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 principle features for selecting controls in open and closed primary-base case-control studies.
7. Explain the principle 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 group of cases and a group of non-cases (*ie* controls), and contrast the frequency of the exposure factor in the cases with the frequency of the exposure factor in the controls (Rothman *et al*, 2008). The study subjects that have developed the disease or outcome of interest are the cases, whereas the study subjects that have not developed the disease or outcome of interest, at the time they are selected, are the controls. It is important to stress that a case-control study is not a comparison between a set of cases and a set of 'healthy' subjects, but between a set of cases and a set of non-case subjects whose exposure to the factor of interest reflects the exposure in the source population. The controls would be included as 'cases' if they had developed the outcome of interest. An overview of key case-control design issues is available elsewhere (Schulz & Grimes, 2002). Although frequently we describe study designs as though an individual animal is the unit of interest, the design also applies to aggregates of individuals such as litters, pens, barns or herds (*eg* see Example 9.1 where the unit of concern is a cage and Example 9.2 where it is a farm). 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 can be studied in a case-control format. (D'Agata, 2005) discusses some limitations (and solutions) when case-control studies are used to study risk factors such as multiple resistant bacteria.

Usually, case-control studies are performed retrospectively since the outcome (usually disease) has occurred when the study begins. It is possible to conduct prospective case-control studies where the cases do not develop until after the study begins and they are enrolled in the study over time (see Example 9.6 and Archer *et al* (2008)).

9.2 THE STUDY BASE

The **study base** is the population from which the cases and 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**. If the study base is one or more steps removed from the actual source population, such as a referral clinic, laboratory or central registry, the source population is referred to as a **secondary study base** or **secondary base**.

In describing the source population in a case-control study, the term **nested** implies that the entire source population from which the cases are drawn has been enumerated. Usually a subsample of the entire population forms the source population with the case series being all, or a known fraction of, the cases from this source (see Example 9.4 where a sample of horses constitutes the source population from which the cases and controls were obtained). In a nested study, the sampling fractions of cases and controls are known and 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. Since the comparison subjects (*ie* the controls) are selected from the non-cases in the same defined population, this process prevents 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.

Variations in the case-control study design 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' and no new subjects are added to the initial population (*eg* all calves born on a dairy farm in a given year); whereas, 'dynamic' or **open** populations can have both additions and losses of subjects during the study period (*eg* a sow herd or kennel over the period of one year). A population is said to be stable if the characteristics of the study subjects, including the exposure, do not change over time. Closed populations are not stable, especially when the follow-up period is long (*eg* they age).

As with cohort studies, closed populations support risk-based case-control designs; 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. Sometimes, for research purposes, it is possible to convert an open population to a closed population. For example, a study of risk factors for mastitis in dairy cows in a particular herd over one calendar year would likely have to contend with new cows being added and cows that were in the herd originally being lost part way through their lactation. However, if the hypothesis of interest is to identify risk factors for first occurrence of mastitis in the initial 60 days of lactation, by following a defined group of cows after they calve for the first 60 days of lactation, we will have created a closed population. Only cows that calve in the herd(s) and are followed for the full 60-day period would be included in the study.

9.3 THE CASE SERIES

Key elements in selecting the case series include identifying the source(s) of the cases, the definition of the disease (the required diagnostic criteria for the outcome), and whether only incident, or both incident and prevalent cases, are to be included. Usually, only the first occurrence of the outcome in each study subject is included in the case series.

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.

A major decision is whether the cases will all be from a primary base, or if they will be obtained from a secondary base such as a veterinary clinic or a specific registry of diseased subjects. Sampling directly from the 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 in veterinary medicine because farms with good records allow complete enumeration of animals and health events (although one might have to deal with both 'owner-diagnosed' and 'veterinary-diagnosed' cases). As noted above, although dairy herds are open populations, the study design might allow these herds to be considered as closed thus allowing a risk-based analysis. In secondary-base studies, a major challenge is to conceptualise the actual source population for the cases such that the selection process ensures that the controls arise from the same source population. In essence, we would like to select controls from that group of subjects that would have gone to the secondary source had they developed the disease of interest; as noted, this population is often difficult to define.

The diagnostic criteria for a subject to become a case should include specific, well-defined manifestational (*ie* clinical) signs and, when possible, other 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 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, a case series of autoimmune disease in dogs 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 'obvious' differences in the disease manifestation, especially if the causes of the different forms of the disease might differ.

9.3.1 Case-control studies with continuous outcomes

Case-control studies are based on outcome-dependent sampling. Typically, the outcome has a dichotomous (diseased/not diseased, or yes/no) scale and can be analysed with a logistic model. When the outcome is measured on a continuous scale (*eg* kg of milk per day, weight gain per day, *etc*), researchers might randomly, or purposively, select study subjects and then compare subjects at the low and high extremes of the outcome distribution. This allows analyses using logistic models but discards information about the outcome. If we desire to use the original continuous outcome, special regression techniques must be used to account for the sampling structure (Jiang *et al*, 2009; Zhou *et al*, 2007). Such models can also 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, then the usual linear models 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 they should be representative of the exposure experience in the population which gave rise to the cases. Controls should be subjects who would have been cases if the outcome occurred. 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; Wacholder *et al*, 1992b; Wacholder *et al*, 1992c) are classic discussions of how best to select control subjects). Grimes & Schulz (2005), provide a more recent discussion of control selection. 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.

The major principles in selecting controls are:

- Controls should come from the same study base (population) as the cases.
- 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 it is also eligible to become a case if the disease should occur.

The implementation of these principles depends on the study design, so we shall begin our discussion with the traditional risk-based design.

9.5 SELECTING CONTROLS IN RISK-BASED DESIGNS

The traditional approach to case-control studies in veterinary medicine has been a risk-based (*ie* cumulative incidence) design. In this approach, the controls are selected from among those animals 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 foodborne outbreak, or bovine respiratory disease occurrence post-arrival in a feedlot—see Example 9.1). Because the risk period has (for practical purposes) ended, the study cases represent virtually all of the cases that would arise from the defined exposure even if the study period were extended. This design assumes censoring is unrelated to exposure (Knol *et al*, 2008).

Controls can also be selected from the population-at-risk each time a case occurs in a risk-based study. If this approach is used, an analysis which accounts for this 'matching' should be used.

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 are those that arose during the study period, whereas the controls are those that remained free of the outcome during the study period. Usually, all or most of the cases (M_1) are included in the study so the sampling fraction (*sf*) among cases approaches one. 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. For example, at the end of the study period there are B_1 exposed non-cases 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 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

Example 9.1 A primary-base risk-based case-control study of infectious salmon anaemia in cages of salmon in Maine

The distribution of infectious salmon anaemia (ISA) in 2004 was examined among 80 cages of salmon from 3 Atlantic grow-out farms in Maine, USA that were stocked with smolts from a single hatchery (the 80 cages constituted the sea-cage source population of interest) (Gustafson L *et al*, 2007). Cage-level ISA disease was defined as one or more moribund fish confirmed positive for ISA virus by 2 laboratory tests. Control cages were all cages that remained ISA-free during 2004. Risk factor information came from company and government records. Cage-level risk factors were screened in univariable models and those significant at $p \leq 0.25$ were explored further using a multivariable logistic regression.

population (*ie* from the group that remains free of the disease at the end of the study period; see Examples 9.1 and 9.2). All available controls were used by Gustafson *et al* (2007) (Example 9.1) and Melendez *et al* (2006) in studies of risk factors for ISA and udder edema, respectively. In contrast, Kung *et al* (2007) chose a subsample of 2 control-farms per case-farm in a study of risk factors for avian influenza (Example 9.2).

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 registry. There is an additional caveat in selecting controls in a secondary-base study—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(s) of interest. As we point out subsequently, most secondary study bases are derived from open populations and a rate-based design should be used to select controls.

In reviewing the literature, we noted a number of studies where a risk-based approach to selecting controls was used when using a secondary base for the source of control subjects. 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, for example, be common in studies of risk factors for diseases in many food-animal species where 'removal' of study subjects is under the owner's control, and some of the diseases that could alter the risk of removal could be related to the exposure of interest.

Example 9.2 A primary-base risk-based case-control study of influenza A on chicken farms in Hong Kong

During 2002, influenza A (H5N1) isolations occurred on 22 of the 146 active chicken farms in Hong Kong (Kung *et al*, 2007). Case farms were defined as farms that had high death rates caused by H5N1 infection or farms where H5N1 was isolated from chickens during the outbreak. Two control-farms per case-farm were selected at the end of the outbreak from farms that remained disease free. The questionnaire on potential risk factors contained 62 closed and 26 open-ended questions, and was pre-tested on 5 chicken farms. Data on geographic location, farm characteristics, stock information, flock health history, farm biosecurity, farm management, and marketing practices were collected by trained interviewers during farm visits. Additional information such as farm area, number of sheds, and incoming day-old chick numbers were obtained from official records held by the Department of Agriculture, Fisheries, and Conservation and used to validate the information collected during on-farm interviews.

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

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*).

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).

9.6 SELECTING CONTROLS IN RATE-BASED DESIGNS

Because the populations we study often are 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
Animal-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 animal-time at risk in each exposure group. 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 group 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 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) in our sample 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.

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

And, therefore, in our sample

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

Given this, 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 controls 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. 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 populations such as herds or flocks with complete records for all animals. For example, in a case-control study of risk factors for bovine leukosis, if herds on milk-recording systems were used for the study, it would be possible to obtain time-at-risk data for each cow and hence, select a sample of non-cases with probability proportional to the time-at-risk. Richardson *et al* (2007) provides instructions on how to program software to achieve valid risk-set sampling when matching on one or more covariates. Olea-Polpelka *et al* (2006) provide an example of this based on selecting control herds for a study of bovine tuberculosis in Ireland. One drawback to this method of selecting controls is that if biological samples need to be collected, some of the subjects selected as controls might not be available when selection occurs.

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-cases eligible to become cases at that time. If we visualise the study population, then 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. 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 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 kept separate and treated as independent in the analysis. If only first incident cases are included in the study, these animals cannot be selected as controls after they have developed the outcome of interest. 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).

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-cases 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. To avoid this bias, we should select control subjects from a variety of non-case diagnostic outcomes that are not associated with exposure. In some specialised, or restricted registries (*eg* reportable human diseases), a high proportion of subjects listed will have diseases that are associated with the exposure of interest (*eg* consuming chicken) and thus, their exposure does not reflect the exposure of non-cases in the source population. Alternative study designs have been proposed for these situations (see Chapter 10; Keogh (2008)).

A key to sampling in secondary bases is to focus on the ‘admission’ and not the subject.

Furthermore, diagnostic category exclusions for controls should only relate to 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 its 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-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 might **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 1 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 (*eg* see Example 9.3). If the exposure level is likely to change over the study period then a matched analysis should be pursued (Example 9.4). 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.

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 procedures can be used in either primary- or secondary-base studies; they include neighbourhood controls, controls identified by random digit dialling (RDD) within the source population, and previously identified population-based controls.

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. Selecting

Example 9.3 A secondary-base case-control study of equine protozoal myeloencephalitis

The study was based on 183 horses with equine protozoal myeloencephalitis **(**EPM) at 11 equine referral hospitals in the United States (Cohen *et al*, 2007). The study used an incidence-density case-control design. Each participating hospital was asked to provide data on at least 10 horses with EPM (cases), 10 horses with non-neurologic disease (non-neurologic controls), and 6 horses with neurologic diseases other than EPM (neurologic controls) each year for 3 years (September 1, 2001 to August 31, 2003). Non-neurologic controls were defined as the next horse >6 months of age that was admitted to the hospital after a case of EPM. Neurologic controls were defined as horses >6 months of age admitted for a neurologic problem other than EPM. Data were compared between the case group and each of the 2 control groups by means of polytomous logistic regression. The time-matched control selection process was viewed as just a convenient way of selecting a control horse so the matching was ignored in the analysis. There was no discussion of why 2 control groups were used, nor of the impact of this on interpretation of results. The results did differ by control group.

neighbours could introduce a bias and might cause overmatching in some studies. For example, in a primary-base study of factors related to *Salmonella* spp in bulk milk tanks on dairy farms, the closest farm was used as a control. However, often these farms were owned by relatives of the case farm owner and many times farm implements and food items were shared between case and non-case farms. Thus, overmatching was likely present (West *et al*, 1988). Non-case animals housed next to cases within a barn might be suitable, spatially matched, controls in some studies.

Random digit dialling can be used to contact potential control subjects (*ie*. animal owners or human subjects). 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.

In order to avoid some of the selection bias issues associated with obtaining controls in secondary-base case-control studies, some researchers select the control subjects directly from

Example 9.4 A nested rate-based case-control study of respiratory disease in horses in the United Kingdom

A matched case-control study was used to determine which infectious agents and other factors were associated with clinically apparent respiratory disease in young racehorses in training in the UK (Newton *et al*, 2003). The case-control study was nested inside a larger longitudinal study in which 10-15 horses in each of 7 training stables were monitored at any given time. Between 1993 and 1996, a total of 170 cases, defined as horses with sudden onset of coughing with nasal discharge or pyrexia, were identified. Up to 4 controls per case were selected from the source population at the time of case occurrence matched by trainer and time period. Horses were eligible as controls if they did not have clinical signs of respiratory disease at the time of sampling, were under the same trainer and had been sampled within 6 weeks of examination of the case (note that some of these horses could later become cases). Factors examined included age, sex, time since entry into the training yard, time since last race and various microbiological agents. Multivariable conditional logistic regression modeling was used to evaluate the risk of being a case for variables after adjustment for other factors.

Example 9.5 A secondary-base rate-based case-control study of risk factors for canine atopic dermatitis in Sweden

This study involved 58 cases of canine atopic dermatitis (CAD) from 12 veterinary practices in Sweden and 61 unaffected controls, matched to cases by breed and year of birth (Nødtvedt *et al*, 2007). Only dogs in the high-risk breeds of boxer, bull terrier and West Highland white terrier were included in the study. The sampling period was set to a maximum of 2 years starting in June 2003. Newly and previously diagnosed cases of CAD were included if they met a series of inclusion and exclusion criteria. As cases occurred, a population control of the same breed and year of birth was randomly selected from the Swedish Kennel Club (SKC) registry. Potential risk factors were screened in univariable models. Because of potential clustering by veterinary practice, the multivariable final model included a term for 'examining veterinarian'. The measures of association included odds ratios and the population attributable fraction for not feeding homemade diets.

the study population (see Examples 9.5 and 9.6). For example, when the registry from which the cases are obtained is largely composed of subjects whose disease(s) is likely related to the exposure of interest, obtaining controls from another source, such as the source population, may be the only way of obtaining a comparison group whose exposure history reflects what is happening in the source population. As one example, (Dore *et al*, 2004) used cases obtained from a provincial reportable disease database. (This database consists mainly of food and waterborne disease cases. Thus, the majority of listed subjects would be associated with a food or waterborne exposure regardless of the actual diagnosis, while the Canadian Provincial Ministry of Health records (each citizen of Canada has a record) were used as their source of controls.) Since the reportable disease database was mainly comprised of people with foodborne diseases, obtaining controls that were independent of food exposures proved to be very difficult (see Chapter 10, Case-case study designs for alternative designs focused specifically on this issue). In other circumstances, selected groups from the study population are available from which to select controls subjects. In Example 9.5, it was likely feasible to obtain controls from non-cases in the 12 clinics; however the authors chose to obtain their controls from Kennel-club members.

One concern about using controls selected from the source population is the potentially low response level and the resultant concern over selection bias. 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.

Example 9.6 A prospective rate-based case-control study of human campylobacteriosis using population based controls

This study of sporadic, domestically acquired campylobacteriosis was conducted in 3 counties in Norway in 1999–2000 (Kapperud *et al*, 2003). After each of the 212 cases occurred, the physician and then the patients were contacted to obtain information on potential risk factors. Criteria for enrolling a case included being a resident of the study area with culture-confirmed campylobacteriosis caused by *C. jejuni* or *C. coli*, that was diagnosed at 1 of the 3 medical microbiologic laboratories during the study period and not having traveled abroad in the 2 weeks prior to onset of illness. At the time each case occurred, 2 randomly selected subjects, matched by age (±5 years), sex and geographic area, from the population registry (the sampling frame) were selected as controls (87% responded positively). Data were analysed using a conditional logistic regression.

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

Some have attempted to balance a perceived bias with one specific control group by using more than one control group (see Example 9.3). 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 (see Example 9.5). 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.

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 extremely helpful 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.

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. In general, 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 breed is a likely confounder, you might include only one breed in the study, usually the dominant one in the source population. This prevents confounding by breed. What we would lose in this approach is the ability to generalise the results to other breeds or to assess interactions with the exposure across the confounder levels (*ie* breeds). All inclusion and exclusion criteria should be stated clearly. **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). Unfortunately, matching often does not work well for either of these objectives in case-control studies (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). 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 estimate disease frequency, 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 below 5-15%. If concurrent sampling is used, then a conditional analysis should be performed (see Section 16.15) and the odds ratio estimates the rate ratio in both closed and open populations. 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 (Knol *et al*, 2008).

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 unexposed 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) indicates 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

von Elm *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 below. As noted earlier, we elaborated these key points in this chapter as they should be used to help plan and report case-control studies, and to help you, the reader, 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

Archer DC, Pinchbeck GL, French NP, Proudman CJ. Risk factors for epiploic foramen entrapment colic in a UK horse population: a prospective case-control study Equine Vet J 2008; 40: 405-10.

Cohen ND, Mackay RJ, Toby E, Andrews FM, Barr BS, Beech J, et al. A multicenter case-control study of risk factors for equine protozoal myeloencephalitis. J Am Vet Med Assoc. 2007;231(12):1857-63.

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

D'Agata EMC. Methodologic issues of case-control studies: a review of established and newly recognized limitations Infect Control Hosp Epidemiol 2005; 26: 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: 771-5.

Dore K, Buxton J, Henry B, Pollari F, Middleton D, Fyfe M, Ahmed R, Michel P, King A, Tinga C, Wilson JB, Multi-Provincial Salmonella Typhimurium Case-Control Study Steering Committee. Risk factors for Salmonella typhimurium DT104 and non-DT104 infection: a Canadian multi-provincial case-control study Epidemiol Infect 2004; 132: 485-93.

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

Grimes DA, Schulz KF. Making sense of odds and odds ratios Obstet Gynec 2008; 111: 423-6.

Gustafson L, Ellis S, Robinson T, Marenghi F, Merrill P, Hawkins L, Giray C, Wagner B. Spatial and non-spatial risk factors associated with cage-level distribution of infectious salmon anaemia at three Atlantic salmon, Salmo salar L., farms in Maine, USA J Fish Dis 2007; 30: 101-9.

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 Epidem 2004; 57: 875-80.

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

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

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

King G, Zeng L. Estimating risk and rate levels, ratios and differences in case-control studies Stat Med 2002; 21: 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: 1073-81.

Kung NY, Morris RS, Perkins NR, Sims LD, Ellis TM, Bissett L, Chow M, Shortridge KF, Guan Y, Peiris MJ. Risk for infection with highly pathogenic influenza A virus (H5N1) in chickens, Hong Kong, 2002 Emerg Infect Dis 2007; 13: 412-8.

Melendez P, Hofer CC, Donovan GA. Risk factors for udder edema and its association with lactation performance on primiparous Holstein cows in a large Florida herd, USA Prev Vet Med 2006; 76: 211-21.

Newton JR, Wood JL, Chanter N. A case control study of factors and infections associated with clinically apparent respiratory disease in UK Thoroughbred racehorses. Prev Vet Med. 2003;60(1):107-32.

Olea-Popelka FJ, Phelan J, White PW, McGrath G, Collins JD, O'keeffe J, Duggan M, Collins DM, Kelton DF, Berke O, More SJ, Martin SW. Quantifying badger exposure and the risk of bovine tuberculosis for cattle herds in county Kilkenny, Ireland Prev Vet Med 2006; 75: 34-46.

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

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

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

Richardson DB. An incidence density sampling program for nested case-control analyses Occup Environ Med. 2004; 61: e59.

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

Rothman K, Greenland S, Lash T. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

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

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

von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies Epidemiology 2007; 18: 800-4.

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles Am J Epidemiol 1992a; 135: 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: 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: 1042-50.

West AM, Martin SW, McEwen SA, Clarke RC, Tamblyn SE. Factors associated with the presence of Salmonella spp. in dairy farm families in southwestern Ontario Can J Public Health 1988; 79: 119-23.

Zhou H, Chen J, Rissanen TH, Korrick SA, Hu H, Salonen JT, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome Epidemiology 2007; 18: 461-8.