The methodology of self-controlled case series studies

Journal Item

The methodology of self-controlled case series studies

Heather J Whitaker, Mounia N. Hocine and C. Paddy Farrington
Department of Statistics, The Open University,
Milton Keynes, MK7 6AA, UK

August 31, 2007

Abstract

The self-controlled case series method is increasingly being used in pharmacoepidemiology, particularly in vaccine safety studies. This method is typically used to evaluate the association between a transient exposure and an acute event, using only cases. We present both parametric and semiparametric models using a motivating example on MMR vaccine and bleeding disorders. We briefly describe approaches for censoring events and a sequential version of the method for prospective surveillance of drug safety. The efficiency of the self-controlled case series method is compared to the that of cohort and case control studies. Some further extensions, to long or indefinite exposures and to bivariate counts, are described.

Corresponding Author: Heather Whitaker, Department of Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK.

Email: h.j.whitaker@open.ac.uk.

1 Introduction

The self-controlled case series method, or case series method for short, can be used to study the temporal association between a time-varying exposure and an adverse event using data on cases only. The key advantages are that it often has high efficiency relative to the cohort method and that it is self-controlled: time-invariant confounders, such as sex, location, genetics, underlying state of health are controlled for implicitly. This paper begins by introducing the model and likelihood, with an example. Key assumptions, study design and model fitting are discussed.

When the method was first introduced in 1995, it was used to study the association between vaccines and adverse events [1] [2]. It gained wider recognition after its use in a study of the association between mumps, measles and rubella (MMR) vaccine and autism and since then has grown steadily in its use [3]. In 2003 Hubbard et al. were the first to use the self-controlled case series method for an exposure that was not a vaccine in a study of antidepressant use and the risk of hip fracture [4].

More recently, the self-controlled case series method has been developed further. A semiparametric approach was proposed by Farrington and Whitaker to avoid mis-specification of the age-specific baseline incidence [5]. This approach is described in Section 4. For events whose occurrence censors the subsequent exposure, such as the occurrence of an adverse event which results in stopping the treatment, the case series method can be adapted to allow for missing exposure data, as described in Section 5. The case series method can also be adapted for use in prospective surveillance of drug or vaccine safety using risk-adjusted SPRT, described in Section 6.

2 Implementation of the case series method

2.1 Model and likelihood

The self-controlled case series method is based on data on cases only, but may be derived from an underlying Poisson cohort model as follows. We begin with a cohort of individuals. For each individual i in this cohort we define an *observation period* $(a_i, b_i]$, over which data on event times and exposure history are available. For each individual, events arise according to a non-homogeneous Poisson process. Each individual's observation period is split into *age groups* j. From the exposure histories, *risk periods* k, $k = 1, 2, ...$ are defined. Risk periods are windows of time either during or after the exposure when individuals are deemed to be at increased (or reduced) risk of the event of interest. Any other times within the observation period, that is before, after or between the risk periods, constitute the control periods, indexed by $k = 0$.

Let n_{ijk} and e_{ijk} denote, respectively, the number of events experienced and the time spent by individual i in age group j and risk period k. The incidence rate, denoted by λ_{ijk}, is assumed to be constant within each such interval and at any one time point is given by the following multiplicative model:

$$\lambda_{ijk} = \exp(\phi_i + \alpha_j + \beta_k) \tag{1}$$

where ϕ_i is an effect for individual i, α_j is an effect for age group j, and β_k is an effect for risk period k. $\alpha_0 = 0$ and $\beta_0 = 0$, so that the baseline incidence λ_{i00} is $exp(\phi_i)$. The number of events n_{ijk} occurring within an interval of length e_{ijk} is assumed to be Poisson with rate $e_{ijk}\lambda_{ijk}$. The exponentiated quantities $\exp(\beta_k)$ are referred to as relative incidences and are a measure of incidence in risk period k relative to the control period $k = 0$.

Assuming that the exposure is not influenced by prior events, we condition on the number of events $n_i = \sum_{j,k} n_{ijk}$ observed for individual i during $(a_i, b_i]$. The kernel of the resulting conditional likelihood is product multinomial. The contibution of an individual i is:

$$L_i(\alpha, \beta) = \prod_{j,k} \left(\frac{e_{ijk} \exp(\alpha_j + \beta_k)}{\sum_{r,s} e_{irs} \exp(\alpha_r + \beta_s)} \right)^{n_{ijk}} . \tag{2}$$

Consequently, the individual effects ϕ_i cancel out, it is in this sense that the method is self-controlled: these individual effects include any time-invariant confounders or random effects. Only age (or other time-dependent covariates) need to be modelled, though it is possible to include the interactions between time-invariant confounders and the exposure effect.

Individuals i with no events ($n_i = 0$) in $(a_i, b_i]$ contribute 1 to the conditional likelihood in equation (2), and hence only cases within the underlying cohort, that is, individuals with $n_i \geq 1$, need be sampled. Suppose that there are n such cases, re-indexed by $i = 1, ..., n$. The self-controlled case series likelihood is then

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j,k} \left(\frac{e_{ijk} \exp(\alpha_j + \beta_k)}{\sum_{r,s} e_{irs} \exp(\alpha_r + \beta_s)} \right)^{n_{ijk}} . \tag{3}$$

Asymptotic and small sample properties of maximum likelihood estimates obtained from this conditional likelihood are discussed in Musonda et al. [6], together with simulation inference methods.

2.2 Motivating example: MMR vaccine and ITP

Idiopathic thrombocytopenic purpura (ITP) is a rare, potentially recurrent autoimmune disorder in which abnormal bleeding into the skin occurs due to low blood platelet count. Miller et al. studied the association between mumps, measles and rubella vaccine (MMR) and hospital admission for ITP within the South East and North East Thames Regions in the UK [7]. ITP cases arising during the period from October 1991 to September 1994 and aged 12-23 months were included in the analysis. These time and age boundaries were used to define the observation period for each case. The data set included a total of 44 admissions experienced by 35 children; 5 of these children were admitted twice and 1 was admitted 5 times. Six age groups were used: 366-426, 427-487, 488-548, 549-609, 610-670 and 671-730 days of age. It was hypothesized that MMR vaccination may, in rare instances, cause ITP. Risk periods covered the 6 week period after MMR vaccination, and three two-week long risk periods 0-14, 15-28 and 29-42 days after vaccination were used. Of the 44 events, 13 occurred within 6 weeks after receipt of the MMR vaccine and 31 of the 35 children were exposed between one and two years of age.

Figure 1 shows time lines of the observation period for the first three individuals $i = 1, 2, 3$ included in the data set. Observation periods ran from ages 454 to 730 days for individual 1 and from ages 366 to 730 days for individuals 2 and 3. Vaccination times and risk periods are shown for individuals 1 and 3, individual 2 was not vaccinated during his or her observation period. Individual 1 was admitted to hospital for ITP during the second post MMR risk period.

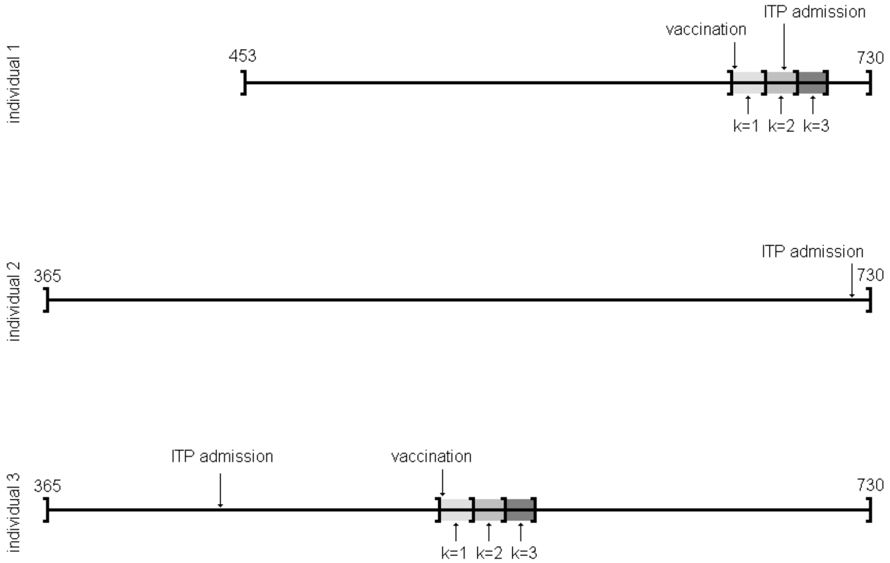


Figure 1: Observation periods for the first 3 individuals in the ITP and MMR data set.

2.3 Fitting the case series model

The multinomial likelihood (3) can be fitted using most statistical software packages as an associated Poisson regression model with log link function. Data need to be formatted with one line for each interval, listing the number of events occurring within that interval n_{ijk}, the length of the interval e_{ijk} and factors for age group j and risk group k; a factor for individual i is also included to ensure that the fitted individual totals equal the observed values. The Poisson model has the number of events n_{ijk} as the response variable and the log interval lengths $\ln(e_{ijk})$ are included as an offset:

$$\begin{aligned} n_{ijk} &\sim Poisson(\lambda_{ijk} e_{ijk}) \\ \log(\lambda_{ijk}) &= \varphi_i + \alpha_j + \beta_k \end{aligned}$$

where φ_i is an individual effect included to ensure that the fitted individual totals equal the observed values. Since there are sometimes large numbers of individuals included in a study and the φ_i are nuisance parameters, it is convenient to fit this as a conditional fixed effects Poisson regression model, with the φ_i as a group variable which is not estimated explicitly [8].

Data are more readily assembled listing one line per event, including the start and end of the observation period, exposure and event dates and an individual identifier. Example files that take data in this format, reformat it as necessary and fit the model are available for STATA, SAS, R, GLIM and GenStat from the self-controlled case series website at http://statistics.open.ac.uk/sccs/.

The relative incidences (RI) and 95% confidence intervals (CI) for each of the three post MMR risk periods for our MMR and ITP example are shown in Table 1.

Table 1: Case series analysis of MMR vaccine and ITP.

risk period (days after MMR)	no. of events	RI (95% CI)
0-14	2	1.31 (0.30, 5.73)
15-28	8	5.95 (2.52, 14.07)
29-42	3	2.60 (0.75, 9.07)
0-42	13	3.23 (1.53, 6.79)

There is a significant increase in the risk of ITP 15-28 days after receipt of the MMR vaccine.

2.4 Advantages and assumptions

The case series method has several advantages. It requires only cases, which reduces the effort and cost of data collection. Perhaps the most attractive feature of the case series method is that it is self-controlled; it implicitly adjusts for all confounders that remain fixed over the observation period, such as genetic and socio-economic factors. Time varying confounders such as age and season can be allowed for in the baseline incidence. Another advantage is that, in some circumstances, the self-controlled case series method has high efficiency relative to the cohort method (see section 3.2). However, the case series method requires the following three key assumptions to be applicable.

Assumption 1: Events arise in a non-homogeneous Poisson process.

In general the self-controlled case series method is suitable for independent recurrent events. It may also be applied to rare non-recurrent events. In this case, λ_{ijk} in equation (1) is a hazard function representing the incidence in individuals who have not experienced the event, rather than a Poisson rate. However the log likelihood (2) is valid in the limit $\phi_i \rightarrow -\infty$. This is because recurrent and non-recurrent events effectively become indistinguishable except in very large samples. Thus rare non-recurrent events may be analysed using the case series model. For more details see Farrington and Whitaker [5].

If recurrent events are not independent yet the occurrence of a first event is rare, then the method can be applied using just the first event. For example in our MMR and ITP data, 5 children had 2, and 1 child had 5 hospital admissions for ITP. It is very likely that these admissions were not independent, and the analysis should exclude any subsequent admissions after the first. The results of this analysis are given in table 2.

Table 2: MMR vaccine and ITP: first event only.

risk period (days after MMR)	no. of 1st events	RI (95% CI)
0-14	2	1.59 (0.36, 7.15)
15-28	8	7.19 (2.92, 17.73)
29-42	3	3.22 (0.89, 11.59)
0-42	13	3.94 (1.78, 8.72)

There is a significant 7-fold increase in risk of ITP 15-28 days after MMR inoculation. The increase in risk is greater than that for the analysis including all events (given in table 1) suggesting that the relative incidences were biased toward the null.

Assumption 2: The occurrence of an event must not alter the probability of subsequent exposure.

The exposure is assumed to be an external time-varying covariate, so that the exposure distribution after time t is independent of the event history prior to t. This is probably the most restrictive assumption of the self-controlled case series method. There are currently two modifications of the method which can be used in some circumstances to correct for event-dependent exposures. Both approaches may be used if the risk periods are not indefinite. The first approach may be used provided individuals receive at most a single exposure, as in the case of single dose vaccines. The observation period is redefined as starting with the age at exposure, while the end of observation is as before. Only individuals who were exposed prior to the event are included in the analysis. The disadvantage of this approach is that cases with events prior to exposure and unexposed cases are excluded from the analysis, reducing study power. The second approach may be used when individuals experience multiple exposures (as with multiple dose vaccines). In this approach, post-event exposures are ignored and the data are analysed using methods for censoring events, described in section 5.

It is sometimes possible to test whether exposures are event-dependent. This is done by including a pre-exposure ‘risk’ period, which we shall call a pre-risk period, and testing whether the relative incidence associated with this additional pre-risk period is significantly different from unity. Such a situation arises in our motivating example: individuals with ITP would not receive the MMR vaccine until they have fully recovered. It follows that there would be fewer cases in the period just before exposure, which can be allowed for by including a pre-risk period. Not allowing for this period of low incidence would inflate the relative incidences in the post exposure risk periods (conversely if exposure is more likely for a period after the event, the relative incidence would be biased toward the null). The typical duration of ITP in children is 2-4 weeks. Including a 28-day pre-risk period gives the results in table 3 (this analysis uses only first events so should be compared with the results in Table 2).

Table 3: MMR vaccine and ITP: case series analysis with pre-risk period

risk period (days after MMR)	RI (95% CI)
0-14	1.28 (0.28, 5.85)
15-28	5.84 (2.32, 14.73)
29-42	2.63 (0.72, 9.61)
0-42	3.20 (1.41, 7.26)

There were no events in the pre-risk period, thus no-one was vaccinated in the 28-day period after the onset of ITP; in this data set the first post-event vaccine dose was 46 days after hospital admission for ITP. Comparing these results with results with those of table 2, there is weak evidence that the relative incidence was inflated when the short term dependence between event and exposure was not taken into account. This method of testing for event-dependent exposures is studied further in Farrington and Whitaker [5].

Assumption 3: The occurrence of the event of interest must not censor or affect the observation period.

Each individual's observation period is usually defined implicitly by combination of calendar time and age limits, but must be independent of the event date.

This assumption may be violated when the event of interest is likely to increase the short-term death rate. A case series method has been developed to deal with the situation where the observation period is completely censored after the event: this will be described in section 5. The impact of event-dependent observation periods on case series estimates remains a topic for further research. Case series studies where the event is likely to result in death have been carried out, such as the study by Smeeth et al. on the association between acute infection or vaccination and myocardial infarction [9]. A simulation study showed that in this study's case, the bias was negligible [5].

2.5 Designing a case series study

A case series design is typically used to study temporal associations for which effect of exposure is confined to a finite risk period, after which the risk returns to the baseline level. Indefinite risk periods can sometimes be used provided there are sufficient unexposed cases, though this results in a loss of efficiency (see subsection 3.2 on relative efficiency and subsection 7.1 on indefinite risk periods). Thus self-controlled case series analyses are best suited to studies on the association between a transient exposure and an acute event, though any event for which it is possible to assign a date of onset, such as date of hospital admission or diagnosis, should be suitable for study.

Event data may come from databases such as hospital admissions or GP data, and exposure histories should be documented using a data source independent of event dates. In particular, spontaneous reporting systems for adverse drug reactions are inappropriate sources of data for the case series method, since reporting is strongly dependent on the time interval between exposure and the event.

Data are usually collected during a predefined study period given in terms of calendar time and possibly age boundaries, typically determined by the availability of database records. These boundaries are usually used to assign an observation period to each case. Typically observation periods may vary between cases. It is advisable to select the study period so as to maximize the chance of exposure. For example, in the study of ITP and MMR, the study period was the second year of life, which corresponds to the age range in which children are most likely to receive primary MMR vaccination. Once observation periods are set in place age groups can be chosen to capture the change in incidence of the event of interest with increasing age (but see section 4). If there are other potential temporal confounders these also need to be allowed for. For example if the event rate varies with season, this could be allowed for using a calendar month effect or a 4 level factor for season.

The risk periods should be chosen a priori. In reality experts often find it difficult to judge how long a potential temporal association between exposure and event may last for, and it may be sensible to choose a small number of contiguous risk periods for each exposure and later combine periods where the relative incidence appears to be similar. Note that if a true association exists and the risk periods do not include the full period for which the temporal association lasts, the relative incidences will be biased toward the null.

Some individuals may experience repeat exposures, for example repeat prescriptions of a drug. If the risk is expected to be the same after each exposure, risk periods after each repeat exposure can be assigned the same factor levels. If the risk is expected to differ each time, different factor levels should apply after each exposure in order to test for a dose effect. See Whitaker et al. [8] for a full discussion of case series studies with multiple risk periods and repeat exposures.

2.6 Sample size

The following formula can be used to calculate the number of events n required to detect an association of relative incidence e^β with power γ and significance level α. This formula ignores any age effects, and assumes that all individuals have observation periods of the same length.

$$\begin{aligned}
A &= 2\frac{p(e^\beta r+1-r)}{1+pr(e^\beta-1)}\left[\beta\left(\frac{e^\beta r}{e^\beta r+1-r}\right)-\log(e^\beta r+1-r)\right] \\
B &= \frac{\beta^2}{A}\left[\frac{p(e^\beta r+1-r)}{1+pr(e^\beta-1)}\right]\left[\frac{e^\beta r(1-r)}{(e^\beta r+1-r)^2}\right] \\
n &= \frac{\left(z_{1-\alpha/2}+z_\gamma\sqrt{B}\right)^2}{A}
\end{aligned}$$

where r is the length of the risk period divided by the length of the observation period and p is the proportion of the population that are exposed at some time during their observation period. Note that the formula calculates the number of events n and not the number of individuals. In fact, events are assumed independent and two events within one individual contribute the same as two individuals with one event. Further details and sample size formulae that take into account age effects can be found in Musonda et al. [10].

3 Comparison with cohort and case-control methods

The examples in section 3.1 compare the results obtained with the case series method with those obtained from cohort and case-control studies. Examples 1 and 2 show the benefit of using the case series method over a full cohort study in reducing the data collection effort and dealing with incomplete information on confounders. Examples 2 and 3 show the advantage of the case series method in controlling for selection and indication bias compared with a case-control study. In section 3.2 the relative efficiency of the three approaches is studied in a simple case. We do not make any comparison with other case-only designs such as the case-crossover method [11], which is a special case of the case-control method. Farrington [12] provides an overall review of case-only methods.

3.1 Comparative study examples

Example 1: Febrile convulsions and MMR vaccine

Barlow *et al.* [13] investigated the association between MMR vaccine and febrile seizures using a large cohort of 679942 children under seven years old, living in the US. The cohort included 716 cases. The relative risk of first febrile seizure in the second week after MMR vaccine ($8-14$ days) was 2.83, with 95% CI $(1.44, 5.55)$. A similar investigation undertaken by Farrington *et al.* [14] using a case series study of 952 cases aged $12-23$ months living in England showed that the MMR vaccine was associated with febrile convulsions with a relative incidence of 3.04, 95% CI $(2.27, 4.07)$. The case series study thus gave an estimate of relative incidence close to the value obtained from the full cohort study, but with a narrower confidence interval. When the exposure risk period is short in comparison to the observation period as in this example, the case series analysis has good efficiency relative to a cohort study. In this instance, the case series method produces greater precision largely because the cohort study, though very large, included fewer events than the case series study.

Example 2: Asthma and influenza vaccine

Kramarz *et al.* [15] used data from a cohort of 70753 asthmatic children aged $1-6$ years during the $1995-1996$ asthma season living on the West Coast of the US to determine whether influenza vaccination precipitates asthma exacerbations. The crude rate ratio of asthma exacerbation within two weeks after influenza vaccination was 3.29, 95% CI $(2.55, 4.15)$, which decreased after adjustment for sex, age, calendar time, indicators of asthma severity and preventive care practices to 1.39, 95% CI $(1.08, 1.77)$. A case series analysis was undertaken in the 2075 children within the cohort who had at least one asthma exacerbation during the influenza season. The relative incidence of asthma exacerbation within the 2 weeks after influenza vaccination was found to be 0.98, 95% CI $(0.76, 1.27)$. One interpretation is that the case series method controls fully for indication bias associated with the confounding effect of underlying asthma severity, as children with more severe conditions are the most likely to receive the vaccine. In the cohort analysis, only partial control of such confounding may have been achieved by adjusting on proxy indicators of severity.

Example 3: Hip fracture and tricyclic antidepressants

Hubbard *et al.* [4] performed case-control and case series analyses to investigate the association between tricyclic antidepressant drugs and the risk of hip fracture. The data set included 16341 cases of hip fracture and 29889 controls drawn from the United Kingdom General Practice Research Database (GPRD). The odds ratio for fracture within the 15 days following a prescription of tricyclic antidepressants was 4.76, 95% CI $(3.06, 7.41)$, whereas the equivalent relative incidence obtained from the case series analysis was 2.30, 95% CI $(1.82, 2.90)$. Both study designs show a positive association between tricyclic antidepressants and hip fracture during the first 15 days of treatment, though the strength of association from the case control study is double that from the case series study. It is thought that the estimates gained from the case series approach may be more accurate because the problems of selection and indication bias in the case-control study are removed. It is also possible that the case series result could be biased toward the null, say if individuals are more likely to take antidepressants after fracturing a hip.

These examples show how useful it can be to use more than one study design, and indeed when suitable data arise we recommend doing a case series analysis as well as a cohort or case-control analysis. If estimates obtained from both study designs agree this strengthens the findings. If results do not agree it can throw light on the sources of bias.

3.2 Relative efficiency

In this subsection an expression is presented for the relative asymptotic efficiency of the case series design, relative to an unmatched case-control design with C controls per case, for estimating the log relative incidence β for an uncommon event in a simple scenario. Provided that the underlying incidence is very small, the log odds ratio coincides with β, and so it makes sense to compare asymptotic variances.

In the scenario we consider, the underlying population comprises T individuals followed up over the same period. Of these individuals, a proportion p experience a risk period of length e_1 and a control period of length e_0, while the remainder experience only a control period, of length $e_0 + e_1$. The ratio of the risk period to the observation period is $r = e_1/(e_0 + e_1)$. There are n cases. In the case-control study, a case with onset at time t is considered exposed if exposure occurred within $(t - e_1, t]$; for the controls a randomly selected interval of length e_1 is used to determine exposure.

As shown in the Appendix, asymptotically as $T \to \infty$ (and hence $n \to \infty$), the relative efficiency of the case series method compared with the case-control method with C controls per case is:

$$RE_C = \frac{1-r}{C\,(1-pr)} . \frac{e^{\beta} + C\left(e^{\beta}pr + 1 - pr\right)^2}{\left(e^{\beta}pr + 1 - pr\right) . \left(e^{\beta}r + 1 - r\right)}.$$

Note that as the number of controls per case increases, the relative efficiency tends to

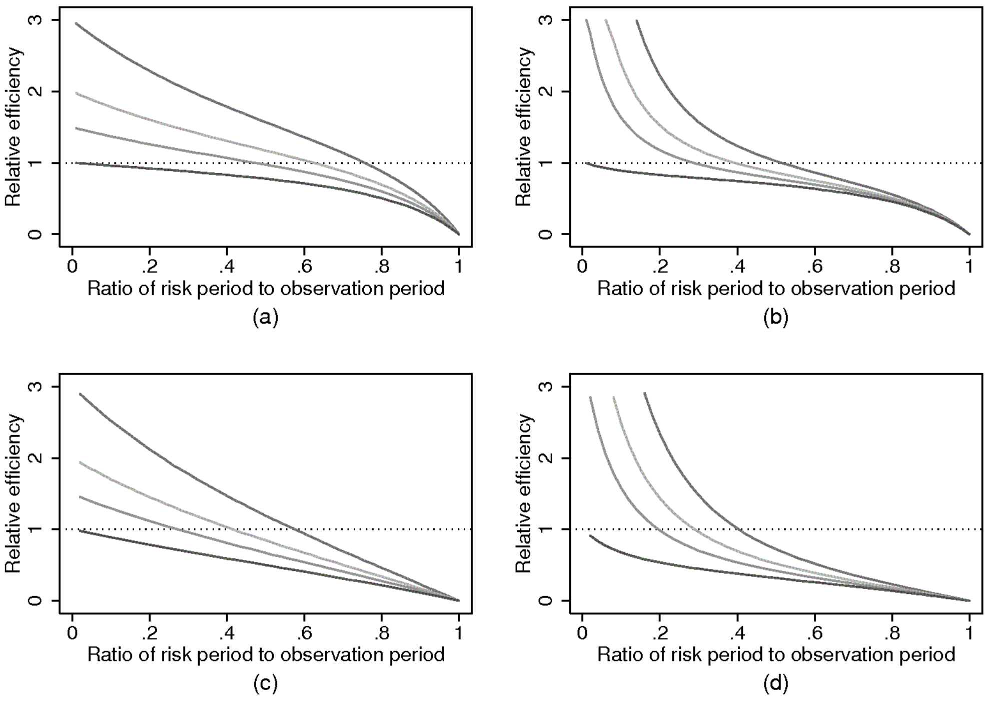
$$RE_{\infty} = \frac{1-r}{1-pr} . \frac{e^{\beta}pr + 1 - pr}{e^{\beta}r + 1 - r}$$

which, as expected, coincides with the relative efficiency of the case series method compared to the cohort method in this simple scenario [1] [5].

Figures 2(a) to (d) show the relative efficiency for $\beta = \log(2)$ and $\beta = \log(10)$, and for exposure prevalences $p = 0.4$ and $p = 0.8$. The case series method is more efficient than the case-control method with C controls per case when r is small, this efficiency advantage reducing as r increases, p declines and β increases. When $p < 1$, then as the number of controls per case increases, the case-control method eventually becomes more efficient than the case series method, equi-efficiency occurring when

$$C = \frac{1-r}{r\,(1-p)\,(e^{\beta}pr + 1 - pr)}.$$

The case series method is less efficient than the cohort method except when $p = 1$, though the loss in efficiency is small when r is low and p is high. A more general discussion of the relative efficiency of the case series method relative to the cohort method may be found in Farrington & Whitaker [5], though the qualitative findings from the simple scenario presented here hold more generally as well.



(a)

(b)

(c)

(d)

Figure 2: Asymptotic relative efficiency (a) $RI = 2$, $p = 0.8$, (b) $RI = 10$, $p = 0.8$, (c) $RI = 2$, $p = 0.4$, (d) $RI = 10$, $p = 0.4$. In each panel the four lines (from top to bottom) are for $C = 1, 2, 4$ and ∞ controls per case.

4 Semiparametric case series approach

In the parametric case series model described in Section 2, the age effect is assumed to be constant in pre-determined age groups and the age-specific relative incidence is then approximatively represented by a step function with steps $\exp(\alpha_j)$ in the j^{th} age group. However, mis-specification of the age groups can produce biased estimates of β. These considerations motivated the development of a semi-parametric model in which the age effect is left unspecified. In this model, the cumulative baseline relative incidence is modelled by a step function with jumps $\exp(\alpha_r)$ at the distinct event times t_r. The contribution to the semi-parametric case series likelihood of a case i who experienced n_i events is:

$$L_i(\beta) = \prod_{j=1}^{n_i} \frac{\exp\left[\alpha_{ij} + \beta x_i(t_{ij})\right]}{\sum_r \exp\left[\alpha_r + \beta x_i(t_r)\right]}$$

where t_{ij} is the j^{th} event time of case i, $x_i(t)$ is the exposure at time t for case i, and α_{ij} is the value α_r corresponding to t_{ij}.

An additional benefit of the semiparametric model is that it implicitly controls for exponential time trends: these are confounded with the age effects but do not influence the exposure effects.

Example: MMR vaccine and bleeding disorders

Using the MMR and ITP data for only the first event and including a 28-day pre-exposure period, the semiparametric model yields the relative incidences given in table 4.

Table 4: MMR vaccine and ITP: semiparametric case series model.

risk period (days after MMR)	RI (95% CI)
0-14	1.34 (0.28, 6.41)
15-28	5.61 (1.97, 15.97)
29-42	2.06 (0.51, 8.24).
0-42	2.90 (1.20, 7.03)

The results are comparable to those in table 3 obtained using the parametric model with 6 age groups.

Example: Hepatitis B vaccination and multiple sclerosis

The hypothesis that hepatitis B vaccination is a risk factor for multiple sclerosis has been discussed at length. Data from a case-control study in France in 1998 were re-analysed using the case series method [16] [17]. Both the event and vaccination were age dependent, and hence age was a potential confounder if not fully allowed for in the model. The relative incidence for the $0-60$ day post vaccination risk period using a case series model with no age effect was 2.11 with 95% confidence interval $(1.06, 4.20)$, this became 2.00 $(0.96, 4.18)$ using a parametric model with 48 1-year age classes, and dropped to 1.68 $(0.77, 3.68)$ using the semiparametric model. Clearly, there is substantial confounding by age which was reduced, but not eliminated, by using 48 age groups. The semiparametric model provided the most reliable estimates in this context, as no prior assumptions were made about the age effect.

5 Self-controlled case series approach for censoring events

Key assumptions of the self-controlled case series method are that the exposure distribution within the observation period, and the observation period itself must be independent of prior event times. These requirements inhibit the use of the case series method for censoring events, that is, events whose occurrence censors subsequent exposures. For example, if the exposure of interest is an intervention which is contra-indicated after occurrence of the event, no exposures can occur after the event, but the individual remains under observation. The most extreme example of a censoring event is death whose occurrence censors the entire post-event exposure history.

5.1 Unique exposures

If the event of interest is of the censoring type, the use of the multinomial likelihood (3) obtained by conditioning on the total number of events and exposure history is not valid since censoring events give rise to missing exposure data. Suppose the boundaries a_i, b_i define the observation period that would have applied had no censoring occurred, and suppose to begin with that all cases experience only one exposure. We proceed by including only exposed cases in the analysis and redefine the observation period for each individual i exposed once at age c_i, to be $(c_i, b_i]$ rather than $(a_i, b_i]$, see Figure 3. Indeed, if there is a unique exposure occurring at age c_i, the subsequent exposure history is always known within $(c_i, b_i]$, since no further exposures occur. Thus, the case series likelihood (3) is valid and the relative incidence can be estimated at the cost of ignoring unexposed cases.

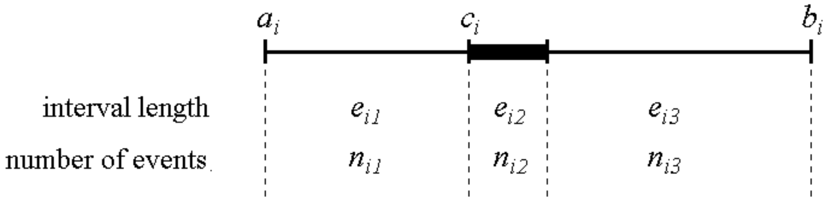


Figure 3: configuration for one exposure

Example: Bupropion and sudden death

Bupropion is an effective smoking cessation therapy but its use in the UK has been limited by concerns that it may increase the risk of sudden death. Data for 9329 individuals who had been prescribed Bupropion were analysed using the self-controlled case series method to estimate the relative incidences of seizures and death during the first 28 days of treatment [19]. The relative incidence of seizures during the first 28 days of treatment, using a standard case series analysis, was 3.62 with 95% CI $(0.87, 15.09)$. As sudden death censors Bupropion therapy, we applied the case series approach for censoring events described in Section 5.1, by defining the individual's observation period to start with the age Bupropion treatment was first started, and to end with age on 11 November 2003. A total of 121 cases of sudden death were included in the analysis, including 2 in the risk period 0-27 days after the start of treatment. The relative incidence was 0.50, 95% CI $(0.12, 2.05)$. These results shows that Bupropion use is probably associated with an increased risk of seizures, but no evidence was found to suggest that the drug is associated with an increased risk of sudden death [18].

5.2 Two or more exposures

In the case of a single exposure and ignoring any age effect, the case series log likelihood contribution of individual i is

$$l_i(\beta) = \begin{cases} n_{i2}\beta - (n_{i1} + n_{i2}) \log\left(e^{\beta} e_{i2} + e_{i3}\right) & \text{if } n_{i1} = 0, \\ 1 & \text{if } n_{i1} = 1 \end{cases}$$

and the elementary score function for β is given by:

$$U(\beta) = n_{i2} - (n_{i2} + n_{i3}) \frac{e_{i2} e^{\beta}}{e_{i2} e^{\beta} + e_{i3}}$$

where n_{id} and e_{id} represent the number of events during, and length of interval d; $d = 1$ during the pre-exposure control period prior to exposure, $d = 2$ during the exposure risk period and $d = 3$ during the post-risk control period as shown in figure 3.

Now consider the case where each individual i is exposed up to two times at ages c_{i1} and c_{i2}, with $c_{i1} \leq c_{i2}$. Let β_1 and β_2 denote the log relative incidences associated with each of the two exposure risk periods ($k = 1$ and $d = 2$, $k = 2$ and $d = 4$), as shown in Figure 4 (note the new d illustrated in the figure). Inference about β_2 can be made using the method described for single exposures, by using the case series likelihood restricted to cases with events in $(c_{i2}, b_i]$. This yields the following score function

$$U_i^1(\beta_2) = n_{i4} - (n_{i4} + n_{i5}) \frac{e^{\beta_2} e_{i4}}{e^{\beta_2} e_{i4} + e_{i5}}.$$

To make inference about β_1, only cases arising in $(c_{i1}, b_i]$ are used. We proceed as if no individual can be exposed a second time at some possibly unobserved age c_{i2}, and denote n_{i4}^* the number of events that would have arisen in the new control period that now replaces the second risk period ($k = 2$ and $d = 4$). If n_{i4}^* were observed, the method for single exposures could be used, applied to cases with events after the first exposure at age c_{i1}. This would then yield the elementary score function

$$U_i^2(\beta_1) = n_{i2} - (n_{i2} + n_{i3} + n_{i4}^* + n_{i5}) \frac{e^{\beta_1} e_{i2}}{e^{\beta_1} e_{i2} + e_{i3} + e_{i4} + e_{i5}}.$$

However, n_{i4}^* is not observed so this score function cannot be evaluated. We therefore replace n_{i4}^* by an unbiased estimator of n_{i4}^*, namely the Horvitz-Thompson-like estimator $n_{i4} e^{-\beta_2}$.

In the general case with more than two exposures and including age groups, the estimating equations and sandwich variance estimators are given in Farrington et al. [18]. In such settings, writing down and solving the estimating equations, and deriving the sandwich variance estimator, becomes extremely cumbersome, and is impractical. An alternative approach more convenient for computation is based on a pseudo-likelihood method. Estimates are obtained by an iterative procedure, in which at each iteration the missing data n_{ijk}^* are replaced by their Horvitz-Thompson-like expected values. The pseudo-likelihood method provides a simple way of obtaining parameter estimates using standard Poisson regression software. It also can be exploited to obtain bootstrap standard errors and interval estimates [18].

Example: Oral polio vaccine and intussusception

Intussusception is a condition where the bowel folds in on itself causing an obstruction. Most children diagnosed with intussusception have an operation and recover completely, so normally this

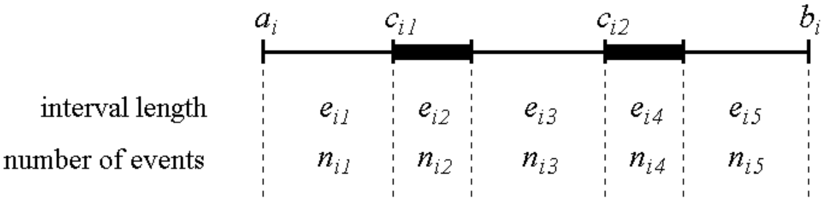


Figure 4: Configuration for two exposures.

is not a censoring event. The data are 456 first hospital admissions for intussusception of children up to age 2 years, collected between 2002 and 2005 in 11 Latin American countries. The vaccination history of each case was collected through an interview with the child's parents at the time of the child's treatment. There was no follow up, so the subsequent vaccination history was censored. The children received up to 5 doses of the oral polio vaccine (OPV). A single risk period $0 - 30$ days after each OPV dose is used, giving 5 risk periods. We used 15 one or three month age groups [18].

Relative incidences and 95% percentile bootstrap confidence intervals are given in Table 5.

Table 5: Censoring event case series analysis of OPV vaccine and intussusception.

risk period 0-30 days after dose:	RI (95% CI)
1	1.25 (0.63, 2.24)
2	1.07 (0.74, 1.53)
3	0.97 (0.68, 1.37)
4	1.00 (0.50, 1.62)
5	1.60 (0.00, 6.13)
any dose	1.05 (0.82, 1.34)

No significant change in incidence of intussusception in the 30-day period after vaccination with OPV relative to the control periods was found.

6 Sequential approach for pharmacovigilance

6.1 The case series SPRT

The case series method is rapid, easy to apply and avoids confounding, this recommends it for use in pharmacovigilance, as well as pharmacoepidemiology. A sequential case series method has been developed for prospectively monitoring the safety of a new drug or a new vaccine by formally testing the null hypothesis of absence of new drug or vaccine effect $H_0 : e^{\beta} = 1$, where e^{β} is the relative incidence, versus the alternative hypothesis $H_1 : e^{\beta} = e^{\beta_A}$ where e^{β_A} is the alternative hypothesis relative incidence value that we wish to detect ($e^{\beta_A} > 1$). To perform this test, a risk-adjusted sequential probability ratio test (SPRT) can be used [20]. This test is applied at the end of successive surveillance intervals, a calendar time period of set length.

At the s^{th} surveillance interval, the age effect is profiled out by maximizing the log likelihood under H_1 and under H_0 (simulations have shown that using profile likelihoods in this way has little bearing on the results) [20]. The logarithm of the case series profile likelihood ratio test statistic Λ_s is calculated, and the value of the sequential test statistic Z_k corresponding to the running total is obtained as follows:

$$Z_s = Z_{s-1} + \Lambda_s, \text{ and } Z_0 = 0.$$

The value of Z_s is compared to pre-determined test boundaries $\log(A)$ and $\log(B)$, conventionally specified in terms of nominal error probabilities α^* and β^*, with $A = \beta^*(1-\alpha^*)^{-1}$ and $B = (1-\beta^*)\alpha^{*-1}$.

If $\log(A) \leq Z_s \leq \log(B)$, then no decision is made and the procedure is repeated at step $s+1$, otherwise, the sequential test terminates at step s. In this case, if $Z_s > \log(B)$ then the alternative hypothesis H_1 is accepted and if $Z_s < \log(A)$ then the null hypothesis H_0 is accepted.

Example: influenza vaccine and Bell's Palsy

After the introduction of an inactivated nasal formulation of the influenza vaccine in Switzerland in October 2000, an increased number of Bell's palsy, an acute facial paralysis affecting the 7^{th} facial nerve were reported. Using the case series method, the relative incidence within the $31-60$ day post-vaccination risk period was estimated to be 35.6, 95% CI $(14.1-89.8)$ [21]. A case series analysis was performed using a total of 2263 episodes of Bell's palsy in the UK after the standard influenza vaccine, recorded in the General Practice Research Database from July 1^{st}, 1992 to 30^{th} June 2005. A non significant relative incidence of Bell's palsy in the 3 months following parenteral inactivated influenza vaccine was found, RI 0.92, 95% CI $(0.78-1.08)$.

We reanalysed these UK data using the case series SPRT, as if they were collected prospectively. We used a six month surveillance interval, $1-60$ day risk period after any dose of influenza vaccine, for two values of the alternative hypothesis relative incidence $e^{\beta_A} = 1.5$ and 5. In view of possible temporal confounding from the highly seasonal administration of influenza vaccine, the analysis was performed using a parametric case series model with 12 one month calendar time periods.

Figure 5(a) shows that the null hypothesis is accepted at the end of the 3^{rd} surveillance interval, that is, after 18 months (for either $\alpha^* = \beta^* = 0.001$ or $\alpha^* = \beta^* = 0.01$). Figure 4(b) shows that if the alternative hypothesis relative incidence is $e^{\beta_A} = 1.5$, then the issue remains undecided after 12 years, or 10.5 years if the upper boundaries correspond to $\alpha^* = \beta^* = 0.001$ or $\alpha^* = \beta^* = 0.01$, respectively. Thus the case series SPRT is sensitive to the choice of e^{β_A}.

6.2 Some recommendations

The case series SPRT is applicable primarily when risk periods are short. A surveillance interval of 6 months is recommended for use, except when events are so rare that there is an appreciable probability of no cases arising within a 6-month period, in which case we suggest using 1-year surveillance intervals. For most vaccination programmes, it is essential to control for age effects or, occasionally, for seasonal effects as in the example on influenza vaccine and Bell's palsy. We recommend that this is done by pre-selecting age groups (or seasonal intervals) in which the effect can reasonably be assumed constant, and using profile likelihoods to eliminate the nuisance parameters. Failing to allow for age and temporal effects may produce erroneous results. See Hocine et al. [17] for further details.

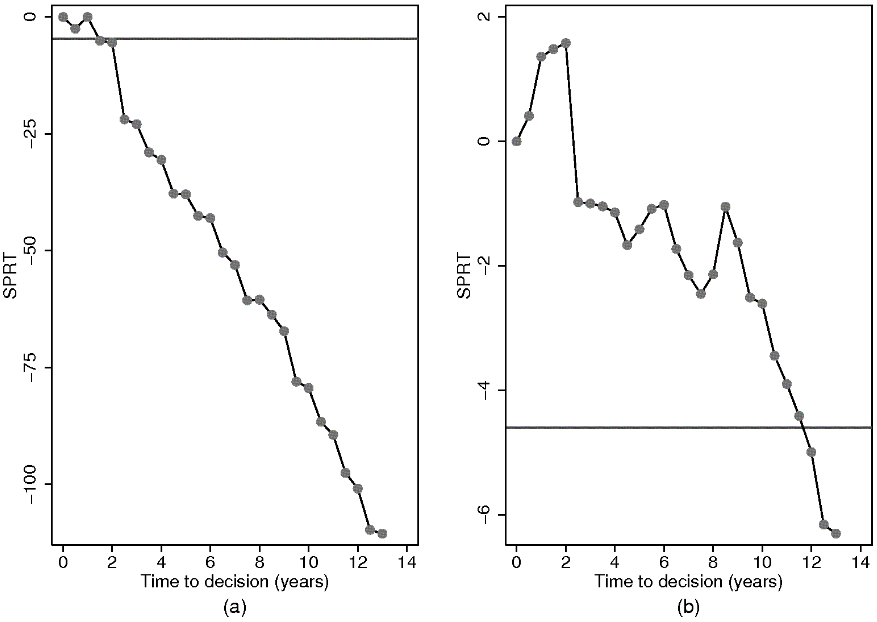


Figure 5: SPRT for 1 - 60 day risk period using Influenza and Bell's Palsy data. Lower boundary $-4.6(\alpha^* = \beta^* = 0.01)$; upper boundary not shown. (a) $e^{\beta_A} = 5$; (b) $e^{\beta_A} = 1.5$.

The SPRT is ideal for monitoring a new drug. A related approach, using the CUSUM, can be used for long-term monitoring of one or more established vaccines assumed to be safe. Indeed, the CUSUM never results in acceptance of the null hypothesis and in this sense is well-suited for long-term monitoring. The aim of such monitoring could be to identify problems resulting from changes in drug production or delivery over time.

7 Some further extensions

7.1 Long and indefinite risk periods

The case series method is most efficient when the risk periods are short in comparison to the observation period. However, it is possible to apply the case series method to non-acute events using long or indefinite risk periods. If risk periods are indefinite this is only recommended if there are some unexposed cases, since age and exposure risk effects may otherwise be confounded with each other. Simulations have shown that relatively few unexposed cases are required to avoid serious confounding with age [6]. The inclusion of unexposed cases is also advisable in studies using long risk periods. We also recommend using the semi-parametric model for such analyses.

Example: MMR vaccine and autism

In the original case series analyses of the association between MMR vaccine and autism, risk of autism diagnosis within 0.5, 1 and 2 years of MMR vaccination were studied; no significant association was found [3]. In a later reanalysis incidence of autism diagnosis both within 5 years and any time after the vaccine was also considered [22]. An appreciable number of autism cases were unvaccinated. The relative incidence of autism diagnosis at any time after MMR vaccination was 1.06, with 95% confidence interval (0.49, 2.30), again, there was no evidence to support an association. The indefinite risk period reanalysis was later performed using the semi-parametric model, which is advisable here. This yielded a slightly lower relative incidence of 0.88, with 95% confidence interval (0.40, 1.95) [5].

7.2 Bivariate counts

In a case series study in which two types of event R and S may occur, the effect of an intermittent exposure on the occurrence of R rather than S, may be evaluated by a conditional relative incidence denoted RR_C. The RR_C is defined as the ratio between RR^R and RR^S, the relative risks of R and S, respectively, and is estimated by the ratio of their estimates obtained separately by maximizing the appropriate likelihood (3). The variance of the RR_C estimate is simply the sum of the variances of the estimates of RR^R and RR^S under the assumption of independence of the counts of R and S [23]. A test for checking such an assumption was proposed by Hocine *et al.* [24]. The test statistic was derived from a conditional likelihood using a bivariate Poisson-generated multinomial model. Robust estimation is described in Hocine *et al* [25].

8 Future developments

This paper gives a general overview of the self-controlled case series method and its extensions available to date. The method is still under development and future work includes looking at replacing risk periods with smooth functions of risk. We also hope to investigate the independence of multiple events and the extent of any bias introduced when the observation period depends on the event of interest.

References

[1] Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* 1995; **51**: 228-35.

[2] Farrington CP, Pugh S, Colville A, Flower A, Nash J, Morgan-Capner P, Rush M and Miller E. A new method for active surveillance of adverse events from diphtheria tetanus pertussis and measles mumps rubella vaccines. *Lancet* 1995; **345**: 567-569.

[3] Taylor B, Miller E, Farrington CP, Petropoulos M-C, Favot-Mayaud I, Li J, Waight PA. Autism and measles, mumps and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 1999; **353**: 2026-2029.

[4] Hubbard R, Farrington P, Smith C, Smeeth L, Tattersfield A. Exposure to Tricyclic and selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. *American Journal of Epidemiology* 2003; **158**: 77-84.

[5] Farrington CP, Whitaker HJ. Semiparametric analysis of case series data (with discussion). *Applied Statistics* 2006; **55**: 553-94.

[6] Musonda P, Hocine MN, Whitaker HJ and Farrington CP. Self-controlled case series analyses: small sample performance. *Computational Statistics and Data Analysis* 2007; in press.

[7] Miller E, Waight P, Farrington P, Andrews N, Stowe J and Taylor B. Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of disease in childhood* 2001; **84**: 227-229.

[8] Whitaker HJ, Farrington CP, Spiessens B and Musonda P. Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine* 2006; **25**: 1768-97.

[9] Smeeth L, Thomas SL, Hall AJ, Hubbard R, Farrington P and Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. *New England Journal of Medicine* 2004; **351**: 2611-2618.

[10] Musonda P, Farrington CP, Whitaker HJ. Sample sizes for self-controlled case series studies. *Statistics in Medicine* 2006; **25**: 2618-2631.

[11] Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* 1991; **133**: 144-153.

[12] Farrington CP. Control without separate controls: evaluation of vaccine safety using case-only methods. *Vaccine* 2006; **22**: 2064-2070.

[13] Barlow WE, Davis RL, Glasser JW, Rhodes PH, Thompson RS, Mullooly JP, Black SB, Shinefield HR, Ward JI, Marcy SM, DeStefano F, Chen RT, for the Centers for Disease Control and Prevention Vaccine Safety Datalink Working Group. The risk of seizures after receipt of whole-cell pertussis or measles, mumps, and rubella vaccine. *The New England Journal of Medicine* 2001; **345**: 656-61.

[14] Farrington CP, Pugh S, Colville A, Flower A, Nash J, Morgan-Capner P, Rush M and Miller E. A new method for active surveillance of adverse events from diphtheria tetanus pertussis and measles mumps rubella vaccines. *Lancet* 1995; **345**: 567-569.

[15] Kramarz P, DeStefano F, Gargiullo PM, Davis RL, Chen RT, Mullooly JP, Black SB, Shinefield HR, Bohlke K, Ward JI, Marcy SM, for the Vaccine Safety Datalink Team. *Archives of Family Medicine* 2000; **9**: 617-23.

[16] Touzé E, Fourrier A, Rue-Fenouche C, Rende-Oustau V, Jeantaud I, Begaud B, et al. HB vaccination and first central nervous system demyelinating event: a case-control study. *Neuroepidemiology* 2002; **21**: 180-6.

[17] Hocine MN, Farrington CP, Touzé E, Whitaker HJ, Fourrier A, Moreau T, Tubert-Bitter P. Hepatitis B vaccination and first central nervous system demyelinating events: Reanalysis of a case-control study using the self-controlled case series method. *Vaccine* 2007; **25**: 5938-43.

[18] Farrington CP, Whitaker HJ and Hocine MN. Case series analyses of censoring events. Technical report 2006; 06/12. (available from http://statistics.open.ac.uk/TechnicalReports/TechnicalReportsIntro.htm)

[19] Hubbard R, Lewis S, West J, Smith C, Godfrey C, Smeeth L, Farrington P and Britton J. Bupropion and the risk of sudden death: a self-controlled case-series analysis using The Health Improvement Network. *Thorax* 2005; **60**: 848-850.

[20] Hocine MN, Musonda P, Andrews NJ and Farrington CP. Sequential case series for pharmacovigilance. Technical report 2007; 07/07. (available from http://statistics.open.ac.uk/TechnicalReports/TechnicalReportsIntro.htm)

[21] Stowe J, Andrews N, Wise L and Miller E. Bell's palsy after parenteral inactivated influenza vaccine. *Human Vaccines* 2006; **2**: 110-112.

[22] Farrington CP, Miller E and Taylor B. MMR and autism: further evidence against a causal association. *Vaccine* 2001; **19**: 3632-3635.

[23] Hocine M, Tubert-Bitter P, Moreau T, Chavance M, Varon E, Guillemot D. A relative risks ratio between antibiotic use and recurrent bacteria colonization in cohort and case-series studies. *Journal of Clinical Epidemiology* 2007; **60**: 361-365.

[24] Hocine M, Guillemot D, Tubert-Bitter P and Moreau T. Testing independence between two Poisson-generated multinomial variables in case-series and cohort studies. *Statistics in Medicine* 2005; **24**: 4035-4044.

[25] Hocine M, Moreau T and Chavance M (2006). Discussion on the paper by Farrington and Whitaker. *Applied Statistics* 2006; **55**: 585-586.

Appendix

This Appendix provides some details of the calculation of the relative efficiency of the case series method relative to the case-control method for rare events. The scenario considered is as described in the main text.

Suppose that n events arise in the T individuals in the underlying population, including m events among individuals who experience the exposure. Of these m events, suppose that m_0 arise in the control period and m_1 in the risk period. The case series log-likelihood is:

$$l_{cs} = m_0 \log \frac{e_0}{e_0 + e_1 e^{\beta}} + m_1 \log \frac{e_1 e^{\beta}}{e_0 + e_1 e^{\beta}},$$

and the asymptotic variance of the log relative incidence $\widehat{\beta}$ in the case series design is given by:

$$var_{cs}(\widehat{\beta}) = \frac{1}{m} \frac{\left(e^{\beta} r + 1 - r\right)^2}{e^{\beta} r(1-r)}.$$

If e^{ϕ} denotes the baseline incidence in the absence of exposure, then

$$E(m) = T e^{\phi} p (e_1 e^{\beta} + e_0).$$

Now consider an unmatched case-control study based on the n cases arising in the same population as described above, and with $N = Cn$ controls sampled from the non-cases. We assume that the underlying incidence $\exp(\phi)$ is very small, so β is approximately equal to the log odds ratio. The case-control log likelihood is:

$$l_{cc} = n\psi + n_1\beta - (n_1 + N_1)\log\left(1 + e^{\psi+\beta}\right) - (n_0 + N_0)\log\left(1 + e^{\psi}\right)$$

where n_1 and n_0 are the number of exposed and unexposed cases, respectively, and N_1 and N_0 are the number of exposed and unexposed controls, respectively, and ψ is an additional parameter that depends on the sampling proportions. Letting π denote the expected proportion of exposed controls,

$$\exp(-\psi) = C(e^{\beta}\pi + 1 - \pi).$$

Since the event is rare, the expected proportion of exposed controls is approximately $\pi = pr$. It follows that the asymptotic case-control variance of $\widehat{\beta}$ for rare events is given by:

$$var_{cc}\left(\widehat{\beta}\right) = \frac{e^{\beta} + C(e^{\beta}pr + 1 - pr)^2}{nCe^{\beta}pr(1 - pr)}.$$

Furthermore,

$$E(n) = Te^{\phi}[p(e_1e^{\beta} + e_0) + (1 - p)(e_1 + e_0)].$$

Hence, asymptotically as $T \to \infty$ (and hence $n \to \infty$), the relative efficiency of the case series method compared with the case-control method for rare events with k controls per case is:

$$RE_C = \lim_{T\to\infty} \frac{var_{cc}(\widehat{\beta})}{var_{cs}(\widehat{\beta})} = \frac{1 - r}{C\,(1 - pr)} . \frac{e^{\beta} + C\left(e^{\beta}pr + 1 - pr\right)^2}{\left(e^{\beta}pr + 1 - pr\right) . \left(e^{\beta}r + 1 - r\right)}.$$