Biostatistics (2009), **10**, 1, *pp.* 3–16
doi:10.1093/biostatistics/kxn013
Advance Access publication on May 21, 2008

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

C. PADDY FARRINGTON*, HEATHER J. WHITAKER, MOUNIA N. HOCINE

Department of Mathematics and Statistics, The Open University,
Milton Keynes MK7 6AA, UK
c.p.farrington@open.ac.uk

Summary

A new method is developed for analyzing 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 is influenced by the event. The method applies to transient point exposures and rare nonrecurrent events. Asymptotic efficiency is studied in some special cases. A computational scheme based on a pseudo-likelihood is proposed to make the computations feasible in complex models. Simulations, a validation study, and 2 applications are described.

Keywords: Censored data; Counterfactual; Endogeneity; Estimating equation; Horvitz–Thompson estimator; Pseudo-likelihood; Self-controlled case series.

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 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 modeling and applications, see Whitaker *and others* (2006).

The model is derived by conditioning on the number of events and the exposure history experienced by each individual over a predetermined 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 the 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

*To whom correspondence should be addressed.

by the application to myocardial infarction discussed in Farrington and Whitaker (2006). Nevertheless, it is desirable to have a method applicable whenever the exposure process or the exposure observation process is 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, applicable to binary exposures, which can be used in such circumstances, provided the postexposure 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 (CIs). The performance of the methods is studied by simulation. Section 6 contains 3 examples including a validation study and 2 applications. 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 predetermined observation period $(a_i, b_i]$, usually defined in terms of age, during which this individual may experience point exposures, at ages $c_{i1}, \ldots, c_{iD}$ say. We assume that $c_{i1} < c_{i2} < \cdots < c_{iD}$ and that following the dth exposure, the incidence of the event of interest is multiplied by a factor e^{β_d} over the interval $(c_{id}, \min\{c_{id} + \tau, c_{id+1}\}]$, which we call a risk period. The intervals during which the individual does not experience an exposure-related risk are called control periods. From now on, to simplify the notation without sacrificing essential generality, we shall assume that $a_i \leqslant c_{id} + \tau \leqslant c_{id+1} \leqslant b_i$. The sequence of exposures for individual i determines $J = 2D + 1$ contiguous nonoverlapping control and risk periods $(a_i, c_{i1}]$, $(c_{i1}, c_{i1} + \tau]$, $(c_{i1} + \tau, c_{i2}]$, …,$(c_{iD} + \tau, b_i]$ indexed by $j = 1, \ldots, J$. The risk periods correspond to even values of j and the control periods to odd values of j. (If $c_{id} + \tau = c_{id+1}$, interval $2d + 1$ is empty.)

Suppose furthermore that age is stratified in $K + 1$ age groups indexed by $k = 0, 1, \ldots, K$, leading to a further subdivision of the J control or risk periods into up to $K + 1$ subintervals. Let E_{ijk} denote the subset of the observation period for individual i lying within the jth risk or control period and the kth age group. E_{ijk} has length e_{ijk}; typically, some $e_{ijk} = 0$. A possible configuration for a single individual i is shown in Figure 1, for which $e_{i11} = e_{i21} = e_{i40} = e_{i50} = 0$.

In the case series method, it is assumed that the exposure is an external time-varying covariate (Kalbfleisch and Prentice, 2002) and that, conditionally on the ages at exposure, events arise in a nonhomogeneous Poisson process with rate

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

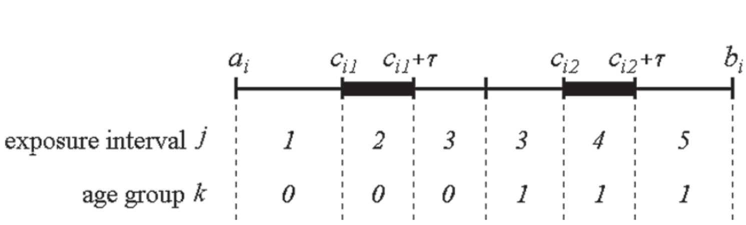

Fig. 1. Configuration for two exposures and two age groups.

This is piecewise constant on the E_{ijk}. The parameters φ_i represent the individual effect, α_k the age effect, and $\beta_{d(j)}$ the exposure effect ($\beta_0 = 0$) with

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

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

$$L_i = \prod_{j,k} \left\{ \frac{e_{ijk} \exp(\alpha_k + \beta_{d(j)})}{\sum_{r,s} e_{irs} \exp(\alpha_s + \beta_{d(r)})} \right\}^{n_{ijk}}. \tag{2.1}$$

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 with which we shall exclusively be concerned in the present paper. In this case, the case series likelihood (2.1) is valid in the limit $\varphi_i \to -\infty$ (see Farrington and 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. It is also assumed that occurrence of an event does not alter the duration of the 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 6.3). Alternatively, the post-event exposure process might only be partially observed. In either situation, the case series likelihood (2.1) 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 contraindication to the drug of interest (as is the case with rotavirus vaccination and intussusception), 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 (2.1) 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 and Whitaker (2006)). In this case, the case series likelihood (2.1) is no longer valid either because the observation period $(a_i, b_i]$ depends on the event time.

Typically, as well as modeling exposures, it is 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 processes are stationary.

2.3 *A way forward*

To make progress, it is useful to think in terms of counterfactuals, 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 processes 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 the case ascertainment can reasonably be regarded as complete. An example is provided in Section 6.1. Hence, in all that follows, we shall assume that b_i, whether factual or counterfactual, is known.

Thus, all 3 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 likelihood. In the myocardial example referred to previously, we imputed counterfactual exposures and showed that not knowing them had little effect on the results (Farrington and 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 analyzing 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 and Thompson, 1952; Levy, 1998) to adjust the event counts to what they would have been, on average, 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 *and others*, 1994, 2000; Bryan *and others*, 2004; Davidian *and others*, 2005). Note, however, that our approach retains the 2 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 assumptions that the exposure is binary, that is, present or absent; the postexposure risk period is short; and the event of interest is an uncommon, nonrecurrent event. Note that the risk is assumed to return 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 special case which will help to motivate the general method. This special case reveals 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 new estimating equations 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.

3.1 *Two exposures and 2 age groups*

We consider a situation in which an individual i can be exposed on up to 2 occasions at ages c_{i1} and c_{i2}. We assume that the observation period $(a_i, b_i]$ that would have applied had no event occurred is known. The exposures partition the observation period $(a_i, b_i]$ into up to $J = 5$ successive control and risk periods, indexed by j. In addition, we assume that there are 2 age groups (so $K = 1$), indexed by k. Thus, the observation period for individual i is partitioned into up to 6 intervals E_{ijk} of length e_{ijk}, in which n_{ijk} events occur. (See Figure 1 for an example.) Note that for all i, $n_{i\cdot\cdot} = 1$ since the event is nonrecurrent and only cases are considered; this fact is essential in what follows.

Let β_1 and β_2 denote the log relative incidences (RIs) associated with the first and second risk period, respectively, and α the RI associated with age group 1 relative to age group 0. We shall take age group 0 to be the earlier one. Now let T_i denote the age at event for an individual i. Generally, the counterfactual exposure history after T_i (i.e. the exposures that would have been observed had the event not occurred) is not known.

There is, however, one key exception: if $T_i \geqslant c_{i2}$, then we know that no further exposures would have taken place. Inference about β_2 may thus proceed using a standard case series analysis with observation period redefined as $(c_{i2}, b_i]$. The log-likelihood for this case series model is

$$l_i(\beta_2, \alpha) = n_{i4\cdot}\beta_2 - (n_{i4\cdot} + n_{i5\cdot}) \log\{\mathrm{e}^{\beta_2}(e_{i40} + \mathrm{e}^{\alpha} e_{i41}) + (e_{i50} + \mathrm{e}^{\alpha} e_{i51})\}$$

and the elementary score function for β_2 is

$$U_{i2}(\beta_2, \alpha) = n_{i4\cdot} - (n_{i4\cdot} + n_{i5\cdot}) \frac{\mathrm{e}^{\beta_2}(e_{i40} + \mathrm{e}^{\alpha} e_{i41})}{\mathrm{e}^{\beta_2}(e_{i40} + \mathrm{e}^{\alpha} e_{i41}) + (e_{i50} + \mathrm{e}^{\alpha} e_{i51})}.$$

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 1 exposure. Instead, we proceed as if no individual experiences a second exposure, and let $n^*_{i4\cdot}$ 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 $n^*_{i4\cdot}$ were observed, we could then estimate β_1 using the elementary score function

$$U^*_{i1}(\beta_1, \alpha) = n_{i2\cdot} - (n_{i2\cdot} + n_{i3\cdot} + n^*_{i4\cdot} + n_{i5\cdot}) \\ \times \frac{\mathrm{e}^{\beta_1}(e_{i20} + \mathrm{e}^{\alpha} e_{i21})}{\mathrm{e}^{\beta_1}(e_{i20} + \mathrm{e}^{\alpha} e_{i21}) + (e_{i30} + \mathrm{e}^{\alpha} e_{i31}) + (e_{i40} + \mathrm{e}^{\alpha} e_{i41}) + (e_{i50} + \mathrm{e}^{\alpha} \mathrm{e}_{i51})}.$$

However, if the event occurs after the second exposure, then $n_{i4\cdot}$ is observed but not $n^*_{i4\cdot}$. In this case, the score function U^*_{i1} cannot be evaluated. We therefore replace $n^*_{i4\cdot}$ by an unbiased estimator of $n^*_{i4\cdot}$, namely the Horvitz–Thompson-like estimator $n_{i4\cdot}\mathrm{e}^{-\beta_2}$. (We call it Horvitz–Thompson like because the adjustment factor is a relative rate not a probability.) Thus, we obtain the elementary estimating function

$$U_{i1}(\beta_1, \beta_2, \alpha) = n_{i2\cdot} - \left(n_{i2\cdot} + n_{i3\cdot} + \frac{n_{i4\cdot}}{\mathrm{e}^{\beta_2}} + n_{i5\cdot}\right) \\ \times \frac{\mathrm{e}^{\beta_1}(e_{i20} + \mathrm{e}^{\alpha} e_{i21})}{\mathrm{e}^{\beta_1}(e_{i20} + \mathrm{e}^{\alpha} e_{i21}) + (e_{i30} + \mathrm{e}^{\alpha} e_{i31}) + (e_{i40} + \mathrm{e}^{\alpha} e_{i41}) + (e_{i50} + \mathrm{e}^{\alpha} e_{i51})}.$$

This estimating function is unbiased conditionally on $n_{i2+} = n_{i2\cdot} + n_{i3\cdot} + n_{i4\cdot} + n_{i5\cdot}$ and on c_{i1} and c_{i2}, the latter possibly being unavailable in the observed realization. Unbiasedness follows because the event is nonrecurrent, so $n_{i2+} = 0$ or 1. The key point is that the estimating function U_{i1} can always be evaluated, even if c_{i2} is unavailable, since in this case $n_{i4\cdot} = n_{i5\cdot} = 0$ and $e_{i3\cdot} + e_{i4\cdot} + e_{i5\cdot} = b_i - (c_{i1} + \tau)$ is known, even though $e_{i3\cdot}$, $e_{i4\cdot}$, and $e_{i5\cdot}$ are not.

We now construct a third unbiased estimating function for α. This comprises 3 components. The case series likelihood restricted to events after the second exposure yields the following elementary score function for α:

$$U_{i3}^{2}(\beta_2,\alpha) = (n_{i41}+n_{i51}) - (n_{i4\cdot}+n_{i5\cdot})\frac{\mathrm{e}^{\alpha}(\mathrm{e}^{\beta_2}e_{i41}+e_{i51})}{(\mathrm{e}^{\beta_2}e_{i40}+e_{i50})+\mathrm{e}^{\alpha}(\mathrm{e}^{\beta_2}e_{i41}+e_{i51})}.$$

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

$$U_{i3}^{1}(\beta_1,\beta_2,\alpha) = \left(n_{i21}+n_{i31}+\frac{n_{i41}}{\mathrm{e}^{\beta_2}}+n_{i51}\right) - \left(n_{i2\cdot}+n_{i3\cdot}+\frac{n_{i4\cdot}}{\mathrm{e}^{\beta_2}}+n_{i5\cdot}\right)$$
$$\times\frac{\mathrm{e}^{\alpha}(\mathrm{e}^{\beta_1}e_{i21}+e_{i31}+e_{i41}+e_{i51})}{(\mathrm{e}^{\beta_1}e_{i20}+e_{i30}+e_{i40}+e_{i50})+\mathrm{e}^{\alpha}(\mathrm{e}^{\beta_1}e_{i21}+e_{i31}+e_{i41}+e_{i51})}.$$

Finally, applying the same argument to all the events in $(a_i, b_i]$, including those occurring prior to the first exposure, yields the following elementary estimating function:

$$U_{i3}^{0}(\beta_1,\beta_2,\alpha) = \left(n_{i11}+\frac{n_{i21}}{\mathrm{e}^{\beta_1}}+n_{i31}+\frac{n_{i41}}{\mathrm{e}^{\beta_2}}+n_{i51}\right)$$
$$-\left(n_{i1\cdot}+\frac{n_{i2\cdot}}{\mathrm{e}^{\beta_1}}+n_{i3\cdot}+\frac{n_{i4\cdot}}{\mathrm{e}^{\beta_2}}+n_{i5\cdot}\right)\frac{\mathrm{e}^{\alpha}e_{i\cdot 1}}{e_{i\cdot 0}+\mathrm{e}^{\alpha}e_{i\cdot 1}}.$$

We shall use as the third elementary estimating function the sum of all 3 elementary estimating functions for α, namely

$$U_{i3}(\beta_1,\beta_2,\alpha) = U_{i3}^{0}(\beta_1,\beta_2,\alpha) + U_{i3}^{1}(\beta_1,\beta_2,\alpha) + U_{i3}^{2}(\beta_2,\alpha).$$

The system $\sum_i U_{i1}$, $\sum_i U_{i2}$, $\sum_i U_{i3}$ provides a triple of unbiased estimating equations (conditionally on each case experiencing a single event) which may be used to estimate β_1, β_2, and α. 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 $\sum_i U_{i1}$ and $\sum_i U_{i2}$ is to take their sum, $\sum_i (U_{i1} + U_{i2})$. Combining estimating equations in this way is convenient for computational reasons that will become apparent in Section 5.

In this subsection, we made the assumption that the maximum number of possible exposures D is known, with $D = 2$. In fact this assumption is not necessary: we can use the method described if the maximum number of pre-event exposures observed is 2, but more could have occurred (but did not in the sample). All this means is that only β_1 and β_2 are estimable. More generally, if maximum D pre-event exposures are observed for each case, then any exposure that might have occurred subsequently can be ignored and only $\beta_1, \ldots, \beta_D$ are estimable.

3.2 *The general case*

The special case described in Section 3.1 helps to motivate the estimating equations for the general case, in which there are up to D distinct observed pre-event exposures at ages $c_{i1}, \ldots, c_{iD}$. We suppose also that there are $K + 1$ age groups indexed by $k = 0, 1, \ldots, K$. To the D risk periods correspond D log RI parameters $\beta_1, \ldots, \beta_D$ (relative to the control periods, for which $\beta_0 = 0$). To the $K + 1$ age groups correspond K parameters $\alpha_1, \ldots, \alpha_K$. These represent log RIs, relative to the 0-indexed age group, so $\alpha_0 = 0$.

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

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

the redundant subscript k being retained for clarity in what follows. The elementary estimating function for β_d, $d = 1, \ldots, D$, is

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

where the subscript $i(2d)k$ represents ijk with $j = 2d$ and $d(j)$ is defined in Section 2.1. The elementary estimating function for α_k, $k = 1, \ldots, K$, is

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

where the subscript $i(D + k)$ represents ir with $r = D + k$, and

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

If the parameters β_d and $\beta_{d'}$ are constrained to be equal, then the 2 elementary estimating functions U_{id} and $U_{id'}$ are replaced by their sum.

3.3 *Sandwich variance estimates*

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

$$V_{rs} = \sum_{i=1}^{n} U_{ir} U_{is},$$

$$H_{rs} = \sum_{i=1}^{n} \frac{\partial U_{ir}}{\partial \theta_s}.$$

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

4. Relative efficiency

In this section, we explore the asymptotic efficiency of $\widehat{\beta}$ (we assume a common β_d and hence suppress the subscripts) estimated using the standard 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 Section 3. 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.

We consider special cases in which individuals experience up to 2 exposures, there are no age effects, and all individuals share the same partition of their observation period. The asymptotic relative efficiencies are derived and discussed in detail in Section 1 of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). More general relative efficiency calculations are possible but unenlightening and are not pursued.

The key messages from these investigations are as follows: The relative efficiency is high for short risk periods and when the preexposure period and inter-exposure control periods are short in proportion to the overall observation time. Relative efficiency is low when there is little postexposure control time. Thus, in designing case series studies of interferent events, it is important where possible (i.e. without unduly reducing the number of events) to select the observation period so as to minimize the amount of time prior to the first exposure and maximize the control time after the last exposure. As expected, the proposed method cannot be used for indefinite risk periods.

5. A pseudo-likelihood method

In practice, there may be many age groups and exposures. In such settings, writing down and solving the estimating equations and deriving the sandwich variance estimator can become very cumbersome. In other circumstances, the asymptotic theory upon which the sandwich variance estimator is based may not be applicable. In this section, we present an alternative approach to derive the estimating equations, which is convenient for computation and which lends itself to bootstrapping. 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 2 age groups*

For a count n and weight w with $0 \leqslant w \leqslant 1$, let the expression $nw \sim P(\mu)$ denote a likelihood contribution proportional to $\mathrm{e}^{-\mu}\mu^{nw}$ 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 and Turner, 2000). Recall from Section 3.1 that the elementary estimating functions U_{i2} and U_{i3}^2 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

$$n_{ijk}w_{ijk}^{(2)} \sim P(\lambda_{ijk}e_{ijk}), \quad j = 4, 5, \ \ k = 0, 1;$$

$$\log(\lambda_{ijk}) = \varphi_i^{(2)} + \beta_2 I(j = 4) + \alpha I(k = 1),$$

where $I(.)$ is the indicator function and $w_{ijk}^{(2)}$ is the weight defined in Section 3.2, which here is 1 for $j \geqslant 4$ and 0 otherwise.

Similarly, U_{i1} and U_{i3}^1 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 $n_{i4k}\,\mathrm{e}^{-\beta_2}$. These estimating functions may equivalently be derived as score equations from the pseudo-Poisson model

$$n_{ijk}w_{ijk}^{(1)} \sim P(\lambda_{ijk}e_{ijk}), \quad j = 2, 3, 4, 5, \ \ k = 0, 1;$$

$$\log(\lambda_{ijk}) = \varphi_i^{(1)} + \beta_1 I(j = 2) + \alpha I(k = 1).$$

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

Finally, U_{i3}^0 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 $n_{i2k}\,\mathrm{e}^{-\beta_1}$, and n_{i4k} with n_{i4k}^*, estimated by $n_{i4k}\,\mathrm{e}^{-\beta_2}$. This estimating function may be derived as a score contribution for α from the pseudo-Poisson model

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

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

Now, stack the 3 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_{i1}, U_{i2}, and U_{i3}.

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, labeled 0 to D. Stack 0 contains all the data to which the following model is to be fitted:

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

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

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

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 Section 3.2.

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 maximizing the pseudo-likelihood with these weights. Update the weights using the new values of the β_d and iterate until convergence. The procedure resembles an Expectation (E)–maximization (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 maximized (the M-step).

5.4 *Bootstrap estimates*

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.

The simplest method is nonparametric bootstrapping in which the stacked data for individuals $i = 1, \ldots, n$ 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 and Hinkley, 1997).

5.5 *Simulations*

The performance of the method and the different approaches to obtain interval estimates were studied by simulation. The simulations and their results are described in Section 2 of the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). The medians of the estimated log RIs 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 95% CIs were also close to 0.95. As expected, the pseudo-likelihood method of Section 5.2 generated the same estimates as the estimating equations. The overall conclusion from these simulations is that the model performs well.

6. Examples

We present 3 examples. The first relates to sudden deaths after a smoking cessation therapy. The other 2 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 United Kingdom, concerns were expressed that starting on Bupropion may increase the risk of sudden death. A study was undertaken within the health improvement network (Hubbard *and others*, 2005). Sudden deaths occurring within a defined ascertainment period ending on November 11, 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 November 11, 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 with individual observation periods stretching from age at which Bupropion treatment was first started and ending with age on November 11, 2003. The risk period was 0–27 days after the start of treatment. There were 121 cases of sudden death including 2 in the risk period. The RI 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 4 weeks, though they are uninformative as to whether there is a long-term risk.

6.2 *Validation study*

This is a reanalysis of data originally described by Andrews *and others* (2001) and is used to evaluate whether there exists an association between 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 in the United Kingdom between January, 1991, and March, 1997. They comprise 207 children aged 28–365 days of whom 10 had one repeat episode that we excluded. The children received up to 3 doses of OPV. In the original analyses, Andrews *and others* (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. We combined these into 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 analyzed were distributed as follows: $n_{\cdot 1 \cdot} = 15$, $n_{\cdot 2 \cdot} = 13$, $n_{\cdot 3 \cdot} = 9$, $n_{\cdot 4 \cdot} = 21$, $n_{\cdot 5 \cdot} = 13$, $n_{\cdot 6 \cdot} = 35$, and $n_{\cdot 7 \cdot} = 101$.

To demonstrate the method described in this paper, the data were analyzed in the following 4 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 and the observation period was redefined to end on the day an intussusception occurred. We analyzed 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 analyzed using the standard case series model ignoring the censoring.

(4) We analyzed the censored data with the new method described in the present paper; using the pseudo-likelihood method described in Sections 5.3 and 5.4.

For all 4 analyses, we obtained bootstrap CIs. We present percentile bootstrap CIs. These were similar to both normal and bias-corrected bootstrap CIs, and where a standard case series model was used, they were also similar to, though a little wider than, the Wald CIs. The results are given in Table 1.

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 the risk of intussusception 14–41 days after the third dose of OPV was found. Analysis 2 violated the assumption that the observation period must not depend on the occurrence of an event. The RIs for the risk periods after each of the 3 doses are all attenuated toward the null, and the CIs 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. RI estimates are substantially biased upward when compared with analysis 1. As a result of censoring, post-event risk periods are classified as control periods, thus biasing the RI upward. Analysis 4 is the correct analysis for censored data. Although the RI estimates are attenuated toward the null when compared with analysis 1, they are

Table 1. *RI and 95% CI for analyses of intussusception and OPV in the United Kingdom*

Dose	Analysis 1: original data	Analysis 2: observation ends at event	Analysis3: censored data	Analysis 4: censored data
	Standard model RI (95% CI)	Standard model RI (95% CI)	Standard model RI (95% CI)	Censoring model RI (95% CI)
1	0.710	0.582	0.850	0.581
	(0.328, 1.408)	(0.048, 3.799)	(0.370, 1.652)	(0.257, 1.170)
2	0.922	0.476	1.431	0.876
	(0.501, 1.639)	(0.174, 1.170)	(0.733, 2.657)	(0.439, 1.629)
3	1.625	1.347	2.913	1.566
	(1.038, 2.589)	(0.485, 5.101)	(1.687, 5.017)	(0.999, 2.519)

Table 2. *RI and 95% CI for analyses of intussusception and OPV in Latin America*

Risk period 0-30days after dose	RI (95% CI)
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)

generally less attenuated than the estimates obtained from analysis 2. In conclusion, our method does correct the bias created by ignoring the dependence of exposures and 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 fourth dose and 12 a fifth dose. Almost all the first 3 doses and the majority of the fourth and fifth doses of OPV were administered during the first year of life. Age at diagnosis peaked within the sixth month.

In our analysis, age was stratified into 15 1- or 3-month age bands. The longer 3-month age groups were used for ages when there were few events: age group 1 contained 1–3 months and the final 3 age groups included 15–17, 18–20, and 21–23 months. 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.

RIs and 95% percentile bootstrap CIs 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.

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 postexposure risk periods, contiguous or otherwise, and the method can readily be adapted for such situations. Suppose for example that there are no age effects (so the subscript k is suppressed) but 2 exposures, at ages c_{i1} and c_{i2}, each giving rise to 2 risk periods, $(c_{ir}, c_{ir} + \tau_1]$ and $(c_{ir} + \tau_2, c_{ir} + \tau_3]$ with $0 < \tau_1 \leqslant \tau_2 < \tau_3$. The observation period $(a_i, b_i]$ is then subdivided into $J = 9$ intervals labeled 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

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

Then, the 4 elementary estimating functions are

$$U_{i1}(\beta_1, \beta_2, \beta_3, \beta_4) = n_{i2} - \left(n_{i2} + \ldots + n_{i5} + \frac{n_{i6}}{\mathrm{e}^{\beta_3}} + n_{i7} + \frac{n_{i8}}{\mathrm{e}^{\beta_4}} + n_{i9}\right) \frac{\mathrm{e}^{\beta_1} e_{i2}}{B_i(\beta_1, \beta_2)},$$
$$U_{i2}(\beta_1, \beta_2, \beta_3, \beta_4) = n_{i4} - \left(n_{i2} + \ldots + n_{i5} + \frac{n_{i6}}{\mathrm{e}^{\beta_3}} + n_{i7} + \frac{n_{i8}}{\mathrm{e}^{\beta_4}} + n_{i9}\right) \frac{\mathrm{e}^{\beta_2} e_{i4}}{B_i(\beta_1, \beta_2)},$$
$$U_{i3}(\beta_3, \beta_4) = n_{i6} - (n_{i6} + \ldots + n_{i9}) \frac{\mathrm{e}^{\beta_3} e_{i6}}{A_i(\beta_3, \beta_4)},$$
$$U_{i4}(\beta_3, \beta_4) = n_{i8} - (n_{i6} + \ldots + n_{i9}) \frac{\mathrm{e}^{\beta_4} e_{i8}}{A_i(\beta_3, \beta_4)}.$$

Other situations are handled in a similar fashion.

Simple variations of the method can also accommodate staggered observation periods for which $c_{id} \leqslant a_i$ or $b_i < c_{id}$ for some i and d.

Recently, a semiparametric case series method has been developed in which the age effect is left unspecified (Farrington and 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. For example, in Section 3.1, we obtained 3 estimating equations for the age effects, which we simply added together. This choice was motivated by computational convenience though it may not be the optimal linear combination.

Finally, we are aware that the method we propose, in which we impose our own counterfactuals to render unobserved exposure histories determinate and then adjust the data where necessary to fit in with these counterfactuals, is derived using arguments very much adapted to the particular circumstances of the case series method. It would be desirable to recast this approach within a more general theoretical framework.

Acknowledgments

We wish to thank Dr Elizabeth Miller of the Health Protection Agency and Dr Thomas Verstraeten of GlaxoSmithKline Biologicals for the intussusception and OPV data sets. We thank the editor, an associate editor, and the 2 referees for very helpful comments. *Conflict of Interest*: None declared.

Funding

Engineering and Physical Sciences Research Council (E02873X); Fondation pour la Recherche Médicale, Wellcome Trust (GR070346).

References

Andrews, N., Miller, E., Waight, P., Farrington, C. P., Crowcroft, N., Stowe, J. and 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. and Turner, R. (2000). Practical maximum pseudo-likelihood for spatial point patterns. *Australia & New Zealand Journal of Statistics* **42**, 283–315.

Bryan, J., Yu, Z. and Van Der Laan, M. J. (2004). Analysis of longitudinal marginal structural models. *Biostatistics* **5**, 361–380.

Davidian, M., Tsiatis, A. A. and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* **20**, 261–301.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

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

Farrington, C. P. and Whitaker, H. J. (2006). Semiparametric analysis of case series data (with discussion). *Applied Statistics* **55**, 553–594.

Horvitz, D. G. and 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. and 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. (1998). Pseudo-likelihood. In: Armitage, P. and Colton, T. (editors), *Encyclopedia of Biostatistics*. Chichester: Wiley, pp. 3566–3568.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, NJ: Wiley.

Levy, P. S. (1998). Horvitz-Thompson estimator. In: Armitage, P. and Colton, T. (editors), *Encyclopedia of Biostatistics*. Chichester: Wiley, pp. 1954–1955.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

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

Whitaker, H. J., Farrington, C. P., Spiessens, B. and Musonda, P. (2006). Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine* **25**, 1768–1797.

[*Received May 14, 2007; revised March 18, 2008; accepted for publication April 22, 2008*]