selected. Nine controls were never located; 34 never answered the letter; 38 refused to participate; 4 were either temporarily abroad and could not be reached or had mental disorders. Thus, to obtain two controls for each case, it was eventually necessary to select 295 controls from the population register. Only 72% of the women with whom contact was sought were interviewed (Meirik et al., 1986).*

Another problem of case–control studies is that accurate measurements of past exposures are usually difficult to obtain, and the degree of accuracy and completeness of these measurements may be different for cases and controls. For instance, recall bias can arise in case–control studies because patients with the disease under study may be inclined to answer questions more carefully than control subjects. Comparison of the exposure histories obtained from

cases and controls with an independent source of information (e.g., medical records) may help to determine whether there was a systematic difference in recall by cases and controls.

Example 9.21. In the study described in the Example 9.20, an introductory letter with a brief description of the aim and scope of the study was sent initially to cases and controls. If they agreed to participate, they were interviewed personally by specially trained professional female interviewers (Meirik et al., 1986).

In Example 9.21, the aim of the investigation was explained to the women involved. This may have increased recall bias, particularly since the study was carried out during a time of great public concern about oral contraceptives and breast cancer. This problem could have been minimized to a certain extent by not disclosing the study hypothesis to the study subjects.

The other potential source of bias in a case–control study is *diagnostic bias*. For instance, if women using oral contraceptives are more likely than non-users to examine their breasts, or to have them examined by a physician or nurse, or to undergo mammography, diagnostic bias may be introduced. Thus, if a positive association between oral contraceptives and breast cancer is found in a study, it may just be due to the fact that oral contraceptive users are more investigated and therefore more likely to be diagnosed with breast cancer than non-users. One way of minimizing diagnostic bias is to obtain information on the frequency of breast examinations for each of the study subjects so that any effects of more frequent surveillance of oral contraceptive users can be controlled for in the analysis.

A well conducted case–control study that has taken into account all the methodological concerns can yield valid and informative results. As discussed earlier in this chapter (see Sections 9.3.3 and 9.5), if cases and controls are selected independently of exposure and controls are sampled randomly from a defined study population from which the cases arose, the results from a case–control study provide an unbiased estimate of the measure of effect that would be obtained from an equivalent cohort study. Nevertheless, it is important to remember that case–control studies always have the potential for bias and that each study should be evaluated individually to determine whether bias influenced the results. Usually, the difficulty lies in the fact that although it is easy to identify potential sources of bias in any particular case–control study, it is rarely possible to estimate the true impact that these biases may have had on the results.

An important limitation of case–control studies is that they cannot provide direct estimates of the incidence of disease in those exposed and in those unexposed (unless they are population-based; see Appendix 16.1). Thus, it is usually not possible to calculate the absolute impact of the exposure on the occurrence of the disease.

Case–control studies are not suitable for studying rare exposures because very few cases will have been exposed, unless a large proportion of the total

Further reading

* A brief history of the development and use of case–control studies in epidemiology is given in Lilienfeld & Lilienfeld (1979).

* Breslow & Day (1980) provides a comprehensive treatment of the analysis of case–control studies in cancer epidemiology.

* Detailed discussions of sampling schemes for selection of controls are given by Miettinen (1976), Greenland & Thomas (1982), Smith *et al*. (1984) and Rodrigues & Kirkwood (1990).

cases of disease are attributable to that particular exposure (i.e., the population excess fraction is high (see Section 16.2.1)). For instance, the prevalence of asbestos exposure is rare in the general population and accounts for a small proportion of lung cancers. Therefore a case–control study would not be appropriate to investigate the relationship between this exposure and lung cancer because very few cases would have been exposed to asbestos. However, this study design would be appropriate to investigate the relation between asbestos and pleural cancer because this exposure is responsible for a large proportion of these cases.

Finally, the temporal sequence between exposure and disease may be difficult to establish. The possibility that the exposure is the result (rather than the cause) of the disease should always be considered (*reverse causality*). For instance, even if an association between diet and stomach cancer is found in a case–control study, there is a possibility that dietary differences are a consequence rather than a cause of the cancer.

Box 9.1. Key issues

- Case–control studies are studies in which a group of people with the condition of interest ('cases') and a group without that condition ('controls') are identified and the prevalence (or level) of the relevant exposure is measured in the two groups and compared.

- The main advantages of these studies are:

  1. They are efficient in time and cost (at least compared with prospective cohort studies)

  2. They provide the possibility to investigate a wide range of possible risk factors.

  3. They are particularly suitable to investigate rare diseases or diseases with a long induction period.

- The main disadvantages of these studies are:

  1. It may be difficult to select an appropriate control group (*selection bias*).

  2. It is difficult to obtain accurate unbiased measures of past exposures (*information bias*).

  3. The temporal sequence between exposure and disease may be difficult to establish (*reverse causality*).

  4. They are not suitable for investigating rare exposures (unless the exposure is responsible for a large proportion of cases, i.e., the population excess fraction is high).

  5. It is not possible to obtain estimates of disease incidence among those exposed and those unexposed to a putative risk factor (except if the study is population-based).