10

HYBRID STUDY DESIGNS

Objectives

After reading this chapter, you should be able to:

1. Describe the key features of each of 6 hybrid study designs (case-crossover, case-series, case-case, case-only, case-cohort, and case-case-control).

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 case-control study.

10.1 Introduction

In this chapter, we describe 6 variants of traditional observational study designs and a 2-stage design. Four of the variants involve cases only, while 2 involve control groups. Each design has its unique advantages and disadvantages.

Studies involving cases only:

- 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. See Section 10.2.
- The case-series design also uses only study subjects with the outcome of interest (*ie* cases), and seeks to identify associations between exposure and outcome using temporal clustering of cases around exposure. Smeeth *et al* (2006) give an excellent overview of these study designs plus numerous examples of their use. See Section 10.5.
- The case-case approach is based on the traditional case-control sampling strategy and typically 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. See Section 10.3.
- The case-only design is used when the level of exposure in a hypothetical control group can be predicted on a biological basis. It allows for inferences about interactions, but not main effects, between an exposure and other risk factors using data from cases only. We note that there can be some confusion in study design terminology because all of the above study types use data from cases only; however, design-wise, this is a specific study design denoted as case-only. See Section 10.7.

Studies involving control groups:

- The case-cohort design includes both cases and non-cases, and incorporates the strengths of the cohort approach with the efficiency of a case-control design. See Section 10.6.
- The case-case-control study is similar to a case-control study, except that it involves 2 distinct case series. It has been used to evaluate factors associated with antibiotic resistance. See Section 10.4.

The last hybrid study we describe here—the two-stage design—is useful as a validation study and also to enhance the cost-effectiveness of the traditional observational study designs. The design allows for collection of readily available data on all subjects and supplementing these data with (the usually) more expensive data on selected covariates from a random sample of the study subjects. These are described in Section 10.8.

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. The level, or fact, of exposure just prior to the outcome case is contrasted to that of exposure in other time periods. 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. Delaney and Suissa (2009) discuss the design in the context of pharmaco-epidemiology. They note that because of the lag effects of many drugs, as well as the impact of the disease on future exposure, the bidirectional design (see Section 10.2.2) where control periods are identified after the incident case often cannot be used. Maclure (2012) describes case-crossover and case-series methods with a focus on pharmaco-epidemiology.

The case-crossover design alleviates many of the problems associated with choosing controls in a case-control study, in that 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 (Navidi and Weinhandl, 2002). Since the same case subject serves as its own control in one or more different time periods, all time-invariant host-related confounders are controlled. 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. The design is used frequently to assess the impact of air pollution or weather on morbidity and mortality (see Carracedo-Martinez *et al* (2010) for a systematic review of its use in this research area).

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. In choosing this risk period, be 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). The risk period for factors such as the potential impact of physical exertion on myocardial infarction might be a few hours, whereas if studying the potential impact of mobile phone use on automobile accidents, it might be 5 minutes (Sato *et al*, 2010). The most common risk period in studies of air pollution impacts on health is one day (Carracedo-Martinez *et al*, 2010); in Example 10.1, the risk period is 6 weeks.

Next, 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

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

Thomas *et al* (2006) studied 92 waterborne disease outbreaks occurring from 1975 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.

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 reside too close in time (*eg* the day previous) to that of the case. The design controls for time invariant variables, but there is an implicit assumption of no trend in exposure prevalence (if binary), or exposure level (if continuous) across the referent window (*ie* the risk period; 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 (Navidi and Weinhandl, 2002; Moller *et al*, 2004).

When this study design was first developed, the selected control periods usually were earlier than the case times. This design has problems with changes in exposure level over time, but 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, such as the distance run in the previous week, with a leg injury, the injury itself would likely alter the subsequent training period). For example, in a study of the potential impact of alcohol consumption on accidental injury, the risk period was set at 6 hours with control periods at the same time the day before and 7 days before the incident injury (Thornley *et al*, 2011; Williams *et al*, 2011). Wang *et al* (2011a) describe two variations in study design—the case-time-control study and the case-case-time-control study—with emphasis on studies in pharmaco-epidemiology where the outcome changes the future exposure. Darrow (2010), recognising the value of bidirectional designs, discusses the selection of control periods in settings where the subject is no longer at risk following the incident event (*eg* if the outcome is death).

More recently, most designs have used symmetric bidirectional designs to counteract potential changes in exposure level or frequency. In this design, a control period is selected both before and after the case occurs (usually equally spaced) in the hope that, if exposure or covariate levels are changing over time, the higher and lower exposure values at these times would cancel each other out. Control periods can be matched to the same day of the week as the case, if confounding by day is likely to occur (Janes *et al*, 2005). Navidi and Weinhandl (2002) had recommended using a semi-symmetric bidirectional design that includes only 1 of the 2 potential control-risk periods (the choice of which is selected randomly); however, recently, Carracedo-Martínez *et al* (2010) have noted that the semi-symmetric design is rarely used for studies of the impact of air pollution, largely because of its decreased power. These same authors provide a step-by-step guide to conducting a symmetric case-crossover study of the impact of air pollutants on health.

One problem with the symmetric approach is that for cases that occur early or late in the study period, 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 (k=1) to the last study day (k=N). To identify the control periods for each case we:

1. Choose a suitable lag time, L, to separate the case occurrence and control periods. For example, it might be L=1 week.
2. Let t_k be the failure time for the j^{th} case.
3. Choose t_{k+L} as the control day for early cases, if $t_k \leq L$.
4. Choose t_{k-L} as the control day for later cases, if $t_k > (N-L)$.
5. Choose control days as t_{k-L} and t_{k+Lf} for most cases, if $t_k > L$ *and* $< (N-L)$.

An extension of this approach is called the full-stratum bidirectional referent selection in which the referents are all time periods (*eg* days) in the follow-up period other than the index day.

Navidi and Weinhandl (2002) describe how a case–crossover analysis with full-stratum bidirectional referents and a shared exposure series (*eg* air pollution exposure) is equivalent to a Poisson regression analysis.

Janes *et al* (2005) suggested an improvement to this design when a database of shared exposures (*eg* air pollution in their example) is available by using a set of time-stratified control-risk periods. In this setting, air pollution data are available for each day and are not (directly) impacted by the occurrence of the incident event. It is nonetheless possible that after a serious asthma attack on a day with high levels of pollution, the affected subjects might change their behaviour and stay indoors. In the time-stratified design, the study period (say, the summer of 2012) could be stratified into months (*a priori*). Then, whenever a case arises, say on a Wednesday in July, all of the other Wednesdays within July would be used as control days (again this fits the setting where exposure data are available at no/low cost for these periods); essentially this approach matches on day and month, and is used instead of selecting the lag time L to identify control periods as shown above. The exposure is 'shared' since the data are available for all cases on a given day. If sampling is used, or if the exposures are 'unshared' (*ie* the exposure is independent across cases, for example, in an athlete'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; in a simple design with just one control period, the data layout is that of a 2X2 table, and McNemar's test can be used to test for an association. With more than one control period, most researchers use conditional logistic regression (Smeeth *et al*, 2006) for analysis. The parameter e^{β} represents the change in the odds of an event associated with a short-term 1-unit increase in exposure. If the exposure changes over time, an extension of the case-crossover design, the case-time-control design (Suissa, 1995) can be used. Kim *et al* (2011) discuss methods to allow for the detection of effect modification by the covariates that are included in the matching process; the usual conditional logistic model cannot assess the effect of the covariates (*eg* day of the week) included in the matching process, but can assess interaction.

When exposure data are available for all risk periods within the study period (*eg* daily 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 and 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 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. Having said this, Carracedo-Martínez *et al* (2010) indicate that most researchers continue to use a logistic model to analyse their data because the Poisson models often have problems with convergence. Wang *et al*

(2011b) demonstrated that when fitting a conditional logistic model to time-stratified data with many cases sharing the same exposure data, it is important to use Breslow's method for handling ties, otherwise biased estimates of exposure effects are likely when using SAS; the bias was not seen when using R or Stata. Examples 10.1 and 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 and Giesecke, 1999)). The design was proposed as optimal 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 would be very difficult because most potential controls have diseases that are associated with the exposure(s) of interest. For similar reasons, Kaye *et al* (2005) suggested this approach for identifying risk factors for antimicrobial resistance. The design has also been described for elaborating risk factors for different molecularly defined subtypes of breast cancer (Martinez *et al*, 2010).

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 similar endemic diseases that have a different specific serovar as the causal agent (*eg* perhaps turkey versus pork as the major source when investigating food-borne *Salmonella* cases). The design also has been applied to outbreak investigations. In this instance, the control subjects have the same 'strain' of causal agent as the case-disease subject, but they do not belong to the outbreak cluster of cases. This application is used to identify exposures associated

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 food-borne outbreak of salmonellosis. Food exposures during the 3-day risk period before onset of illness were compared with 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 note 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 physiological or immune status. In this study, all control-risk periods were prior to the case-risk periods since the outcome, as well as the passage of time, would alter the food items consumed by these patients.

with being in the set of outbreak cases rather than a sporadic case of salmonellosis (see Example 10.4).

Similar to case-crossover studies, this study design is best suited to situations where the risk factors (*eg* contaminated food) have only a short incubation or induction period before they produce their effect. Similar to secondary-base case-control studies, it is best to select the comparison 'cases' randomly from subjects who 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 calendar-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. It can, however, help identify specific risk factors that relate to the risk of having clinical disease from a particular organism. Examples 10.3 and 10.4 demonstrate the utility of this design.

Wilson *et al* (2008) discussed some limitations of case-case studies including selection bias (tendency for only the more serious cases of the disease to be reported); information bias (most of the data are collected and recorded by people who know the specific outcome); confounding (because of the lack of information on confounders in most surveillance databases); and a lack of detail on exposure. Nonetheless, they found good agreement between the results of case-case studies and other methods applied to routine surveillance data.

10.3.3 Analysis of case-case studies

The data from a case-case study can be analysed by the same techniques as risk-based, case-control studies; namely logistic regression. Note that since the exposure in the control-cases does not estimate the level of exposure in the source population, the odds ratio is not a true risk measure. Rather, it reflects the relative differences in exposure level between two subtypes of one disease.

10.4 CASE-CASE-CONTROL STUDIES

This study design was developed to overcome limitations of traditional case-control studies when applied to the study of risk factors for antimicrobial resistant organisms (Kaye *et al*, 2005). The example used by these authors was the study of risk factors for vancomycin-

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

Gillespie *et al* (2002) describe 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 is 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 one 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 χ^2 test and the Student's 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 one butcher in the district.

susceptible *Enterococcus* (VSE) and vancomycin-resistant *Enterococcus* (VRE). This organism is common in hospital settings.

10.4.1 Design issues

This design is similar to a traditional case-control study, except that 2 sets of cases are used—one the subjects with VSE, the other subjects with VRE. In a hospital setting, the control group would be chosen from other hospitalised patients who have either VSE or VRE. The authors recommend that only the first positive culture result for a patient be included, and if nosocomial infections are of interest, the cases should include only the first positive culture taken more than 48 hours after admission. Control subjects should come from the same source population as the cases; thus one must consider whether or not to use other patients, or other subjects attending a local primary care facility, or randomly choosing people from the presumed source population. For nosocomial infections, the controls should be selected from other patients that have been hospitalised more than 48 hours. In either instance, it would be important to know that the controls were culture negative for both VSE and VRE.

10.4.2 Analysis

Analysis of case-case-control studies can be accomplished similar to a risk-based case-control study by using logistic regression applied to each case series separately (but using the single control series). Because there is only one control set, it is difficult to use restricted sampling or matching when selecting controls. Thus, control of confounders is by multivariable modelling using unconditional logistic regression. Inferences about factors leading to VRE are made by comparison of the findings in the VSE and VRE models (Kaye *et al*, 2005). Category A variables are those present only in the model for the resistant phenotype of the target organism. These variables represent unique risk factors for resistant cases. Category B includes risk factors present only in the model for the susceptible phenotype of the target organism. These variables represent unique risk factors for the susceptible phenotype.

Category C variables are present in both models and represent risk factors for the target organism in general. The authors argue that this design is better than a case-case design, which would contrast the exposure history of VSE subjects directly with that of VRE subjects. Their basis for this argument is that VRE appears to arise from external sources and "the likelihood of de novo emergence of vancomycin resistance in a susceptible endogenous strain of *Enterococcus* is negligible."

As another example, Melo and Fortaleza (2009) studied nasopharyngeal colonisation with methicillin-resistant *Staphylococcus aureus* (MRSA). To identify risk factors for MRSA colonisation, they conducted a case-case-control study, enrolling 122 patients admitted to a medical-surgical intensive care unit (ICU). All patients had been screened for nasopharyngeal colonisation with *S. aureus* upon admission and weekly thereafter. The 2 sets of cases used patients who acquired colonisation with MRSA and methicillin-susceptible *S. aureus* (MSSA), respectively. Control subjects were patients in whom colonisation was not detected during an ICU stay.

10.5 Case-series studies

10.5.1 Basis

Recently, a new study design called the self-controlled case-series, or just 'case-series', design has been published (Whitaker *et al*, 2006; 2009). 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 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* this often ranges from 6–35 days for febrile conditions post-vaccination depending on the specific vaccine components). 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.

10.5.2 Design issues

The case-series design has been used to study associations between vaccination and a host of untoward health events (see Weldeselassie *et al* (2011)) for a thorough review of this study design).

One of the first design considerations is to define the outcome of interest. Then, we need to specify the (usually) calendar period in which the subjects will be observed (the observation period) for the outcome event and the source population for cases. Once these are established, data on the case-series can be obtained in either a retrospective or a prospective manner. Obviously, it is important to clearly define what is meant by exposure and the outcome of interest; for example, vaccination with a measles, mumps, and rubella vaccine. The vaccination date (more generally the specified exposure) of each case is used to define one or more risk periods, during which individuals are hypothesised to be at increased (or reduced) risk of the event of interest after (or, for reasons to be discussed later, before) vaccination. All other time

within an individual's observation period, that does not fall within a risk period, is included in that individual's control period.

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, and include, the presumed 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 subgroupings (for example a 3-month risk period could be subdivided into 3, 1-month risk periods). If the total observation 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* (2009), and this (or related) publication should be studied for further details on design issues.

A basic assumption is that the occurrence of the outcome does not alter the probability of future exposure. Whitaker *et al* (2009) describe methods for coping with this assumption if it is unlikely to be valid. One strategy is to ignore all post-outcome exposures (*ie* second vaccinations). Also, 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* (2009) cite other studies that suggest that the bias from violating this assumption may not be great. Multiple occurrences per study subject can be included provided they are independent of each other. If this is unlikely to be a valid assumption, then only first events should be included (see Example 10.5).

10.5.3 Analysis

Estimation of parameters is most readily achieved by fitting a conditional Poisson regression model. The parameter of interest is the relative incidence, which is the incidence in a risk period relative to the control periods. A tutorial is available with full practical details and worked examples (Whitaker *et al*, 2006).

Whitaker *et al* (2009) provide other examples for structuring and analysing the data. The analysis uses the Poisson regression model where the outcome is the number of events per risk and control time interval and the log of the length of each time interval is used as the offset. The

Example 10.5 Risk of falls associated with anti-hypertensive medication: A case-series study

Gribbin *et al* (2011) used data from a database containing the diagnostic and prescription data recorded by primary-care physicians from 386 general practices who used a specific practice management system in the United Kingdom. Cases of falling (n=9862) in the years 2003–2006 in patients 60 years or older were obtained. Based on an analysis of elapsed time between prescriptions for a particular class of antihypertensive, the authors calculated episodes of continuous exposure of not more than 60 days. After the prescription was initiated for each patient, periods of exposure were defined (and subdivided into day 0, days 1–21 and day 22–60). All remaining person-time was used as the baseline (unexposed) comparison period.

Poisson regression was used to estimate incidence rate ratios (*IR*s) for the different periods of exposure.

measure of association is the relative *IR* (see Chapter 18). Specific codings for the analysis are available at http://statistics.open.ac.uk/sccs/.

10.6 Case-cohort studies

10.6.1 Basis

The case-cohort design has the same advantages as a full cohort study, but it has the additional advantage of being an efficient study design when disease is infrequent, and the cost of obtaining covariate (including exposure) 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' and cases arise from these subjects. Since most diseases are infrequent, there would be insufficient cases in the control cohort to provide reasonable power to the study. Thus, 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. The exposure and covariate data in the case subjects are compared with those of the study subjects in the control cohort who did not develop the outcome of interest (risk-based design), or had not developed the outcome at the time of case occurrence (rate-based design) (Kulathinal *et al*, 2007). The design also can be modified when the outcome is not rare by sampling only some of the cases from the full cohort. 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. The design is especially efficient if biological samples can be obtained from the control-cohort at the study outset and stored for later analysis.

10.6.2 Design issues

Initially, we need to define the eligible cohort based on such information as having a health history, being willing to provide details on personal and lifestyle characteristics (age, race, sex, weight, smoking status, nutritional survey *etc*) and be willing to provide essential tissue samples (*eg* blood sample). The subcohort can be drawn from this eligible group using simple random sampling without replacement, or the eligible cohort can be classified based on a few key variables that might confound the study results and the subcohort drawn using stratified random sampling.

If the original full cohort is a closed population (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. The essential information alluded to above regarding 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. As an example of a risk-based study, Matsuda *et al* (2011) conducted an analysis of factors associated with placental abruption and placenta previa, based on a subcohort of 5,036 of 242,715 births in Japan. A multivariable unconditional logistic model was used for analysis with the odds ratio being an estimate of the relative risk.

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 baseline characteristics of the control cohort would be obtained and this group would be followed, usually by regular surveys to update exposure and covariate data. For example, Agalliu *et al* (2011) followed a subcohort of 1,979 men from their enrolment date (between 1995 and 1998) to the end of 2003 using food intake and nutritional-supplement surveys in a study of risk factors for prostate cancer. Men in the full cohort with existing prostate cancer at enrolment were excluded. All cases arising in the full cohort and in the subcohort were identified, and their exposure and covariate information at the time of becoming a case recorded. If the disease is 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 (this arises naturally in a rate-based proportional hazard model providing the data are structured correctly).

If exposure status can change during the study period, depending on the nature of exposure and how it is measured, additional data maybe required to establish the exposure status of subjects in the control-cohort at the time the cases occur. As noted previously, consideration needs to be given to the requirements for obtaining exposure data or biological specimens from study subjects; only subjects likely to agree to these requirements should be considered eligible for the subcohort. Serially stored tissue specimens allow for the detection of exposure changes. For example, Pfeiffer *et al* (2005) used stored dust samples taken at intervals throughout the study period for endotoxin levels in a study of childhood asthma. In other situations, 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 most studies, self-declared exposure and covariate status are updated regularly throughout the follow-up period.

Note that if several outcomes are to be assessed, exposure and covariate data are needed on each of the cases as well as all members of the control-cohort. 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 adults have a higher risk of the outcome than older adults, the control cohort can be selected in a manner to ensure that the majority of study subjects in the control cohort will be young adults.

Kubota and Wakana (2011) give sample size formulae for case-cohort studies.

10.6.3 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, we can combine (*ie* add) the 2 types of case together, and the data can be analysed in a 2X2 table using a case-control format with the *OR* as the measure of association. Logistic regression can be used to control for additional covariates.

Many researchers with open-population studies use Cox survival methods and hence, hazard ratios, for analysis (Kulathinal *et al*, 2007); these models need to be weighted usually inversely to the probability of being sampled. For example, if a 20% random sample of the cohort is used for the subcohort, typical weights are 1 for cases and 1/0.2=5 for controls in the subcohort. If a stratified sample was selected, the weights should reflect the (inverse of the) sampling probabilities in each of the strata. If non-response and exclusion criteria exclude very many of the initially sampled individuals, the sampling probabilities should reflect the actual number who agreed to participate, not the initial sampling probabilities (so if only 80% of the cases agree to participate, the inverse weighting would be 1.2 for cases, not 1). 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). The use of robust standard errors is also recommended. Other analytic methods are available when not all cases from the full cohort are used in the study (Pfeiffer *et al*, 2005).

Breslow *et al* (2009) suggest using the entire non-case population as the full control cohort when at least some of the important covariate values are known for all subjects. The subcohort is selected and the appropriate laboratory tests, or surveys, run to establish the levels or existence of key exposure variables. Using the observed outcome, linear (if the outcome is continuous) or logistic (if the outcome is binary) models are used to develop a predictive model of the outcome in the subcohort based on the covariates that are known for the full cohort. Then, the imputed values of the exposure are used for all members of the source population. This allows an analysis of the exposure outcome association in the full cohort. The delta-beta values from the population model can then be used to recalibrate the model based on the subcohort. Although this approach appears valid, it ignores the uncertainty concerning imputation model parameters and the values to impute according to a given model. This led Marti and Chavance (2011) to develop a method of analysis based on multiple imputations.

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. Zhang *et al* (2011) describe how to adjust for clustered data in case-cohort studies. The need for this occurs if many of the cases are diagnosed/treated at the same clinic. One can use frailty models for this purpose, but Zhang *et al* focus on a marginal model which uses adjusted variances to account for within-cluster correlation. Li *et al* (2008) suggest using a weighted-likelihood method to adjust for clustering. Example 10.6 describes a case-cohort study.

10.7 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 genetic exposure in the 'controls' is derived from theoretical grounds such as the blood-type

Example 10.6 A case-cohort study of drinking water quality and risk of stomach cancer

Auvinen *et al* (2005) evaluated 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. An 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 1, suggesting a sparing effect of radon levels on stomach cancer.

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. The design allows for the identification of interactions 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), but not the main effects. Schwartz (2005) provides a good introduction to the basic design and analysis of case-only studies. VanderWeele (2011) describes how to determine the appropriate sample for a case-only study.

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 socioeconomic class on the risk of mortality. For example, the design has been used to assess if personal characteristics interact with extreme weather (Medina-Ramon *et al*, 2006) and if socioeconomic class interacts with weather to modify the risk of death (Armstrong, 2003).

Clarke and Morris (2010) discuss sample size for case-only studies of gene-environment interactions.

10.7.1 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 of 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:

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

We could use this model to create a 2X2 table of expected outcome event counts according to the 4 combinations of exposure and sex (Schwartz, 2005). 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 (a covariate of interest), as a function of the exposure in subjects experiencing the outcome:

$$\text{logit}(sex = 1) = \beta_0 + \beta_1(exposure)$$

If β_1 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. Tchetgen and Robins (2010) propose a semi-parametric approach to analysis. Examples 10.7 and 10.8 outline typical case-only studies.

10.8 Two-stage sampling designs

A 2-stage sampling design can be applied to the traditional cohort, case-control, or cross-sectional study designs (Hagel, 2011). 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 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; 2005).

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 of a 2-stage study, 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.

Example 10.7 Case-only study of potential effect modifiers of risk of death in humans

Schwartz (2005) investigated whether sex, non-white 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 were investigated: one focused on a single day, and the other 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.

Example 10.8 A case-only analysis of the health impacts of heat waves in 5 regions of New South Wales, Australia

Khalaj *et al* (2010) used a case-only design to identify underlying health conditions that increased the risk of hospital admission during heat waves in Australia. Daily hospital admissions were obtained from NSW databases during September 1 to February 28 of each year, from 1998 to 2006 in 5 regions. Data from Sydney weather stations were used for exposure data. The authors fitted logistic regression models of the presence or absence of each primary diagnosis as the outcome and an extreme heat indicator as the predictor. The analysis was repeated using this temperature indicator on the day of hospitalisation (lag0), the day before hospitalisation (lag1), and for the 3-day average ending on the day of hospital admission. If the proposed modifier (*ie* a primary diagnosis) of the effect of extreme temperature was a modifier of season (*eg* if diabetics have a stronger seasonal pattern than other diseases), a confounding with the interaction of interest could occur. Therefore, all models also included a sine and cosine term with a 365, 24-day period to capture interactions between season and the characteristic being investigated.

Cain and Breslow (1988) developed the methodology to analyse 2-stage data using logistic regression followed by Flanders and Greenland (1991). 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). Chen *et al* (2009) describe 2-stage case-only studies. Fears and Gail (2000) describe 2-stage studies with cluster sampling of controls. Example 10.9 describes a 2-stage study design.

Example 10.9 A 2-stage case-control study of determinants of the incidence of childhood asthma in Quebec

Martel *et al* (2009) conducted a case-control study with a 2-stage sampling strategy using data from 3 interlinked administrative health databases in Quebec, Canada. From the databases, a cohort of pregnant women and their children was formed. It consisted of all asthmatic women (n=8226) and a sample of non-asthmatic women (n=18039) who had had at least one singleton pregnancy ending in a live birth between 1990 and 2002. If a woman had had more than one pregnancy during the study period, only the latest pregnancy was retained in the cohort. The first stage of the study consisted of a case-control study, nested in the cohort of children. 5,226 asthmatic children (cases) and 20 non-asthmatic children per case (selected using density sampling matching to the time of case occurrence) were selected. Covariate information was available in the original database. The second stage of the study used a questionnaire mailed to a subsample of mothers of cases and controls to obtain more information on variables not available in databases. Balanced sampling of cells of the first-stage exposure-outcome cross-table was performed. This allowed the over-representation of small cells and an increase in statistical power. Crude rates of childhood asthma for children of asthmatic and non-asthmatic mothers were estimated from the cohort. For the first stage of the study, the authors obtained crude and adjusted rate ratios using conditional logistic regression. For the analysis using the subsample of cases and controls, crude and adjusted odds ratios were obtained using unconditional logistic regression. Final corrected estimates were obtained for maternal asthma using sampling fractions and maternal asthma estimates from the first stage of the study. See Collet *et al* (1998) for variance adjustment in the second-stage sample.

REFERENCES

Agalliu I, Kirsh VA, Kreiger N, Soskolne CL, Rohan TE. Oxidative balance score and risk of prostate cancer: results from a case-cohort study. Cancer Epidemiol. 2011;35(4):353-61.

Armstrong BG. Fixed factors that modify the effects of time-varying factors: applying the case-only approach. Epidemiol. 2003;14(4):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(1):109-13.

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol. 2009;169(11):1398-405.

Cai J, Zeng D. Power calculation for case-cohort studies with nonrare events. Biometrics. 2007;63(4):1288-95.

Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol. 1988;128(6):1198-206.

Carracedo-Martinez E, Taracido M, Tobias A, Saez M, Figueiras A. Case-crossover analysis of air pollution health effects: a systematic review of methodology and application. Environ Health Persp. 2010;118(8):1173-82.

Chen YH, Lin HW, Liu H. Two-stage analysis for gene-environment interaction utilizing both case-only and family-based analysis. Genetic epidemiology. 2009;33(2):95-104.

Clarke GM, Morris AP. A comparison of sample size and power in case-only association studies of gene-environment interaction. Am J Epidemiol. 2010;171(4):498-505.

Darrow LA. Invited commentary: application of case-crossover methods to investigate triggers of preterm birth. Am J Epidemiol. 2010;172(10):1118-20; discussion 21-2.

Delaney JA, Suissa S. The case-crossover study design in pharmacoepidemiology. Stat Meth Med Res. 2009;18(1):53-65.

Fears TR, Gail MH. Analysis of a two-stage case-control study with cluster sampling of controls: application to nonmelanoma skin cancer. Biometrics. 2000;56(1):190-8.

Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med. 1991;10(5):739-47.

Gillespie IA, O'Brien SJ, Frost JA, Adak GK, Horby P, Swan AV, et al. A case-case comparison of Campylobacter coli and Campylobacter jejuni infection: a tool for generating hypotheses. Emerg Inf Dis. 2002;8(9):937-42.

Gribbin J, Hubbard R, Gladman J, Smith C, Lewis S. Risk of falls associated with antihypertensive medication: self-controlled case series. Pharmacoepidemiology and drug safety. 2011;20(8):879-84.

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(2):809-13.

Hagel BE. Emergency department injury surveillance and aetiological research: bridging the gap with the two-stage case-control study design. Injury Prev. 2011;17(2):114-8.

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(12):1225-34.

Janes H, Sheppard L, Lumley T. Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. Epidemiol. 2005;16(6):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 Cont and Hosp Epidemiol. 2005;26(4):346-51.

Khalaj B, Lloyd G, Sheppeard V, Dear K. The health impacts of heat waves in five regions of New South Wales, Australia: a case-only analysis. Int Arch Occup Environ Halth. 2010;83(7):833-42.

Kim I, Cheong HK, Kim H. Semiparametric regression models for detecting effect modification in matched case-crossover studies. Stat Med. 2011;30(15):1837-51.

Kim MY, Xue X, Du Y. Approaches for calculating power for case-cohort studies. Biometrics. 2006;62(3):929-33; discussion 33.

Krumkamp R, Reintjes R, Dirksen-Fischer M. Case-case study of a Salmonella outbreak: an epidemiologic method to analyse surveillance data. Int J Hygiene and Environ Health. 2008;211(1-2):163-7.

Kubota K, Wakana A. Sample-size formula for case-cohort studies. Epidemiol. 2011;22(2):279.

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

Li Z, Gilbert P, Nan B. Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. Biometrics. 2008;64(4):1247-55.

Lu Y, Symons JM, Geyh AS, Zeger SL. An approach to checking case-crossover analyses based on equivalence with time-series methods. Epidemiol. 2008;19(2):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(8):850-3.

Maclure M, Fireman B, Nelson JC, Hua W, Shoaibi A, Paredes A, et al. When should case-only designs be used for safety monitoring of medical products? Pharmacoepidemiology and drug safety. 2012;21 Suppl 1:50-61.

Martel MJ, Rey E, Malo JL, Perreault S, Beauchesne MF, Forget A, et al. Determinants of the incidence of childhood asthma: a two-stage case-control study. Am J Epidemiol. 2009;169(2):195-205.

Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. Stat Med. 2011;30(13):1595-607.

Martinez ME, Cruz GI, Brewster AM, Bondy ML, Thompson PA. What can we learn about

disease etiology from case-case analyses? Lessons from breast cancer. Cancer Epidemiol Biomark & Prev. 2010;19(11):2710-4.

Matsuda Y, Hayashi K, Shiozaki A, Kawamichi Y, Satoh S, Saito S. Comparison of risk factors for placental abruption and placenta previa: case-cohort study. J Obstet and Gynecol Res. 2011;37(6):538-46.

McCarthy N, Giesecke J. Case-case comparisons to study causation of common infectious diseases. Int J Epidemiol. 1999;28(4):764-8.

McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. Stat Med. 2002;21(23):3609-25.

McNamee R. Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. Biostatistics. 2005;6(4):590-603.

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. Environ Health Persp. 2006;114(9):1331-6.

Melo EC, Fortaleza CM. Case-case-control study of risk factors for nasopharyngeal colonization with methicillin-resistant Staphylococcus aureus in a medical-surgical intensive care unit. Braz J Inf Dis. 2009;13(6):398-402.

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

Moller J, Hessen-Soderman AC, Hallqvist J. Differential misclassification of exposure in case-crossover studies. Epidemiol. 2004;15(5):589-96.

Navidi W, Weinhandl E. Risk set sampling for case-crossover designs. Epidemiol. 2002;13(1):100-5.

Onland-Moret NC, van der ADL, van der Schouw YT, Buschers W, Elias SG, van Gils CH, et al. Analysis of case-cohort data: a comparison of different methods. J Clin Epidemiol. 2007;60(4):350-5.

Pfeiffer RM, Ryan L, Litonjua A, Pee D. A case-cohort design for assessing covariate effects in longitudinal studies. Biometrics. 2005;61(4):982-91.

Rosenbaum PR. The case-only odds ratio as a causal parameter. Biometrics. 2004;60(1):233-40.

Sato Y, Akiba S, Kubo O, Yamaguchi N. A case-case study of mobile phone use and acoustic neuroma risk in Japan. Bioelectromagnetics. 2010.

Schwartz J. Who is sensitive to extremes of temperature?: A case-only analysis. Epidemiol. 2005;16(1):67-72.

Smeeth L, Donnan PT, Cook DG. The use of primary care databases: case-control and case-only designs. Family practice. 2006;23(5):597-604.

Suissa S. The case-time-control design. Epidemiol. 1995;6(3):248-53.

Tchetgen EJ, Robins J. The semiparametric case-only estimator. Biometrics. 2010;66(4):1138-

44.

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(3):167-80.

Thornley S, Kool B, Robinson E, Marshall R, Smith GS, Ameratunga S. Alcohol and risk of admission to hospital for unintentional cutting or piercing injuries at home: a population-based case-crossover study. BMC Public Health. 2011;11:852.

VanderWeele TJ. Sample size and power calculations for case-only interaction studies. Epidemiol. 2011;22(6):873-4.

Wang S, Linkletter C, Maclure M, Dore D, Mor V, Buka S, et al. Future cases as present controls to adjust for exposure trend bias in case-only studies. Epidemiol. 2011a;22(4):568-74.

Wang SV, Coull BA, Schwartz J, Mittleman MA, Wellenius GA. Potential for bias in case-crossover studies with shared exposures analyzed using SAS. Am J Epidemiol. 2011b;174(1):118-24.

Weldeselassie YG, Whitaker HJ, Farrington CP. Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. Epidemiol and Inf. 2011;139(12):1805-17.

Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. Stat Med. 2006;25(10):1768-97.

Whitaker HJ, Hocine MN, Farrington CP. The methodology of self-controlled case series studies. Stat Meth Med Res. 2009;18(1):7-26.

Williams M, Mohsin M, Weber D, Jalaludin B, Crozier J. Alcohol consumption and injury risk: a case-crossover study in Sydney, Australia. Drug and Alcohol Review. 2011;30(4):344-54.

Wilson N, Baker M, Edwards R, Simmons G. Case-case analysis of enteric diseases with routine surveillance data: Potential use and example results. Epidemiol Persp & Innov. 2008;5:6.

Zhang H, Schaubel DE, Kalbfleisch JD. Proportional hazards regression for the analysis of clustered survival data from case-cohort studies. Biometrics. 2011;67(1):18-28.