Chapter 5
Case-Control Studies

Qian Wu

Key Points

- The case-control study population consisted of a case group selected from those with the disease of interest and a control group selected from those who did not have the disease.
- Case-control studies belong to observational studies. It set up a control group.
- In case-control studies, Odds Ratio was used to estimate the strength of the association between disease and exposure factors.
- Selection bias, information bias, and confounding bias are major sources of bias in case-control studies.

5.1 Overview of Case-Control Studies

The purpose of the case-control study is to evaluate the relationship between the disease and the exposure factors suspected of causing the disease. Both cohort and case-control studies are analytical studies, their main difference lies in the selection of the study population. In a cohort study, the subjects do not have the disease when entering the study and are classified according to their exposure to putative risk factors, in contrast, subjects in case-control studies are grouped according to the presence or absence of the disease of interest. Case-control studies are relatively easy to conduct and are increasingly being applied to explore the causes of disease, especially rare diseases. Case-control studies are used to estimate the relative risk of disease caused by a specific factor. When the disease is rare, case control study may be the only research method.

Q. Wu (✉)
School of Public Health, Xi'an Jiaotong University, Xi'an, China
e-mail: wuqian@xjtu.edu.cn

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_5

5.1.1 History

Case control study has a long history. In 1843, Guy compared male occupations with lung diseases with those with other diseases. But it was not until 1926 that Janet Lane Claypon first proposed a case-control study in a breast cancer research. Richard Doll's research on smoking and lung cancer in the 1950s gave a great impetus to the applications of case-control study. Since then, case-control studies have become more prominent in biomedical literature, and their design, implementation, and analysis have become more standardized in methodology.

5.1.2 Definition

A case-control study involves cases from those individuals with disease of interest and controls from those who are without the disease. Previous exposure histories of case and control subjects were examined to evaluate the relationship between exposure and diseases. If the exposure history of the case group and the control group is different, it is possible to infer that the exposure may be related to the disease. The difference in exposure between the case and control group helps to identify potential risk factors. The purpose is to explore whether there are factors related to the disease. The basic principle of a case-control study is shown in Fig. 5.1.

A case-control study is called a retrospective study because researchers need to investigate the exposure factors of the subjects before the occurrence of the disease. Sometimes retrospective studies are used to represent case-control studies. It may be confusing because the terms retrospective and perspicacity are also used to describe the time of data collection related to the current date. In this sense, case-control

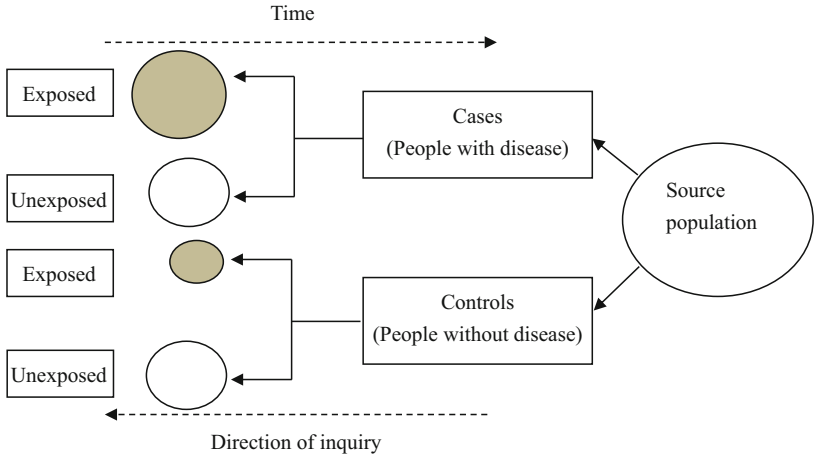

Fig. 5.1 Design of case-control study

studies can be retrospective, when all data are related to the past; it can also be forward-looking, in which data collection continues over time. Therefore, retrospective study is not the essential characteristics of case-control study. The essence of a case-control study is to divide the subjects into case and control groups according to the presence or absence of the disease of interest.

Example Some researchers surveyed the relationship between plasma metal concentration and the incidence rate of coronary heart disease (CHD) [Yu Yuan, Yang Xiao, Wei Feng, et al. Plasma Metal Concentrations and Incident Coronary Heart Disease in Chinese Adults: The Dongfeng-Tongji Cohort. Environ Health Perspect. 2017,125(10): 107007.]. The researchers compared 1621 CHD cases with 1621 controls free of CHD in Shiyan City, Hubei Province, China, in 2013. All of the participants were retired. Concentrations of aluminum, arsenic and barium, were significantly higher in cases (57.41, 2.32, 40.53 μg/L) than controls (48.95, 1.96, 35.47 μg/L). The study presented the concentrations of aluminum; arsenic and barium were higher in the cases than in the controls, indicating that circulating metals were associated with an increased incidence of CHD.

For example, information of participants' disease and their plasma metal was extracted from previous studies. In 2013, investigators according to the interest disease divided retirement employees into two groups. The case group is retirement employees with CHD, while control group is free of major cardiovascular disease. The researchers explored metal concentrations in plasma of participants from 2008 to 2013.

Firstly, the case-control study recruited patients according to their current disease status. Exposure history was inquired for in each case and control. Data were mostly collected after disease occurred, thus case-control study was considered retrospective, which was a limitation. Compared with cohort design, case-control study design has weak support for causal hypothesis. However, it provides more powerful evidence than cross-sectional studies in analyzing and interpreting the results. Case-control study is one of the commonly used research designs. The reason is that the implementation of case-control study is relatively simple and convenient compared with other study designs.

5.1.3 Type of Design Case-Control Studies

There are three kinds of case-control studies. First is the *traditional case-control design*. In this type, cases and controls are recruited from population. The case group is assumed to include all cases that occurred in that hypothetical cohort up to the time when the study is conducted. Control group is selected from those without the disease of interest throughout the study period. There are three subgroups of traditional case-control, which are unmatched, frequency matching, and individual matching case-control studies. Next is the nest case-control design which is conducted in a cohort population. At the beginning of nest case-control study (t_0),

members of the cohort are collected exposure factors. Cases and controls are identified subsequently at time t_1. The control group is selected from the cohort members who do not meet the case definition at t_1. Third is *a case-cohort design*. in the first step, a population was identified as the cohort for the study, and a sample within that cohort was selected as the control group using a randomized method. In the whole cohort, all cases of the disease to be studied were collected as a case group. Finally, the two groups were compared and analyzed to explore the factors affecting disease onset, disease survival time, and prognosis.

5.1.4 Characteristics of Case-Control Study

Case-control studies belong to *observational study*. Case-control study draws inferences from a sample to a population where the independent variable is not under the intervention of the investigator because of ethical concerns or logistical constraints.

Case-control study set up a *control group*. The differences of exposure were compared between case and control group.

Case-control studies are a special type of *retrospective study*. Investigators look back in time and access prior exposure status between two groups.

The relative risk (RR) cannot be calculated directly in case-control studies, and the Odds Ratio (OR) can be used to estimate the RR.

5.1.5 Application

Case-control studies are suitable for investigating rare diseases or diseases with a long latency period, as subjects are selected from the outset based on their outcome status. Therefore, compared to cohort studies, case-control studies are faster and relatively less expensive to implement, require relatively fewer subjects, and allow for multiple exposures or risk factors to be assessed for a single outcome.

5.1.5.1 Example of a Case-Control Study

In October 1989, physicians in the United States reported three patients with a newly recognized disease characterized by marked peripheral eosinophilia with features of scleroderma. After reporting this obvious association, more cases were found in the United States and Europe. To illustrate a possible link between EMS and the tryptophan manufacturing process, they conducted case-control studies to assess potential risk factors, including the use of tryptophan from different manufacturers. In early November 1989, they carried out a case-control study that demonstrated an epidemiologic association between the consumption of tryptophan products and the eosinophilia-myalgia syndrome (EMS). The case-control studies were used to

evaluate potential risk factors, including the use of tryptophan from different manufacturers. The investigators analyzed the tryptophan samples using high-performance liquid chromatography to determine the other chemical component. The results found that 29 of 30 case patients (97%) and 21 of 35 controls (60%) of the subjects using tryptophan had consumed tryptophan produced by one company. The EMS outbreak in 1989 was due to the ingestion of a chemical ingredient that was associated with a specific tryptophan manufacturing condition in one company.

This study suggests several important characteristics of case-control studies. Firstly, the design provides a suitable research method for studying this rare disease of EMS. Case-control study is applicable to the etiology of rare diseases. Secondly, case-control studies allow researchers to investigate several risk factors at the same time. In this research, researchers explored the effect of tryptophan and other factors on EMS. Finally, a case-control study usually does not "prove" causality, but it can suggest a hypothesis. The researchers believe that more research is necessary to identify the composition of the chemicals that trigger EMS and to clarify the pathogenesis of the syndromes. Follow-up revealed that the removal of tryptophan-containing products from the market resulted in the near elimination of reported cases of EMS.

5.2 Design of Case-Control Studies

Case-control study is the most commonly used method of analytical epidemiology. In its implementation, the selection of research objects is crucial. Especially the selection of control group is difficult to master. It is usually required that the control should represent the source population that generated the case.

The case-control study determines whether the subjects are case group or control group according to the status at the beginning of the investigation. This status is considered as the outcome variable of the study. The outcome may be whether the subject has been diagnosed with a certain disease or has experienced a complication. Once outcome status is identified and subjects are categorized as cases or controls. Then, information on exposure to one or several risk factors is then collected retrospectively, usually through interviews or surveys.

5.2.1 Basic Principles

There are three principles of case-control study design. First, it is the study population, also called a source population. The source population may produce the cases and controls. The selection of the control group should not be influenced by exposure factors. Overall, the key issue is for the control group to be representative of the population that generated the cases. The second is de-confounding principle. De-confounding address issues that arise when the exposure of concern is associated

with other possible risk factors. Confounding factors can be eliminated by getting rid of the variability of that factor. For example, if gender is a possible confounding factor, selecting only males would eliminate gender variability altogether. Finally, the principle of comparability was introduced in the two investigation processes. The precision of the exposure measurements was consistent between the control group and the case group. For example, in studies on the effects of smoking on lung cancer, researchers have used nicotine levels in urine to measure smoking in the case group, while questionnaires to measure the controls group, which is inappropriate. Bias due to different measurement methods between cases and controls should be eliminated.

The selection of controls and cases was determined based on the presence or absence of interested disease and could not be influenced by exposure status. Cases and controls do not have to be representative of everyone; in fact, they can be restricted to any specific subgroup, such as elderly, male, or female.

5.2.2 Selection of Cases

Case groups for case-control studies should be representative of all cases in a population. Case selection is based on interested disease and does not have to consider exposure. Cases were available at the beginning of the study. Cases may include new cases, existing cases, and deaths.

New cases are preferred when selecting cases to avoid the influence of survival factors related to the etiology of the disease. Cases found in one clinic or treated by a physician are alternative cases for case-control studies. The source population of cases treated at a clinic is all those who may be seen at that clinic. Reviewing previous studies, many case-control studies were conducted using one or a small group of hospitals or clinics. This will help to obtain cases in a timely manner and increase the possibility of cooperation, thus limiting selection bias. At the same time, however, there may be problems in the definition of the population from who the case originated.

Community-based population disease registries, particularly for cancer and birth defects, are generally considered to be the best source of cases. This is because the population at risk may be clearly defined by geographic or administrative boundaries.

5.2.3 Selection of Controls

The most difficult task in case-control studies is the selection of the control group. The control group should be selected from the population that generated the cases with interested disease. Controls are personals without the disease. A key and difficult aspect of population-based case-control studies is to identify a control group in a more efficient way. Otherwise, it would be necessary to demonstrate

that the population providing the control group had the same exposure distribution as the population that was the source of the cases, a very stringent requirement that can rarely be demonstrated. The control group should be selected independent of their exposure status. There are four types of controls in case-control studies.

5.2.3.1 Population Controls

The best control group ensure that controls are random sample of all noncases in the same population that produced the cases. Another way to ensure that cases and controls are comparable is to draw from the same cohort which is called a nested case-control study. The approach, relative to simply analyzing the data as a cohort study, is that analyses are more efficient.

A control group is selected from the same institution or community. Neighbors or friends were controls, and if these individuals showed results of interest, they would be classified as cases. Selecting a control from a neighbor or friend of the case is also a more feasible method. All households in the area surrounding the case were censored and approached in random order until a suitable control was found. It is important to note that the control was present while the case was being diagnosed. The same difficulty is faced with the use of friend control, i.e., random selection from the census of friends provided in each case. The main advantage of friend control is the low level of non-response.

5.2.3.2 Hospital or Disease Registry Controls

The method of selecting controls from hospitals or clinics is more feasible, but it is hardly representative of the source population. For example, a case-control study investigates the relationship between depression and social and economic factors. A particular clinic may be known to have the best depression specialists in a particular area. If both cases and controls are selected from that clinic, then the depression cases may represent the entire region, while the controls represent only the local neighborhood. Cases and controls may then have different social and economic characteristics. Therefore, cases and controls should be selected from multiple diagnosis and treatment institutions to improve their representativeness.

Controls from a medical practice may be more appropriate than controls from hospitals in an urban health center study. The control may have the same high response level as the case. In the medical practice, they may be interviewed in the hospital, which is an advantage from the perspective of the principle of comparable accuracy. The likelihood of patients going to different hospitals varies. If a patient has the disease being studied, the likelihood of going to a specific hospital will be different from the likelihood of going to that hospital for patients with other diseases. In addition, the exposure may be related to the diseases of some controls. Hospital-based case-control studies generally believe that the disease of the control has not associated with exposure. It is hoped that controls for these diseases will effectively

form the basis of the study in a randomized sample. Because there is little certainty about the independence of exposure and disease diagnosis, the standard recommendation is to select controls with multiple diagnoses to ensure that failure of any of them to meet the criteria will not affect the study. If a diagnosis is found to be related to exposure, these controls can be excluded.

5.2.4 Matching

In case-control studies, matching is a common method to control confounding factors. Matching means that the control group is similar to the case group in some characteristics (such as age and sex).

The goal of matching is to control confounders and increase the efficiency of study. If the factors used for matching are related to exposure, the matched control sample usually has a more case-like exposure distribution than the unmatched control sample. Matching eliminates differences in the distribution of certain confounding factors between cases and controls, thus improving the efficiency of the study. In this way, studies can achieve a strong statistical power with a smaller sample size.

Matching begins with the identification of the case group. The investigator then selects a control group from the source population. Matching is divided into two types, depending on whether it is performed at the individual or group level.

5.2.4.1 Matching Type

Matching can be performed on a group of subjects, which is called group matching, or on a subject-by-subject basis, which is called individual matching.

Group Matching

Group matching means that the matching factors are in the same proportion in the case and control groups and is also referred to as frequency matching. For example, the percentage of women in the case group was 45%, so we chose the control group with 45% women as well. Keeping the control group and case group have the same characteristics (e.g., proportion of male participants). Such that, a group of controls is matched to a group of cases on a particular characteristic (e.g., gender).

Individual Matching

Investigators select a specific control for each case by matching variables. For instance, if the first case enrolled in a study is a 40-year-old black woman, we will

seek a 40-year-old back female control. Each case can be matched with more than one control group. However, the ratio of controls to cases rarely exceeds 4:1, as the higher the ratio the increasing difficulty of implementation.

5.2.4.2 Overmatching

If more variables are matched, it may be difficult to find appropriate controls. And we were unable to explore possible associations of the disease with any of the variables already matched in the cases and controls. In this way, overmatching may happen.

An overmatch is a match that causes a loss of information in the study. There are two types of overmatching. The first type is a match that impairs statistical efficiency, such as a variable related to exposure but not to disease being matched. The second type is a match that impairs validity, such as an intermediate variable between exposure and disease being matched. If the investigator happens to match on a factor that is itself related to the exposure, overmatching will appear. For example, in a particular study of NSAIDs and renal failure, if arthritis symptoms were matched in cases and controls, and arthritis symptoms were usually treated with NSAIDs. Matching for arthritis may then affect NSAIDs. This overmatching can decrease the association between exposure and disease.

5.2.5 Exposure

An important element of case-control studies is to determine the difference in past exposure to a factor between cases and controls. The validity of case-control studies also depends on measuring exposure. In the case-control design, the exposure status of the case is usually investigated after the occurrence of the disease, usually by asking the patient or relatives or friends. The purpose of measuring exposure is to assess the extent of the subject's exposure over a period of time prior to the onset of the disease. The method of collecting exposure data should be the same for cases and controls.

Most case-control studies use questionnaires or interviews to determine the exposure of subjects. The validity of this information will depend in part on the attitude of the subject. People are able to remember well some constant information, such as where they lived in the past and what they did for a living. However, the long-term memory of subjects for specific dietary information may be less reliable. Exposure is sometimes measured by biochemical tests (e.g., calcium in the blood) and may not accurately reflect relevant past exposures if not designed in advance. For instance, lead in the blood of children at age 6 years is not a good indicator of exposure at age 1–2 years. This problem can be avoided if exposure is estimated from established record systems (routine blood tests or stored results from

employment records) or if information is collected prospectively for case-control studies so that exposure data can be collected before disease occurs.

Exposure information can sometimes be determined from historical records. For example, a case-control study on the relationship between sinusitis and multiple sclerosis determined their contact history by searching the general practitioner records of patients and control groups. As long as the records are reasonable and complete, this method is usually more accurate than the method relying on memory.

5.2.6 Sample Size

The sample size was calculated to ensure confidence in the findings and conclusions of the study. Every researcher wants to complete a meaningful scientific study. The estimation of the sample size is a necessary consideration in the study design. Should an applicant receive funding from a funding agency if a sufficient number of subjects are not enrolled in the study, resulting in no chance of finding a statistically significant difference? Most funding agencies are concerned about sample size and power in the studies they support and do not fund studies that would waste limited resources.

There is also a problem with too large sample size. If the number of samples recruited exceeds the required amount, the duration of the study will be extended. Excessive sample size will also affect the quality of the investigation work and increase the burden and cost of research.

Recognize that sample size is essential to ensure scientifically meaningful results and proper management of financial, organizational, material, and human resources. Let's review how to determine statistical capacity and sampling size. Statistical power is calculated with regard to a particular set of hypotheses.

Statistical power is calculated based on a set of assumptions. Epidemiological hypothesis usually compares the observed proportion or ratio with the assumed value. Statistical power refers to the probability that the null hypothesis will be rejected if the specific alternative hypothesis is true. ß denotes the Type II error, i.e., the probability of not rejecting the null hypothesis when the alternative is true. A study should be at least 80% power, and typically studies are designed to have 90–95% power to detect an outcome. What factors affect the power of a study? There are α, β, effect size, variability, and n.

α is the probability of type I error, also known as the significance level of the test hypothesis. This is often determined to be 5% or 1%, implying that the researcher is willing to accept the risk of making a mistake in the alternative hypothesis.

Statistical power is related to effect size, sample size, and significance level. All other factors being equal, an increase in effect size, sample size, or significance level will yield more statistical power.

The sample size of case-control study is calculated according to Formula 5.1.

$$n = \left(\frac{\left[Z_\alpha \sqrt{(1+m)\overline{p}'(1-\overline{p}')} + Z_\beta \sqrt{p_1(1-p_1) + mp_0(1-p_0)} \right]^2}{(p_1 - p_0)^2} \right. \tag{5.1}$$

$$\overline{p}' = \frac{p_1 + p_0/m}{1 + 1/m} p_1 = \frac{p_0 \mathrm{OR}}{1 + p_0(\mathrm{OR} - 1)}$$

Here n is that needed individuals in each group, α = alpha, $\beta = 1 -$ power. OR is the odds ratio which is the ratio of the exposure ratio between cases and controls. "m" is ratio of the sample size of the control group to the sample size of the case group. "p_1"—probability of exposure in case, p_0 can be estimated as prevalence of exposure in the control group.

The formula gives the minimum number of cases needed to detect true odds ratio or case exposure with power and bilateral type *I* error probability α.

Calculation of sample size for individual matched case-control studies.

The estimated case sample size for paired matched case-control studies was calculated according to Eq. 5.2, and the control sample size was $r \times n$.

$$n = [Z_{1-\alpha/2} \sqrt{(1+1/\mathrm{r})\bar{p}(1-\bar{p})} + Z_\beta \sqrt{p_1(1-p_1)/r + p_0(1-p_0)}]^2 / (p_1 - p_0)^2 \tag{5.2}$$

$$p_1 = (\mathrm{OR} \times P_0)/(1 - P_0 + \mathrm{OR} \times P_0)$$

$$\overline{P} = (P_1 + rP_0)/(1 + r)$$

Where α = alpha, $\beta = 1 -$ power, P_1, P_0 denote the estimated exposure rates of the case and control groups in the target population, respectively.

5.3 Data Collection and Analysis

When researchers have determined the outcomes (disease or health status) of interest in the case-control study and the factors to be studied, they can develop methods for collecting information. The data should include information about research outcomes and factors. Data analysis involves two parts job. First is descriptive data. Next is statistical inference and measure of association. The odds ratio represents an indicator of the association between the disease and each factor of interest.

Researchers often consider data analysis to be the most enjoyable part of epidemiological research. Because after all the hard work and waiting, they have a chance to gain answers. The basic method of analysis in case-control studies is to compare the proportion of exposure in the case and control groups and to calculate the OR.

5.3.1 Main Analysis Objectives

Assess and refine data quality. Describe the study population and its relationship to the target population. Assess potential bias. Estimate the frequency of exposure. Estimate the strength of the association between exposure factors and disease.

A quality data analysis consists of three phases. In the first stage, the analyst should review the recorded data for accuracy and completeness. Next, the analyst should summarize the data in a concise form and perform descriptive analyses, such as classifying observations according to key factors, using a contingency table. Finally, the summarized data are used to estimate epidemiologic measures of interest, usually expressed in terms of strength of association with appropriate confidence intervals.

5.3.2 Descriptive Analysis

The number of study subjects and the composition of the various characteristics are described. The exploration of the data reports the frequencies. These measures will provide the basis for important subgroups. Standardization or other adjustment procedures may be required to account for differences in age and other risk factor distributions, duration of follow-up, etc. Compare whether certain basic characteristics are similar between case and control groups.

5.3.3 Statistical Inference

The indicator that indicates the strength of the association between disease and exposure in case-control studies is the odds ratio (OR). Data analysis included calculating odds ratios as a measure of the association between the disease and the interested factors. When analyzing data on the relationship between exposure and disease variables, we usually have to make statistical inferences about relationship. Several means were employed to avoid random errors, such as p-value and confidence interval (CI) tests. But we should understand that the role of statistically significant is limited. Statistical significance is usually based on the *P*-value: depending on whether the *P*-value is less than or greater than the critical value, usually 0.05. The critical value is then referred to as the alpha level of the test, and the result is considered "significant" or "insignificant."

The type of analysis used in case-control studies depends on whether controls are sampled in an unmatched or matched manner. Different analysis methods are used for different matching methods.

5.3.3.1 Unmatched (Frequency Matching) Design

In case-control studies, researchers attempt to assess the strength of the association between disease and study factors. The investigators analyzed the proportion of exposure in the case and control groups. Data from unmatching or frequency matching case-control studies are summarized in Table 5.1. For better understanding, only two levels of exposure are discussed here. Each object can be divided into four basic cells, which are defined by disease and prior exposure status.

A simple unmatched case-control study, such as that in Table 5.1, can be analyzed by using OR (odds ratio) for association. In case-control studies, groupings are made according to the presence or absence of disease. Therefore, we can't measure health outcomes or disease incidence rate. The proportion of persons in the study who have the disease is no longer determined by risk of developing the disease, but rather by the choice of investigator. So, investigators could not calculate RR (relative risk). Investigators can obtain valid estimates of risk ratios by using OR. When the disease interested is a rare disease, the odds ratio approximates the risk ratio or RR. However, this is not always the case, researcher should be careful taken to interpret the odds ratio appropriately.

χ^2 test and statistical inference (formula 5.3)

$$\chi^2 = \frac{\left(|ad - bc| - \frac{N}{2}\right)^2 N}{n_1 n_0 m_1 m_0} \tag{5.3}$$

Odds Ratio

The odds ratio (OR) is an index of the association between exposure and disease or outcome. The odds ratio is the ratio of exposure in the case group divided by the ratio of exposure in the control group. With the notation in Table 5.1, the odds of exposure for case represent the probability that a case was exposed divided by the probability that a case was not exposed. The odds are estimated by the following formula.

$$\text{Odds of case exposure} = \frac{\text{Exposed cases}}{\text{All cases}} \Big/ \frac{\text{Unexposed cases}}{\text{All cases}} = \frac{a}{a+b} \Big/ \frac{b}{a+b} = \frac{a}{b}$$

Similarly, the odds of exposure among controls are estimated by the following formula:

Table 5.1 The result of case-control study

	Case	Control	Total
Exposed	a	b	$a + b$ (m_1)
Unexposed	c	d	$c + d$ (m_0)
Total	$a + c$ (n_1)	$b + d$ (n_0)	$a + b + c + d$ (n)

$$\text{Odds of control exposure} = \frac{c}{d}$$

The odds of exposure for cases divided by the odds of exposure for the controls are expressed as the OR. Substituting from the preceding equations, the OR is estimated by formula 5.4

$$\text{OR} = \frac{\text{odds of case exposure}}{\text{odds of control exposure}} = \frac{a}{b} / \frac{c}{d} = \frac{a \times d}{c \times b} \tag{5.4}$$

OR indicated "How many times more exposed are cases than no-case exposed?" Since OR have a different scale of measurement than RR, the answer to this question can sometimes differ from the answer to the corresponding question about RR. However, case-control studies are concerned with rare diseases, for which RR and OR are very similar.

Interpreting the Odds Ratio

A case-control study comparing the smoking habits of 58 lung cancer cases with 93 controls showed the following results (Table 5.2).

$$\text{OR} = \frac{a \times d}{b \times c} = \frac{22 \times 86}{7 \times 36} = 7.5$$

The proportion of lung cancer cases exposed to smoking was 7.5 times greater than the proportion of controls who smoked. It is suggested that there is a strong association between lung cancer and smoking. Smoking could thus be a factor that increases the probability of having lung cancer.

As can be seen, we can determine the risk factors by calculating the OR. It is important to recognize that case-control studies are comparing the odds of exposure $[(a/c)/(b/d)]$ between cases and controls. Conceptually, this is very different from comparing the odds of illness $[(a/b)/(c/d)]$ between exposed and unexposed individuals, which is the result we are really interested in.

Fortunately, in rare disease studies, the ratio $[(a/c)/(b/d)]$ of the ratio of cases and controls with exposure is equal to ad/bc. It can also be seen that the odds ratio $[(a/b)/(c/d)]$ in favor of disease in exposed and unexposed populations is also equal to ad/bc.

Table 5.2 Results of a case-control study of lung cancer and smoking

	Individuals with lung cancer (cases)	Individuals without lung cancer (controls)
Smokers	22 (*a*)	7 (*b*)
Nonsmokers	36 (*c*)	86 (*d*)
Total	58	93

Table 5.3 Study on the association between obesity and eating vegetables

	Obese individuals (Cases)	Non-obese individuals (controls)
Eat vegetables	121	171
Do not eat vegetables	129	79
Total	250	250

Table 5.4 Results of a study on depression and eating vegetables

	Individuals with depression (cases)	Individuals without depression (controls)
Eat vegetables	80	80
Do not eat vegetables	120	120
Total	200	200

Sometimes, the factors studied would reduce the probability of developing the disease. Such factors are known as protective factors of the disease. For instance, 250 obese individuals (cases) in a case-control study were compared to 250 non-obese individuals (controls) in terms of vegetable consumption in their diet. The results are shown below (Table 5.3).

$$\mathrm{OR} = \frac{\mathrm{a} \times \mathrm{d}}{\mathrm{b} \times \mathrm{c}} = \frac{121 \times 79}{129 \times 171} = 0.43$$

The proportion of cases eating vegetables was 0.43 times greater than the proportion exposed in the control group. Therefore, the proportion of eating vegetables in the case group was 48% lower than the exposure proportion in the control group was 68%. The results of the case-control study showed that compared with the control group, the case group were less likely to eat vegetables. Eating vegetables may be a protective factor in reducing obesity.

Sometimes case-control studies did not find an association between study factors and outcomes. In this case, the OR for the strength of the association between factors and disease in the case-control study was 1.0. For example, in a case-control study, 200 people with depression were compared with 200 people without depression regarding their vegetable consumption (Table 5.4).

$$\mathrm{OR} = \frac{a \times d}{b \times c} = \frac{80 \times 120}{80 \times 120} = 1.00$$

The odds of eating vegetables among depressed patients were the same as the odds in the control group. An OR of 1.00 was calculated, indicating a lack of association between depression and eating vegetables. The results of the study did not show an association between eating vegetables and suffering from depression.

In summary, OR $>$ 1 indicates that the factor may increase the risk of disease, OR $<$ 1 indicates that the factor may attenuate the risk of disease, and OR $=$ 1 indicates no association.

Confidence Interval Estimation of Odds Ratio

An OR is a point value estimate, which may have a random error. The OR confidence interval gives the range of estimates of the OR. The range of estimates is calculated based on a given set of sample data. The OR confidence interval reduces the random error generated by a single study. Ninety-five percent confidence interval (CI) means a 95% probability which the interval includes the true OR. If 95% CI range includes "1," it is not statistically significant since it could be either a risk factor (OR≧1) or a protective factor (OR ≦ 1). If 95% CI range is greater than 1, the exposure is a significant risk factor (OR ≧ 1) with a probability of higher than 95%.

An approximate 95% CI around the point estimate of OR for an unmatched case-control study can be calculated using the formula (5.5).

$$OR95\%\mathrm{CI} = (\mathrm{OR})\exp\left[\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right] \tag{5.5}$$

Where exp. is the natural logarithm, and a, b, c, and d represent the numerical entries into the summary format in Table 5.1.

$$95\%CI = (7.5)\exp\left[\pm 1.96\sqrt{\frac{1}{22}+\frac{1}{36}+\frac{1}{7}+\frac{1}{86}}\right] = (7.5)\exp(\pm 1.96\times 0.477)$$

$$\text{Lower bound} = (7.5)\exp(-1.96\times 0.477) = (7.5)\ \exp(-0.94) = 2.9$$

$$\text{Up bound} = (7.5)\exp(+1.96\times 0.477) = (7.5)\exp(+0.94) = 19.1$$

The CI provides two values, low (L) and high (U), with a specific confidence level between these two values for the population parameter. A 95% confidence interval means that if we conduct a study, there is a 95% probability that the results will fall within the confidence interval. The above example illustrates that the interval between 2.9 and 19.1 includes a probability of 0.95 for the true OR value.

5.3.3.2 Matched Design

In individually matched case-control studies, the analysis must take into account the matched sampling scheme. When a control is matched to one case, summary data in the format shown in Table 5.5 can appear. This table is different from the one that we

Table 5.5 A 1:1 matched case-control study

	Control exposed	Control unexposed	Total
Case exposed	*a*	*c*	*a* + *c*
Case unexposed	*b*	*d*	*b* + *d*
Total	*a* + *b*	*c* + *d*	*a* + *b* + *c* + *d*

introduced in our previous group matching analysis. Each cell in Table 5.5 represents not one subject but a pair (one case and one control). Each case-control pair can be classified as one of the exposure states. Just as Table 5.5, "*a*" means numbers of pairs that both case and control exposed while "*c*" means numbers of pairs that case exposed but control unexposed. "*b*" means numbers of pairs that case unexposed but control exposed. "*d*" means number of pairs that both case and control unexposed. In the analysis of individual matching studies, only pairs with inconsistent exposure were used. Inconsistent pairs of exposures occur when the exposure status of the case differs from that of the control group.

2 × 2 Table

χ^2 Test and Statistical Inference

$$\chi^2 = \frac{(|b-c|-0.5)^2}{b+c}$$

OR and 95%CI

The OR of individual matched case-control study is calculated by simple ratio.

$$\mathrm{OR} = \frac{c}{b} \tag{5.6}$$

$$OR95\%\mathrm{CI} = (\mathrm{OR})\exp\left[\pm 1.96\sqrt{\frac{1}{b}+\frac{1}{c}}\right] \tag{5.7}$$

The significance of individual matching OR is the same as that of group matching case-control study. Endometrial cancer and estrogen are used as examples to illustrate the procedure for calculating OR in individual matched case-control studies. The 390 pairs consisted of 390 patients with endometrial cancer and 390 controls,

Table 5.6 Hypothetical matched case-control study

	Control exposed	Control unexposed	Total
Case exposed	84	96	180
Case unexposed	48	162	210
Total	132	258	390

Table 5.6. Exposure is defined as women who have ever taken any estrogen. The OR form the study is as below.

$$\text{OR} = \frac{c}{b} = \frac{96}{48} = 2.00 \qquad OR95\%\text{CI} = \ (1.40 \text{ to } 2.89)$$

The calculation method of OR 95% confidence interval (CI) of individual matched case-control study is the same as that of group matched case-control study. Formula 5.7 gives the formula for calculating the OR 95% confidence interval of individual matched case-control study. The approximate 95% CI for the OR is 1.40 to 2.89. This individually matched case-control study showed a moderate association between endometrial cancer and estrogen use.

5.4 Common Bias and Controlling

A case-control study is an observational study in which subjects are enrolled based on the presence or absence of the disease of interest. The exposure history of both groups is then evaluated to determine the strength of the association between disease and exposure. Case-control studies are susceptible to observational epidemiological study bias. These biases include selection, information, or confounding biases.

5.4.1 Selection Bias

Selection bias is the most common bias in case-control studies. Selection bias may exist if the control group is not from the source population that generated the cases. For example, to study asthma, cases of asthma are drawn from high school students, while people without asthma are drawn from the elderly population to form a control group. The fact that the control and case groups are not a source population has the potential to introduce serious bias. The factors that cause asthma may be different in younger and older people. Thus, based on studies of such mismatched cases and controls, many of the factors that may be found to be associated with asthma may simply be due to the different ages of the two populations.

Sampling of controls and cases can sometimes be stratified, e.g., by sex and age group. In addition to this, there should be randomization in subgroups of subjects

with and without disease. However, researchers are often not randomly sampled, and selection bias arises. This bias poses a significant impact on the validity of case-control studies.

Bias does occur when the sampling fractions depend jointly on exposure and disease, usually because exposed controls are more or less likely to be sampled than non-exposed controls. When hospital patients are utilized as cases and controls, the control is not a random sample of the target population because the control is a subset of hospital patients. Cases in case group are only part patients in the hospital. Patients and hospitals are mutually selective. The systematic differences in some characteristics between the case group and the control group are unavoidable, resulting in an admission rate bias. This is also known as Berkson bias.

The following factors contribute to selection bias.

5.4.1.1 Prevalence-Incidence Bias

More information might have been obtained if the survey respondents had chosen existing cases, but much of this information was only relevant to survival and may have overestimated the etiologic role of certain exposure factors. In addition, survivors of a disease change their habits so as to reduce the level of a risk factor or distort their pre-morbid habits when they are investigated, resulting in the association of a factor with the disease being incorrectly estimated. This type of bias is usually referred to as prevalence-incidence bias. Therefore, new cases should be included in the investigation as much as possible to avoid the effect of prevalence-incidence bias.

5.4.1.2 Unmasking Bias

Patients often seek medical attention for certain symptoms unrelated to the causative agent, thereby increasing the detection rate of early cases and leading to an overestimation of exposure. This systematic error is then referred to as unmasking bias.

5.4.1.3 Subject Refuses Participation

In case-control studies, the most common reason is that subjects refuse to participate, either by actively refusing to sign a consent form or by passively not returning questionnaires or failing to attend laboratory tests at the specified time. Cases tended to be highly motivated to participate, while controls selected from the population were not willing to participate. Participation rates in the control group tended to depend on a number of factors related. For example, rejection rates for telephone surveys are higher for people who are older, less socially connected, less educated, and have lower incomes.

5.4.2 Information Bias

Information bias is a systematic bias in the process of collecting and organizing information due to flaws in the methods used to measure exposure and outcome. Even if the classification of subjects' exposure and outcome is completely accurate, bias may be introduced due to different choices in case-control studies. More commonly, subjects are incorrectly classified in terms of exposure status or outcome, and estimates of association can be biased. These errors are often referred to as *misclassification*. Misclassification can be classified as differential misclassification or non-differential misclassification.

Differential misclassification is also referred to as "recall bias." Recall bias may arise when cases remember past exposures more completely than controls. This often happens because cases tend to try to find out the cause of their disease. As a result, when they are interviewed, they tend to report more information about the past. Control do not deliberately report information about past exposures.

The second type of information bias is non-differentiated misclassification. Non-differential error classification means that the frequency of errors is similar in the case and control groups. Misclassification of exposure status is more serious than misclassification of outcome. However, both misclassifications can bias a study. For example, a case-control study was conducted to explore the relationship between a high-fat diet and coronary artery disease. Subjects with heart disease and controls without heart disease were recruited and asked to fill out a questionnaire about their dietary habits. Then they were determined whether to consume a high fat diet. It is difficult to accurately assess the amount of fat in the diet from questionnaires. Therefore, it would not be surprising if there were errors in the classification of exposure. In such cases, misclassification may occur regardless of the final disease status. When exposed is qualitative variables, non-differential misclassification always favors the null. Or, if there is an association, whether positive or negative, it tends to minimize it. For example, the OR between a high-fat diet and coronary heart disease is 5.0, but a biased estimate might give an OR is 2.4 if about 20% of exposed subjects are misclassified as "non-exposed" in both disease and control. This implies that the bias tends towards the null.

If there are multiple exposure levels, non-differential misclassification may bias the estimate toward or away from the null, which rely on the category to which the subject was misclassified.

5.4.3 Confounding Bias

Confounding is that the relationship between exposure factors and outcomes is distorted by external variables. The systematic error generated by this distortion is the confounding bias. Confounding factors usually have three characteristics. One is a variable associated with the exposure and independent of that exposure, and the

third is a risk factor for the disease. The distortion introduced by confounding factors can be significant, and it can even change the direction of the effect. However, confounding bias can be adjusted for in the analysis, which is different from selection and information bias. For example, the crude death rate in city A may be higher than the crude death rate in city B, but after adjusting for age, there is no difference in the adjusted death rate between cities A and B. The age-induced deviation in crude death rates in two cities is known as the confounding bias.

There are two strategies for controlling confounding. Prevent confounding bias from occurring in the first place, which can be done by limiting or matching during the study design phase. Next is to deal with it when it occurs by using analytic techniques such as stratification and statistical model. The effectiveness of all of these strategies except randomization depends on the ability to identify and measure any confounders accurately.

5.5 Strengths and Weaknesses of Case-Control Studies

5.5.1 Advantage of the Case-Control Study

Case-control studies save time, cost less, and are the most effective design. Case-control studies are the preferred choice for rare disease research. This is because in a cohort design, studies of rare diseases must follow many people to identify those with outcomes. Case-control studies, on the other hand, do not have to worry about no outcomes occurring. Case-control studies are also advantageous in studying diseases with longer latency periods.

In addition, case-control studies have several other advantages. First, occurrence of exposure in subjects retrospectively investigated in case-control studies. Investigators do not have to follow study subjects over time as in cohort studies. Investigators do not have to follow study subjects over time to collect exposure and disease information as they do in cohort studies. Finally, the sample size of the case-control study was small. Compared to cohort studies and experimental studies, case-control studies are easier to implement. (Table 5.7).

Table 5.7 Advantages and disadvantage of case-control studies

Advantages	Disadvantages
Suitable for research on rare diseases	The relative risk of disease cannot be directly estimated
Suitable for long latency chronic disease studies	Not suitable for studying rare exposures
Smaller sample size required compared to other types of studies	More susceptible to selection bias than alternative designs
Less expensive than alternative designs	Information on exposure may be less accurate than that available in alternative designs.
Save time over other types of study designs	

5.5.2 Disadvantage of the Case-Control Study

Case-control studies are divided into case and control groups according to the presence or absence of the disease of interest. Therefore, incidence rates could not be calculated for either group. Without knowing the incidence, it is not possible to calculate the relative risk in case-control studies. One can calculate the OR in a case-control study, which is a measure of association that approximates relative risk under certain condition.

The temporal sequence of exposure and disease may be difficult to determine in a case-control study, so it may not be possible to know whether the exposure occurred before the disease. For example, A case-control study of asthma in high school students suggests an association between asthma and cat ownership. However, it may be difficult to know whether high school students had cats first or whether they had asthma attacks first. People usually choose newly diagnosed cases to overcome this drawback.

Although case-control studies have advantages in studying rare diseases, they are not suitable for studying rare exposures (Table 5.7). For example, we would like to study the risk of asthma associated with working in a nuclear submarine shipyard and would probably not prefer a case-control study because only a small percentage of people with asthma would be exposed to this environmental factor.

Case-control studies are grouped by study disease, so they can only be used to study one disease. However, it is possible to study the association between a disease and multiple factors. If want to study more than one disease, you can consider a cohort study design.

In conclusion, case-control studies are a more efficient research method, but the results are susceptible to the influence of known and unknown confounding variables. Case-control studies are suitable for investigating the association between diseases and factors, and the etiology of diseases. When there is limited evidence on a topic, there are cost-effective ways to raise and investigate hypotheses before conducting larger and more expensive studies. Sometime, they are often the only choice of research method, especially when cohort studies or randomized controlled trials are impractical. Case-control studies investigated information about each subject's exposure up to a certain time period. Case-control studies require first defining the case, then identifying the source population that generated the case, and finally identifying the case group and control group. The studies have some strong characteristics such as being cheap, efficient.