9

CASE-CONTROL STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Describe the major design features of risk-based and rate-based case-control studies.
2. Identify hypotheses and population types that are consistent with risk-based case-control studies.
3. Identify hypotheses and population types that are consistent with rate-based case-control studies.
4. Differentiate between open and closed primary-base and secondary-base case-control studies.
5. Elaborate the principles used to select and define the case series.
6. Explain the principal features for selecting controls in open and closed primary-base case-control studies.
7. Explain the principal features for selecting controls in open secondary-base case-control studies.
8. Design and implement a valid case-control study to meet specific study objectives.

9.1 Introduction

The basis of the case-control study design is to select a number of individuals that have (newly) developed the disease or outcome of interest (the cases) and, as a comparison, a number that have not developed the disease or outcome of interest, at the time of selection (the **controls***)*. We then contrast the frequency of the exposure factors in the cases with the frequency of exposure factors in the controls (Schulz and Grimes, 2002). It is important to stress that a case-control study is not a comparison between a set of cases (*ie* people with a specific disease) and a set of 'healthy' subjects, but between a set of cases and a set of non-case subjects (*ie* people who have not developed the specific disease but may have other diseases) whose exposure to the factors of interest reflects the exposure in the source population (Grimes and Schulz, 2005). In this process, we assume that the controls would be included as cases if they had developed the outcome (disease) of interest. Overviews of key case-control design issues are available (Schulz and Grimes, 2002; Thrift, 2010; van Stralen *et al*, 2010). Most frequently individual people are the units of interest, but the design also applies to aggregates of individuals. For generality, the unit of interest here is denoted as a study subject. In most instances, the outcome of interest is a specific disease, or mortality from a specific cause; however, a variety of outcomes such as quality of life measures can be studied in a case-control format.

Usually case-control studies are performed retrospectively since the outcome (usually disease) has occurred when the study begins. It is possible to conduct case-control studies prospectively; in these, the cases have not developed until after the study begins so the cases are enrolled in the study over time (see Example 9.1). Focused introductions to the use and design of case-control studies in a variety of disciplines are given by Caro and Huybrechts (2009) (pharmaco-economic focus), Fletcher (2010) (ophthalmology), Levin (2006) (dental), Mayo and Goldberg (2009a, 2009b) (rehabilitation science), and Busse and Obremskey (2009) (orthopedic). Singh and Mahmud (2009) describe traditional case-control designs and their variants used in cancer research. Goldberg (2010) describes the application of case-control studies in environmental research. D'Agata (2005) discusses some limitations (and solutions) when case-control studies are used to study risk factors for developing or having multiple resistant bacteria as outcomes. Walter (2003) discusses the application of case-control studies to assessing screening programs (specifically mammographic screening). Sun *et al* (2010) describe multiple events (more than one case series) case-control studies; this design will not be pursued further here.

9.2 The study base

We have previously defined the target and source populations. Here we define a new grouping termed the **study base** which is the population from which the cases and possibly the controls are obtained. If the study base is a well-defined **source population** for which there is, or could be, an explicit listing of sampling units (*ie* potential study subjects), this population is denoted as a **primary study base** or **primary base** (see Example 9.2).

If the study base is one or more steps removed from the actual source population, such as people who are members, or patients, at a referral clinic, laboratory or central registry, the source population is referred to as a **secondary study base** or **secondary base** (see Example 9.3).

In describing the source population in a case-control study, the term **nested** implies that the entire source population which gives rise to the cases has been, or could be, enumerated (see

Example 9.1 A prospective risk-based case-control study of serum estradiol and testosterone as risk factors for breast cancer

This secondary-base case-control study used serum samples came from the Columbia, MO, Serum Bank (Dorgan *et al*, 2010). Participants were volunteers identified primarily through the Breast Cancer Detection Demonstration Project (BCDDP) at the University of Missouri Hospital and Ellis Fischel Cancer Center in Columbia. A total of 6,915 women who were free of cancer, other than non-melanoma skin cancer, donated blood to the bank on one or more occasions between December 1977 and June 1989. Serum specimens and clinical data were collected. Cancer diagnoses were ascertained by self-report and by searching the Missouri Cancer Registry, BCDDP Cohort files, and National Death Index (NDI). Of the 6,720 women included in the extended follow-up, 91% were located, 1,751 of these were identified as deceased with confirmation of causes of death provided via NDI.

Cases were diagnosed with confirmed in situ or invasive breast cancer. They were excluded if they were taking exogenous estrogens or progestins at the time they donated blood. For each of the 117 potential cases, 2 potential controls who were alive at the age of the case diagnosis and who remained cancer free up until the end of the study period were randomly selected; this is a risk-based sampling strategy. Controls were matched to the case on age (±2 years), date (±1 year), and menstrual cycle day (±2 days). After exclusion of postmenopausal women, 98 cases and 168 matched controls remained.

The association of serum estradiol and testosterone with breast cancer risk was evaluated using conditional logistic regression. Geometric mean concentrations for each hormone in cases and controls were calculated and compared by testing the statistical significance of the trend of the log-transformed variable entered as a continuous term in conditional logistic regression models.

Example 9.4). In a nested study, a subsample of this population forms the study group with the case series being all, or a known fraction of, the cases from this source and the controls being a subsample of the non-cases (those not having developed the specific outcome of interest) in that source population. Since the sampling fractions of cases and controls can be obtained, this allows us to estimate the frequency of disease by exposure status, a feature that is absent in almost all other types of case-control study.

Regardless of whether the study is nested, selecting the control subjects from the same defined population (study base) as the cases, helps to prevent selection bias (Hak *et al*, 2004). Rundle *et al* (2005) demonstrate that the nested design is better than a case-cohort design (see Chapter 10) if there is a need to collect and analyse biological specimens to determine exposure. The key issues they identify include: accounting for the effects of analytic batch, of long-term storage, and of freeze-thaw cycles on biomarkers. Whether or not the study is truly nested in an explicitly definable population, it is useful to think of all case-control designs in this context because it aids in the valid selection of control subjects. Wacholder (2009) defends the validity of nested designs in response to criticisms of their merits, and discusses and restresses the need for appropriate sampling of controls to prevent bias. Wacholder notes, "As long as exposure is measured only up to the time of the event, the particular choice of exposure summary cannot introduce bias in comparing cases and controls." Baksh *et al* (2005) discuss the design and analysis of sequential case-control studies; these can be advantageous to minimise the costs of testing biological specimens, but will not be pursued further here.

Example 9.2 A primary-base case-control study

This rate-based study on risk factors for *Salmonella* Typhimurium was conducted in the provinces of Alberta, British Columbia, Ontario and Saskatchewan, Canada and covered the study period between December 1, 1999 and November 30, 2000 (Dore *et al*, 2004). Eligible cases were individuals with diarrheal illness who had *S.* Typhimurium isolated from stool samples. Cases were excluded if their primary residence was outside the study province. The questionnaire usually was completed within 30 days after onset of diarrhea. Controls, matched 1:1 on the cases' age and province of residence, were randomly selected from provincial Ministry of Health registered persons databases (*S.* Typhimurium is a laboratory notifiable disease in Canada, and all Canadians are registered in the healthcare system). Potential controls were contacted by telephone within 7 days of the matched case interview. If the initial control was excluded, the next eligible name on the client registry list was selected. In total, 138 case-control pairs met the criteria for the *S.* Typhimurium DT104 risk-factor analyses and 258 pairs for the *S.* Typhimurium non-DT104 analyses. Overall 61% of cases participated; the percentage of potential controls participating was not specified. Cases and controls were interviewed by telephone using a pre-tested, standardised questionnaire adapted from similar, previously validated, survey instruments. Data collected included demographics; health history including previous medication use; recent travel history, and animal contact. Matched bivariable analyses were performed for each potential risk factor using McNemar's test for dichotomous variables and paired t-tests for continuous variables. Final analysis was conducted using a conditional multivariable logistic regression model.

Variations in the design of a case-control study, especially the manner of selecting controls, are necessary depending on whether one is conducting the study in an open- or closed-source population. As noted in Section 4.4.1, a **closed** source population refers to a population whose members are 'fixed'; no new subjects are added to the initial population, and in this instance all subjects are observed for the full risk period of the outcome. In a 'dynamic' or **open** population study subjects can leave the study part-way through the risk period for the outcome of interest and new subjects can enter during the risk period. A population is said to be stable if its characteristics (*eg* average age, weight, and sex distribution *etc*), including the level of exposure, do not change over time. Closed populations are not stable, especially when the follow-up period is long (*eg* the subjects age over the study period).

As with cohort studies, we can use simpler risk-based case-control designs in closed

Example 9.3 Retrospective, secondary-base case-control study of hypercholesterolemia as risk factor for prostate cancer

In this risk-based study, cases were patients newly diagnosed with prostate cancer between 2004 and 2006 at Meritcare hospital (Magura *et al*, 2008). Controls were identified from the primary-care database of the same hospital. Data on personal characteristics, family history, prostate-specific antigen (PSA) level, co-morbidities, and multivitamin use were abstracted. Covariate information was obtained for the period prior to diagnosis for cases and prior to examination for controls. Inclusion criteria for cases were men with incident, histologically confirmed prostate cancer as a primary site with cancer diagnosed between 2004 and 2006, aged between 50 and 74, and having a lipid profile test within a year prior to the diagnosis. Exclusion criteria included a diagnosis of any other cancer and race other than Caucasian. Inclusion criteria for controls were men without cancer who had an annual physical exam at the same hospital between 2004 and 2006, aged between 50 and 74, and lipid profile tests within a year of their annual physical exam. The authors used a widely accepted definition of hypercholesterolemia (total cholesterol >5.17 (mmol/l). Odds ratios (*OR*) and 95% confidence intervals (CI) were estimated using multiple logistic regression to control confounding. All two-way interactions involving hypercholesterolemia were assessed.

Example 9.4 A nested, rate-based, case-control study of risk factors for gastroenteritis

This community-based (*ie* primary-base), case-control study was nested within a larger randomised controlled trial conducted in South Australia (Rodrigo *et al*, 2011). For this trial, 300 households having no less than four members with at least two aged 1–15 years and using untreated rainwater as their usual drinking-water source were recruited. Participating households received either an active or a sham water-treatment unit (WTU) for filtering of all water intended for drinking or cooking. An adult reporting participant ensured that a health diary was completed weekly for each participant. The outcome—highly credible gastroenteritis (HCG)—was defined as any of the following in a 24-h period: two or more loose stools; two or more episodes of vomiting; one loose stool together with abdominal pain or nausea or vomiting; or one episode of vomiting with abdominal pain or nausea. A person could be a case on more than one occasion during the study period. Controls were matched with cases according to the study week in which illness was experienced by the case. All people in the cohort without symptoms of gastroenteritis in the week of or the week prior to onset of illness of a case were eligible to be selected as the control for that case. One control was selected randomly for each case.

A total of 769 episodes of HCG from 501 individuals were identified from the health diaries. Participants were successfully interviewed for 298 (36.5%) episodes; cross-checking revealed that 281 of these were valid HCG cases. The 281 HCG cases occurred in 215 persons: 171 persons had 1 HCG episode; 31 had 2 episodes; and 13 had 3 or more episodes. Of the 297 control interviews, 35 individuals were interviewed twice and 3 individuals were interviewed 3 times. A total of 51 people who were a HCG case at some point in time during the study were interviewed as controls for other cases.

Data collection was by diary and by telephone interview. Cases and controls completed a structured questionnaire that addressed risk factors in the 7-day period before onset of symptoms. Potential risk factors investigated were consumption of chicken, beef, organ meat (offal), fish, shellfish, raw vegetables, salad, fresh fruit, rice, milk products including cheese, eggs (runny, cooked, or raw), and take-away fast-food from any source. Other factors assessed were the presence of a child in diapers; changing and/or washing diapers; and contact with pets through having animals in the household or other sources (farm, zoo). Logistic regression was used to study univariable associations of potential risk factors with HCG and robust standard errors were calculated to allow for familial clustering and repeated observations on individuals (as cases or controls on multiple occasions). Food exposures were adjusted for age, gender, location and treatment group (active or sham). The population attributable fractions for factors with statistically significant odds ratios >1 were calculated.

populations, whereas open populations require a rate-based design. Populations are more likely to be closed if the risk period for the outcome is of limited duration such as occurs in many food-borne outbreaks. Sometimes, for research purposes (*eg* a study to identify risk factors for postpartum diseases in women), it is possible to convert an open population (the population of pregnant women eligible to give birth in a defined place or clinic in a year) to a closed population by following a sample, or all of, the group of women who actually gave birth in the year for the first hundred days after their date of giving birth to identify postpartum health issues and their causes.

9.3 THE CASE SERIES

Key elements in selecting the case series include: specifying the disease (including the required diagnostic criteria for the outcome); identifying the source(s) of the cases; deciding whether only incident or both incident and prevalent cases are to be included; and estimating the

required number of cases and total sample size. Edwards (2001) provides formulae for estimating sample size with 1:1 and 1:m controls, with and without matching). Peng *et al* (2010) describe sample size estimation in genetic studies. The issue of selecting incident versus prevalent cases seems fairly clear as there is virtually unanimous agreement that, when possible, only incident cases should be used for the study. There are specific circumstances in which the inclusion of prevalent cases may be justified, but this would be the exception, not the rule. The problems that arise from using prevalent cases have been discussed in Chapter 7. Usually, only the first occurrence of the outcome in each study subject is included in the case series (see Examples 9.1 and 9.3; however, multiple occurrences of the same disease can be included (see Example 9.4)

A major decision is whether the cases will be derived from a primary base such as a specific registry that contains virtually all cases for a defined population (*eg* provincial or state disease registries), or if the cases will be obtained from a secondary base such as a physician's clinic, or one or more hospitals. Sampling, or taking a census of cases directly from the primary source population, has the advantage that it avoids a number of potential selection biases, but it may be more difficult to implement and more costly than using a secondary base. In a primary-base study, every effort should be made to obtain complete case ascertainment. Primary-base designs are moderately common because provincial or state records allow complete enumeration of people and their health events. In secondary-base studies, (see Example 9.3) a major challenge is to conceptualise the actual source population from which the cases arose so that the selection process for controls can help ensure that the controls arise from the same source. In essence, we would like to select controls from that group of subjects that would have gone to the secondary source of the cases had they developed the disease of interest; as noted, this population is often difficult to define. A common solution is to select the controls from records at the same source (*eg* hospital; see Example 9.3) as the cases.

As explained in Example 9.3, the diagnostic criteria for a subject to become a case should include specific, well-defined manifestational (*ie* clinical) signs where appropriate and, when possible, clearly documented diagnostic criteria (*eg* laboratory test results) that can be applied to all study subjects in a uniform manner. Some care is needed in imposing detailed diagnostic criteria for the cases because the set of cases from a tertiary care facility could become increasingly different from the majority of cases of that disease in the source population, if high cost or time commitment is required to complete the diagnostic work-up. Thus, the nature of a case series of autoimmune disease obtained from a referral hospital might differ from the majority of autoimmune cases seen in private practice. Nonetheless, there is merit in using a set of very specific diagnostic criteria for the cases because preventing false positives will reduce any bias in the measure of association caused by lack of sensitivity in the detection of cases (Orenstein *et al* (2007); see also Chapter 12). In some instances, it might be desirable to subdivide the case series into one or more subgroups based on differences in the disease characteristics, especially if the causes of the different forms of the disease might differ.

9.3.1 Case-control studies with continuous outcomes

All case-control studies are based on outcome-dependent sampling. Typically, the outcome has a dichotomous (diseased/not diseased, or yes/no) scale and the distribution of risk factors in the two groups can be analysed with a logistic or survival model. When the outcome is naturally expressed or measured on a continuous scale (*eg* birth weight), researchers might randomly, or purposively, select study subjects at the low and high extremes of the outcome distribution and

then compare the level, or presence, of exposure in these groups. This allows valid analyses using logistic models, but discards information about the outcome on the continuous scale. If we desire to use the original continuous outcome, using linear models, special techniques must be used to account for the sampling structure (Jiang *et al*, 2009; Zhou *et al*, 2007). Such models also can convert the outcome from the linear scale to a logistic-like approach. These techniques will not be pursued here. Suffice it to say that, if outcome dependent sampling is used and the outcome is measured on a continuous scale, then the usual linear models (Chapter 14) cannot be used to analyse the data.

9.4 Principles of control selection

The selection of appropriate controls is often one of the most difficult aspects of a case-control design. The key guideline for valid control selection is that controls should be representative of the exposure experience in the population which gave rise to the cases. Also, controls should be subjects who would have been cases if the outcome occurred during the study period. Hence, the more explicitly the source population can be defined, the easier it is to design a valid method for selection of controls. Wacholder *et al* (1992a; 1992b; 1992c) provide classic discussions of how best to select control subjects. Grimes & Schulz (2005) provide a more recent discussion of control subject selection.

The major principles in selecting controls are:

- Controls should come from the same study base (population) as the cases. If this is not done, the authors must defend their source of controls on a study-by-study basis (*eg* Palmer *et al,* 2010).
- In closed populations, controls should be representative of the source population with respect to exposure.
- In open populations, controls should mirror the exposure-time distribution of the non-case subgroup in the population.
- The time period during which a non-case subject is eligible for selection as a control is the same time period in which that subject is eligible to become a case if the disease should occur.

In general, random population control subjects may be more difficult to locate and less motivated to take part in the study than patient-related control subjects such as partners, friends, neighbours, or (unaffected) family members (Pomp *et al*, 2010). Population control subjects potentially have the drawback of recall bias and selective participation; their motivation to recall past events is likely to be different from that of case subjects. Asking patient-related control subjects has the risk of overmatching on the study exposure because of joint exposures. Hospital control subjects are readily accessible, usually cooperative and more likely to have the same recall ability as the cases, but always pose the problem whether exposure is unrelated to the disease leading to the hospitalisation of the control.

The implementation of these principles depends on the study design, so we shall begin our discussion with issues related to selecting controls in risk-based designs.

9.5 Selecting controls and data layout in risk-based designs

The traditional approach to case-control studies has been a risk-based (*ie* cumulative incidence) design. In this approach, the controls are selected from among the people that did not become

cases by the end of the study period. A subject can be selected as a control only once. This design is appropriate if the population is closed and is most informative if the risk period for the outcome in a subject has ended before subject selection begins. It fits situations such as outbreaks from infectious or toxic agents where the risk period for the disease is short and essentially all cases that would arise from that exposure will have occurred within the defined study period (*eg* a point-source food-borne outbreak; Gutierrez *et al* (2011)). Because the risk period has ended, for practical purposes, the study cases represent virtually all of the cases that would arise from the defined exposure even if the study period was extended. This design also assumes censoring is unrelated to exposure (Knol *et al*, 2008). If censoring of study subjects is not independent of exposure, a rate-based sampling approach (see Section 9.6), coupled with the usual unmatched risk (odds ratio) calculations, will provide a more consistent estimator of the risk ratio than sampling from the non-case group at the end of the risk period. Non-independent censoring might occur in studies of risk factors that could alter the risk of study-subject losses and also be related to the exposure of interest.

Suissa *et al* (2011) describe what they term a time-window bias in case-control studies. The source of this bias arises from the methods used to select controls and to measure their exposure. For example suppose the study population was observed for 67 months, from October 1, 1998 until June 1, 2004. The observation period was necessarily less than 67 months for the cases occurring over this period, while likely closer to 67 months for the controls. As a result, controls had more time to be 'exposed' to a number of risk factors than cases (see Strand *et al* (2011) for a general discussion). Using incidence density sampling (described in Section 9.6.1), choosing a number of controls each time a case occurs would circumvent this problem. Similarly Barnett (2011) describes what is termed the fixed cohort bias. For example, in a study of stillbirths occurring between January 1, 2001 and December 31, 2007, pregnancies starting (conception) in June 2000, stillbirths may or may not have been included in the cohort depending on whether or not they occurred before January 2001. In contrast, all live births from this time would have made it into the cohort. This means that first trimester exposures during June 2000 may look remarkably protective, as the number of stillbirths would be very small. The bias for a study of other exposures such as air pollution would then depend on what exposure had occurred in June 2000 and what the true association is. The bias can also occur at the end of the cohort, with the longer pregnancies missed and the shorter pregnancies captured. This potential bias is not prevented by a sampling strategy, but in this example at least, a way to avoid the bias is by excluding case and control subjects with estimated conception dates 20 weeks (shortest gestation) before the study period started or 43 weeks before it ended (assuming a longest gestation time of 43 weeks). This ensures that the exposures examined during any gestation period could equally apply to cases and controls.

The closed-source population can be categorised with respect to exposure and outcome as shown below (upper-case letters denote the population, lower case the sample):

	Exposed	**Non-exposed**	**Total**
Cases	A_1	A_0	M_1
Non-cases	B_1	B_0	M_0
Total	N_1	N_0	N

The cases (M_1) are those that arose during the study period, whereas the controls (M_0) are those that remained free of the outcome during the study period. Usually, all or most of the cases are included in the study so the sampling fraction (*sf*) among cases approaches 1. Usually, only a

small fraction of the non-cases are selected as controls, and the controls should be selected independently of exposure status so that there is an equal *sf* in exposed and non-exposed controls. At the end of the study period, there are B_1 exposed and B_0 non-exposed non-cases in the source population from which we select our study control subjects b_1 and b_0. Since we want to select the controls, without regard to their exposure status, from the list of non-case subjects at the end of the follow-up period, the sampling fractions in the 2 exposure-groups of non-cases should be equal. Hence,

the number of exposed controls in the sample is $b_1=sf(B_1)$, and
the number of non-exposed controls in the sample is $b_0=sf(B_0)$.

In a primary-base, closed-population, case-control study, an equal sampling fraction among controls would be obtained by random selection of a fixed number, or proportion, of study subjects from the non-case population (*ie* from the group that remains free of the disease at the end of the study period) (see Example 9.1).

In a secondary-base study, equal sampling with respect to exposure could be achieved by selecting controls randomly from the listing of non-case subjects recorded in the hospital records or registry. There is an additional caveat in selecting controls, from other non-cases in a secondary-base study; namely, in order to obtain a valid estimate of the frequency of exposure in the study population, we should sample controls from non-case subjects that have diagnostic outcomes that are not associated with the exposure of interest. However, since the cases in most secondary study bases are derived from open populations, a rate-based design is generally preferred for control selection (see Section 9.6)

In risk-based studies, the measure of association we use to contrast the odds of exposure in the cases to the odds of exposure in the controls is the odds ratio (*OR*).

$$\frac{a_1/a_o}{b_1/b_0}=\frac{a_1*b_o}{b_1*a_0} \qquad \textit{Eq 9.1}$$

The *OR* is a valid measure of association in its own right, and it also estimates the ratio of risks (*RR*) if the outcome is relatively infrequent (*eg* <5%) in the source population (see Chapter 6).

Knol *et al* (2008) clarify that although the odds ratio is the central measure of association in case-control studies, whether or not it approximates other measures of association (*eg* rate ratio) depends on the study design and assumptions about the source population.

9.6 Sampling controls and data layout in rate-based designs

Because the populations we study are often open, the case-control designs for these populations should use a rate-based approach (*ie* incidence density sampling), which seeks to ensure that the time-at-risk is taken into account when the control subjects are selected.

We can visualise the classification of the open-source population with respect to the number of cases and the cumulative time-at-risk in each of the exposure levels in the population as shown below (in this section, upper-case letters denote the population, lower case the sample):

	Exposed	Non-exposed	Total
Cases	A_1	A_0	M_1
Person-time at risk	T_1	T_0	T

To help understand rate-based, case-control designs, it is useful to think about how the 2 key rates of interest would be measured, and what subjects would be included in a cohort study of the same source population. Recall that, in a cohort study, if we wanted to study the association between exposure and the rate of the outcome, the 2 rates of interest at the end of the follow-up period would be:

$$I_1 = A_1 / T_1 \quad \text{and} \quad I_0 = A_0 / T_0 \qquad \textit{Eq 9.2}$$

where A represents the number of incident cases and T the cumulative person-time at risk in each exposure category. Note that, at the start of the follow-up period, all study subjects are non-cases and accumulate time-at-risk in either the exposed or non-exposed categories until they develop the outcome, or they are selected as controls, or the study period ends. The drawback to the cohort study design is that all subjects in the study population must be followed and, when the outcome is infrequent, this often means following a very large number of non-diseased subjects. The advantage of the case-control study design is that the much smaller (numerically) control series is used to reflect the subject-time exposure experience without the full enumeration of the population or the time at risk. Thus, in a rate-based case-control study, the cases are those subjects that would experience the outcome in the hypothetical cohort study. The controls are selected from non-case subjects such that the number of exposed and non-exposed control subjects reflects the relative magnitude of the T_1 and T_0 denominators without actually knowing their values.

To achieve this, we select controls using a sampling rate (*sr*) that is equal in the exposed and non-exposed non-case populations. More specifically, the ratio of the number of exposed controls (b_1) divided by the exposed population subject-time equals the number of non-exposed controls (b_0) in our sample divided by the non-exposed population subject-time.

$$sr = \frac{b_1}{T_1} \approx \frac{b_0}{T_0} \qquad \textit{Eq 9.3}$$

Therefore, in our sample the ratio of exposed to unexposed controls equals the ratio of the cumulative exposed and unexposed subject times.

$$\frac{b_1}{b_0} \approx \frac{T_1}{T_0} \qquad \textit{Eq 9.4}$$

Thus, the summary table has the same appearance as that of a risk-based table shown in Section 9.5. The ratio of the exposed cases to exposed controls divided by the ratio of the non-exposed cases to non-exposed controls in the study population estimates the ratio of the incidence rates (*IR*) in exposed and non-exposed subjects in the source population.

$$\frac{a_1 / b_1}{a_0 / b_0} \approx \frac{A_1 / T_1}{A_0 / T_0} \qquad \textit{Eq 9.5}$$

This ratio can also be viewed as the odds of exposure in the cases compared with the odds of exposure in the controls which, as we have seen, is called the cross-product ratio or odds ratio (*OR*). In this design, the *OR* estimates the *IR* (from a cohort study) and no assumption about rarity of outcome is necessary for a valid estimate.

9.6.1 Sampling from a primary-base open population

If the population is stable (*eg* the exposure does not vary over the study period), one way to ensure valid selection of controls is to randomly select controls from the source population at the end of the study period, provided the probability of selecting each potential control subject is proportional to the total time-at-risk for the outcome (*ie* regardless of exposure status). This proviso is needed because it is the amount of time-at-risk in the exposed and non-exposed groups that we should mirror in the controls. See Benson *et al* (2010) for an example of a primary-base study; virtually all cases of Type I diabetes were registered in the provincial database and all residents are also listed in the provincial medical registry. If time-at-risk data are available, controls can be selected at the end of the study period using the time-at-risk to weight the probability of their selection. Since every study subject is a non-case for at least part of the follow-up period, every study subject has some non-zero probability of being selected as a control, even those subjects that become a case subsequently. Time-at-risk would be known in well-defined population such as employees at a specific business. For example, in a nested case-control study, the study base might use recorded data for the employees whose time at risk of the outcome was known. Hence, we could select a sample of non-cases with probability proportional to the time-at-risk. Richardson (2004) provides instructions on how to program software to achieve valid risk-set sampling when matching on one or more covariates.

In the more common situation where the time-at-risk of individual subjects in the source population is not known, controls can be selected at fixed time points throughout the study period from the **risk set** (those non-cases in the source population eligible to become cases at that point in time). This approach is suitable if the level of exposure is unlikely to vary during the study period and if there is ongoing monitoring of the membership of the source population to identify the 'at-risk' population. The number of controls to be selected at each time point can vary and need not have a constant ratio to cases. As noted previously, if the exposure and covariate characteristics of the population do not change over the study period (*ie* the source population is stable), the sample *OR* from a logistic model estimates the *IR*.

The most common method of obtaining controls is by selecting a specified number of non-cases from the risk set **matched**, time-wise, to the occurrence of each case. This is called **incidence density sampling** and has the advantage that we do not need to know the time-at-risk for potential controls, nor do we need to assume that the population is stable**.** A number of controls is randomly selected at the time the case arises from those non-case subjects eligible to become cases at that time. Thus, at each time a subject develops the outcome, we choose a number of controls (*ie b*) from the non-case subjects (*ie B*) that exist in the source population at that point. The number of controls per case can vary and need not have a constant ratio over time. Using incidence density sampling is particularly well-suited to situations when the level of exposure might vary with calendar time, and in this instance, the data from the matched design should be analysed as such. However, if the level of exposure is unlikely to change over time (*ie* a stable population), the temporal-matching can be treated as just a convenient way of identifying when to select controls and the data can be analysed by unmatched procedures. When the temporal-matching design is used, the *OR* estimates the *IR* whether or not the population is stable.

In rate-based designs, subjects initially identified as controls can subsequently become cases. Since the period of time in which a subject is eligible to be a control should be the same as that in which it is eligible to be a case, should that event occur, controls can subsequently become cases. Their data are treated as independent in the analysis. If only first incident cases are

included in the study, the people who become cases cannot be selected as controls after they have developed the outcome of interest. If multiple occurrences per person of the case outcome are included, cases can subsequently be selected as controls and cases are at risk of being recurrent cases (see Example 9.4). The data for controls reflects their exposure and covariate status at the time they were selected as a control. The exposure and covariate status of the cases relate to the time at which the subject became a case. The process of selecting controls in open populations also means that the same subject can be selected as a control more than once. Note that because we are sampling directly from the source population, there should be no exclusions of potential controls because of exposure status (*ie* any subject in the source population that has not been a case at the time of sampling is eligible as a control, even those subjects with prior diseases that are associated with exposure).

One concern about using controls selected from the source population is the potentially low response level and the resultant concern over selection bias. Neupane *et al* (2010) make a strong case for using population-derived controls for hospitalised cases if reasonable response rates are likely. Kalton and Piesse (2007) discuss the selection of controls from the source population in both primary-base and secondary-base studies, and the appropriate analysis which might need to be used to account for a complex sampling design.

9.6.2 Sampling controls from a secondary base

When a clinic, laboratory, or other registry is the source of the cases, we have a secondary-base study. In such studies, selecting **non-cases** from the same registry is preferable to obtaining them from other sources. As before, the basic tenet is that the controls should reflect the exposure distribution in the population of potential cases that would have entered that registry had they developed the disease or outcome of interest. The problem is to know whether having the exposure of interest alters the probability that non-case subjects will be included in the registry; if it did, the exposure of the controls would not be a valid estimate of exposure levels in the source population. For example, it is well-documented that smoking increases the risk of hospitalisation for a number of diseases. Hence, if controls are randomly selected from all non-case subjects, the level of smoking in the controls will be higher than in the source population and this will bias the measure of association (*eg* odds ratio) toward 1. To avoid this bias, we should select control subjects from a variety of non-case diagnostic outcomes that are not associated with the exposure of interest. However, in some specialised, or restricted registries (*eg* reportable infectious human diseases), a high proportion of subjects listed will have diseases that are associated with the exposure of interest (*eg* consuming chicken that has been improperly stored or cooked as a common source of food-borne pathogens) and thus, their exposure does not reflect the exposure to chicken of non-cases in the source population. Alternative study designs have been proposed for these situations (see Chapter 10). It is our observation, that perhaps because of the difficulty in knowing if the reason for admission to the hospital or registry is independent of exposure, many researchers are choosing the control subjects either from patients at a primary-care clinic, instead of the specialised-care hospital, or from a primary base judged to be the source of the majority of patients with the outcome of interest (see Examples 9.5 and 9.6). However, one also needs to consider the potential for selection bias (refusal rates), confounding, and accuracy of information when selecting the source for control subjects (Fang *et al*, 2011; Neupane *et al*, 2010).

A key to sampling controls from a secondary base is to focus on the ‘admission’ and not the subject. Furthermore, diagnostic category exclusions for controls should only relate to

Example 9.5 A secondary-base case-control study with population controls

Patients (n=218) with symptom onset and confirmation of Crohn's disease in 1999-2004 were recruited from 9 hospitals in 5 regions of England (Abubakar *et al*, 2007). The *a priori* design was a matched study with a sample size estimate of 104 cases. In 2 regions, community controls (n=812) were recruited on the basis of the catchment area for each hospital, via general practices. For each Crohn's patient 2 general practitioners were randomly selected and asked to identify 5 randomly selected controls from their practice list, frequency matched by age (±1 year) and gender. In the remaining 3 regions, a geographic catchment area was allotted to each general practice and 2 postal codes in each area were randomly selected, within which they randomly selected 10 controls of the same gender as the index Crohn's patient. Ultimately, 63% of cases and 38% of selected non-cases participated.

Each participant's water source was identified and 6 different proxy measures of contamination were obtained. All participants received a short, validated self-administered questionnaire blinded to the study hypotheses. The questionnaire asked about occupation; types of vacation; diet, including quantities of water and dairy products consumed; family history of Crohn's disease, plus other potential risk factors. Because the matching on age could not be completed using the electoral register the data were analysed using both conditional and unconditional logistic regression (the latter including statistical control of age and gender). The authors also noted that the choice of control group had little impact on their results (*ie* general practice controls, population controls from voter list, or combined).

admissions during the study period time frame, and not to previous admissions (if the subject was admitted for a condition related to exposure before the study period, that subject should still be eligible as a control in the study period provided the reason for hospitalisation, at this time, is deemed to be independent of exposure). Some recommend that control subjects should only be selected from those diagnostic categories for which data exist to show that they are not related to the exposure of interest. However, most researchers have tended to use less stringent exclusion criteria for independence and select control subjects from diagnostic outcome categories that are not known or suspected to be associated with exposure.

Similar to primary-base studies, one method of selecting controls is to select them randomly from all the non-case admissions up to the end of the study period, having excluded those non-

Example 9.6 A case-control study using both primary-care-derived controls and population-based controls

This risk-based, case-control study was conducted in the greater Toronto area to evaluate potential lung cancer risk factors including environmental tobacco smoke (ETS) exposure, family history of cancer, indoor air pollution, workplace exposures and history of previous respiratory diseases with special consideration given to never smokers (Brenner *et al*, 2010). Controls were residents of metropolitan Toronto who did not have cancer at the time of recruitment. Population-based controls were randomly sampled from property tax assessment files (n=425). Hospital based controls were sampled from patients seen in the Mount Sinai Hospital Family Medicine Clinic (n=523), which is a non-specialty, family medicine practice situated within the hospital where recruitment into the study was conducted independent of reason for visit to the clinic. In total, 156 cases of lung cancer in people between the ages of 20 and 84 who had never-smoked, were identified through 4 major tertiary care hospitals in metropolitan Toronto between 1997 and 2002 and were frequency-matched on sex and ethnicity with 425 population controls and 523 hospital controls. Unconditional logistic regression models were used to estimate adjusted odds ratios (*OR*) and 95% confidence intervals (CI) for the associations between exposures and lung cancer risk. Since the emphasis was on non-smoking exposures, a separate analysis based on the 156 non-smoking cases and 466 non-smoking controls was conducted. The source of controls did not alter the main findings of the study.

case categories that are associated with the specified exposure(s). This might seem like a risk-based sampling strategy but in this instance, the sampling unit is 'the admission' not the subject. Since non-case subjects can be listed in the registry numerous times because of admission for the same, or different, non-case diseases, using 'the admission' as the sampling unit is an attempt to reflect their time at risk (*ie* those non-cases that are in the source population for longer periods will, on average, have more admissions for non-case diseases).

It is also possible to select controls randomly from the non-cases in the registry at regular intervals throughout the study period. Thus, if a 3-year study period was used and 300 controls were to be selected, 8 or 9 subjects would be selected each month, from all the non-case admissions listed in the registry during that month. If the population is stable, the sample *OR* estimates the *IR*. If the exposure level in the source population(s) is likely to vary with calendar time, then when fixed-time sampling is used, we should stratify on time in the analysis to prevent bias.

Alternatively, we can use incidence density sampling in which we **match** for 'time at risk' by selecting a specified number of non-cases that are admitted to the registry immediately after each case was admitted (or randomly from subjects admitted with appropriate non-case diseases within a defined period, such as one month). If the exposure level is likely to be constant over the study period, an unmatched analysis can be performed and the temporal-matching treated as just a convenient way of identifying control subjects. If the exposure level is likely to change over the study period, then a matched analysis should be pursued. Keogh (2008) discusses a variety of ways of selecting matched controls, including inverse sampling when most controls are expected to have the same exposure status as the case. This approach was designed to overcome the problem that if most study subjects are likely to be exposed (or unexposed), since it is only those case-control pairs whose exposure differs that contribute useful information in the analysis, the number of these discordant pairs might be very small, so the study would lack power.

In all instances, if a subject's exposure can change, the classification of that subject's exposure is based on the exposure of the subject at the time that subject became a case, or at the time of selection, if the subject is a control.

9.7 Other sources of controls

The following sources of non-cases can be used to select controls in either primary- or secondary-base studies; they include neighbourhood controls, friend controls, partner controls, and controls identified by random-digit-dialling (RDD) within the source population. Bunin *et al* (2011) found using friend controls was convenient, but did lead to potentially biased estimates of association because of over-matching.

When random sampling of controls is not possible, choosing neighbours of cases might suffice but their suitability needs to be established according to the study context. This means that a matched analysis should be conducted if neighbourhood is related to exposure. As with friend controls, selecting neighbours could introduce a bias and cause overmatching.

Random-digit-dialling can be used to contact potential control subjects (see Example 9.5). For example, the telephone number of potential controls might be matched to that of cases by area code. There are numerous hidden problems with this approach including time of calling, business versus home phone *etc*. If used, then the 'matching' should be accounted for in the

analysis if there is any chance that matching process is related to the exposure. DiGaetano and Waksberg (2002) discuss the selection of controls using RDD in comparison to planned in-person screening of the study population, as well as the use of clustered RDD. Randomly selecting households based on tax assessment roles is another method of selecting non-cases from the population as controls (see Example 9.6).

Selecting partners of patients has advantages and disadvantages (Pomp *et al*, 2010). As a result of selecting partners of patients, who usually are of the opposite sex, the age-sex distribution of the partner controls showed some peculiarities. Pomp *et al* studied risk factors for venous thrombosis. They found there was only a small group of young men affected, while there was a relatively large group of young women with venous thrombosis (due to pregnancy and oral-contraceptive use). The small group of men yielded an even smaller control group of female partners, which made women-specific risk factors difficult to analyse, due to a relative lack of control subjects. Moreover, not all patients had a partner, so there were fewer available partners than patients. In addition, individuals with a partner may be different than those without a partner. One might expect 'friend controls' to pose similar, but less-extreme issues than partners.

9.8 The number of controls per case

Most studies use a 1:1 case-control ratio; however, other than being statistically efficient, there is nothing magical about having just one control per case. Indeed, if the information on the covariates and exposure is already recorded (*ie* in a sense, exposure data are free), one might use all of the qualifying non-cases in the registry as controls to avoid issues of sampling. In addition, when the number of cases is small, the precision of association measures can be improved by selecting more than one control per case. There are formal approaches for deciding on the optimal number, but usually the benefit of increasing the number of controls per case is small; often 3–4 controls per case is the practical maximum.

9.9 The number of control groups

Pomp *et al* (2010) review the history of using multiple controls groups and comment on their experience with two control groups in a recent study. Some researchers have attempted to balance a perceived bias with one specific control group by using more than one control group (see Examples 9.5 and 9.6). However, if this is done, it needs to be very clearly defined as to what biases are likely to be present in each control group and how one will interpret the results especially if they differ dramatically from one control group to another. The use of more than one control group also adds complexity to the analyses. If we choose more than one control group, the different control groups should be compared with respect to exposure. If they do not differ significantly, it ensures that, if a bias is present, the control groups may have the same net bias. However, if they differ, we often are not sure which one is the correct group to use. The general experience is that the value of more than one control group is very limited.

Suissa *et al* (2010) note that the conventional approach to improve precision of the odds ratio in a case-control study is to increase the number of controls per case. With time-varying exposures, an alternative is to increase the number of observations per control. This design uses multiple control person-moments (*eg* days, months) of exposure within each control subject. The point and variance estimators of the odds ratio need to be corrected for within-subject correlation when using this approach.

9.10 Exposure and covariate assessment

Most case-control studies are retrospective and record-searching replaces the follow-up period that would be present in a prospective study. Because of this, a concise, specific, workable definition of 'exposure' (and also of the confounders) is needed when implementing the study design. When ascertaining exposure status and information on confounders, it is preferable to obtain the greatest accuracy possible. Failing that, the process of ascertaining exposure history should have comparable accuracy in both groups. Usually this can be achieved by using the same process for obtaining exposure and confounder data in both cases and controls and, where possible, having the data collectors blinded to case status. Achieving this becomes difficult if the case data are obtained from hospital records and the control exposure data are obtained from randomly selected subjects in the source population.

Many times the exposures that are studied are not permanent and can change over time. If a subject's exposure history changes during the follow-up period, care is needed to document the change and when it occurred. As a general rule, the exposure status of cases should be the exposure category that existed at the time of outcome occurrence. For controls, their exposure status reflects their exposure situation at the time of their selection.

9.11 Keeping the cases and controls comparable

In order to obtain unbiased estimates of any association between exposure and the outcome, it is important that covariates that are related to both the outcome and the exposure have a similar distribution in the case and control series. Both **exclusion** and **inclusion** criteria can be used to reduce the number of extraneous factors that can adversely affect the study results; the criteria used should apply to both cases and potential controls. For example, if race is a likely confounder, include only one race in the study, usually the dominant one in the source population (see Example 9.3). This prevents confounding by race. What we would lose in this approach is the ability to generalise the results to other races or to assess interactions with the exposure across the confounder levels (*ie* races). All inclusion and exclusion criteria should be stated clearly in the study design (Examples 9.1 and 9.3).

Matching on known confounders is a second strategy frequently used to prevent confounding and, to a lesser extent, to increase efficiency (*ie* power of the study) (Examples 9.1, 9.2 and 9.5). Unfortunately, matching often does not work well for either of these objectives in case-control studies (see Section 13.3). If matching is to be used, how it is to be implemented should be described, and a conditional analysis of the data will be required (Section 16.15). When there is a large number of potential confounders, it is not practical to control confounding using restricted sampling and matching. Thus, analytic control is the approach most often relied upon. However, similar to using propensity scores in cohort studies, Epstein *et al* (2007) and Allen and Satten (2011) propose using the stratification score to assess the balance of covariates in the case and control groups. The stratification score for a case-control study is the probability of disease modelled as a function of potential confounders. By standardising the distribution of co-variates based on the stratification score we cannot only control confounding, but we can also assess how much of the original crude association between the exposure and disease is explainable by the confounders. Sinha and Mukherjee (2006) describe a score test for analysis and sample-size estimation of matched case-control studies with a polychotomous exposure.

The third approach to preventing confounding is **analytic control**. Here we measure the

confounders and use **multivariable** techniques to prevent confounding. Often, this is our preferred choice, sometimes working in concert with restricted sampling (see Chapter 13 for more detail).

9.12 Analysis of case-control data

The data format for case-control studies is shown below, and analysis of both risk-based and rate-based, case-control sampling designs proceeds in a similar manner. We will assume that in our study group we observe a_1 exposed cases and b_1 exposed controls, and a_0 non-exposed cases and b_0 non-exposed controls. There are m_1 cases and m_0 controls. Remember that we cannot directly estimate disease frequency (unless the study is nested within an enumerated source population)—overall or by exposure level—because the m_1:m_0 ratio was fixed by sampling design. In a 2X2 table the format is:

	Exposed	Non-exposed	Total
Cases	a_1	a_0	m_1
Controls	b_1	b_0	m_0

Chapter 6 outlines the analysis of these data including hypothesis-testing, estimating the odds ratio, and developing confidence intervals for the odds ratio. Grimes and Schulz (2008) reiterate the interpretation and uses of the odds ratio. Rauscher and Poole (2006) discuss different methods of combining categorical covariates so that a common referent category for the odds ratio is achieved (they believe this is the most appropriate way to perform the analysis). Recall that whether or not the odds ratio estimates the risk ratio or rate ratio depends on the study design.

- With risk-based designs, and sampling of controls at the end of the follow-up period, the odds ratio estimates the risk ratio if the frequency of disease in the source population is low (*eg* below 10%), and censoring is unrelated to exposure.
- If concurrent sampling (*ie* incidence density sampling) is used, the odds ratio estimates the rate ratio in both closed and open populations—for validity, stability of exposure is needed in the closed population but not in the open population. If matching is ignored in the analysis of data from a closed population, the odds ratio is just that, an odds ratio.
- When controls are selected from an open population without concurrent sampling of controls with the occurrence of cases, the odds ratio estimates the rate ratio only if the population is stable, otherwise it is just the odds ratio
- If matching is used to select the controls but is ignored in the analysis, the impact depends on the extent of exposure changes during the study period (see earlier comments in Section 9.6.1) (Knol *et al*, 2008).

Niccolai *et al* (2007) describe the analysis of matched data and show how to assess the impact of time since vaccination and age at vaccination on vaccine efficacy. Liu *et al* (2010) describe how to use a proportional hazards analysis when the exposure is time-varying. Mandrekar and Mandrekar (2008) reiterate that if matching is used to select controls, the analysis must take that into account to avoid bias in the estimates.

King and Zeng (2002) and Richardson (2004) note that often the odds ratio is not the association measure of most interest; however, historically it is the only feasible association measure we can estimate unless disease frequency data are available in the exposed and non-

exposed subsets of the source population. Using a parameter they denote as the τ fraction of exposed individuals in the source who experience the outcome and its estimated upper and lower bounds, they show (and provide software code for) how to estimate risk and rate differences (with confidence intervals) from case-control data. Similarly, Cox (2006) and Lui (2005) indicate how to estimate attributable fractions.

Sometimes, the data from one case-control study can be used validly for a second study. Reilly *et al* (2005) demonstrate how to analyse the data when a former exposure variable becomes the outcome for a second study (in their example the original study used cancer as the outcome with *Heliobacter pylori* (*Hp*) as the exposure. Later, it was desired to use the same data to assess potential risk factors for the presence of *Hp*). Similarly, Richardson *et al* (2007) describe how to analyse case-control data for an outcome different from the one used in the original study.

9.13 Reporting guidelines for case-control studies

Vandenbroucke *et al* (2007) have described the key elements of case-control studies that should be reported (Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)). The complete listing is shown in Table 7.3; those specific to case-control studies are included in Table 9.1. As noted, we elaborated these key points in this chapter as they should be used to help plan and report case-control studies, and to help assess the validity of published case-control studies.

Table 9.1 The STROBE checklist of items specific to case-control studies that should be addressed in reporting of results (see Table 7.3 for complete listing)

Methods	6a Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls
	6b Case-control study—For matched studies, give matching criteria and the number of controls per case
	12 Case-control study—If applicable, explain how matching of cases and controls was addressed
Results	15 Case-control study—Report numbers in each exposure category, or summary measures of exposure

References

Abubakar I, Myhill DJ, Hart AR, Lake IR, Harvey I, Rhodes JM, *et al*. A case-control study of drinking water and dairy products in Crohn's Disease—further investigation of the possible role of Mycobacterium avium paratuberculosis. Am J Epidemiol. 2007;165(7):776-83.

Allen AS, Satten GA. Control for confounding in case-control studies using the stratification score, a retrospective balancing score. Am J Epidemiol. 2011;173(7):752-60.

Baksh MF, Todd S, Whitehead J, Lucini MM. Design considerations in the sequential analysis of matched case-control data. Stat Med. 2005;24(6):853-67.

Barnett AG. Time-dependent exposures and the fixed-cohort bias. Environ Health Persp. 2011;119(10):A422-3; author reply A3.

Benson VS, Vanleeuwen JA, Taylor J, Somers GS, McKinney PA, Van Til L. Type 1 diabetes mellitus and components in drinking water and diet: a population-based, case-control study in Prince Edward Island, Canada. J Am Coll Nutrition. 2010;29(6):612-24.

Brenner DR, Hung RJ, Tsao MS, Shepherd FA, Johnston MR, Narod S, et al. Lung cancer risk in never-smokers: a population-based case-control study of epidemiologic risk factors. BMC Cancer. 2010;10:285.

Bunin GR, Vardhanabhuti S, Lin A, Anschuetz GL, Mitra N. Practical and analytical aspects of using friend controls in case-control studies: experience from a case-control study of childhood cancer. Paediatric and Perinatal Epidemiology. 2011;25(5):402-12.

Busse JW, Obremskey WT. Principles of designing an orthopaedic case-control study. J Bone and Joint Surg. 2009;91 Suppl 3:15-20.

Caro JJ, Huybrechts KF. Case-control studies in pharmacoeconomic research: an overview. Pharmacoeconomics. 2009;27(8):627-34.

Cox C. Model-based estimation of the attributable risk in case-control and cohort studies. Stat Methods Med Res. 2006;15(6):611-25.

D'Agata EM. Methodologic issues of case-control studies: a review of established and newly recognized limitations. Inf Control and Hosp Epidemiol. 2005;26(4):338-41.

DiGaetano R, Waksberg J. Commentary: Trade-offs in the development of a sample design for case-control studies. Am J Epidemiol. 2002;155(8):771-5.

Dore K, Buxton J, Henry B, Pollari F, Middleton D, Fyfe M, et al. Risk factors for *Salmonella* typhimurium DT104 and non-DT104 infection: a Canadian multi-provincial case-control study. Epidemiol and Inf. 2004;132(3):485-93.

Dorgan JF, Stanczyk FZ, Kahle LL, Brinton LA. Prospective case-control study of premenopausal serum estradiol and testosterone levels and breast cancer risk. Breast Cancer Res 2010;12(6):R98.

Edwards MD. Sample size requirements for case-control study designs. BMC Med Res Meth. 2001;1:11.

Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. Am J Human Genetics. 2007;80(5):921-30.

Fang R, Le N, Band P. Identification of occupational cancer risks in British Columbia, Canada: a population-based case-control study of 1,155 cases of colon cancer. Int J Envrion Res and Public Health. 2011;8(10):3821-43.

Fletcher AE. Case-control design: making the case. Am J Opthalmol. 2010;149(4):540-2.

Goldberg MS. The intersection of sound principles of environmental epidemiologic research and ethical guidelines and review: an example from Canada of an environmental case-control study. Rev Environm Health. 2010;25(2):147-60.

Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. Lancet. 2005;365(9468):1429-33.

Grimes DA, Schulz KF. Making sense of odds and odds ratios. Obstetrics and gynecology.

2008;111(2 Pt 1):423-6.

Gutierrez Garitano I, Naranjo M, Forier A, Hendriks R, De Schrijver K, Bertrand S, et al. Shigellosis outbreak linked to canteen-food consumption in a public institution: a matched case-control study. Epidemiol and Inf. 2011;139(12):1956-64.

Hak E, Wei F, Grobbee DE, Nichol KL. A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. J Clin Epidemiol. 2004;57(9):875-80.

Jiang Y, Scott A, Wild CJ. Case-control analysis with a continuous outcome variable. Stat Med. 2009;28(2):194-204.

Kalton G, Piesse A. Survey research methods in evaluation and case-control studies. Stat Med. 2007;26(8):1675-87.

Keogh RH. Inverse sampling of controls in a matched case control study. Biostatistics. 2008;9(1):152-8.

King G, Zeng L. Estimating risk and rate levels, ratios and differences in case-control studies. Stat Med. 2002;21(10):1409-27.

Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol. 2008;168(9):1073-81.

Levin KA. Study design V. Case-control studies. Evidence-based Dent. 2006;7(3):83-4.

Liu M, Lu W, Shore RE, Zeleniuch-Jacquotte A. Cox regression model with time-varying coefficients in nested case-control studies. Biostatistics. 2010;11(4):693-706.

Lui KJ. Estimation of attributable risk for case-control studies with multiple matching. Stat Med. 2005;24(19):2953-62.

Magura L, Blanchard R, Hope B, Beal JR, Schwartz GG, Sahmoun AE. Hypercholesterolemia and prostate cancer: a hospital-based case-control study. Cancer Causes & Control. 2008;19(10):1259-66.

Mandrekar JN, Mandrekar SJ. Case-control study design: what, when, and why? Journal Thoracic Oncology. 2008;3(12):1371-2.

Mayo NE, Goldberg MS. When is a case-control study a case-control study? J Rehab Med. 2009a;41(4):217-22.

Mayo NE, Goldberg MS. When is a case-control study not a case-control study? J Rehab Med. 2009b;41(4):209-16.

Neupane B, Walter SD, Krueger P, Loeb M. Community controls were preferred to hospital controls in a case-control study where the cases are derived from the hospital. J Clin Epidemiol. 2010;63(8):926-31.

Niccolai LM, Ogden LG, Muehlenbein CE, Dziura JD, Vázquez M, Shapiro ED. Methodological issues in design and analysis of a matched case-control study of a vaccine's effectiveness. J Clin Epidemiol. 2007;60(11):1127-31.

Orenstein EW, De Serres G, Haber MJ, Shay DK, Bridges CB, Gargiullo P, et al. Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. Int J Epidemiol. 2007;36(3):623-31.

Palmer KT, Kim M, Coggon D. Bypassing the selection rule in choosing controls for a case-control study. Occup and Envrionm Med. 2010;67(12):872-7.

Peng B, Li B, Han Y, Amos CI. Power analysis for case-control association studies of samples with known family histories. Human Genetics. 2010;127(6):699-704.

Pomp ER, Van Stralen KJ, Le Cessie S, Vandenbroucke JP, Rosendaal FR, Doggen CJ. Experience with multiple control groups in a large population-based case-control study on genetic and environmental risk factors. Europ J Epidemiol. 2010;25(7):459-66.

Rauscher GH, Poole C. Common referent versus shifting referent methods when using case-control data to examine patterns of incidence across multiple exposure variables. Annals Epidemiol. 2006;16(10):743-8.

Reilly M, Torrang A, Klint A. Re-use of case-control data for analysis of new outcome variables. Stat Med. 2005;24(24):4009-19.

Richardson DB. An incidence density sampling program for nested case-control analyses. Occup and Envrionm Med. 2004;61(12):e59.

Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. Epidemiol. 2007;18(4):441-5.

Rodrigo S, Sinclair M, Wolfe R, Leder K. Risk factors for gastroenteritis: a nested case-control study. Epidemiol and Inf. 2011;139(4):552-9.

Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol Biomark & Prev. 2005;14(8):1899-907.

Schulz KF, Grimes DA. Case-control studies: research in reverse. Lancet. 2002;359(9304):431-4.

Singh H, Mahmud SM. Different study designs in the epidemiology of cancer: case-control vs. cohort studies. Methods Mol Biol. 2009;471:217-25.

Sinha S, Mukherjee B. A score test for determining sample size in matched case-control studies with categorical exposure. Biometrical J. 2006;48(1):35-53.

Strand LB, Barnett AG, Tong S. Methodological challenges when estimating the effects of season and seasonal exposures on birth outcomes. BMC Med Res Meth. 2011;11:49.

Suissa S, Dell'aniello S, Martinez C. The multitime case-control design for time-varying exposures. Epidemiol. 2010;21(6):876-83.

Suissa S, Dell'aniello S, Vahey S, Renoux C. Time-window bias in case-control studies: statins and lung cancer. Epidemiol. 2011;22(2):228-31.

Sun W, Joffe MM, Chen J, Brunelli SM. Design and analysis of multiple events case-control studies. Biometrics. 2010;66(4):1220-9.

Thrift AG. Case-control studies: the importance of design and conduct. Neuroepidemiology.

2010;34(4):264-6.

van Stralen KJ, Dekker FW, Zoccali C, Jager KJ. Case-control studies--an efficient observational study design. NephronClinical practice. 2010;114(1):c1-4.

Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007;18(6):805-35.

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol. 1992a;135(9):1019-28.

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. Am J Epidemiol. 1992b;135(9):1029-41.

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. Am J Epidemiol. 1992c;135(9):1042-50.

Wacholder S. Bias in full cohort and nested case-control studies? Epidemiol. 2009;20(3):339-40.

Walter SD. Mammographic screening: case-control studies. Annals of Oncology. 2003;14(8):1190-2.

Zhou H, Chen J, Rissanen TH, Korrick SA, Hu H, Salonen JT, et al. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. Epidemiol. 2007;18(4):461-8.