Case series analysis for censored, perturbed or curtailed post-event exposures

C. P. FARRINGTON, H. J. WHITAKER,

AND M. N. HOCINE

Department of Mathematics and Statistics, The Open University,

Milton Keynes, MK7 6AA, UK.

December 6, 2007

SUMMARY

A new method is developed for analysing case series data in situations where occurrence of the event censors, curtails or otherwise affects post-event exposures. Unbiased estimating equations derived from the self-controlled case series model are adapted to allow for exposures whose occurrence or observation are influenced by the event. The method applies to transient point exposures and rare non-recurrent events. The asymptotic efficiency of the effect estimator is studied in some special cases. A computational scheme based on a pseudo-likelihood is proposed to make the computations feasible in complex models. The methods are evaluated by simulation. A validation study and an application are described.

Key words: censored data, counterfactual, endogeneity, estimating equation, Horvitz-Thompson estimator, pseudo-likelihood, self-controlled case series

Address for correspondence: Paddy Farrington, Department of Mathematics and Statistics,

The Open University, Walton Hall, Milton Keynes MK7 6AA, UK. Email: c.p.farrington@open.ac.uk.

1 INTRODUCTION

The self-controlled case series method, or case series method for short, was developed to investigate the association between time-varying exposures and outcome events using data on cases, that is, individuals who have experienced the event. Its advantages are that only cases need be sampled, and that it is self-matched, so that time-invariant multiplicative confounders are necessarily adjusted. The method was originally described in Farrington (1995). For a review of modelling and applications, see Whitaker et al. (2006).

The model is derived by conditioning on the number of events and the exposure history experienced by each individual over a pre-determined observation period. The main limiting assumption is that both the exposure distribution and the observation period must be independent of event times. These requirements inhibit use of the case series method when occurrence of an event alters in some way the subsequent exposure process, or the observations made of that process. This occurs for exposures whose distribution depends on the event history. It also occurs for terminal events, by virtue of the fact that follow-up, and hence the exposure history, is curtailed by the event. Similarly, the case series method cannot be used if observation of the exposure process is censored or otherwise disrupted by occurrence of an event. In some circumstances, violation of the assumptions does not result in severe bias, as illustrated for example by the application to myocardial infarction discussed in Farrington & Whitaker (2006). Nevertheless, it is desirable to have a method applicable whenever the exposure process, or the exposure observation process, are affected by occurrence of an event, and whose validity does not depend on robustness to failure of assumptions.

In this paper we derive a case series method for binary exposures which can be used in such circumstances, provided the post-exposure risk period is short. In Section 2 we briefly review the standard case series method, describe some of the situations in which the assumptions it requires might fail, and outline our proposed approach. In Section 3 we derive a set of unbiased estimating

equations, applicable in such situations. These are based on counterfactuals in which post-event exposures do not take place. The asymptotic efficiency of the method is discussed in Section 4. In Section 5 we present a pseudo-likelihood formulation that leads to a straightforward method for estimating the parameters and calculating bootstrap confidence intervals. The performance of the methods is studied by simulation. Section 6 contains three examples, including a validation study and an application. We end with a brief discussion of some further issues in Section 7.

2 THE CASE SERIES METHOD

2.1 The case series likelihood

We begin by introducing the case series method and relevant notation for use in the paper. We suppose that an individual i is observed over a pre-determined observation period $(a_i, b_i]$, usually defined in terms of age, during which this individual may experience point exposures, at ages $c_{i1}, ..., c_{iD}$ say. We assume that $c_{i1} < c_{i2} < ... < c_{iD}$ and that following the d^{th} exposure the incidence of the event of interest is multiplied by a factor e^{β_d} over the period $(c_{id}, \min\{c_{id} + \tau, c_{id+1}\}]$, which we call a risk interval. From now on, to simplify the notation without sacrificing essential generality, we shall assume that there are no overlaps and that $a_i \leq c_{id} + \tau \leq c_{id+1} \leq b_i$. The intervals during which the individual does not experience an exposure-related risk are called control intervals. The sequence of exposures for individual i thus determines $J = 2D + 1$ contiguous non-overlapping control and risk intervals $(a_i, c_{i1}]$, $(c_{i1}, c_{i1} + \tau]$, $(c_{i1} + \tau, c_{i2}]$, ...,$(c_{iD} + \tau, b_i]$ indexed by $j = 1, ..., J$ and of length e_{ij}. If in fact there are overlaps, the intervals are truncated as required, giving precedence to most recent exposures; some intervals may thus be empty.

Suppose furthermore that age is stratified in $K + 1$ age groups indexed by $k = 0, 1, ..., K$, leading to a further subdivision of each of the J control or risk intervals into $K+1$ sub-intervals.

Let E_{ijk} denote the subset of the observation period lying within the k^{th} age group and j^{th} risk or control interval for individual i, of length e_{ijk}. Thus $e_{ij} = e_{ij\cdot} = \sum_{k=0}^{K} e_{ijk}$. Typically, some E_{ijk} will have $e_{ijk} = 0$. A possible configuration for a single individual i is shown in Figure 1.

[Figure 1 about here.]

In the case series method, the exposure is assumed to be an external time-varying covariate (Kalbfleisch & Prentice 2002), and, conditionally on the ages at exposure, events arise in a non-homogeneous Poisson process with piecewise constant rate

$$\lambda_{ijk} = \exp(\varphi_i + \alpha_k + \beta_{d(j)}),$$

where φ_i is an individual effect, α_k an age effect and $\beta_{d(j)}$ an exposure effect ($\beta_0 = 0$), with

$$d(j) = \begin{cases} d & \text{if } j = 2d,\ d = 1, ..., D, \\ 0 & \text{otherwise.} \end{cases}$$

Let n_{ijk} denote the number of events arising in E_{ijk}. The case series likelihood is obtained by conditioning on both the exposure history $\{c_{i1}, ..., c_{iD}\}$ and on the total number of events observed during the observation period, $n_{i\cdot\cdot} = \sum_{j,k} n_{ijk}$. The likelihood contribution of individual i is

$$L_i = \prod_{j,k} \left\{ \frac{e_{ijk} \exp(\alpha_k + \beta_{d(j)})}{\sum_{r,s} e_{irs} \exp(\alpha_s + \beta_{d(r)})} \right\}^{n_{ijk}}. \tag{1}$$

The overall likelihood is product multinomial. Note that the individual effects φ_i factor out. It follows that the method is self-matched and controls implicitly for all fixed multiplicative confounders.

The above derivation of the method applies to recurrent events. However, the method also applies for rare unique events, which we shall exclusively be concerned with in the present paper. In this case the case series likelihood (1) is valid in the limit $\varphi_i \to -\infty$ (see Farrington & Whitaker (2006) for details).

2.2 How key assumptions might fail

A key assumption of the case series model is that the exposure is an external time-varying covariate: equivalently, the occurrence of an event does not alter an individual's subsequent exposure history. It is also assumed that occurrence of an event does not alter the subsequent observation period. Furthermore, complete information on exposure status throughout the observation period of each individual is required. We briefly review situations where these assumptions and requirements might fail.

Censored or partially observed post-event exposures. The exposure process and observation period are unaffected by the occurrence of an event, but the observation of the exposure process is affected by such an occurrence. This arises typically when the exposure data collection occurs at time of event, so that post-event exposures are undocumented: in effect, the event censors subsequent exposures (for an example see Section 6). Alternatively, the post-event exposure process might only be partially observed. In either situation, case series likelihood (1) is valid but its denominators cannot be evaluated owing to missing exposure data.

Curtailed or perturbed post-event exposures. The individual remains under observation after the event, but the exposure process is stopped or perturbed by the event. This might occur in pharmacoepidemiology, for example, when the event of interest is a contra-indication to the drug of interest (as is the case with rotavirus vaccination and intussusception, or myocardial infarction and some anti-smoking therapies), in which case the indication for the drug changes after an event has occurred. This violates the assumption that the exposure is an external variable, and hence that its distribution is unaffected by the event history. In this scenario, the case series likelihood (1) is no longer valid.

Event-dependent observation periods. The end of the observation period is a random variable which is not independent of the event process. The most important case relates to death, either because the event of interest is death, or because it increases the mortality rate (as is the case

with myocardial infarctions, discussed in Farrington & Whitaker (2006)). In this case, the case series likelihood (1) is no longer valid either, because the observation period $(a_i, b_i]$ depends on the event time.

Typically, as well as modelling exposures, it is also necessary to take into account age effects. This is particularly important in studies in children and the elderly, for whom it can seldom be assumed that event and exposure rates are stationary.

2.3 A way forward

To make progress, it is useful to think in terms of counterfactual outcomes, in which the event did not occur and the exposure process (and observation of it) unfolded unperturbed (Rubin 1976). The case series likelihood may validly be derived using the counterfactual (i.e. event-free) end b_i of the observation period, and the counterfactual exposure process up to b_i. As it happens, even when the observation periods are event-dependent, the end of the observation period b_i that would have applied had the event not taken place is often known. This is typically the case when observation periods are determined by calendar time and age boundaries and case ascertainment can reasonably be regarded as complete. An example is provided in Section 6. Hence, in all that follows we shall assume that b_i, whether factual or counterfactual, is known.

Thus all three scenarios described above share the following essential characteristic: the observed post-event exposure history is not that which would have been observed had an event not occurred. The post-event exposure history that would have been observed, had the event not occurred, is not known, because the event has interfered with the subsequent exposure process or our observation of it. (As noted above, such interference ranges from censoring to termination of follow-up.) For simplicity, in order to deal with all the possibilities together, we shall henceforth refer to such events as *interferent* events.

One approach for dealing with interferent events might be to model the event-free exposure process in some way, and either impute post-event exposures or integrate them out of the like-

lihood. In the myocardial example referred to previously, we imputed counterfactual exposures and showed that not knowing them had little effect on the results (Farrington & Whitaker 2006). In most cases, however, there is no empirical basis upon which to build a reliable model for the exposures.

We therefore propose a different approach, which requires no assumptions about the event-free exposure process. This is achieved by analysing the data for each exposure as if there could be no subsequent exposures: thus, we impose our own counterfactual. Where such exposures do in fact occur, we apply Horvitz-Thompson-like estimators (Horvitz & Thompson 1952, Levy 1998) to adjust the event counts to what they would have been under our counterfactual. This strategy enables us to derive a set of unbiased estimating equations. Related approaches using reweighted estimating equations have been used in longitudinal data analysis (Robins, Rotnitzky & Zhao 1994, Robins, Hernan & Brumback 2000, Bryan, Yu & Van der Laan 2004, Davidian, Tsiatis & Leon 2005). Note, however, that our approach retains the two essential features of the standard case series method: it requires only cases, and all time-invariant multiplicative confounders − whether measured or not − are controlled.

Throughout, we make the assumption that the exposure is binary i.e. present or absent, that the post-exposure risk period is short and that the event of interest is an uncommon, non-recurrent event. The short risk period assumption implies that the risk returns to an age-related baseline level at the end of each risk period: this assumption is essential.

3 AN ESTIMATING EQUATION APPROACH

The estimating equations we propose for arbitrary numbers of exposures and age groups are rather obscure. We therefore lead up to them by describing a few special cases which will help to motivate the general method. This sequence of special cases is intended to reveal the key recursive principle that lies at the heart of the method: we start with the last observed pre-

event exposure, and work back through the exposures, deriving a new estimating equation at each stage. The process gets started by virtue of the fact that a valid case series analysis is possible for the final observed pre-event exposure. We can then work back step by step through earlier exposures, for which the case series score equations cannot be evaluated, and adjust them using estimates derived previously, in such a way as to produce estimating equations that can be evaluated.

3.1 Unique exposures

We first consider the simplest situation in which there is at most a single exposure, as is the case with single-dose vaccination schedules. Then it is possible to undertake a case series analysis for an interferent event, by re-defining the observation period as starting at the age of exposure. This is analogous to one of the approaches to deal with event-dependent exposures in section 6.4 of Whitaker et al. (2006). This opens a chink in the armour through which we shall attack more complicated settings in the next subsections. To keep matters simple we will assume here that there is no age effect.

Let T_i denote the age at event for individual i. Thus, since it is an interferent event, we only observe exposures in $(a_i, T_i]$. We assume that the observation period $(a_i, b_i]$ that would have applied had no event occurred is known. If an exposure did occur prior to the event at T_i, the timing of the post-exposure risk period $(c_i, c_i + \tau]$ is known. If no exposure took place before T_i, the subsequent exposure history is censored, but we shall assume that some unobserved exposure at age $c_i > T_i$ would have occurred, had an event not preceded it. Values $c_i > b_i$ represent exposures that occurred after the end of the observation period, and $c_i = \infty$ can if required stand for 'never exposed'.

Let n_{i1} denote the number of events arising in the period $(a_i, c_i]$ of length e_{i1}, n_{i2} the number of events in the risk period $(c_i, c_i + \tau]$ of length e_{i2}, and n_{i3} the number of events in $(c_i + \tau_i, b_i]$, of length e_{i3}; $n_{i1} + n_{i2} + n_{i3} = 1$ since the event is non-recurrent and only cases are considered.

A possible configuration is shown in Figure 2.

[Figure 2 about here.]

If the event were not of the interferent type, then c_i would always be observed and the case series log likelihood contribution from individual i, obtained by conditioning on the occurrence of one event in $(a_i, b_i]$, would be

$$l_i^1(\beta_1) = n_{i2}\beta_1 - n_{i\cdot}\log\left(e_{i1} + e^{\beta_1}e_{i2} + e_{i3}\right)$$

where $n_{i\cdot} = n_{i1} + n_{i2} + n_{i3} = 1$ by assumption. However, for interferent events, c_i and thus the interval lengths e_{ij} are not observed when the event occurs prior to any exposure $c_i > T_i$ and the likelihood for these individuals cannot be evaluated. One way round this is to redefine the observation period to be $(c_i, b_i]$ and only use cases whose events occur in this interval. This works because we use only cases for whom the age at exposure is known. Thus the case series log likelihood contribution of individual i, obtained by conditioning on the number of events in $(c_i, b_i]$, becomes

$$l_i^2(\beta_1) = \begin{cases} n_{i2}\beta_1 - (n_{i2} + n_{i3})\log\left(e^{\beta_1}e_{i2} + e_{i3}\right) & \text{if } n_{i1} = 0, \\ 0 & \text{if } n_{i1} = 1 \end{cases}$$

and the elementary score function for β_1 is

$$U_i^1(\beta_1) = n_{i2} - (n_{i2} + n_{i3})\frac{e^{\beta_1}e_{i2}}{e^{\beta_1}e_{i2} + e_{i3}}.$$

This score function can be evaluated for all individuals. Thus β_1 can be estimated, but at the cost of ignoring events (and control periods) which occur prior to exposure. This is inevitable unless it is possible to infer or impute the counterfactual exposure age c_i is in cases for whom the event precedes exposure. Our aim is to develop an analysis method which avoids such imputation.

3.2 Two exposures

Consider now a more general setting in which the exposure is not unique. For simplicity, suppose an individual i can be exposed up to two times at ages c_{i1} and c_{i2}, with $c_{i1} \leq c_{i2}$, the inequality always being strict unless $c_{i1} = c_{i2} = \infty$. There are now up to five periods of lengths $e_{ij} \geq 0$: a control period of length e_{i1} prior to c_{i1}, the first risk period of length e_{i2}, a further control period of length e_{i3} prior to c_{i2}, the second risk period of length e_{i4}, and a final control period of length e_{i5}. The event count for individual i in period j is n_{ij}. A possible configuration is shown in Figure 3.

[Figure 3 about here.]

Let β_1 and β_2 denote the log relative incidences associated with each risk period. Inference about β_2 can be made using the method described in the previous subsection, by using the case series likelihood with observation period $(c_{i2}, b_i]$ and restricting the analysis to only those cases who experienced the second exposure. This yields the elementary score function

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

Suppose that we now try to apply the same method for inference about β_1, using only cases with events arising in $(c_{i1}, b_i]$. Unfortunately this will not work, since the age at second exposure is unavailable for cases whose interferent event occurs after experiencing just one exposure. Instead, we proceed as if no individual experiences a second exposure, and let n_{i4}^* denote the number of events that would have arisen in the new control period that now replaces the second (possibly unobserved) risk period $(c_{i2}, c_{i2} + \tau]$. If n_{i4}^* were observed, the method of the previous subsection 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, if the event occurs after the second exposure, we only observe n_{i4}, the number of events arising in the risk period following the second exposure; n_{i4}^*, the number of events that would have arisen had this been a control period, 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 $n_{i4}e^{-\beta_2}$. (We call it Horvitz-Thompson-*like* because the adjustment factor is a relative rate, not a probability.) Thus we obtain the elementary estimating function

$$U_i^2(\beta_1, \beta_2) = n_{i2} - (n_{i2} + n_{i3} + \frac{n_{i4}}{e^{\beta_2}} + n_{i5})\frac{e^{\beta_1}e_{i2}}{e^{\beta_1}e_{i2} + e_{i3} + e_{i4} + e_{i5}}.$$

This estimating function is unbiased, conditionally on $n_{i2+} = n_{i2} + n_{i3} + n_{i4} + n_{i5}$ and on c_{i1} and c_{i2}, the latter possibly being unavailable in the observed realization. Unbiasedness follows because the event is non-recurrent, so $n_{i2+} = 0$ or 1, and

$$\begin{aligned} E(n_{i2}|n_{i2+}, c_{i1}, c_{i2}) &= \frac{n_{i2+}e^{\beta_1}e_{i2}}{e^{\beta_1}e_{i2} + e_{i3} + e^{\beta_2}e_{i4} + e_{i5}}, \\ E(n_{i2} + n_{i3} + \frac{n_{i4}}{e^{\beta_2}} + n_{i5}|n_{i2+}, c_{i1}, c_{i2}) &= \frac{n_{i2+}(e^{\beta_1}e_{i2} + e_{i3} + e_{i4} + e_{i5})}{e^{\beta_1}e_{i2} + e_{i3} + e^{\beta_2}e_{i4} + e_{i5}}. \end{aligned}$$

The key point is that the estimating function U_i^2 can always be evaluated, even if c_{i2} is unavailable, since in this case $n_{i4} = n_{i5} = 0$ and $e_{i3} + e_{i4} + e_{i5} = b_i - (c_{i1} + \tau)$ is known, even though e_{i3}, e_{i4} and e_{i5} are not.

The system $\Sigma_i U_i^1, \Sigma_i U_i^2$ thus provides a pair of unbiased (conditionally on each case experiencing a single event) estimating equations which may be used to estimate β_1 and β_2, using cases with events occurring after the first exposure. Once again events prior to the first exposure are not used.

So far we have assumed that β_1 and β_2 are distinct. In many circumstances it will make sense to assume that $\beta_1 = \beta_2 = \beta$. The simplest way of combining the estimating functions is to take their sum, $\Sigma_i(U_i^1 + U_i^2)$. This choice is convenient for computational reasons that will become apparent in Section 5.

3.3 Two exposures and two age groups

We now expand the method described in the previous subsection to allow for two age groups. Thus there is one additional parameter α, the log relative incidence associated with age group 1 relative to age group 0. We shall take age group 0 to be the earlier one. The notation of the previous subsection is expanded to include an extra subscript k denoting the age group. Thus, for example, e_{ijk} is the time spent by individual i in period j: $j = 1$, control period before c_{i1}, $j = 2$ risk period after c_{i1}, $j = 3$ control period between c_{i1} and c_{i2}, etc... and in age group k. The unbiased estimating function U_i^1 becomes

$$U_i^1(\beta_2, \alpha) = n_{i4\cdot} - (n_{i4\cdot} + n_{i5\cdot}) \frac{e^{\beta_2}(e_{i40} + e^{\alpha} e_{i41})}{e^{\beta_2}(e_{i40} + e^{\alpha} e_{i41}) + (e_{i50} + e^{\alpha} e_{i51})}.$$

Similarly, the unbiased estimating function U_i^2 becomes

$$\begin{aligned} U_i^2(\beta_1, \beta_2, \alpha) &= n_{i2\cdot} - (n_{i2\cdot} + n_{i3\cdot} + \frac{n_{i4\cdot}}{e^{\beta_2}} + n_{i5\cdot}) \\ &\quad \times \frac{e^{\beta_1}(e_{i20} + e^{\alpha} e_{i21})}{e^{\beta_1}(e_{i20} + e^{\alpha} e_{i21}) + (e_{i30} + e^{\alpha} e_{i31}) + (e_{i40} + e^{\alpha} e_{i41}) + (e_{i50} + e^{\alpha} e_{i51})}. \end{aligned}$$

In order to obtain a third unbiased estimating function, we apply a similar argument to that used in the previous subsection to obtain the unbiased estimating function U_i^2, this time using all events including those occurring prior to the first exposure. To do this, we assume that no individuals can be exposed, and apply the case series method with n_{i2k}^* and n_{i4k}^* to denote the numbers of events that would occur in the newly defined control periods E_{i2k} and E_{i4k}. We obtain the corresponding elementary score function for α and replace n_{i2k}^* with $n_{i2k} e^{-\beta_1}$ and n_{i4k}^* with $n_{i4k} e^{-\beta_2}$. This yields the following elementary estimating function

$$\begin{aligned} U_i^{30}(\beta_1, \beta_2, \alpha) &= (n_{i11} + \frac{n_{i21}}{e^{\beta_1}} + n_{i31} + \frac{n_{i41}}{e^{\beta_2}} + n_{i51}) \\ &\quad -(n_{i1\cdot} + \frac{n_{i2\cdot}}{e^{\beta_1}} + n_{i3\cdot} + \frac{n_{i4\cdot}}{e^{\beta_2}} + n_{i5\cdot}) \frac{e^{\alpha} e_{i\cdot 1}}{e_{i\cdot 0} + e^{\alpha} e_{i\cdot 1}}. \end{aligned}$$

Note that this estimating function now uses the full data set. More generally, whenever age parameters are included, all the data can be used. The reason is that we always know the age the person would have been had the event not occurred.

In what follows we shall in fact use a third estimating function other than U_i^{30}. The case series likelihood restricted to events after the second exposure yields the following elementary score function for α:

$$U_i^{32}(\beta_2, \alpha) = (n_{i41} + n_{i51}) - (n_{i4\cdot} + n_{i5\cdot}) \frac{e^{\alpha}(e^{\beta_2} e_{i41} + e_{i51})}{(e^{\beta_2} e_{i40} + e_{i50}) + e^{\alpha}(e^{\beta_2} e_{i41} + e_{i51})}.$$

Similarly, the case series likelihood restricted to events after the first exposure, assuming no further exposures take place, yields the following unbiased elementary estimating function for α:

$$\begin{aligned} U_i^{31}(\beta_1, \beta_2, \alpha) &= (n_{i21} + n_{i31} + \frac{n_{i41}}{e^{\beta_2}} + n_{i51}) - (n_{i2\cdot} + n_{i3\cdot} + \frac{n_{i4\cdot}}{e^{\beta_2}} + n_{i5\cdot}) \\ &\quad \times \frac{e^{\alpha}(e^{\beta_1} e_{i21} + e_{i31} + e_{i41} + e_{i51})}{(e^{\beta_1} e_{i20} + e_{i30} + e_{i40} + e_{i50}) + e^{\alpha}(e^{\beta_1} e_{i21} + e_{i31} + e_{i41} + e_{i51})}. \end{aligned}$$

We shall use as third elementary estimating function the sum of all three elementary estimating functions for α, namely

$$U_i^3(\beta_1, \beta_2, \alpha) = U_i^{30}(\beta_1, \beta_2, \alpha) + U_i^{31}(\beta_1, \beta_2, \alpha) + U_i^{32}(\beta_2, \alpha).$$

Once again, this choice turns out to be convenient for computational reasons to be described in Section 5.

3.4 Any number of possible exposures

In the preceding subsections we have assumed that the total possible number of exposures is fixed and known. In fact this assumption is not necessary. We show that the method works if we just assume that there can be no more exposures than the maximum observed for any individual. For simplicity we assume there is no age effect and that a maximum of one exposure is observed.

Consider an individual i with an event after the first exposure. For all we know, this individual might have experienced further exposures at ages $c_{i2}, c_{i3}, ...$ had the event not interfered

with the subsequent exposure history. The elementary unbiased estimating function for β_1 is then

$$U_i^1(\beta_1) = n_{i2} - (n_{i2} + n_{i3} + \frac{n_{i4}}{e^{\beta_2}} + n_{i5} + \frac{n_{i6}}{e^{\beta_3}} + n_{i7} + ...)\frac{e^{\beta_1} e_{i2}}{e^{\beta_1} e_{i2} + e_{i3} + e_{i4} + ...}.$$

This can always be evaluated because $e_{i3} + e_{i4} + ... = b_i - (c_{i1} + \tau)$ is known, and since exactly one exposure is observed we know that $n_{i4} = n_{i5} = n_{i6} = ... = 0$ and hence

$$n_{i2} + n_{i3} + \frac{n_{i4}}{e^{\beta_2}} + n_{i5} + \frac{n_{i6}}{e^{\beta_3}} + n_{i7} + ... = n_i. \quad .$$

More generally, if a maximum D exposures are observed for each case, then only $\beta_1, ..., \beta_D$ are estimable, and any exposure that might have occurred subsequently can be ignored.

3.5 The general case

The special cases described in the previous subsections help to motivate the estimating equations for the general case, in which there are up to D distinct observed exposures at ages $c_{i1}, ..., c_{iD}$. We assume that the c_{id} are properly ordered with age for increasing d unless $c_{id} = \infty$ for some d. There are thus $J = 2D + 1$ contiguous exposure (D) or control ($D + 1$) periods indexed by $j = 1, ..., J$, of length e_{ij}, some of which may be empty so that $e_{ij} = 0$. The risk periods correspond to even values of j, the control periods to odd values of j. We suppose also that there are $K + 1$ age groups indexed by $k = 0, 1, ..., K$. Let e_{ijk} denote the length of that part of the j^{th} exposure or control period that lies within age group k, so that $e_{ij} = \sum_{k=0}^{K} e_{ijk}$.

To the D exposure periods correspond D log relative incidence parameters $\beta_1, ..., \beta_D$ (relative to the control periods, for which $\beta_0 = 0$). To the $K + 1$ age groups correspond K parameters $\alpha_1, ..., \alpha_K$. These represent log relative incidences, relative to the 0-indexed age group, so $\alpha_0 = 0$.

Now define, for $d = 0, 1, ..., D$ (note the inclusion of $d = 0$ here),

$$w_{ijk}^{(d)} = \begin{cases} 0 & \text{if } j < 2d, \\ 1 & \text{if } \ j = 2d \ \text{ or } \ j = 2d' + 1, d' \geq d, \\ \exp(-\beta_{d'}) & \text{if } \ j = 2d', d' > d, \end{cases}$$

and

$$d(j) = \begin{cases} d & \text{if } j = 2d,\ d = 0, 1, ..., D, \\ 0 & \text{otherwise.} \end{cases}$$

The elementary estimating function for β_d, $d = 1, ..., D$, is:

$$U_{id}: \qquad \sum_{k=0}^{K} n_{i(2d)k} - \left(\sum_{k=0}^{K} \sum_{j=1}^{J} w_{ijk}^{(d)} n_{ijk} \right) \frac{\sum_{k=0}^{K} w_{i(2d)k}^{(d)} e^{\beta_d + \alpha_k} e_{i(2d)k}}{\sum_{k=0}^{K} \sum_{j=1}^{J} w_{ijk}^{(d)} e^{\beta_{d(j)} + \alpha_k} e_{ijk}}.$$

where the subscript $i(2d)j$ represents ijk with $j = 2d$. The elementary estimating function for α_k, $k = 1, ..., K$, is

$$U_{i(D+k)} = \sum_{d=0}^{D} U_i^{(D+k)d}$$

where the subscript $i(D+k)$ and the superscript $(D+k)d$ represent ir and rd respectively, with $r = D + k$, and

$$U_i^{(D+k)d}: \qquad \sum_{j=1}^{J} w_{ijk}^{(d)} n_{ijk} - \left(\sum_{k=0}^{K} \sum_{j=1}^{J} w_{ijk}^{(d)} n_{ijk} \right) \frac{\sum_{j=1}^{J} w_{ijk}^{(d)} e^{\beta_{d(j)} + \alpha_k} e_{ijk}}{\sum_{k=0}^{K} \sum_{j=1}^{J} w_{ijk}^{(d)} e^{\beta_{d(j)} + \alpha_k} e_{ijk}}.$$

In the special cases considered above, these estimating functions reduce to those derived earlier. If the parameters β_d and $\beta_{d'}$ are constrained to be equal, then the two elementary estimating functions U_{id} and $U_{id'}$ are replaced by their sum.

3.6 Sandwich variance estimates

Let θ denote the parameter vector $(\beta_1, ..., \beta_D, \alpha_1, ..., \alpha_K)$ and $U_{i1}, ..., U_{i(D+K)}$ the set of elementary estimating functions. Denote V the observed covariance matrix and D the Jacobian of the $\Sigma_i U_{ij}$, with (r, s) elements

$$\begin{aligned} V_{rs} &= \sum_{i=1}^{n} U_{ir} U_{is} \\ D_{rs} &= \sum_{i=1}^{n} \frac{\partial U_{ir}}{\partial \theta_s}. \end{aligned}$$

Then the asymptotic sandwich estimator of $\text{cov}(\widehat{\theta})$ is $D^{-1} V D^{-1T}$. This can in turn be used to obtain Wald confidence intervals for θ.

4 RELATIVE EFFICIENCY

In this section, we explore the asymptotic efficiency of $\widehat{\beta}$ (we shall always assume a common β_d and hence suppress the subscripts) estimated using the case series model with the complete exposure data on $(a_i, b_i]$, relative to the efficiency of $\widehat{\beta}$ estimated without post-event exposures, as described in the previous section. Our purpose in doing this is to quantify the loss in efficiency resulting from incomplete observation of the underlying, event-free exposure history, and hence help guide the choice of observation periods in practical applications.

For simplicity we consider two special cases, in which we assume that there are no age effects, and that all individuals share the same partition of their observation period. Thus, for all cases $i = 1, ..., n$, $e_{ij} = e_j$. Without loss of generality, we shall assume that the total observation time $\sum_{j=1}^{J} e_j = 1$. Thus, for example, e_1 is the proportion of total observation time prior to the first exposure. The asymptotic relative efficiencies are derived in Section 1 of the supplementary material (http://www.biostatistics.oxfordjournals.org).

4.1 Unique exposure

Suppose first that every individual has a unique exposure, as in Subsection 3.1. The asymptotic relative efficiency of the case series method is

$$ARE_1 = \frac{\left(e_1 + e^{\beta} e_2 + e_3\right) e_3}{\left(e^{\beta} e_2 + e_3\right)\left(e_1 + e_3\right)}.$$

As the risk period e_2 tends to zero, then ARE_1 tends to 1. Note also that ARE_1 tends to zero as e_3 tends to zero (indeed $\text{var}(\widehat{\beta})$ under the proposed model increases without bound). Thus, the method we propose only works when the risk period is finite, since otherwise $e_3 = 0$.

Figure 4 shows the relative asymptotic efficiency when the post exposure risk period as a proportion of the overall risk period is $e_2 = 0.3$. The relative asymptotic efficiency declines as e_1, the proportion of the observation period preceding the exposure, and the relative incidence $RI = e^{\beta}$ increase.

[Figure 4 about here.]

4.2 Two exposures

Now suppose there are two exposures, as in Subsection 3.2. The asymptotic efficiency of the proposed method relative to the case series method for complete data is

$$\begin{aligned} ARE_2 &= \frac{e_1 + e_2 e^{\beta} + e_3 + e_4 e^{\beta} + e_5}{(e_2 + e_4)(e_1 + e_3 + e_5)} \\ &\quad \times \frac{\{Ae_2(e_3 + e_5) + Be_4e_5\}^2}{Ae_2(e_3 + e_5) + Be_4e_5 + Ae_2e_4\{Ae_2(1 - e^{\beta}) + 2B(e_4 + e_5)e^{\beta} - 1\}} \end{aligned}$$

where $A = (e^{\beta}e_2 + e_3 + e_4 + e_5)^{-1}$ and $B = (e^{\beta}e_4 + e_5)^{-1}$.

As the risk periods e_2 and e_4 tend to zero, then ARE_2 tends to 1. As $e_3 + e_5$ tends to zero, ARE_2 tends to zero ($\text{var}(\widehat{\beta})$ under the proposed model increases without bound).

Figure 5 shows the asymptotic relative efficiency for $e_2 = e_4 = 0.1$. The relative efficiency decreases as the proportion of observation preceding the first exposure e_1 and the relative incidence RI increase. It declines also as the proportion e_3 of the observation period between exposures increases; left panel: $e_3 = 0.05$ and right panel: $e_3 = 0.25$.

[Figure 5 about here.]

More general relative efficiency calculations are possible but unenlightening, and are not presented here. The key message from these investigations is that the relative efficiency is high for short risk periods and when the pre-exposure period and inter-exposure periods are small as a proportion of the overall observation time. Relative efficiency is low when there is little unexposed post-exposure time. Thus, in designing case series studies of interferent events, it is important where possible to select the observation period so as to minimise pre-exposure times, and maximise unexposed post-exposure times. As expected, the proposed method cannot be used for indefinite risk periods.

5 A PSEUDO-LIKELIHOOD METHOD

In practice, there are typically many more than two age groups, and often more than two exposures. In such more complicated settings, writing down and solving the estimating equations, and deriving the sandwich variance estimator, becomes extremely cumbersome, and is impractical. In this section we present an alternative approach to deriving the estimating equations, which is convenient for computation. The trick is to view the estimating equations derived above as pseudo-score equations resulting from a particular pseudo-likelihood (Kalbfleisch 1998). As in Section 3 we develop the argument first in a special case, before moving on to the general case.

5.1 Two exposures and two age groups

For a count n and weight w with $0 \leq w \leq 1$, let the expression $nw \sim P(\mu)$ denote a likelihood contribution proportional to $e^{-\mu}\mu^{nw}$, when $w \neq 0$, and equal to 1 when $w = 0$. We shall refer to this as a pseudo-Poisson likelihood; similar pseudo-likelihoods appear in the literature on spatial point patterns (Baddeley & Turner 2000). Recall from subsection 3.3 that the elementary estimating functions U_i^1 and U_i^{32} were score contributions obtained from the case series likelihood restricted to events after the second exposure. These may equivalently be obtained as score contributions from the pseudo-Poisson model

$$\begin{aligned} n_{ijk}w_{ijk}^{(2)} &\sim P(\lambda_{ijk}e_{ijk}), \quad j = 4, 5; k = 0, 1 \\ \log(\lambda_{ijk}) &= \varphi_i^{(2)} + \beta_2 I(j = 4) + \alpha I(k = 1) \end{aligned}$$

where $I(.)$ is the indicator function and $w_{ijk}^{(2)}$ is the weight defined in Subsection 3.5, which here is 1 for $j \geq 4$ and 0 otherwise.

Similarly, U_i^2 and U_i^{31} were elementary estimating functions obtained from the case series likelihood restricted to events after the first exposure, assuming that there are no subsequent exposures, and replacing n_{i4k} with the unobserved n_{i4k}^* which was estimated by $n_{i4k}e^{-\beta_2}$. These

estimating functions may equivalently be derived as score equations from the pseudo-Poisson model

$$\begin{aligned} n_{ijk} w_{ijk}^{(1)} &\sim P(\lambda_{ijk} e_{ijk}), \quad j = 2,3,4,5; k = 0,1 \\ \log(\lambda_{ijk}) &= \varphi_i^{(1)} + \beta_1 I(j=2) + \alpha I(k=1). \end{aligned}$$

Here, the weights $w_{ijk}^{(1)}$ are $e^{-\beta_2}$ for counts of events in the risk period of the second exposure, 0 for $j = 1$, and 1 elsewhere.

Finally, U_i^{30} was an elementary estimating function obtained from the case series likelihood assuming that there were no exposures at all, replacing n_{i2k} with n_{i2k}^*, estimated by $n_{i2k}e^{-\beta_1}$ and n_{i4k} with the n_{i4k}^*, estimated by $n_{i4k}e^{-\beta_2}$. This estimating function may be derived as a score contribution for α from the pseudo-Poisson model:

$$\begin{aligned} n_{ijk} w_{ijk}^{(0)} &\sim P(\lambda_{ijk} e_{ijk}), \quad j = 1,2,3,4,5; k = 0,1 \\ \log(\lambda_{ijk}) &= \varphi_i^{(0)} + \alpha I(k=1). \end{aligned}$$

In this case the weights $w_{ijk}^{(0)}$ are $e^{-\beta_1}$ for counts of events in the first risk period, $e^{-\beta_2}$ for counts of events in the second risk period, and 1 elsewhere.

Now stack the three sets of data for individual i and the corresponding models, duplicating the counts where required. Thus, the counts n_{i1k} will occur once, the counts n_{i2k} and n_{i3k} will occur twice, and the counts n_{i4k} and n_{i5k} will occur 3 times in the stacked data. The pseudo-Poisson likelihood for the stacked data constitutes a pseudo-likelihood for individual i, the pseudo-score contributions of which are exactly the elementary estimating functions U_i^1, U_i^2 and U_i^3.

5.2 The general case

With D risk periods and $K+1$ age groups, up to D exposure parameters β_d and K age parameters α_k, the method requires $D+1$ stacked data sets, labelled 0 to D. Stack 0 contains all the data,

to which the following model is to be fitted:

$$\begin{aligned} n_{ijk}w_{ijk}^{(0)} &\sim P(\lambda_{ijk}e_{ijk}), \quad j = 1, ..., 2D+1; k = 0, ..., K \\ \log(\lambda_{ijk}) &= \varphi_i^{(0)} + \alpha_1 I(k=1) + ... + \alpha_K I(k=K). \end{aligned}$$

Stack $d, d = 1, ..., D$, contains the data for periods $2d, 2d+1, ..., 2D+1$, to which the following model is to be fitted:

$$\begin{aligned} n_{ijk}w_{ijk}^{(d)} &\sim P(\lambda_{ijk}e_{ijk}), \quad j = 2d, ..., 2D+1; k = 0, ..., K \\ \log(\lambda_{ijk}) &= \varphi_i^{(d)} + \beta_d I(j=2d) + \alpha_1 I(k=1) + ... + \alpha_K I(k=K). \end{aligned}$$

These models are fitted together to the stacked data, or rather, pseudo-data $n_{ijk}w_{ijk}^{(d)}$, as a whole. Thus, the parameters α_k are estimated from all levels of the stack. The pseudo-Poisson likelihood for the stacked data yields pseudo-score equations which reproduce exactly the estimating equations based on the elementary terms U_{id} and $U_{i(D+k)}$ described in Subsection 3.5.

5.3 A fitting algorithm

Estimates are obtained by an iterative procedure. Choose initial values of the β_d, for example 0, and calculate the weights $w_{ijk}^{(d)}$. Then obtain estimates of the parameters β_d and α_k by maximising the pseudo-likelihood with these weights. Update the weights using the new values of the β_d, and iterate until convergence.

The procedure is closely related to an E-M algorithm, in which at each iteration the missing data n_{ijk}^* are replaced by their Horvitz-Thompson-like expected values: the E-step, and the resulting pseudo-likelihood is then maximised: the M-step.

5.4 Bootstrap estimates

The pseudo-likelihood method described in the previous subsection 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.

The simplest method is non-parametric bootstrapping, in which the stacked data for individuals $i = 1, ..., n$ are resampled with replacement. More precisely, what is resampled is not the counts themselves but blocks of counts corresponding to individuals. Bootstrap estimates may then be obtained in the usual way (Davison & Hinkley 1997).

5.5 Simulations

The performance of the method, and different approaches to obtaining interval estimates, were studied by simulation. The simulations and detailed results are reported in Section 2 of the supplementary material (http://www.biostatistics.oxfordjournals.org). The medians of the estimated log relative incidences associated with exposure and age were close to their true values, improving in accuracy and precision as the sample size increased, as expected. Coverage probabilities of the the 95% confidence intervals were also close to 0.95. As expected, the pseudo-likelihood method of Section 5 generated the same estimates as the estimating equations. The overall conclusion from these simulations is that the model performs well.

6 EXAMPLES

We present three examples. The first relates to sudden deaths after a smoking cessation therapy. The other two include a validation study and an application, both relating to a putative association between vaccination with the oral polio vaccine (OPV) and intussusception in infants. The data and STATA program used to apply the proposed method in the validation study are available from the self-controlled case series website (http://statistics.open.ac.uk/sccs).

6.1 Bupropion and sudden death

Bupropion is an effective smoking cessation therapy. However, soon after its introduction in the UK, concerns were expressed that starting on Bupropion may increase the risk of sudden

death. A study was undertaken within the health improvement network (Hubbard et al 2005). Sudden deaths occurring within a defined ascertainment period ending on 11 November 2003 were documented. Clearly, individual observation periods and exposure histories are curtailed following a sudden death. However, in this study, it is not unreasonable to suppose that, had an individual died of a sudden death at any time within the ascertainment period, they would have been captured by the ascertainment process. Thus, we can take each individual's counterfactual end of observation b_i as being their age on 11 November 2003.

The question of interest is whether there is a risk associated with the initiation of Bupropion. In this analysis, only individuals who died following Bupropion were included, so the analysis proceeds as described in Subsection 3.1, with individual observation periods stretching from age at which Bupropion treatment was first started, and ending with age on 11 November 2003. The risk period was $0-27$ days after start of treatment. There were 121 cases of sudden death, including 2 in the risk period. The relative incidence was 0.50, 95% CI $(0.12, 2.05)$. These results provide no evidence that initiation of Bupropion is associated with an increased risk of sudden death within the first four weeks, though they are uninformative as to whether there is a long-term risk.

6.2 Validation study

This is a re-analysis of data originally described by Andrews et al. (2001) and used to evaluate whether there exists an association between oral polio vaccine (OPV) and intussusception. Intussusception is a condition where the bowel folds in on itself, obstructing the intestine. Most children diagnosed with intussusception have an operation and recover completely, so normally this is not an interferent event. The data are hospital episode statistics collected between January 1991 and March 1997. They comprise 207 children aged 28-365 days, of which 10 had one repeat episode that we excluded. The children received up to 3 doses of OPV. In the original analyses, Andrews et al. (2001) used risk periods 14-27 and 28-41 days after each dose, so that

there were a total of 6 post-OPV risk periods. For simplicity we took a single risk period 14-41 days after each of the 3 OPV doses. We used 11 monthly age groups. In this data set, $D = 3$, $J = 7$ and $K = 11$, the 207 events analysed were distributed as follows: $n_{.1} = 15$, $n_{.2} = 13$, $n_{.3} = 9$, $n_{.4} = 21$, $n_{.5} = 13$, $n_{.6} = 35$ and $n_{.7} = 101$.

To demonstrate the method described in this paper, the data were analysed in four ways:

1. A standard case series analysis with likelihood given in Section 2, using the full exposure information.

2. The post-intussusception exposures were censored, the observation period was redefined to end on the day an intussusception event occurred. We analysed the censored data by the standard case series method, without taking into account the fact that the exposures and the observation periods were censored.

3. The observation periods ended as in the original data, but post-intussusception exposures were censored. Data were analysed using the standard case series model, ignoring the censoring.

4. We analysed the censored data using the new method described in the present paper, using the pseudo-likelihood method described in sections 5.3 and 5.4.

For all four analyses, we obtained bootstrap confidence intervals. We present percentile bootstrap confidence intervals. These were similar to both normal and bias corrected bootstrap confidence intervals, and where a standard case series model was used, they were also similar to, though a little wider than, the Wald confidence intervals. The results are given in Table 1.

[Table 1 about here.]

Analysis 1 represents the gold standard case series analysis using data where the full exposure history is known. In this analysis a significant increase in intussusception 14-41 days after the 3rd dose of OPV was found. Analysis 2 violated the assumption that the observation period

must not depend on the occurrence of an event. The relative incidences for the risk periods after each of the 3 doses are all attenuated towards the null and the confidence intervals for doses 1 and 3 are considerably wider when compared with analysis 1. The estimated age effects are not shown, but were very biased, especially for older age groups. Analysis 3 ignored the assumption that exposures must not depend on previous events. Relative incidence estimates are substantially biased upwards when compared with analysis 1: as a result of censoring, post-event risk periods are classified as control periods, thus biasing the relative incidence upwards. Analysis 4 is the correct analysis for censored data. Although the relative incidence estimates are attenuated toward the null when compared with analysis 1, they are generally less attenuated than the estimates obtained from analysis 2. In conclusion, our method does correct the bias created by ignoring the dependence of the observation period on the event of interest to some extent, though cannot fully make up for the loss of data.

6.3 Polio vaccine and intussusception in Latin America

We present an analysis of data on intussusception in several Latin American countries. The data consist of 456 confirmed diagnoses of first intussusception of children aged up to 2 years who attended a participating hospital during the study period. Hospitals within 11 countries participated in the study: Argentina, Brazil, Chile, Costa Rica, Honduras, Mexico, Nicaragua, Panama, Peru, Dominican Republic and Columbia. Study periods for 10 countries spanned approximately 2 years, the other just 1 year, starting between January 2002 and September 2003. 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.

Of the 456 cases, 26 were unvaccinated, over half received at least 3 doses of OPV, 86 were given a 4th dose and 12 a 5th dose. Almost all of the first 3 doses, and the majority of the 4th and 5th doses of OPV, were administered during the 1st year of life. Age at diagnosis peaked

within the 6th month.

In our analysis age was stratified into 15 one or three month age bands. The longer 3-month age groups were used for ages when there were few events: age group 1 contained months 1-3, and the final 3 age groups included months 15-17, 18-20 and 21-23. Risk periods were taken to be 0-30 days after each of the 5 doses of OPV, so that there were a total of 5 risk periods, thus $D = 5$, $J = 11$ and $K = 14$. Analyses were carried out both with separate exposure effects β_d for each dose and with a common parameter β for all doses.

Relative incidences and 95% percentile bootstrap confidence intervals are given in Table 2. No significant change in incidence of intussusception in the post-OPV risk periods in comparison to all other periods was found.

[Table 2 about here.]

7 FURTHER POINTS

For simplicity it has been assumed throughout that there is a single risk period after each exposure. More generally, there may be several post-exposure risk periods, contiguous or otherwise, and the method can readily be adapted for such situations. Suppose for example that there are no age effects but two exposures, at ages c_{i1} and c_{i2}, each giving rise to two risk periods, $(c_{ir}, c_{ir} + \tau_1]$ and $(c_{ir} + \tau_2, c_{ir} + \tau_3]$ with $0 < \tau_1 \leq \tau_2 < \tau_3$. The observation period $(a_i, b_i]$ is then subdivided into $J = 9$ intervals labelled 1 to 9 in increasing order. The baseline incidence is multiplied by the factor e^{β_1} in interval 2, by e^{β_2} in interval 4, by e^{β_3} in interval 6, and by e^{β_4} in interval 8. Let

$$\begin{aligned} A_i(\beta_3, \beta_4) &= e^{\beta_3} e_{i6} + e_{i7} + e^{\beta_4} e_{i8} + e_{i9}, \\ B_i(\beta_1, \beta_2) &= e^{\beta_1} e_{i2} + e_{i3} + e^{\beta_2} e_{i4} + e_{i5} + e_{i6} + e_{i7} + e_{i8} + e_{i9}. \end{aligned}$$

Then the four elementary estimating functions are:

$$\begin{aligned}
U_i^{11}(\beta_3,\beta_4) &= n_{i6} - (n_{i6} + ... + n_{i9})\frac{e^{\beta_3} e_{i6}}{A_i(\beta_3,\beta_4)}, \\
U_i^{12}(\beta_3,\beta_4) &= n_{i8} - (n_{i6} + ... + n_{i9})\frac{e^{\beta_4} e_{i8}}{A_i(\beta_3,\beta_4)}, \\
U_i^{21}(\beta_1,\beta_2,\beta_3,\beta_4) &= n_{i2} - (n_{i2} + ... + n_{i5} + \frac{n_{i6}}{e^{\beta_3}} + n_{i7} + \frac{n_{i8}}{e^{\beta_4}} + n_{i9})\frac{e^{\beta_1} e_{i2}}{B_i(\beta_1,\beta_2)}, \\
U_i^{22}(\beta_1,\beta_2,\beta_3,\beta_4) &= n_{i4} - (n_{i2} + ... + n_{i5} + \frac{n_{i6}}{e^{\beta_3}} + n_{i7} + \frac{n_{i8}}{e^{\beta_4}} + n_{i9})\frac{e^{\beta_2} e_{i4}}{B_i(\beta_1,\beta_2)}.
\end{aligned}$$

Other situations are handled in a similar fashion.

Recently, a semi-parametric case series method has been developed, in which the age effect is left unspecified (Farrington & Whitaker 2006). Similar ideas could be applied to the analysis of interferent events. In practice, this means subdividing observation periods into large numbers of short intervals of unit length and ignoring those in which no event occurs. The asymptotics of such a scheme require further study.

The estimation method we propose could perhaps be improved by a more judicious combination of estimation equations. Thus, in Subsection 3.3 we obtain three estimating equations for the age effects, which we simply add together. This choice was motivated by computational reasons, though it may not be the optimal linear combination.

ACKNOWLEDGEMENTS

We wish to thank Dr Elizabeth Miller of the Health Protection Agency and Dr Thomas Verstraeten of GlaxoSmithKline Biologicals for the intussusception and OPV datasets. This research was supported by the Wellcome Trust, the EPSRC and the Fondation pour la Recherche Médicale.

References

ANDREWS, N., MILLER, E., WAIGHT, P., FARRINGTON, C.P., CROWCROFT, N., STOWE, J. & TAYLOR, B. (2001). Does oral polio vaccine cause intussusception in infants? Evidence from a sequence of three self-controlled case series studies in the United Kingdom. *European Journal of Epidemiology* **17**, 701-706.

BADDELEY, A. & TURNER, R. (2000). Practical maximum pseudo-likelihood for spatial point patterns. *Australia & New Zealand Journal of Statistics* **42**, 283-315.

BRYAN, J., YU, Z. & VAN DER LAAN, M.J. (2004). Analysis of longitudinal marginal structural models. *Biostatistics* **5**, 361-380.

DAVIDIAN, M., TSIATIS, A.A. & LEON, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* **20**, 261-301.

DAVISON, A.C. & HINKLEY, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

HORVITZ, D.G. & THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

HUBBARD, R., LEWIS, S., WEST, J., SMITH, C., GODFREY, C., SMEETH, L., FARRINGTON, P. & BRITTON, J. (2005). Bupropion and the risk of sudden death: a self-controlled case-series analysis using The Health Improvement Network. *Thorax* **60**, 848-850.

KALBFLEISCH, J.D. & PRENTICE, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition. Hoboken, New Jersey: Wiley.

KALBFLEISCH, J.D. (1998). *Pseudo-Likelihood* in *Encyclopedia of Biostatistics* ed. Armitage, P. and Colton, T. Chichester: Wiley. pp3566-3568.

FARRINGTON, C.P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **51**, 228-235.

FARRINGTON, C.P. & WHITAKER, H.J. (2006). Semiparametric analysis of case series data (with Discussion). *Applied Statistics* **55**, 553-594.

LEVY, P.S. (1998). *Horvitz-Thompson Estimator* in *Encyclopedia of Biostatistics* ed. Armitage, P. and Colton, T. Chichester: Wiley. pp1954-1955.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846-866.

ROBINS, J.M,. HERNAN, M.A. & BRUMBACK, B. (2000). Marginal structural models and causal infertence in epidemiology. *Epidemiology* **11**, 550-560.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

WHITAKER, H.J., FARRINGTON, C.P., SPIESSENS, B. & MUSONDA, P. (2006). Tutorial in Biostatistics: The self-controlled case series method. *Statistics in Medicine* **25**, 1768-1797.

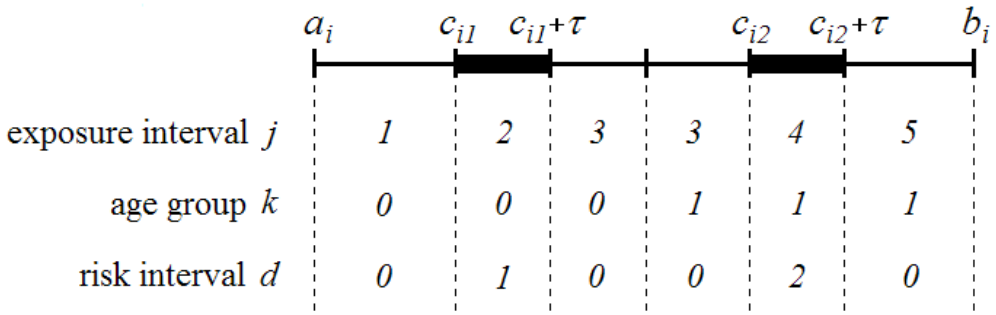

Figure 1: Configuration for two exposures and two age groups.

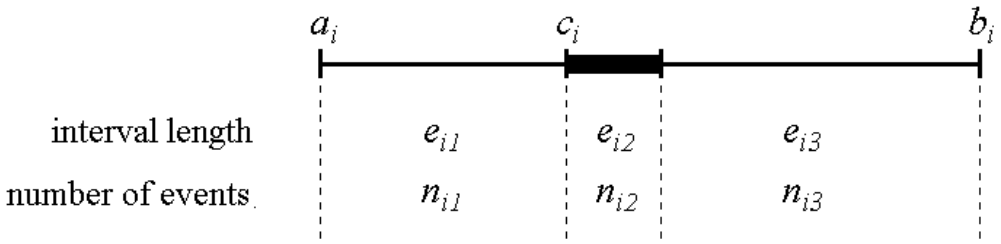

Figure 2: Configuration for a unique exposure.

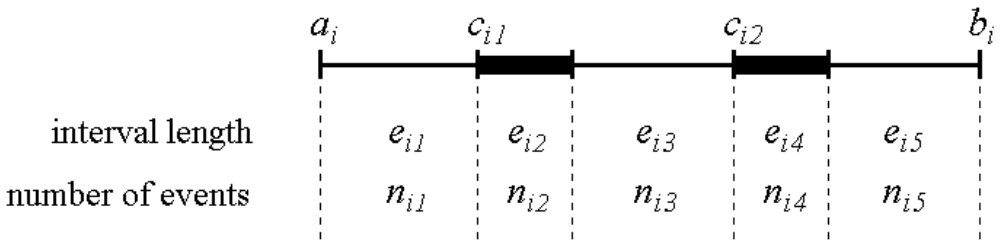

Figure 3: Configuration for two exposures.

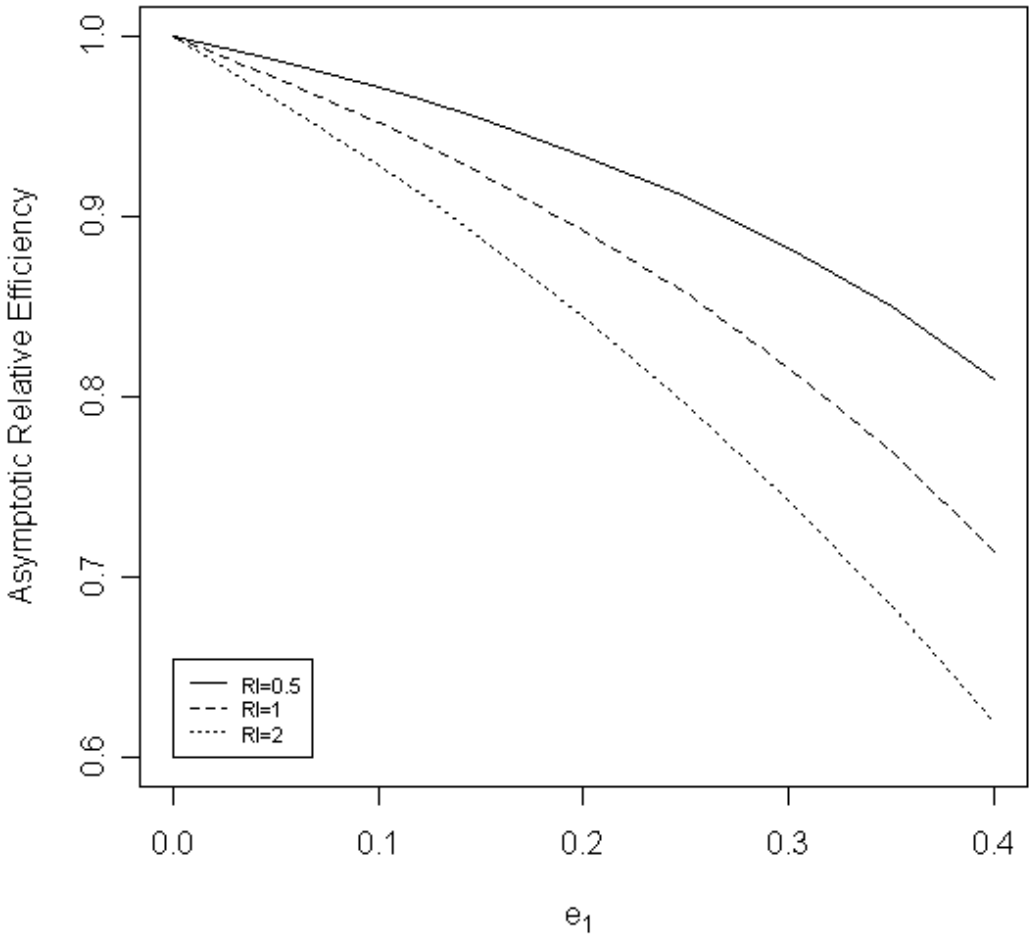

Figure 4: Asymptotic relative efficiency with one exposure as a function of e_1 for different values of the relative incidence (RI).

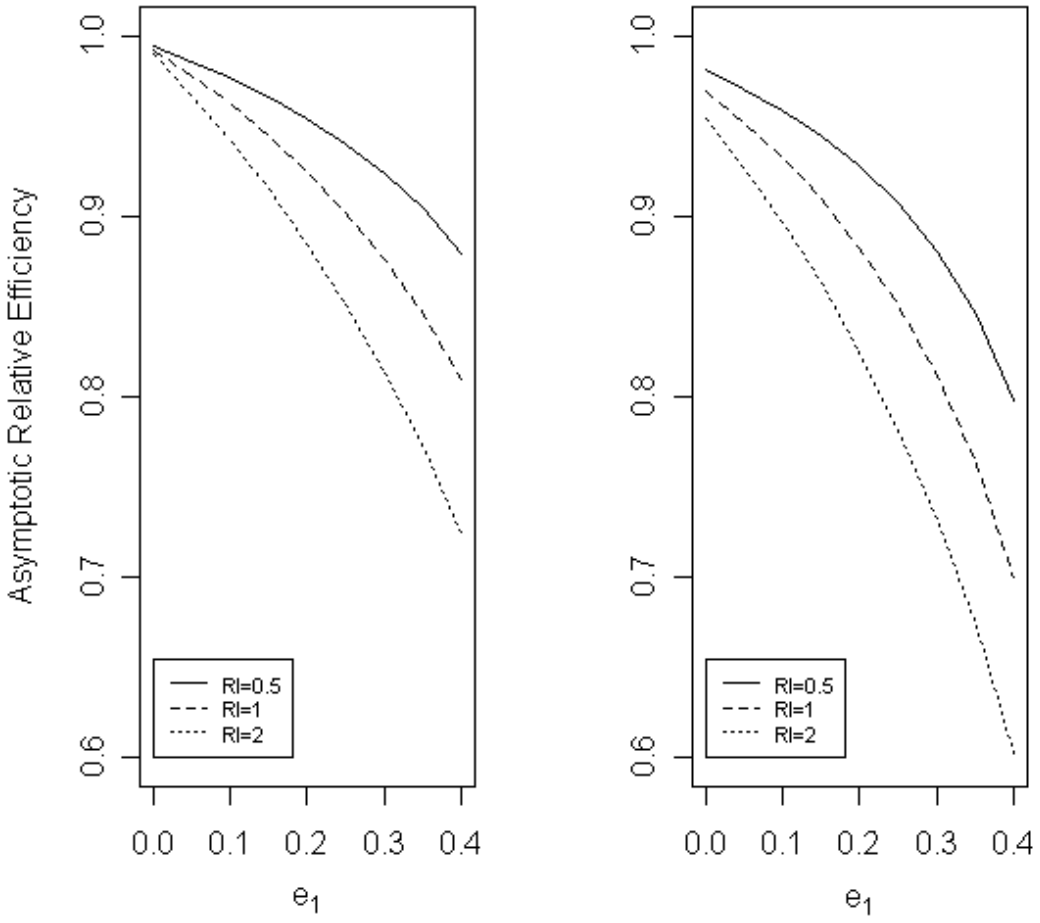

Figure 5: Asymptotic relative efficiency with two exposures as a function of e_1 for different values of the relative incidence (RI). Left panel: $e_3 = 0.05$; right panel: $e_3 = 0.25$.

Table 1: Relative incidence (RI) and 95% confidence interval (CI) for analyses of intussusception and OPV

dose	analysis 1: original data standard model RI (95% CI)	analysis 2: observation ends at event standard model RI (95% CI)	analysis3: censored data standard model RI (95% CI)	analysis 4: censored data censoring model RI (95% CI)
1	0.710 (0.328, 1.408)	0.582 (0.048, 3.799)	0.850 (0.370, 1.652)	0.581 (0.257, 1.170)
2	0.922 (0.501, 1.639)	0.476 (0.174, 1.170)	1.431 (0.733, 2.657)	0.876 (0.439, 1.629)
3	1.625 (1.038, 2.589)	1.347 (0.485, 5.101)	2.913 (1.687, 5.017)	1.566 (0.999, 2.519)

Table 2: Relative incidence (RI) and 95% confidence interval (CI) for analyses of intussusception and OPV in Latin America

risk period 0-30days after dose:	RI (95% CI) 1
1	1.253 (0.629, 2.235)
2	1.069 (0.741, 1.533)
3	0.970 (0.677, 1.365)
4	0.998 (0.500, 1.618)
5	1.599 (0.000, 6.125)
all doses	1.054 (0.820, 1.342)