10

HYBRID STUDY DESIGNS

OBJECTIVES

After reading this chapter, you should be able to:

1. Describe the key features of each of 5 hybrid study designs (case-cohort, case-crossover, case-case, case-series and case-only).

2. Identify source population characteristics, including types of exposure and outcome, for which these designs are appropriate.

3. Describe 2-stage study designs and identify situations in which the traditional cross-sectional, cohort and case-control studies can benefit from a 2-stage design.

4. Design the basic sampling strategy for a specific 2-stage study.

10.1 INTRODUCTION

In this chapter, we describe 5 variants of traditional observational study designs and a 2-stage design. Each design has its own unique advantages and disadvantages and, although they have been used relatively infrequently, researchers should be aware of the potential of these study designs. The case-crossover is an elaboration on the crossover experimental design that allows the researcher to use only cases in the study by contrasting their exposure in 2 different time periods. The case-case approach uses the traditional case-control sampling strategy and contrasts the exposure of study subjects with an etiologically defined disease with the exposure of other subjects (*ie* controls) with a related etiologically defined disease. The case-series design uses only study subjects with the outcome of interest (*ie* cases) and seeks to identify associations between exposure and outcome using temporal clustering. The case-cohort design incorporates the strengths of the cohort approach with the efficiency of a case-control design. Finally, the case-only study design allows for inferences about interaction (but not main effects) between an exposure and other factors to be made from studies in which only data from cases are available.

Two-stage designs are useful as validation studies and also to enhance the cost-effectiveness of the traditional observational study designs. They allow for collection of readily available data on all subjects and supplement this with more expensive data on covariates which is collected on a random sample of the study subjects.

10.2 CASE-CROSSOVER STUDIES

10.2.1 Basis

This is the observational study analogue to the crossover experimental design where the case serves as its own control. It is most suited for the situation where the exposure is well-defined and transient and the outcome is almost immediate (*ie* the outcome will happen temporally close to the exposure, if the exposure was the cause of the outcome). For validity, the design needs to meet the same assumptions about lag effects (*ie* none or time limited) and duration of disease (*ie* short duration) as in crossover experiments or crossover clinical trials.

The case-crossover design alleviates many of the problems associated with choosing controls in a case-control study. In the case-crossover design, the exposure status of the case just prior to the time of the event occurrence is compared with the exposure status of the same individual at other times. Only subjects that develop the outcome need to be followed; hence, all time-invariant host-related confounders are controlled. This design is only applicable to situations where the exposure status of study subjects (individuals or groups) can change over time. Effect estimates are based on comparing exposure levels just prior to case development with exposure levels at other (control or referent) times (Navidi & Weinhandl, 2002). Maclure (2007), has characterised case-crossover studies as answering the ‘why now’ question, as distinct from the ‘why me’ question which is answered by traditional case-control studies.

10.2.2 Design issues

Initially, and dependent on the biology of the disease and suspected risk factors, we need to

identify the case-risk time. This is the period during which the outcome would likely occur if the association with the exposure was causal. In the context of this study design, it is the time period during which the case subject would be exposed to the suspect causal factor. As an example, we might assume that stray voltage in cattle housing is the exposure of interest and decreased milk production (DMP) is the outcome of interest. For our purposes, we will assume a decrease of 4 or more kg of milk per day is classified as DMP, and that a monitoring mechanism is available to detect stray voltage (in this sense this scenario would be similar to investigations of air pollution as a cause of increased mortality in humans). We will assume that if stray voltage (however defined) is transient and if it were to impact milk production, one would expect that impact to occur within 24 hours of the stray voltage incident. In choosing this risk period, we are being mindful that shortening the length of the risk period to the most reasonable induction period for a specified exposure and outcome will reduce the false detection of exposure-outcome associations (Mittleman, 2005). We will assume that the duration of any potential production impact would also be short (say, 1 or 2 days).

Now, we need to consider the referent or control period selection strategy. Normally, we want the control periods to be temporally close to the time of the index case (this minimises the effects of long-term temporal changes in exposures). However, if there is likely to be a high correlation of exposure level from day to day, then it is best not to choose control periods that are close in time (*eg* the next day) to that of the case. The design controls for time invariant variables, but there is an implicit assumption of no trend in exposure (if binary), or exposure level (if continuous) across the referent window (the length of time between the earliest and latest point at which exposure would be measured for each case). How best to resolve this issue has been the major focus of controversy (Moller *et al*, 2004; Navidi & Weinhandl, 2002).

When this study design was first developed, the selected control periods were earlier than the case times. This is an acceptable approach for obtaining exposure data if the occurrence of the case event might affect subsequent exposures (*eg* if one is studying the impact of a training schedule on an outcome, such as a leg injury, the injury would likely alter the subsequent training period; also see Example 10.2). However, this design is subject to bias from temporal changes in the level of exposure. For example, early studies of the number of speed training sessions in the week prior to limb injuries in horses had to use control periods that pre-dated the suspected exposure. Given an increasing frequency of speed training sessions as the racing season progresses, there is a potential bias associated with using the earlier control period training schedule as, on average, the number of speed training sessions would be lower than experienced subsequently. Thus, more recent designs used symmetric bidirectional designs, especially when environmental exposures such as weather or air pollution were studied. In the symmetric bidirectional protocol, a control-period was selected both before and after the case risk period (usually equally spaced) in the hope that, if exposure or covariate levels were changing over time, the higher and lower exposure values at these times would cancel each other out. When selecting control-periods, they can be matched to the same day of the week as the case, if confounding by day is likely to occur. This has been the most commonly used case-crossover design to date (Janes *et al*, 2005). Nonetheless, Navidi and Weinhandl (2002) recommend using a semi-symmetric bidirectional design that includes only 1 of the 2 potential control-risk periods (the choice of which is selected randomly—see below). A problem with this approach is that for cases that occur early, or late, only 1 of the 2 risk periods may be feasible. The implementation of the suggested selection method is as follows:

Suppose that a case might occur at any time (t_k) in a defined study period from the first day of follow-up (k=1) to the last study day (k=N). To identify the control period for each case, we:

1. Choose a short lag time (L). For example, in the stray voltage example above, it might be 2 days. If day of the week might affect DMP, we could use 7 days. The choice of these would denote the referent window and we assume there is no time-trend in exposure within this window. Decisions about the appropriate lag time are made on a context, and disease-specific basis as shown in Examples 10.2-10.4.
2. Let t_k be the failure time for the j^{th} case.
3. For early cases, if $t_k \leq L$, choose t_{k+L} as the control day.
4. For later cases, if $t_k > (N\text{-}L)$, choose t_{k-L} as the control day.
5. For all other values of t_k, randomly choose half of the control days from before the case time (*ie* t_{k-L}) and half from after (*ie* t_{k+L}).

Note that depending on the context, the duration of the control-period can vary from a single day to a grouping of days (*eg* a week).

Janes *et al* (2005), suggested an improvement to this design when a database of shared exposures (*eg* air pollution in their example; or daily recordings of transient voltages in our example) is available by using a set of time-stratified control-risk periods. In the time-stratified design, the study period (say, the April-June in year 20XX for studying transient voltage and mil production effects) is stratified into weeks . Then, whenever a case arises, all, or a sample, of the remaining days within that week would be used as control days. This leads to unbiased estimates because all cases that arise in a specific stratum (*eg* week) use the same time-window for control periods. Without *a priori* stratification, a one-week time window around the case event time uses different but potentially overlapping time windows for the control period; the extent of overlap depends on the event times and L. For studies of exposures such as stray voltage effects (or air pollution impacts) on health, exposure data for all of the days, in the stratum, can be used if these data are available. Since the exposure is 'shared' the data are available for all cases. If sampling is used, or if the exposures are 'unshared' (*ie* the exposure times are independent across cases as for example in a horse's training pattern) more than one control period (or day) within the referent time window can be selected for each case. More control periods increase statistical power, but, of course, this may demand more detailed follow-up to obtain the data on exposure.

10.2.3 Analysis

The case-crossover design reduces the chance that unmeasured confounding will bias the results. Hence, data can be analysed as a matched case-control study using conditional logistic regression. It turns out that conditional logistic regression gives slightly biased estimates if the bidirectional referent sampling is used (conditional logistic regression with an offset of log(2) for cases with only one referent, and zero otherwise, will produce unbiased estimates (Janes *et al*, 2005). It is unbiased if the time-stratified approach is used.

When exposure data are available for all times within the study period (*eg* exposure data for a 3 month period), we could use exposure data for all of the days in the observational period except for the risk period for each index case occurrence as referent days. In this instance, the case count on each day could be modelled as a Poisson random variable whose mean is a function of the exposure level on that day (Janes *et al*, 2005; Navidi & Weinhandl, 2002). This approach also allows adjustments for overdispersion and autocorrelation in the data. Lu *et al* (2008) and Janes *et al* (2005), make the linkages between conditional logistic regression analysis using multiple time-matched referents in case-crossover studies and Poisson time-series explicit. The

Example 10.1 A case-crossover study of weather events and waterborne disease outbreaks

Thomas *et al* (2006), reported on a study of 92 waterborne disease outbreaks occurring from 1974 to 2001 in Canada. The authors hypothesised that extreme rainfall and spring weather conditions might influence the occurrence of these outbreaks. Data on these exposures were obtained from Environment Canada. Each outbreak of waterborne disease was considered a case and the case-risk time was the six weeks prior to the date of onset of the outbreak. For analysis, the 27 year period was stratified into 6 mutually exclusive time periods. A 6-week control-risk period was selected from each of the remaining 5 non-case periods and matched by month, day and ecozone (describing the location of the outbreak). The data were analysed using forward stepwise conditional logistic regression analysis; ecozone was forced into all models. Two-way interactions involving ecozone and the environmental exposures were considered based on biological plausibility. Warmer temperatures and extreme rainfall were identified as possible contributing factors to the outbreaks.

advantages of using the Poisson approach are that it allows for overdispersion, the fit of the model can be checked using standardised residuals, and influential cases can be identified using standard Poisson model diagnostics (see Section 18.5). Removing influential cases can often change the model results considerably. Examples 10.1-10.2 describe 2 case-crossover studies.

10.3 CASE-CASE STUDIES

10.3.1 Basis

Case-case studies are a variant of case-control studies where the control subjects have the same 'disease' as the case (*eg* the cases might be subjects with *Salmonella typhimurium*, whereas the controls could be subjects with *Salmonella heidelberg* (McCarthy & Giesecke, 1999)). The design was proposed as an optimal study design for identifying risk factors for disease when using data from ongoing surveillance systems for focused subsets of disease (*eg* reportable food and waterborne disease). Since all subjects whose data are in the surveillance database have undergone a similar selection experience, and all subjects have a somewhat similar clinical experience, the design should minimise both selection and recall bias. In this situation, trying to choose a valid set of controls to use in a traditional case-control study approach is very difficult because most potential controls have diseases that are associated with the exposure of interest.

Example 10.2 A case-crossover study within a common source epidemic

Haegebaert *et al* (2003) used a case-crossover design to identify risk factors in a common source foodborne outbreak of salmonellosis. Food exposures during the 3-day risk period before onset of illness were compared to those of a control-period of 3 days that preceded the case-risk period by 2 days. Thirty five confirmed cases, most of whom lived in chronic care institutions, with complete records of food consumption during these periods were identified. The relative risk for each meat product in the diet was estimated using the Mantel-Haenszel odds ratio for matched pairs. The authors discuss the pros and cons of the case-crossover study in this context and noted that the design had the advantage of not requiring the selection of control subjects many of whom might have eaten the same foods but not developed illness because of their physiologic or immune status. In this study, all control risk periods were prior to the case risk periods since the outcome would alter the food intake, and possibly the survival of patients.

For similar reasons, Kaye *et al* (2005) suggest this approach for identifying risk factors for antimicrobial resistance.

10.3.2 Design Issues

In most situations where this design has been used, the controls have the same family (genus) of agent (*eg Salmonella*) but perhaps a different serovar. This design allows us to identify risk factors for diseases that differ by specific serovar as the causal agent (*eg* perhaps turkey versus pork as the major source when investigating foodborne *Salmonella* cases). The design also has been applied to outbreak investigations. In this instance, the control has the same 'strain' of causal agent as the case-disease subject, but does not belong to the outbreak cluster of cases. This application is used to identify exposures associated with being in the set of outbreak cases.

Similar to case-crossover studies, this study design is best suited to situations where the risk factors (*eg* contaminated food) have only a small lag period before they produce their effect. And, similar to secondary-base case-control studies, it is best to select the comparison 'cases' randomly from subjects that have one of a variety of other serotypes, or strains, of the same agent. Control cases also should have entered the surveillance database during the same time period. In general, the design will not identify global risk factors for the onset of disease such as patient characteristics or surrogate risk factors such as 'food item' or 'water source' since many of the subjects in the surveillance system will share these in common. Examples 10.3 and 10.4 demonstrate the utility of this design.

10.4 CASE-SERIES STUDIES

10.4.1 Basis

Recently, a new study design called the self-controlled case-series, or just 'case-series', design has been published (Whitaker, 2008). This design (which might be viewed as a variant of the case-crossover design) can be used to study the temporal association between a time varying exposure and an adverse outcome using only study subjects who experience that outcome. For example, assume we have a defined cohort of study subjects; each study subject will have an observation period during which time the exposure history and outcome events can be observed. Given the knowledge of the potential effects of the exposure, a risk period for each

Example 10.3 A case-case study of 2 *Campylobacter* species

Gillespie *et al* (2002) described a study in which the exposure history of people with *Campylobacter coli* infection was compared with that of cases of *Campylobacter jejuni* infection. Although the former species was much less common it was deemed important to differentiate the risk factors for *C coli* from those for *C jejuni.* Many previous studies tended to examine risk factors for just 1 of the *Campylobacter* species or risk factors for undifferentiated *Campylobacter*. Data were obtained from a population-based surveillance system in England and Wales. Exposure history was obtained from the standard structured questionnaire used as part of the surveillance system. Differences in demographic characteristics in exposure history were assessed using Pearson's chi-square test and the Student t-test. Backward stepwise logistic regression was used to model multiple characteristics and exposures and to investigate potential interactions among the main effects. As we have mentioned, the authors noted that exposures common to both species of *Campylobacter* would not be identified using this study design.

Example 10.4 A case-case study of a Salmonella outbreak

Krumkamp *et al* (2008) investigated a salmonella outbreak that occurred in June and July 2003 in Germany. Data for the affected district were obtained from a routine salmonella surveillance system. Exposure history was collected via telephone interviews 6 weeks after the last outbreak case was notified. There were 10 cases in the outbreak group of S*almonella* strain 1.4.5.12:i-. Two hundred and fifteen other salmonella cases (mostly *Salmonella enteritis* and a variety of less frequent sporadic strains) were reported in the same geographic area in 2003. Ninety-seven control cases were obtained from these 215 cases, the remaining potential control cases had either incomplete information or could not be contacted for the telephone interview. Fisher's exact test and odds ratios were used for analyses. The major and only risk factor identified was meat sold from 1 butcher in the district.

study subject will be defined. The risk period denotes periods during, or after, exposure when the study subjects are deemed to be at increased (or decreased) risk of the outcome (*eg* recall our use of 24 hours for transient voltage). All other times within the observation period constitute control periods. The design is based on using the number of cases arising in the risk period compared with the number of cases arising in the remainder of the observation period after adjusting for the duration of these periods. The advantages of this study design include the fact that only cases need to be studied in detail and all time invariant factors are controlled (*ie* they are not confounders) by the design. Depending on the context, one characteristic that may need control, however, is the age of the study subject; similarly, if the outcome is influenced by factors that vary with season, then season should also be controlled.

10.4.2 Design Issues

The case-series design has been used to study associations between vaccination and untoward events such as idiopathic thrombocytopenic purpura. Given the interest in potential adverse effects of vaccination in veterinary medicine, the application of this design may prove to be of great value in that context also.

Obviously, it is important to clearly define what is meant by exposure and the outcome of interest. Once this is done, data on the case-series can be obtained in either a retrospective or a prospective manner. The design is best suited for studying outcomes that only occur once per study subject; however, multiple outcomes per study subject can be studied provided the outcomes are independent of each other (see comments later). The observation period usually is selected to coincide with the high risk period of the outcome. If age of subject should be controlled, age groups, within which there is unlikely to be confounding by age, should be specified. Similarly, the length of the risk period should be decided. It is possible to subdivide the total risk period into smaller sub-groupings (for example a 3-month risk period could be subdivided into monthly risk periods). If the total risk period used in the study does not include the full-time interval during which the risk of the outcome is altered by the exposure, any resulting association between the exposure and the outcome will be biased toward the null. Formulae for determining sample size are given in Whitaker et al, 2008, and this (or related) publications should be studied for further details on design issues.

Multiple occurrences per study subject are assumed to be independent of each other. If this is unlikely to be a valid assumption, then only first events should be included. It is also assumed that the occurrence of the outcome does not alter the probability of future exposure. Whitaker *et al* (2008) describe methods for coping with this assumption if it is unlikely to be valid. One

strategy is to ignore all post-outcome exposures. A third assumption is that the outcome event should not censor or affect the observation period after its occurrence. That is, it should not alter the survival of the study subject or their participation in the study. Whitaker *et al*, 2008 cite other studies that suggest that the bias from violating this assumption may not be great.

10.4.3 Analysis

Whitaker *et al* (2008) provide examples for structuring and analysing the data. The analysis uses the Poisson regression model where the outcome is the number of events per time interval and the log of the length of the time interval is used as the offset. The measure of association is the *IR* (see Chapter 18). Specific codings for the analysis are available at http://statistics.open.ac.uk/sccs/

10.5 CASE-COHORT STUDIES

10.5.1 Basis

The case-cohort design has the same advantages and disadvantages as a full cohort study, but it allows for an efficient study design when disease is infrequent, and the cost of obtaining covariate information is expensive. The basis of the design is that a random sample of all subjects in the full cohort is obtained at the start of the study; this serves as the 'control-cohort'. The full cohort is observed for the study period and all cases arising in the full cohort (including the control-cohort) are included in the study; their exposure and covariates are compared with those of the study subjects in the control cohort who did not develop the outcome of interest (Kulathinal *et al*, 2007). The design also can be modified when the outcome(s) is not rare by sampling only some of the cases from the full cohort. The case-cohort design allows us to study a number of different outcomes from a specific exposure, and it is especially efficient if biological samples can be obtained from the control-cohort at the study outset and stored for later analysis.

As noted, a major advantage of the case-cohort approach is that the one control-cohort can provide the basis of comparison for a series of outcomes, thus allowing the investigation of associations among more than one disease (or different definitions of the same disease) and a defined exposure (as in a regular cohort study), but without having to follow the entire population at risk. The disease frequency can be estimated using the data from the control-cohort.

10.5.2 Design issues

If the original full cohort is closed (see section 8.7.1), then a risk-based design, which is particularly suited to studying permanent risk factors, can be used. In this design, the control-cohort is selected from the at-risk members of the full cohort at the start of the study using random sampling (without replacement) and the subjects in this sample that do not become cases during the study period serve as the control series. Information about covariates and exposure status is obtained from cases arising outside of the control cohort. If the outcome frequency is high, a significant proportion of the subjects in the control cohort will become cases; hence, the number initially sampled for the control-cohort should be adjusted upward to

compensate for this. For valid inferences, if significant losses to follow-up are present, we must demonstrate that the reasons for loss are not related to the risk of developing the outcome(s) of interest.

If the original cohort is open, the control-cohort is selected from the at-risk members of the full cohort at the start of the study using random sampling without replacement. The full-cohort and the subjects in this sample are observed repeatedly during the study period either at fixed intervals (*eg* every 3 months) or, more commonly, at the point when cases occur. When the timing of control selection is matched to the time the case occurred, only one outcome can be studied. In the original case-cohort design, all cases in the full cohort would be observed and their characteristics compared with all, or a sample of, those not having developed the disease by that point (either the fixed time point or when each case occurred) in the control-cohort. If the disease is more common, only a sample of cases from outside of the control-cohort need be included in the study (Pfeiffer *et al*, 2005). If the exposure and covariates are permanent, the status of the cases can be assessed as of the time of occurrence, whereas the status of members of the control-cohort can be assessed at the start of the study. All members of the control-cohort who have not developed the outcome at the time the case occurred are eligible for inclusion as control subjects, and all, or a sample of them, can be used in the analysis.

The case-cohort design is not optimal if exposure status can change during the study period, since additional data are required to establish the exposure status of subjects in the control-cohort at the time the outcome occurred. Sometimes serially stored specimens are available for this, or data from external sources can be used, For example in a study of the effects of air pollution on health, historical records of air pollution levels might suffice to establish the exposure of cases and members of the control-cohort at different points in time during the study period.

In this design, consideration needs to be given to the requirements for obtaining exposure data or biological specimens from study subjects. Only subjects likely to meet these requirements should be considered at-risk. Note that if several outcomes will be assessed, exposure and covariate data are needed on each of the cases as well as all members of the control-cohort (Kulathinal *et al*, 2007). When selecting the original control-cohort, the subjects can be sampled using stratified sampling to ensure that the covariate patterns of the control-cohort are similar to those of the anticipated (future) cases (Kulathinal *et al*, 2007). For example, if young animals have a higher risk of the outcome than older animals, the control-cohort can be selected in a manner to ensure that the majority of study subjects in the control-cohort will be young animals.

10.5.3 Analysis

At the end of the study period, there will be records of the number of cases arising from within the control-cohort, the number of cases arising outside the control-cohort and the remaining number of non-cases in the control-cohort. If a risk-based design is appropriate, you can combine (*ie* add) the 2 types of cases together, and the data can be analysed in a 2X2 table using a case-control format with the odds ratio (*OR*) as the measure of association. If direct estimates of risk are required, then the 2 types of case need to be differentiated.

The analysis is more complex if exposure data for cases and controls is obtained at the time each case occurred. Many researchers with open-population studies use survival methods and hence hazard ratios, for analyses (Kulathinal *et al*, 2007). Historically, authors have proposed 3

different weighting schemes in the Cox model that account for whether the cases come from the full or control-cohort (Onland-Moret *et al*, 2007) and the choice of these weights is available in modern computing packages (Prentice's method provides estimates that most closely resemble estimates from the full cohort). Other analytic methods are available when not all cases from the full cohort are used in the study (Pfeiffer *et al*, 2005). Cai and Zeng (2007) provide methods for determining power when subsampling of cases is used and in the simpler situation when all cases are used in the analyses. Kim *et al* (2006) show that using the case-control approach to estimating sample size works well and is simple to implement. Example 10.5 describes a case-cohort study.

10.6 CASE-ONLY STUDIES

This design was originally conceived for use when the exposure status of the 'controls' could be predicted without having an explicit control group (*eg* in genetic studies, the distribution of exposure in the 'controls' is derived from theoretical grounds such as the blood-type distribution in the source population). Underlying the design, which is highly efficient relative to case-control designs, lies a strong assumption about independence between the gene frequency and other environmental factors. Specifically, the genes being studied need to be inherited and not mutations which might be caused by the environmental exposures. Thus, the case-only design may be used to study interaction between a covariate (not necessarily a genetic factor) and an exposure provided the exposure and the covariate of interest are independent of each other (Rosenbaum, 2004). Schwartz (2005) provides a good introduction to the basic design and analysis of case-only studies (see also Example 10.6).

Recently, the design has been applied to the study of the effects of non-genetic risk factors such as personal-level risk factors (*eg* age, race, behaviours) and factors related to socio-economic class on the risk of mortality. As noted, because the design does not include control-subjects or control-times, analyses are limited to identifying interactions; main effects cannot be determined. For example, the design has been used to assess if personal characteristics interact

Example 10.5 A case-cohort study of drinking water quality and risk of stomach cancer

Auvinen *et al* (2005) reported on a study of radon and other radionuclides in drinking water and the risk of stomach cancer. The subjects of interest were those who obtained their drinking water from drilled wells and this comprised a base population of over 144,000 people during the presumed exposure period from 1967 to 1980. A initial control-cohort of 4,590 subjects was selected as the referent group using random sampling after stratifying by age and sex. However, most of these subjects were not long-term users of drilled well water; only 371 subjects had used drinking water from drilled wells prior to 1981. These became the effective control-cohort of interest for the study. The occurrence of stomach cancer up to January 1, 1996 was identified through a cancer registry; 107 cases using drinking water from drilled wells prior to 1981 were identified; none were from the control-cohort.

Information on the characteristics of wells was obtained directly from the study subjects, proxy respondents, current residents of the dwellings, and local health authorities. Water samples were collected blindly with regard to case status between July and November 1996 and analysed for radon and other radionuclides; about 80 percent of the cases and the effective control-cohort subjects had water samples tested. Data analysis was based on a proportional hazard model. This approach takes account of how long each study subject was exposed to a particular level of radon each time a case occurred. All statistically significant hazard ratios were below one suggesting a sparing effect of radon levels on stomach cancer.

Example 10.6 Case-only study of potential effect modifiers of risk of death in humans

Schwartz (2005) investigated whether sex, nonwhite status, or age greater than 85 years were modifiers of the effect of temperature extremes on the number of deaths in Wayne County, Michigan. Data on weather were obtained from a near-by meteorological station and the days with excessive hot and cold weather were identified. Two periods of time were investigated; one focused on a single day, and the second on a 3-day average of events. Data on the potential effect modifiers were obtained from medical records of people who died. Separate models for excessive heat and cold were developed. The results indicated that depending on the temperature extreme, all 3 covariates interacted with the temperature extreme and affected the number of deaths.

with extreme weather (Medina-Ramon *et al*, 2006) and if socio-economic class interacts with weather to modify the risk of death (Armstrong, 2003).

10.6.1 Analysis

Armstrong (2003) describes the analytical approach, and how the choice of model depends on the nature of the potential interacting variable of interest.

Assume that we can use a Poisson model to investigate the association of the number subjects experiencing the outcome (Y) as a function of a binary exposure and a binary covariate (*eg* sex). The model, including the potential interaction between *exposure* and *sex*, might look like:

$$\ln \mathrm{E}(Y) = \beta_0 + \beta_1(exposure) + \beta_2(sex) + \beta_3(exposure * sex)$$

and we could use this model to create a 2X2 table of expected outcome event counts according to the 4 combinations of exposure and sex. In turn, we could then create an odds ratio of these counts which would reflect any interaction between the exposure and sex (*ie* the β_3 term). It turns out that this is equivalent to a logistic model of sex as a function of the exposure

$$\mathrm{logit}(sex = 1) = \beta_0 + \beta_3(exposure)$$

If β_3 is significant, it indicates that sex is an effect modifier for the exposure in terms of the outcome of interest. This is the basis of testing for interaction in case-only studies.

Example 10.6 outlines a typical case-only study while Example 10.7 is an (atypical?) case-only study of risk factors for dog bites.

10.7 TWO-STAGE SAMPLING DESIGNS

A 2-stage sampling design can be applied to the traditional cohort, case-control or cross-sectional study designs. There are numerous uses of the term '2-stage' but herein it refers to studies where information on the exposure and outcome of concern is gathered on an appropriate number of first-stage subjects (*ie* the number of subjects based on sample-size estimates) and then, a sample of the study subjects is selected for a second-stage study in which more detailed information (and often more expensive exposure or covariate data) is collected. This approach is very efficient when the cost of obtaining the data on covariates is high. The design also fits the situation where a valid measure of the exposure of interest is very expensive to obtain, but an inexpensive surrogate measure is available. The surrogate measure is applied to all study subjects, then a more detailed work-up is performed on a subsample of the study

Example 10.7 Potential risk factors for dog bites in Greece

A total of 2,642 cases of dog bite reported at emergency centres in Greece occurred between May 1, 1996 and December 21, 1999 (Frangakis & Petridou, 2003). Of the total cases, 61% were males, the average age was 26 and one-third of the cases was below 11 years of age. Putative risk factors included week day vs weekend, season and time period since the last full moon. Later, gender and age were included in the model. Given a stable population size, inferences about the main effect of week day and season could be made. The time between being bitten and last full moon was included and compared with the length of time between adjacent full moons to identify clustering close to the full moon. The average length of the moon-biting period was 17 days, so no clustering was detected. Males and youth had an excess risk. The inclusion of gender and age allowed inferences about their relative impact given the population of bitten subjects (*ie* they were not necessarily risk factors for being bitten but may have modified the effect of other factors). No interaction terms were included in the model.

This case-only study differs from the majority of case-only studies (explained in the text; Section 10.6) whose main aim is to identify interactions between covariates of interest and environmental exposures. We also note that a case-series approach could have been used for this study.

subjects to more accurately determine the true exposure status. The approach also can be used to obtain data on variables for which there are numerous missing values. Instead of assuming that the data are missing at random, the study subjects with missing data can be the subject of a second-stage data collection effort. As discussed in Section 12.8, the 2-stage approach is the basis of validation substudies (McNamee, 2002).

A key question in a 2-stage design is what sample size should be used for the second stage? There are a number of approaches but, as Hanley *et al* (2005) noted, the tools available have not been greatly improved in the past decade. In cohort studies, we can take a fixed number of exposed and non-exposed subjects. In a case-control study, we could take a fixed number of cases and controls. However, for optimal efficiency, it is better to stratify on the 4 exposure-disease categories (present in a 2X2 table) and take an approximately equal number of subjects from each of the 4 categories. This might involve taking all of the subjects in certain exposure-disease categories and a sample of subjects in others.

Cain and Breslow (1988) developed the methodology to analyse 2-stage data using logistic regression. Hanley *et al* (2005) give a worked example of calculating the adjusted odds ratio and its variance. Essentially, one uses the adjusted odds ratio from stage 2 as the adjusted estimate of association between the exposure and disease. The variance of the estimate is based on the variance of the stage 2 odds ratio with adjustments for the sample sizes used in each stage. The approach to obtaining correct variance estimates is somewhat more complex, with multiple confounders, but is relatively simple to implement if the data are all dichotomous (see Hanley *et al* (2005) for details). Example 10.8 describes a 2-stage study design.

Example 10.8 A 2-stage case-control study of risk factors for diarrhea and *Cryptosporidium* in dairy calves

Trotz-Williams *et al* (2007), report on a study investigating calf-level risk factors for diarrhea and shedding of *Cryptosporidium parvum* in dairy calves in Ontario. Over 900 calves on 11 dairy farms were enrolled in the study. At the first stage, all calves were examined for both *C. parvum* and diarrhea. At the second stage, 25% of the fecal specimens were selected for bacterial and viral pathogen isolation; approximately equal numbers of calves were selected from the 4 categories of calves formed by cross-tabulating the presence or absence of *C. parvum* with the presence or absence of diarrhea. The method of Cain and Breslow (1988) was used to estimate the odds ratio reflecting the association between *C. parvum* and diarrhea after adjustment for the effects of other potential pathogens that might confound the association.

REFERENCES

Armstrong BG. Fixed factors that modify the effects of time-varying factors: applying the case-only approach Epidemiology 2003; 14: 467-72.

Auvinen A, Salonen L, Pekkanen J, Pukkala E, Ilus T, Kurttio P. Radon and other natural radionuclides in drinking water and risk of stomach cancer: a case-cohort study in Finland Int J Cancer 2005; 114: 109-13.

Cai J, Zeng D. Power calculation for case-cohort studies with nonrare events Biometrics 2007; 63: 1288-95.

Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies Am J Epidemiol 1988; 128: 1198-206.

Frangakis CE, Petridou E. Modelling risk factors for injuries from dog bites in Greece: a case-only design and analysis Accident Analysis and Prev 2003; 35: 435-8.

Gillespie IA, O'Brien SJ, Frost JA, Adak GK, Horby P, Swan AV, Painter MJ, Neal KR, *Campylobacter* Sentinel Surveillance Scheme Collaborators. A case-case comparison of *Campylobacter coli* and *Campylobacter jejuni* infection: a tool for generating hypotheses Emerg Inf Dis 2002; 8: 937-42.

Haegebaert S, Duche L, Desenclos JC. The use of the case-crossover design in a continuous common source food-borne outbreak Epidemiol and Inf 2003; 131: 809-13.

Hanley JA, Csizmadi I, Collet JP. Two-stage case-control studies: precision of parameter estimates and considerations in selecting sample size Am J Epidemiol 2005; 162: 1225-34.

Janes H, Sheppard L, Lumley T. Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias Epidemiology 2005; 16: 717-26.

Kaye KS, Harris AD, Samore M, Carmeli Y. The case-case-control study design: addressing the limitations of risk factor studies for antimicrobial resistance. Inf Control and Hosp Epidemiol 2005; 26: 346-51.

Kim MY, Xue X, Du Y. Approaches for calculating power for case-cohort studies Biometrics 2006; 62: 929-33.

Krumkamp R, Reintjes R, Dirksen-Fischer M. Case-case study of a Salmonella outbreak: an epidemiologic method to analyse surveillance data Int J Hyg and Env Health 2008; 211: 163-7.

Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K, the MORGAM Project. Case-cohort design in practice - experiences from the MORGAM Project Epidemiol Perspect Innov 2007; 4: 15.

Lu Y, Symons JM, Geyh AS, Zeger SL. An approach to checking case-crossover analyses based on equivalence with time-series methods Epidemiology 2008; 19: 169-75.

Maclure M. 'Why me?' versus 'why now?'—differences between operational hypotheses in case-control versus case-crossover studies Pharmacoepidemiology and drug safety 2007; 16: 850-3.

McCarthy N, Giesecke J. Case-case comparisons to study causation of common infectious diseases Int J Epidemiol 1999; 28: 764-8.

McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value Stat Med 2002; 21: 3609-25.

Medina-Ramon M, Zanobetti A, Cavanagh DP, Schwartz J. Extreme temperatures and mortality: assessing effect modification by personal characteristics and specific cause of death in a multi-city case-only analysis Environmental health perspectives 2006; 114: 1331-6.

Mittleman MA. Optimal referent selection strategies in case-crossover studies: a settled issue Epidemiology 2005; 16: 715-6.

Moller J, Hessen-Soderman AC, Hallqvist J. Differential misclassification of exposure in case-crossover studies Epidemiology 2004; 15: 589-96.

Navidi W, Weinhandl E. Risk set sampling for case-crossover designs Epidemiology 2002; 13: 100-5.

Onland-Moret NC, van der ADL, van der Schouw YT, Buschers W, Elias SG, van Gils CH, Koerselman J, Roest M, Grobbee DE, Peeters PH. Analysis of case-cohort data: a comparison of different methods J Clin Epidemiol 2007; 60: 350-5.

Pfeiffer RM, Ryan L, Litonjua A, Pee D. A case-cohort design for assessing covariate effects in longitudinal studies Biometrics 2005; 61: 982-91.

Rosenbaum PR. The case-only odds ratio as a causal parameter Biometrics 2004; 60: 233-40.

Schwartz J. Who is sensitive to extremes of temperature?: A case-only analysis Epidemiology 2005; 16: 67-72.

Thomas KM, Charron DF, Waltner-Toews D, Schuster C, Maarouf AR, Holt JD. A role of high impact weather events in waterborne disease outbreaks in Canada, 1975-2001 Int J Env Health Res 2006; 16: 167-80.

Trotz-Williams LA, Martin SW, Leslie KE, Duffield T, Nydam DV, Peregrine AS. Calf-level risk factors for neonatal diarrhea and shedding of *Cryptosporidium parvum* in Ontario dairy calves. Prev Vet Med 2007; 82: 12-28.

Whitaker H. The self controlled case series method BMJ 2008; 337: a1069.