The use of primary care databases: case–control and case-only designs
Liam Smeetha, Peter T Donnanb and Derek G Cookc


Study designs based on the identification of cases are frequently utilized when undertaking epidemiological research. Traditionally these have been mainly based on identification of cases from hospital records. This paper discusses the use of study designs based on the identification of cases focusing on their application to research data derived from primary care. The designs are discussed in the context of using computerized clinical data derived from primary care. The traditional case–control design is considered, with emphasis on the identification of cases and the selection of controls. A common problem when using primary care research databases is that information about potential confounding variables is often limited. Case-only designs, specifically the case-crossover and the within-person case-series, offer alternative designs that aim to overcome problems with confounding. The principles underlying these case-only designs are presented along with examples of their use. The advantages and limitations of the different designs are discussed.

Keywords. Primary care databases, case–control, case-only, methodology.

Introduction
Research, which takes as its starting point cases of disease, has always been attractive to clinicians. In seeking risk factors, or causes of disease, one naturally examined cases to assess the presence of risk factors or exposures. What surprised, and thus suggested a possible cause, was when the prevalence of an exposure amongst cases was markedly different from the clinician’s impression of the prevalence in the wider community. The need to provide objective information on the community prevalence led to the need to identify controls who did not have disease. Thus the classical case–control study was born. Primary care is no different from other clinical specialties in focussing on diseases and problems and a case–control approach might seem natural. However, the GP, unlike the hospital clinician is in a better position to take a population perspective, particularly in the UK where virtually all patients are registered with GPs. The GP is thus in the position to view the patients with disease and controls without disease within a particular study as being nested within a study of all registered patients. The same is true of those who analyse data from large PCDs (Table 1).

In this article we explore the uses of the case–control design within PCDs, focusing on the use of PCDs available in the UK. We discuss the advantages and disadvantages of using a case–control or cohort design using the same dataset, before moving on to discuss the merits of more recent innovations in the form of case-only designs in which exposure information within a subject is compared in different periods when they were, or were not, about to develop a disease. Such designs have the merit of overcoming specific confounding problems common to both cohort and case–control study designs.

The case–control design

Basics of a case–control study
Case–control studies are classically based on asking the question ‘Do persons with a disease have a
Thus Mckeever used a cohort approach to examine the risk of developing atopic diseases in a birth cohort of children born into practices registered with GPRD (the General Practice Research Database; Table 1), while Bremner et al.4,5 used a case–control approach to study early life risk factors for development of hayfever in both GPRD and DIN (the Doctor’s Independent Network database; Table 1). Bremner explicitly identified all children born into DIN and GPRD (registered within 3 months of birth) and who remained registered with the same practice for at least 5 years. He then selected all cases of hayfever and matched each case with a control who did not develop hayfever.5 The cohort and case–control approaches produced very similar findings.

Advantages of case–control approach
The usual advantage of case–control studies is that they are efficient in two important respects: (i) they can yield important findings in a relatively short space of time since it is not necessary to wait for subjects to develop disease as in a cohort study—this is especially important in chronic diseases with long latency periods; (ii) the total number of individuals required to obtain adequate power in a case–control study is often considerably less than in a cohort study. As a result case–control studies can be much cheaper. For rare outcomes, cohort studies or randomized trials would often require unfeasibly large numbers of participants in order to have adequate power. Historically case–control studies have been thought of as weaker than cohort studies, mainly because of the difficulty in selecting representative samples of cases and controls and of assessing exposure retrospectively. More recently it has been recognized that well-designed case–control studies can provide evidence as strong as cohort studies.2

Case–control studies using primary care databases
Since the data on exposure and outcome already exist in electronic form for the population of subjects registered with a practice it can be argued that it is as easy to carry out a cohort analysis as to set up a case–control study within a Primary Care Database (PCD); both should give the same result since both cases and controls can be selected randomly from the set of all possible cases and all possible controls within the database. In essence any case–control study, but especially those carried out using PCDs, can be viewed as a case–control study nested within a cohort study.2 Thus Mckeever used a cohort approach to examine the

### Table 1  Abbreviations used with explanatory comments

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full text</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCD</td>
<td>Primary Care Database</td>
<td>Used in the paper as a generic term</td>
</tr>
<tr>
<td>GPRD</td>
<td>General Practice Research Database</td>
<td>The most widely used primary care research database. UK based</td>
</tr>
<tr>
<td>DIN</td>
<td>Doctor’s Independent Network database</td>
<td>Another UK based primary care research database: smaller than GPRD</td>
</tr>
<tr>
<td>ACORN</td>
<td>A Classification Of Residential Neighbourhoods</td>
<td>A small area based socioeconomic index available in DIN</td>
</tr>
</tbody>
</table>

characteristic/exposure more frequently than those without the disease? Thus case–control study groups are defined on the basis of having the disease or not; information on past or present exposure is then compared in these two groups. Because they start from the outcome, they are often thought of as doing research in reverse.1 This is in contrast to cohort studies where populations of exposed and non-exposed subjects are followed forward in time and the proportion developing disease in the two groups compared.

**Identifying and selecting cases and controls**
In PCDs defining a case is largely a question of identifying the relevant diagnostic codes likely to be used by GPs to record a particular problem. Effort invested at this stage avoids later problems. In particular it is important to identify all possible codes and to decide whether they should be included. These are likely to include diagnostic and symptom codes, but may also take account of drugs prescribed if these are specific, given that these are more reliably recorded. Experience suggests it may be important to include a practising GP in this process.

In the early days of GPRD there was considerable concern over the accuracy with which cases and controls could be identified from electronic GP data. A system whereby the accuracy of case status in selected cases could be checked against paper records held in the practice was, therefore, set up and this validation was carried out in many of the studies carried out by Jick’s group.5 The system still exists in GPRD but is expensive to use. In theory absence of the disease could also be validated in controls, but for all except very common diseases this is largely a waste of time, since most controls will lack disease by chance. Moreover for serious conditions it always seemed unlikely that subjects recorded as having a diagnosis for a serious disease would not have the disease—and this proved to be the case. Even if significant numbers of cases are not recorded as such it will have little effect on any aetiological analysis since all the cases will be cases and almost none of the controls will have the
The use of primary care databases

599
disease. Overall, while errors in identifying cases may lead to bias in estimating prevalence rates, they will usually have little impact on odds ratios estimated from case–control studies. For example, take a disease with a prevalence of 1% for which 20% of cases are not recorded as having the disease. These 20% could be wrongly selected as healthy controls. This would mean that for every 1000 controls sampled, on average 2 of the controls would in fact have the disease—such an error rate would have negligible impact on any measures of effect.

For some conditions it may be possible to identify cases with varying degrees of certainty; thus, Bremner distinguished between certain cases of hay fever who had diagnoses and/or treatment in at least two hay fever seasons and those who were less certain who were diagnosed/treated in one season only.5

In his study Bremner selected cases as those subjects who developed hay fever during the course of the follow up available—5–10 years from birth. Controls were selected from those who did not develop hay fever during follow-up. Subjects who had some evidence of allergic rhinitis, but who did not satisfy the definition of being a case, were excluded from both groups. Clearly some subjects who had not developed hay fever, and thus were eligible as controls, will have gone on to develop hay fever. This design is typical of most case–control studies but is a concept which confuses many. Alternatives exist, including selecting controls from amongst subjects without disease at the time the case becomes a case, rather than from amongst those who never develop disease during available follow-up. Such an approach will sacrifice power if the disease is common (because some of the controls will subsequently become cases) but has the merit of being equivalent to what happens in a cohort study where risk ratios at any point in time are based on comparisons of those who have developed disease out of all those who have not yet done so but are at risk of doing so. Rothman calls this ‘risk set sampling’.2

Exposure information
A limited range of exposures are available in PCDs. The vast majority of case–control studies have focussed on drug therapies. While prescriptions issued are well recorded in PCDs, whether they are dispensed or taken is another matter. Any interpretation needs to consider such issues discussed in more details elsewhere in this series.

How many cases and controls?
A simple case–control approach would select equal numbers of cases and controls—typically matched on a 1–1 basis. However, for a given number of cases power for detecting an effect can be increased by selecting more controls than cases. Power increases as one moves from a 1–1 to a 1–2 matching but the benefits of additional controls rapidly decreases above four controls per case.7 There is an argument for using a many to many matching so that if either a case or control is excluded in a sensitivity analysis the whole matched set is not discarded.2

Measuring and allowing for confounding
Within PCDs only a limited number of confounding variables are available. Age and sex are often potential confounding factors and are almost always available to researchers. However, practices or other primary care groupings that may be used as units of analysis may also be important given that recording quality varies markedly by practice as do process measures such as diagnostic practices and prescribing. It is, therefore, crucial that practice be controlled for in analysis, if not matched for in the design. In Bremner’s study controls were matched within practice. Another key confounder, which is linked to practice but also varies markedly between individuals, is consultation frequency. Subjects who consult more frequently are more likely to be diagnosed with any problems they have and to be treated. Thus Davey found that the association between migraine and asthma was explained by control for consultation frequency,8 and Bremner reported a marked reduction in odds ratios of hay fever in relation to antibiotic use during early life when consultation frequency was allowed for.5 Individual characteristics such as smoking status may be inconsistently recorded in PCDs, while social confounding has been little examined. Sometimes different confounding variables are available in different PCDs. Thus, in DIN a social indicator, the ACORN index (the abbreviation stands for A Classification Of Residential Neighbourhoods; Table 1) has been linked in to individual patient records at post code level, allowing the role of social confounding to be investigated in DIN. In GPRD a family linkage variable is available allowing analyses that control for number of older and younger sibs.5

Matching controls to cases
Matching refers to deliberately identifying controls who are similar to cases for one or more matching criteria. Matching can improve statistical efficiency, and thus increase power, as well as being a method of controlling for confounding by a matching factor. However, some caution is needed. Matching must be allowed for in the analysis, or biased estimates can result. The number of strata in a matched analysis rises exponentially with the number of factors matched on, such that matching on several factors simultaneously will often lead to major analytical problems. Matching on a factor strongly associated with exposure, but not independently associated with the outcome, may be referred to as ‘over-matching’. This
can seriously reduce efficiency and may lead to a biased result. Matching on a variable that is on the causal pathway between exposure and disease will also lead to a biased result. While matching on age and sex is often worthwhile, and is unlikely to lead to problems, care is needed when matching on other factors. A better approach is often to adjust in the analysis, rather than to risk over matching. These issues are discussed in detail elsewhere.9,10

Methods of statistical analysis

Analysis of case–control studies is usually a dichotomous outcome in relation to a categorical exposure variable; in its simplest form the results can be expressed as a $2 \times 2$ table and the odds ratio calculated as the cross-product ratio. Multiple logistic regression can be used to adjust for confounding variables. Where matching of cases and controls has been carried out either on a 1–1 or 1–many basis, the matching must be taken into account usually by using conditional logistic regression. If matching is ignored, then the relative risk estimate will be closer to 1 than it should be: the degree of dilution depending on the strength of confounding due to the matched variables.2 Sensitivity analyses may be useful in examining possible biases. Typically this takes the form of omitting subjects and observing what happens to the estimated odds ratio. Thus Bremner carried out analyses omitting pairs where: (i) cases had a less certain diagnosis of hay fever; (ii) controls may have moved without deregistering before the case was diagnosed.5

Why use a case–control design?

Given that it is always possible to carry out a cohort analysis using a PCD, it is reasonable to ask why use a case–control approach rather than a cohort. Historically one advantage has been the reduction in computing power needed for analyses, though at the expense of more data preparation. A second is that it is often easier to control for some factors such as practice, by matching rather than by adjusting in the analysis. A third, which remains important when one is paying for the data on a per subject basis, is cost. This may arise if it is necessary to validate cases and or controls by checking information with practices. It might also arise if one was looking to link additional information, perhaps on confounding variables or to refine case definition, to both cases and controls. Historically PCDs have not been used in this way because of the need to maintain anonymity, but conceptually they could be used in such a manner, and the evidence is that many subjects are happy for additional information collected by questionnaire to be linked to their primary care records.11,12 A potential disadvantage may be the need to select controls separately for each case-series if interested in different outcomes.

Case-only designs

A problem common to both case–control and cohort designs is that information about confounding factors may often not be available within PCDs, particularly for characterizing individual exposures and behaviour. Recent innovations have resulted in two study designs in which cases act as their own controls and thus overcome such confounding: these are the case-series design and the case-crossover design. The case-crossover design is based on identifying the outcome of interest and assessing exposure to the risk factor of interest in a chosen time period preceding the outcome. One or more control time periods are then selected, and exposure to the risk factor of interest in the time period preceding the outcome is compared with exposure during these control periods in the same individual. Thus the case-crossover design is analogous to a case–control design. The case-series design starts with identifying exposure to the risk factor of interest. The likelihood of the outcome of interest occurring in a time period following exposure is compared with the likelihood of the outcome of occurring in ‘unexposed’ periods, with the comparison again being in the same individual. The case-series design is, therefore, more similar to a cohort design. The case-crossover design is conceptually more intuitive and somewhat simpler statistically and analytically. However, the case-series design has some advantages. It can allow for changes in the risk of exposure with time and is well suited to the study of recurrent outcomes.

The case-crossover design

The case-crossover design was originally proposed by Maclure,13 with the aim of eliminating control selection bias and confounding by constant within-subject characteristics in observational studies. The case-crossover design has been utilized to assess the short-term immediate triggers of events in a variety of areas such as myocardial infarction (MI), drug safety, road traffic accidents, air pollution and mass media campaigns. In this design only cases who have experienced the outcome of interest are considered. Each case acts as his/her own control. A risk period of interest immediately prior to the outcome is defined, and ‘control’ or comparison time period(s) selected (Fig 1).

Exposure is then compared in the risk period of interest immediately prior to the event with one or more time periods in the past. At its simplest, with one time period for comparison, analysis involves constructing a $2 \times 2$ table and implementing McNemar’s test (Table 2).

Then the statistic \( (b - c)^2/(b + c) \) has a Chi-square distribution with one degree of freedom.
With N time periods used as ‘controls’ conditional logistic regression can be used to assess the relationship between events and exposure with 1-to-N matching. The effects of constant within-subject factors on the results are removed by the design. However, other risk factors that change over time such as changes in co-medication can be added as covariates in these models. Changes over time in dose can be incorporated with an extension of the case-crossover design, the case-time-control design as advocated by Suissa.14 The design resembles a retrospective cohort study with crossover between exposure and non-exposure, although conceptually and statistically it is closer to a case–control design. It also resembles an experimental crossover design, except that the order of exposure is not randomized. The design is not suitable for studying chronic conditions nor in situations where the exposure is constant throughout the observation period. Its strength lies in eliminating control selection bias and in its ability to assess acute events that are transient following intermittent exposure.

A series of studies incorporating the case-crossover design considered several potential triggers of MI. These demonstrated strong associations of MI with physical exertion,15 although mainly in those with a sedentary lifestyle, as it was shown that regular physical exertion was protective.16 Further studies indicated increased risk with anger17 and that this risk declined with increasing educational attainment.18 Use of cocaine was also found to be an abrupt and transient trigger for MI in those who otherwise were considered to be at low risk.19 A landmark US study demonstrated a significant relationship between prior use of mobile phones and subsequent road traffic accidents.20 The design has been used extensively in solving the problem of confounding by indication in pharmaco-epidemiological studies, which arises whenever the indication for the drug is also associated with the outcome.21,22 In this case the indication and drug exposure are confounded and if the indication is a constant within-subject factor it is eliminated.

The case-crossover design was utilized in assessing the acute risks of collision associated with the primary care prescribing of anti-anxiolytics, such as benzodiazepines.23 Information on all road traffic accidents occurring in Tayside, Scotland were obtained for the period 1 August, 1992 to 30 June 1995. Dispensed community prescribing is routinely entered in the database of the Tayside Medicines Monitoring Unit (MEMO)24 and a case-crossover study was initiated to assess associations between benzodiazepine use and the risk of a road traffic accident by linking these databases. The use of long half-life benzodiazepines and short half-life zopiclone were found to be significantly associated with an increased risk of an accident. A recent study of the effect of mass media campaigns on consulting behaviour with GPs compared recall of receiving media health messages in the week before contact with a GP with previous control weeks.25 No significant relationship was found suggesting media campaigns rarely trigger visits to the GP.

An important consideration in the case-crossover design is how to choose an appropriate size of the time period or exposure window, as this is under the control of the researchers. In previous studies this has varied from 10 minutes prior to the event, such as that used in the study of the risk of a collision associated with cellular phones,20 to the whole year or longer prior to the event. There is the potential for bias whenever the time period is particularly large and so a sensitivity analysis is often implemented by repeating the analysis with varying sizes of the exposure window.15–19 A second important consideration is the number of control time periods sampled per case. Mittleman26 found that the precision of the effect size improved with the number of time periods selected, with the greatest efficiency achieved by using the whole year prior to the event. However, there is a trade-off with
bias as the greater and longer the number of time periods the greater the potential for confounding. Further details of the case-crossover design can be found in the overview by Maclure and Mittleman.27

The self-controlled case-series design

The case-series method uses within-person comparisons in a population of individuals who have experienced the outcome of interest. The incidence of the outcome in defined risk intervals after an exposure are compared with the incidence of the outcome during all other observed time periods for each person.28,29 Only cases are included: there is no separate control group. The major advantage of the case-series is that confounding due to differences between people studied is eliminated. The method is illustrated in Figure 2.

The case-series method was originally developed to study adverse events occurring within a specified time period following vaccination. Examples of its use with vaccines include studies of the relative incidence of febrile convulsions following DTP and MMR vaccines;30 the onset of autism following MMR vaccine;31 asthma following influenza vaccine;32 and intussusception following rotavirus vaccination.33

The design was used recently to assess the risk of cardiovascular events in time periods following infection using the GPRD.34 The background to this study was the evidence that chronic inflammation enhances atherosclerotic disease, but the effects of short-term fluctuations in systemic inflammation on vascular risk is less clear. Because people with and without diagnosed infections would be likely to differ in ways that are difficult to measure and control for, the self-controlled case-series method was used. The null hypothesis was that vascular event rates remain constant from day to day and are not affected by an acute exposure to infection. The exposed period was defined as up to 91 days after a diagnosed infection and was subdivided as follows: 1–3 days, 4–7 days, 8–14 days, 15–28 days and 29–91 days after exposure. All other observation time was taken as the baseline (unexposed) period. The risk of both MI and of stroke events were substantially raised following a diagnosis of respiratory tract infection and were highest in the first 3 days following exposure: incidence ratio for MI 4.95 (95% CI 4.43–5.53); incidence ratio for stroke 3.19 (95% CI 2.81–3.62). The risk then gradually fell over the following weeks. Risks were significantly raised but to a lesser degree following a diagnosis of urinary tract infection.

In another example, the risk of hip fracture following the initiation of different antidepressants was studied.35 A large case–control study had found a higher risk of hip-fracture associated with selective serotonin reuptake inhibitors (SSRIs) than with tricyclic antidepressants (TCAs).36 However, because of the perception that SSRIs have less adverse effects such as sedation, it is likely that more frail patients could have been selectively prescribed SSRIs by physicians. The higher risk observed with SSRIs compared with TCAs could, therefore, have been due to selection bias and confounding. The case-series design allowed an estimate of the effect of different antidepressants on hip fracture risk while eliminating confounding between individuals. Using the GPRD, both classical case–control and case-series studies were performed assessing the risk of hip fracture in the first 15 days after starting antidepressants.35 First, this study showed that the risks observed using the case–control design were much higher, suggesting that in spite of controlling for a wide range of risk factors, there was residual confounding in the case–control study. Second, the

![Figure 2](attachment:image.png)  
**Figure 2**  Pictorial representation of the case-series analysis. The risk of an event during the high-risk exposed period is compared with the risk of an event during unexposed or baseline time periods.
case-series method showed that if anything, the risk associated with TCAs was higher than for SSRIs, suggesting that selective prescribing of SSRIs to people at higher risk of fracture may explain the previously observed higher risk with SSRIs.

There are two specific methodological issues that can arise when using the case-series method. Risk factors for an outcome that change with time for a particular individual, could, if also associated with the chance of exposure, produce within-person confounding. For example, existing cardiovascular disease is a major risk factor for MI. Cardiovascular disease is an indication for influenza vaccine, so the probability of being vaccinated may be temporally associated with the risk of MI if someone develops cardiovascular disease during their observation period. To overcome this, in the study of cardiovascular events and vaccination described above, the time before first recorded influenza vaccine was not included in the analysis thus ensuring minimal variation in the chance of being vaccinated during follow-up.34 The second issue is about choice of time periods to be compared which may vary for fatal and non-fatal events. In a study of the risk of seizures and sudden death following prescription of bupropion, for seizures the whole observation period was utilized. However, because people must be alive to receive a prescription (i.e. not have suffered the outcome), for fatal events the analysis was based solely on assessing the risk of death in time periods following the first prescription for bupropion.37

For anyone considering using the case-series method, the best source of practical information is a website run by the statistician who developed the method: http://statistics.open.ac.uk/scscs/.

This includes a tutorial38 and allows users to download files to run the case-series in GLIM, Stata, SAS and GenStat.

Discussion

When attempting to assess the association between an exposure and an outcome, the case–control design is often used and offers particular advantages for rare outcomes. The previous experience of exposure to the risk factor of interest is compared between cases of the outcome of interest and a control group that does not have the outcome of interest. However, associations observed between exposures and outcomes are not necessarily causal but may be due to confounding factors that are associated with both the exposure and outcome of interest. Such confounding can be a particular problem when using primary care data because the availability of information about some potential confounding factors may be limited. Case-only approaches can overcome problems of confounding by only using within individual comparisons. It is important to recognize that such approaches deal well with controlling for characteristics of individuals, which are stable or fixed, but they may introduce problems for temporally changing confounding variables since the periods chosen for comparison may be related to such variables (e.g. in case-crossover studies of mobile phone use in relation to car accidents, roads may be more often wet in the accident period than in the control period). In case-only approaches, exposure experiences in different time periods are related to the timing of the outcome of interest. Such methods depend on there being variation in exposure over time. Case-only approaches are most suitable for outcomes with an acute onset. The case-crossover and the self-controlled case-series designs are the most commonly used case-only designs. Statistically, the case-series method is derived from the modelling of the rate of outcome using Poisson modelling and is thus analogous to a cohort model. The case-crossover method is derived from the modelling of odds of an outcome and is thus more analogous to a case–control method.29 The case-series design is more suitable for recurrent outcomes. Recent work has suggested that the case-crossover method requires an assumption of exchangeability of the control periods.39 No such assumption is required of the case-series method. However the case-crossover design is more intuitive and the analysis often much simpler—in fact a simple $2 \times 2$ table will often suffice. The case-series is computationally intensive, and for large studies can present a formidable analytical challenge. The use of case-only approaches is likely to increase in the future as our knowledge improves.

Acknowledgements

We would like to thank Frank Sullivan for useful comments on the paper.

Declaration

Ethical approval: not applicable
Funding: LS is supported by a Medical Research Council Clinician Scientist Fellowship
Conflicts of interest: none

References
