Self-Controlled Case Series Analysis with Event-Dependent Observation Periods

C. Paddy Farrington[1] Karim Anaya Heather J. Whitaker

Mounia N. Hocine Ian Douglas Liam Smeeth

[1]Paddy Farrington (c.p.farrington@open.ac.uk), Karim Anaya, and Heather Whitaker are statisticians at the Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, United Kingdom. Mounia Hocine is statistician at the Conservatoire National des Arts et Métiers, Chaire Hygiène et Sécurité, Paris 75003, France. Ian Douglas and Liam Smeeth are epidemiologists at the Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom. This research was supported by funding from the UK Engineering and Physical Sciences Research Council.

Abstract

The self-controlled case series method may be used to study the association between a time-varying exposure and a health event. It is based only on cases, and it controls for fixed confounders. Exposure and event histories are collected for each case over a pre-defined observation period. The method requires that observation periods should be independent of event times. This requirement is violated when events increase the mortality rate, since censoring of the observation periods is then event-dependent. In this paper, the method is extended to remove this independence assumption, thus introducing an additional term in the likelihood that depends on the censoring process. In order to remain within the case series framework in which only cases are sampled, the model is reparameterized so that this additional term becomes estimable from the distribution of intervals from event to end of observation. The exposure effect of primary interest may be estimated unbiasedly. The age effect, however, takes on a new interpretation, incorporating the effect of censoring. The model may be fitted in standard loglinear modelling software; this yields conservative standard errors. We describe a detailed application to the study of antipsychotics and stroke. The estimates obtained from the standard case series model are shown to be biased when event-dependent observation periods are ignored. When they are allowed for, antipsychotic use remains strongly positively associated with stroke in patients with dementia, but not in patients without dementia. A detailed simulation study is included as Supplemental Material.

Key words: Censoring at random; Censoring completely at random; Conditional Poisson regression; Epidemiologic study.

1 INTRODUCTION

The self-controlled case series method, or case series method for short, is used to study the association between a time-varying exposure and a health event (Farrington 1995; Farrington and Whitaker 2006). The method is derived from a risk-interval Poisson cohort model, by conditioning on a pre-defined observation period, the exposure history and the number of events observed over this observation period. A key feature of the method is that only cases, namely individuals with one or more events, need be sampled. A further property of the method is that it is self-controlled, adjusting implicitly for all fixed confounders acting multiplicatively on the hazard. The method has been used in pharmacoepidemiology and in general epidemiology; for a review of recent developments see Whitaker, Hocine and Farrington (2009).

The observation period for a given individual is an interval $(a, b]$ over which exposure status at each age $t \in (a, b]$ and all events are recorded. (The case series method also works for uncommon non-recurrent events, so in fact only the first event need be recorded.) The observation period is generally determined using pre-specified age and calendar time boundaries. In practice, observation may be censored at some point $c \leq b$, so that the individual is only observed over $(a, c] \subseteqq (a, b]$. If the individual is not censored before b, we formally set $c > b$ and observe the individual over $(a, b]$. Let $b \wedge c$ denote the minimum of b and c. We call $(a, b \wedge c]$ the actual observation period and $(a, b]$ the nominal observation period.

A key assumption of the case series method is that the actual observation period for each individual is independent of event times. Thus censoring of observation is independent of the individual's event history, so that the subsequent history is missing completely at random: we call this *censoring completely at random* (CCAR). There are many circumstances in which this assumption is violated. The most extreme violation is when the event is death: then the event time is equal to c. In

less extreme situations, the event may increase mortality in the short term - as with myocardial infarction (MI) and stroke, for example. Thus, some individuals who become cases may die some time after the event, prior to the end of the nominal observation period, as a result of the event having occurred. For these cases the individual's actual observation period is event-dependent. The subsequent history is missing at random, but not completely at random: we call this *censoring at random* (CAR). Clearly, it is desirable to weaken the CCAR assumption so that event-dependent observation periods may be accommodated.

One approach to handling event-dependent censoring is to impute the missing post-event exposures. This may be done when there is good information on individual-specific exposure histories. This was the approach taken by Farrington and Whitaker (2006) in an application to MI, in which missing infections were imputed for individuals whose follow-up was curtailed shortly after an MI (the case series method was found to be robust to failure of the CCAR assumption in this instance). In most circumstances, however, it is not possible to impute exposure histories reliably. Recently, a new self-controlled case series method has been developed which avoids any imputation, and may be used provided that the risk period associated with the exposure is not indefinite (Farrington, Whitaker and Hocine 2009). This method works even when the event of interest is death. However, if the event merely increases the risk of death, this method is likely to be inefficient, as all post-event exposures are ignored. In this paper we propose a third method, in which conditioning on the actual observation period is made explicit. This method is closely related to the conditional Poisson modelling methods of Rathouz (2004) and Roy, Alderson, Hogan and Tachima (2006). However, unlike these methods, it cannot be assumed that data on the whole cohort are available.

In section 2 we derive the likelihood for the new model, which explicitly involves

the time interval between the event and end of the actual observation period. In section 3 we discuss inference about the parameters of interest. Section 4 describes the implementation of semi-parametric and parametric models using standard software. In section 5 we describe an application to stroke, in which event-dependent censoring is likely to be a problem. We present a simulation study and a detailed application. Section 6 contains some final remarks.

2 A MODIFIED CASE SERIES METHOD

Suppose that individuals, indexed by i, experience exposures $x_i(t)$ at age t and events which arise according to a non-homogeneous Poisson process of rate $\lambda_i(t|x_i(t))$. We suppose that observation starts at age a_i and may last to the pre-determined age b_i. We assume that the event rate is low, so that recurrences may be ignored (see Farrington and Whitaker 2006 for details of the approximation involved if in fact recurrences cannot occur).

Observation is censored by a process of hazard $\mu_i(t|h(t))$, where $h(t)$ denotes the event history to time t. It is assumed that $\mu_i(t|h(t))$ does not depend on the exposure process $x_i(t)$. If there are no events before t, then $\mu_i(t|h(t)) \equiv \mu_i(t)$. If there is an event at $s < t$ then $\mu_i(t|h(t)) \equiv \mu_i(t|s)$. The history $h(t)$ introduces possible dependence of the censoring process on the event history.

Suppose that individual i experiences $n_i = 0$ or 1 event, at time t_i (if the event occurs) in the actual observation period $(a_i, b_i \wedge c_i]$. The likelihood for individual i,

conditional on observation starting at a_i, is

$$\begin{aligned} L_i(t_i, c_i) &= \{\lambda_i(t_i|x_i(t_i))\}^{I(n_i=1)} \exp - \int_{a_i}^{b_i \wedge c_i} \lambda_i(s|x_i(s))\mathrm{d}s \\ &\quad \times \{\mu_i(c_i|h_i(c_i))\}^{I(c_i<b_i)} \exp - \int_{a_i}^{b_i \wedge c_i} \mu_i(s|h_i(s))\mathrm{d}s \end{aligned}$$

where $I(P) = 1$ if P is true and 0 if P is false. We require that the exposure process $x_i(t)$ is exogenous: see Farrington *et al* (2009) for a discussion of this key assumption.

Now condition on the value of n_i and on the actual observation period. The conditional likelihood is

$$L_i^c(t_i, c_i) = \frac{L_i(t_i, c_i)}{\int_{a_i}^{b_i \wedge c_i} L_i(t, c_i)\mathrm{d}t}.$$

If $n_i = 0$ then

$$L_i(t_i, c_i) = \exp - \int_{a_i}^{b_i \wedge c_i} \lambda_i(s|x_i(s))\mathrm{d}s \times \{\mu_i(c_i)\}^{I(c_i<b_i)} \exp - \int_{a_i}^{b_i \wedge c_i} \mu_i(s)\mathrm{d}s,$$

which does not depend on t_i. Hence the conditional likelihood $L_i^c(t_i, c_i)$ is the constant $(b_i \wedge c_i - a_i)^{-1}$. Thus, as with the standard case series method, non-cases need not be sampled. If $n_i = 1$ then

$$\begin{aligned} L_i(t_i, c_i) &= \lambda_i(t_i|x_i(t_i)) \exp - \int_{a_i}^{b_i \wedge c_i} \lambda_i(s|x_i(s))\mathrm{d}s \times \{\mu_i(c_i|t_i)\}^{I(c_i<b_i)} \\ &\quad \times \exp - \left\{ \int_{a_i}^{t_i} \mu_i(s)ds + \int_{t_i}^{b_i \wedge c_i} \mu_i(s|t_i)\mathrm{d}s \right\} \end{aligned}$$

and the conditional likelihood is

$$L_i^c(t_i, c_i) = \frac{\lambda_i(t_i|x_i(t_i)) \times \{\mu_i(c_i|t_i)\}^{I(c_i<b_i)} \exp - \left\{ \int_{a_i}^{t_i} \mu_i(s)\mathrm{d}s + \int_{t_i}^{b_i \wedge c_i} \mu_i(s|t_i)\mathrm{d}s \right\}}{\int_{a_i}^{b_i \wedge c_i} \lambda_i(t|x_i(t)) \times \{\mu_i(c_i|t)\}^{I(c_i<b_i)} \exp - \left\{ \int_{a_i}^{t} \mu_i(s)\mathrm{d}s + \int_{t}^{b_i \wedge c_i} \mu_i(s|t)\mathrm{d}s \right\} \mathrm{d}t}. \quad (1)$$

The integrals $\int_{a_i}^{t} \mu_i(s)\mathrm{d}s$ in (1) involve the censoring hazard $\mu_i(s)$ before events arise. They cannot be evaluated from cases alone: information on the entire cohort, including non-cases, is required. To get round this problem, and thus remain within the case series framework, we set

$$\lambda_i^*(t|x_i(t)) = \lambda_i(t|x_i(t)) \times \exp - \int_0^t \mu_i(s)\mathrm{d}s. \tag{2}$$

Since the terms $\exp - \int_0^{a_i} \mu_i(s)ds$ cancel out, we can rewrite (1) as

$$L_i^c(t_i, c_i) = \frac{\lambda_i^*(t_i|x_i(t_i)) \times \{\mu_i(c_i|t_i)\}^{I(c_i<b_i)} \exp - \int_{t_i}^{b_i \wedge c_i} \mu_i(s|t_i)\mathrm{d}s}{\int_{a_i}^{b_i \wedge c_i} \lambda_i^*(t|x_i(t)) \times \{\mu_i(c_i|t)\}^{I(c_i<b_i)} \left\{\exp - \int_t^{b_i \wedge c_i} \mu_i(s|t)\mathrm{d}s\right\} \mathrm{d}t}$$

or alternatively as

$$L_i^c(t_i, c_i) = \frac{\lambda_i^*(t_i|x_i(t_i)) \times w_i(c_i; t_i)}{\int_{a_i}^{b_i \wedge c_i} \lambda_i^*(t|x_i(t)) \times w_i(c_i; t)\mathrm{d}t} \tag{3}$$

where

$$w_i(c; t) = \{\mu_i(c|t)\}^{I(c<b_i)} \left\{\exp - \int_t^{b_i \wedge c} \mu_i(s|t)\mathrm{d}s\right\} \tag{4}$$

is the density of the censoring time c conditional on an event at t, defined on the truncated support $(t, b_i]$, with $w_i(b; t) = \exp - \int_t^b \mu_i(s|t)\mathrm{d}s$, the survivor function at b conditional on an event at t. The rate $\lambda_i^*(t_i|x_i(t_i))$ is now no longer the individual event rate conditional on survival to t_i, but incorporates the thinning effect of censoring: it is truly an incidence, rather than a hazard. Expression (3) has the form of a weighted case series likelihood. Note that if all censoring is caused by events, then $\mu_i(s) = 0$ and hence $\lambda_i^* \equiv \lambda_i$.

3 MODELLING EVENT-DEPENDENT OBSERVATION PERIODS

In this section we discuss how the modified case series model described above may be used to provide unbiased estimates of the exposure effect.

3.1 Exposure and age effects

In the standard case series model, the event rate $\lambda_i(t|x_i(t))$ is assumed to take the following multiplicative form:

$$\lambda_i(t|x_i(t)) = \exp(\gamma_i + \beta^T x_i(t) + \alpha(t)).$$

The parameter γ_i represents an individual effect, and factors out of the conditional likelihood (1). The parameter β represents the exposure effect; this is the focus of inference. The parameter $\alpha(t)$ represents the (relative) effect of age, and is assumed common to all individuals. Suppose first that the censoring hazard $\mu_i(t)$ is also common to all individuals, with $\mu_i(t) = \mu(t)$. Then the rate (2) may be written

$$\lambda_i^*(t|x_i(t)) = \exp(\gamma_i + \beta^T x_i(t) + \alpha^*(t)), \qquad \alpha^*(t) = \alpha(t) - \int_0^t \mu(s)\mathrm{d}s.$$

Thus the interpretation of the exposure effect β is the same as before, and only the (relative) age effect is affected. In practice, of course, it is likely that the censoring hazard may vary between individuals. This may be allowed for by stratifying the age effects according to relevant subgroups defined by fixed covariates y_i, so that

$$\lambda_i^*(t|x_i(t), y_i) = \exp(\gamma_i + \beta^T x_i(t) + \alpha^*(t; y_i)).$$

Generally, it is sensible to stratify in this way by any fixed covariates that have been used to model the density of censoring times c_i. As with the standard case

series model, we can also study effect modification of the exposure effect by fixed covariates y_i by inclusion of interactions $x_i(t) \times y_i$.

3.2 ESTIMATION

The major novelty in the modified case series model is that it involves the weights $w_i(c,t)$ specified in equation (4). They are obtained from the densities of the censoring times for each individual i. We make the key (but reasonable) assumption that these densities do not involve the parameters in the rate function $\lambda_i^*(t|x_i(t))$. In consequence, in a first step the weights $w_i(c;t)$ may be estimated from data on ages at event and censoring, and these estimates are then plugged in to (3). This estimation procedure yields consistent estimates (Robins, Rotnizky and Zhao 1995, Rathouz 2004, Roy *et al* 2006).

Note that if the event is death, then the weights $w_i(c;t) = \delta(c-t)$, the Dirac delta function. It follows that the likelihood (3) is degenerate and equal to 1. In this situation, conditioning on censoring times is the same as conditioning on event times, and estimation becomes impossible. In such circumstances the methods of Farrington *et al* (2009) must be used.

3.3 STANDARD ERRORS

Let $\zeta = (\beta, \alpha)$ and let η denote the parameters in the censoring model (4); the parameters in ζ and η are assumed to be distinct. Let $U_i = \partial \log L_i^c / \partial \zeta$ and $S_i = \partial \log w_i / \partial \eta$, evaluated at the mles, where L_i^c and w_i are defined in (3) and (4), respectively. It may be verified directly that $E_i(\partial \log L_i^c / \partial \eta) = -E_i(U_i S_i^T)$, where E_i denotes expectation with respect to the event time density in the i^{th} case, conditional on the actual observation period. The calculation of standard errors proceeds as described in (Robins et al 1995, Rathouz 2004, Roy *et al* (2006), with expected

information replaced by observed information to allow for different follow-up times between individuals. The variance of $\widehat{\zeta}$ is consistently estimated by

$$\widehat{V}(\widehat{\zeta}) = \Gamma^{-1}(\Sigma_i \widetilde{U}_i \widetilde{U}_i^T)\Gamma^{-1}, \tag{5}$$

where $\Gamma = \Sigma_i \partial^2 \log L_i^c / \partial\zeta\partial\zeta^T$ is the observed information for ζ, and

$$\widetilde{U}_i = U_i - (\Sigma_j U_j S_j^T)(\Sigma_j S_j S_j^T)S_i. \tag{6}$$

Since $U = \Sigma_i U_i$ and $S = \Sigma_i S_i$ are likelihood scores, the asymptotic variance may be written

$$V(\widehat{\zeta}) = V_0 - V_0 D W D V_0$$

where V_0 is the variance of $\widehat{\zeta}$ obtained by maximising (3) with the w_i evaluated at the mle $\widehat{\eta}$, regarded as known. W is the asymptotic variance of $\widehat{\eta}$, and D is the asymptotic covariance of U and S. As noted by Rathouz (2004), estimating η can reduce the variance of $\widehat{\zeta}$, even when η is known. Thus the standard errors obtained from the case series model with estimated weights $w_i(c_i, t_i)$ are conservative. As shown in the next section, this model may be fitted using standard log-linear modelling software. This greatly simplifies the fitting procedure. Accordingly, it is of interest to evaluate the resulting loss of efficiency. In practice, interest usually focuses on β rather than ζ. Let $r = \dim(\beta)$ and $q = \dim(\eta)$, and write

$$V_0 = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}, \quad D = \begin{pmatrix} A \\ B \end{pmatrix}$$

where V_{11} is $r \times r$, and A is $r \times q$. Then

$$V(\widehat{\beta}) = V_{11} - \left\{ V_{11} A W A V_{11}^T + V_{12} B W A^T V_{11}^T + V_{11} A W B^T V_{12}^T + V_{12} B W B^T V_{12}^T \right\}.$$

Thus the loss of efficiency in estimation of $\widehat{\beta}$ is small when $nV_{12} = o_p(1)$ and $n^{-1}A = o_p(1)$. This arises when there is little correlation between exposure status at event

and age at event, and between exposure status at event and age at end of observation. This arises when exposure is not confounded with age. Thus, in the absence of such confounding, there is likely to be little loss in efficiency by using V_{11} to estimate the variance of $\widehat{\beta}$. Simulation studies reported below suggest that the loss of efficiency is small more generally.

4 IMPLEMENTATION

The fitting of case series models is described in Whitaker, Farrington, Spiessens and Musonda (2006). If the age effect $\alpha^*(t)$ is to be modelled semiparametrically (Farrington & Whitaker 2006), with jumps only at the event times, then the integral in the denominator of (3) becomes the sum

$$\sum_{j=1}^{m_i} \lambda_i^*(s_j|x_i(s_j)) \times w_i(c_i; s_j)$$

where $s_1, ..., s_{m_i}$ are the distinct event times in $(a_i, b_i \wedge c_i]$. The model can then be fitted using standard log-linear modelling software, with the quantities $\log(w_i(c_i; s_j))$ as offsets in the loglinear model.

However, for large datasets, the semiparametric model may be too computationally demanding to fit. In this situation the age effect $\alpha^*(t)$ may be modelled parametrically as constant on intervals. Thus, $\lambda_i^*(t|x_i(t))$ takes constant values, say λ_{ij}^*, on intervals $(a_{ij}, a_{ij+1}]$, $j = 0, ..., k_i - 1$ with $a_{i0} = a_i$ and $a_{ik_i} = b_i \wedge c_i$. Suppose that the event age t_i occurs in interval s_i, that is, $(a_{is_i}, a_{is_i+1}]$. Since the weights $w_i(c, t)$ do not depend on the parameters in λ_{ij}^*, the likelihood contribution (3) of case i may be replaced by

$$L_i'^c = \frac{\lambda_{is_i}^* W_{is_i}(c_i)}{\sum_{j=0}^{k_i-1} \lambda_{ij}^* W_{ij}(c_i)} \tag{7}$$

where

$$W_{ij}(c) = \int_{a_{ij}}^{a_{ij+1}} w_i(c;t)\mathrm{d}t. \tag{8}$$

Note that expression (7) is similar to that of the standard parametric case series model, with the interval lengths e_{ij} replaced by $W_{ij}(c_i)$. The standard model is obtained if we set $w_i(c;t) = 1$. The integral in (8) may need to be evaluated numerically. The log-likelihood $\Sigma_i \log L_i'^c$ can be maximised using a standard log-linear model, in exactly the same way as the standard case series model, the only difference being that the offsets are the terms $\log W_{ij}(c_i \wedge b_i)$.

5 APPLICATION AND SIMULATIONS

5.1 Antipsychotics and stroke

We reanalyse data from a standard case series analysis of antipsychotics and stroke (Douglas and Smeeth, 2008). The aim of the study was to investigate whether taking an antipsychotic drug alters the risk of stroke. Briefly, data on 6790 patients with at least one prescription for an antipsychotic drug and a recorded incident stroke between January 1988 and the end of 2002 were obtained from the UK General Practice Research Database.

The authors carried out a self-controlled case series analysis of these data. The observation periods started at the earliest of 12 months after the start of the GPRD record within this time interval and first antipsychotic prescription, and ended at the earliest of end of the GPRD record and end of 2002. Patients were considered to be at risk while they were taking antipsychotics; each risk period, the length of which varied between patients and prescriptions, was followed by up to five contiguous 35-day washout periods. Age effects were controlled in five-year bands. The authors

found a raised relative incidence in the risk period, $RI = 1.73$, 95% CI $(1.60, 1.87)$, declining gradually towards unity over the subsequent washout periods. The relative risk was higher, $RI = 2.32$ 95% CI $(1.73, 3.10)$, for patients taking atypical antipsychotics, and in patients with dementia, $RI = 3.50$, 95% CI $(2.97, 4.12)$.

Figure 1 shows a plot of the intervals from stroke to end of observation. There is a large peak for intervals of one year or less, followed by a long tail; 1571 of the GPRD records were current, that is, the GPRD record had not ended by the end of observation.

Accurate information on reasons for termination of the GPRD record or precise date of death was not always available. The large peak in Figure 1 suggests that some strokes may have cut short the subsequent observation period, as might be expected since stroke substantially increases short-term mortality. The observation periods are thus likely to be event-dependent, violating the CCAR assumption.

5.2 Simulations

We undertook simulations based on these data to evaluate the likely impact of event-dependent observation periods on the estimated association between antipsychotics and stroke, and to evaluate the adjustment method described above. The simulation scheme, detailed results, and full discussion are in the Supplemental Material for this paper.

Briefly, the simulations were based on the stroke data, with stroke times generated conditionally on the actual exposures and observation periods, assuming exponentially distributed times from stroke to termination of the GPRD record with mean ranging from 50 to 200 days.

The simulations show that the standard case series method can produce substantially biased results when observation periods are subject to event-dependent

censoring. For example, when the true value of the relative incidence is $\exp(0) = 1$ then the bias is positive and the estimated relative incidence is typically about 1.3. The bias is likely to be less for relative risks slightly less than 5. The simulations show good performance for the proposed adjusted method, both when τ is known and when it is estimated. The loss in efficiency from using the loglinear modelling framework (in which, in effect, τ is fixed at its estimated value) is negligible.

5.3 Modelling survival

In view of the potential for bias if the standard case series model is used, we apply our method to these data. The first step is to model the density $f(c|t, x)$ of age $C = c$ at the end of the GPRD record, conditionally on the occurrence of stroke at age $T = t$ and individual characteristics x. It is natural to think of the density of $C \geq t$ as a mixture of two components. The first component may be expressed as the density of the intervals $C - t$ and describes the short-term impact on duration of the GPRD record of a stroke at age t. The second component reflects both the underlying process by which GPRD records terminate, and the possible longer term impact on this process of a stroke at t. Thus, a natural model for the density f is of the form:

$$f(c|t, x) = \pi(t, x)g(c - t) + (1 - \pi(t, x))h(c|t, x) \tag{9}$$

where g is a density on $(0, \infty)$ and h is a density on (t, ∞). Model (9) can further be specialised according to whether the second component models C or $C - t$. In the first instance, we have

$$h(c|t, x) \equiv \frac{h'(c|x)}{\int_t^\infty h'(u|x)du}$$

and in the second,

$$h(c|t, x) \equiv h'(c - t|t, x)$$

for some density h'. We refer to the first form as the age model and the second as the interval model.

We chose $g(z) = \rho^{-1}\exp(-z/\rho)$ to be an exponential density with (small) mean ρ. For h we tried both Weibull and Gamma densities, expressed in both age and interval forms. This yields four distinct mixture models. Let i denote an individual with age at stroke t_i, gender x_{1i} (0 = male, 1 = female), prior dementia x_{2i} (0= absent, 1 = present), age at start of observation a_i, age at end of study b_i, and age at end of GPRD record c_i. The four models were as follows.

5.3.1 Exponential - Weibull (age) mixture model (EWA).

The likelihood contribution for an individual i is

$$L_i^s(c_i|t_i,a_i,x_i) = f(c_i|t_i,a_i,x_i)^{I(c_i<b_i)} \times P(C>b_i|t_i,a_i,x_i)^{1-I(c_i<b_i)} \qquad (10)$$

where

$$\begin{aligned} f(c|t,a,x) &= \frac{\pi(t,x)}{\rho(x)}e^{-(c-t)/\rho(x)} + (1-\pi(t,x))\frac{\nu(a,x)}{\mu(a,x)}\left(\frac{c}{\mu(a,x)}\right)^{\nu(a,x)-1} \\ &\quad \times \exp-\left\{\left(\frac{c}{\mu(a,x)}\right)^{\nu(a,x)} - \left(\frac{t}{\mu(a,x)}\right)^{\nu(a,x)}\right\}, \\ P(C>b|t,a,x) &= \pi(t,x)e^{-(b-t)/\rho(x)} + (1-\pi(t,x)) \\ &\quad \times \exp-\left\{\left(\frac{c}{\mu(a,x)}\right)^{\nu(a,x)} - \left(\frac{t}{\mu(a,x)}\right)^{\nu(a,x)}\right\}. \end{aligned}$$

The regression models for the parameters are:

$$\begin{aligned} \rho(x) &= \rho_x, \\ \text{logit}\{\pi(t,x)\} &= \gamma_x + \delta_x t, \\ \log\{\mu(a,x)\} &= \zeta_x + \eta_x a, \\ \log\{\nu(a,x)\} &= \theta_x + \xi_x a, \end{aligned} \qquad (11)$$

where the notation ρ_x, for example, indicates four different parameters according to the values taken by $x = (x_1, x_2)$.

5.3.2 Exponential - Weibull (interval) mixture model (EWI).

The likelihood contribution for an individual i is

$$L_i^s(c_i|t_i,x_i) = f(c_i|t_i,x_i)^{I(c_i<b_i)} \times P(C>b_i|t_i,x_i)^{1-I(c_i<b_i)} \tag{12}$$

where

$$\begin{aligned} f(c|t,x) &= \frac{\pi(t,x)}{\rho(x)} e^{-(c-t)/\rho(x)} + (1-\pi(t,x)) \frac{\nu(t,x)}{\mu(t,x)} \left(\frac{c-t}{\mu(t,x)}\right)^{\nu(t,x)-1} \\ &\quad \times \exp - \left\{ \left(\frac{c-t}{\mu(t,x)}\right)^{\nu(t,x)} \right\}, \\ P(C>b|t,x) &= \pi(t,x) e^{-(b-t)/\rho(x)} + (1-\pi(t,x)) \\ &\quad \times \exp - \left\{ \left(\frac{c-t}{\mu(t,x)}\right)^{\nu(t,x)} \right\}. \end{aligned}$$

The regression models for the parameters are

$$\begin{aligned} \rho(x) &= \rho_x \\ \text{logit}\{\pi(t,x)\} &= \gamma_x + \delta_x t, \\ \log\{\mu(t,x)\} &= \zeta_x + \eta_x t, \\ \log\{\nu(t,x)\} &= \theta_x + \xi_x t. \end{aligned} \tag{13}$$

5.3.3 Exponential - Gamma (age) mixture model (EGA).

The likelihood contribution for an individual i has the same form as (10), with

$$\begin{aligned} f(c|t,a,x) &= \frac{\pi(t,x)}{\rho(x)} e^{-(c-t)/\rho(x)} + (1-\pi(t,x)) \frac{1}{\Gamma(\nu(a,x))} \frac{1}{\mu(a,x)} \\ &\quad \times \left(\frac{c}{\mu(a,x)}\right)^{\nu(a,x)-1} \exp\left(-\frac{\nu(a,x)\,c}{\mu(a,x)}\right) \; S(t|a,x)^{-1}, \\ P(C>b|t,a,x) &= \pi(t,x) e^{-(b-t)/\rho(x)} + (1-\pi(t,x)) \; S(b|a,x) \; S(t|a,x)^{-1} \end{aligned}$$

where

$$S(t|a,x) = \int_t^\infty \frac{1}{\Gamma(\nu(a,x))} \frac{1}{\mu(a,x)} \left(\frac{z}{\mu(a,x)}\right)^{\nu(a,x)-1} \exp\left(-\frac{\nu(a,x)\,z}{\mu(a,x)}\right) \mathrm{d}z.$$

The regression models for the parameters are as in (11).

5.3.4 Exponential - Gamma (interval) mixture model (EGI)

The likelihood contribution for an individual i has the same form as (12), with

$$\begin{aligned} f(c|t,x) &= \frac{\pi(t,x)}{\rho(x)} e^{-(c-t)/\rho(x)} + (1-\pi(t,x)) \frac{1}{\Gamma(\nu(t,x))} \frac{1}{\mu(t,x)} \\ &\quad \times \left(\frac{c-t}{\mu(t,x)} \right)^{\nu(t,x)-1} \exp\left(-\frac{\nu(t,x)\,(c-t)}{\mu(t,x)} \right), \\ P(C > b|t,a,x) &= \pi(t,x) e^{-(b-t)/\rho(x)} + (1-\pi(t,x))\; S(b-t|t,x) \end{aligned}$$

where

$$S(b-t|t,x) = \int_{b-t}^{\infty} \frac{1}{\Gamma(\nu(t,x))} \frac{1}{\mu(t,x)} \left(\frac{z}{\mu(t,x)} \right)^{\nu(t,x)-1} \exp\left(-\frac{\nu(t,x)\,z}{\mu(t,x)} \right) \mathrm{d}z.$$

The regression models are as in (13).

5.4 ESTIMATING THE WEIGHTS

Gender had little effect on the exponential mean ρ, so only the effect of prior dementia was retained on this parameter. The maximised loglikelihoods and AIC for the four models, each with 26 parameters, are shown in Table 1. The lowest AIC is achieved by the exponential - Weibull (age) model EWA.

If the model is correct, then for each t the cumulative hazards $H(x|t)$, $x > t$ constitute a sample from a unit exponential. So the empirical survival function for the fitted cumulative hazards should approximate that of a unit exponential. Figure 2 shows the log of the empirical survival function of the fitted cumulative hazards, together with a plot of unstandardized residuals (interval from stroke to end of observation minus expected value) against expected values, both obtained using the EWA model. The plots suggest a satisfactory model fit, the boundaries in the residual plot resulting from the non-negativity and boundedness of the observed intervals.

Figure 3 shows some of the features of the model: in particular, the mixing probability $\pi(t, x)$ increases with age at stroke t. This suggests that the risk of dying of a stroke (the most obvious reason for censoring of the GPRD record) increases with age.

The weights $W_{ij}(c_i)$ were estimated using all four models described above, by numerical integration with respect to age at stroke t as per equation (8).

5.5 FITTING THE MODIFIED CASE SERIES MODEL

We fitted model (7) to the stroke data using the same risk periods and age groups as used by Douglas and Smeeth (2008), with the values $\log(W_{ij}(c_i))$ described in the previous subsection using model EWA, and also using the standard case series model. The latter corresponds to setting the W_{ij} equal to the interval lengths.

Figure 4 contrasts the age effects obtained with the two versions of the model. This illustrates the different interpretations of the age effects in the two models. In the standard model, the age effect corresponds to the relative age-specific incidence, conditional on remaining under observation within the GPRD. In the new model, the age effect is not conditional on remaining under observation.

Further investigations suggested that, in view of the very marked increase in stroke incidence with age, it was advisable to use narrower age groups, at least where data were abundant. We therefore used 45 age bands of differing lengths, including 1-year age bands in the range 58 - 96 years where most strokes occurred. The results for the four models EWA, EWI, EGA and EGI, along with those for the standard case series model, are shown in Table 2. In all models, underlying age effects have been stratified according to gender, prior dementia, and type of antipsychotics used (such stratification improves model fit but has little impact on the estimates of relative incidence).

Table 2 shows that, in this example, ignoring event-dependent observation periods results in an upward bias in the relative incidences. The choice of model for the weights has little impact on the estimates. The confidence intervals reported in Table 2 are calculated from the loglinear model.

As in the published analysis, we found a big interaction with prior dementia, the relative incidence being much higher in persons with prior dementia, but no interaction with antipsychotic drug type. Table 3 shows the results for the exposed risk period for the new and standard models. Once again, the results are not unduly sensitive to the choice of model for the age at end of GPRD record. All four adjusted models give similar results. These differ from those obtained with the standard case series model in one key respect: whereas the standard model indicated an increased relative incidence associated with the use of antipsychoptics of any type in patients without dementia, the analyses adjusted for the event-dependent censoring of the observation periods suggest that there is no increased risk in patients without dementia.

In models EWA and EGA, the dependence between age at end of observation and age at stroke occurs explicitly only via the first component of the mixture model. Setting the parameters $\pi(t, x)$ to zero eliminates this dependence. Thus, for these two models, the validity of the second component can be checked by dropping the first, since models with weights calculated from the second component only should give the same results as the uncorrected case series model. We checked the validity of the age models EWA and EGA by recalculating the weights W_{ij} with the parameters $\pi(t, x)$ set to zero, all other parameters in the densities of age at censoring retaining their previously estimated values. The results are shown in Table 4, along with the results for the standard case series model (reproduced from Table 3). The maximised loglikelihoods for the three models are similar, as

are the parameter estimates. This suggests that both the Weibull and the gamma distributions provide adequate descriptions of the distribution of age at end of GPRD record in individuals who do not die of stroke. The contrast between these results for models EWA and EGA and those shown in Table 3 provides a direct demonstration of the impact of stroke-induced censoring of observation periods on the parameter estimates.

Further analyses (not presented) were undertaken in which the exponential stroke-associated component was modelled as Weibull or gamma. For each model the estimated shape parameter was close to 1 (corresponding to the exponential density) and the parameter estimates were virtually identical to those obtained with an exponential component.

5.6 Interpretation

These analyses suggest that there is little evidence of an increased risk of stroke after taking antipsychotic drugs for persons without dementia, irrespective of the type of antipsychotic. For persons with dementia, however, the relative incidence is increased during the prescription period by about three-fold for both typical and atypical antipsychotics, during the prescription period. Thereafter the relative incidence falls to 1 gradually over the five washout periods for patients with dementia (results not shown). For patients without dementia there is a suggestion of a reduction in stroke risk during the five washout periods (results not shown). Whether this decrease is causally related to taking antipsychotics or coming off them, is artefactual, or is a chance finding, remains unclear. It could, for example, reflect an increased propensity to use antipsychotics following a stroke, in which case the relative incidences associated with exposure and washout periods will have been underestimated.

The model used for the censoring process was a mixture model depending on age at start of observation, age, gender, and dementia status. The first component described the short-term impact of stroke on censoring; the second the combined effect of the longer term impact of stroke and the underlying censoring process. Results were insensitive to this second component. In two of the models, the validity of the second component was tested by dropping the first, and demonstrating that the results of the standard analysis are thereby retrieved. This suggests that any bias in the standard analysis is due only to the short-term censoring induced by stroke. The validity of the first component was tested by embedding it within broader model classes, with no difference in results.

The method makes the assumption that the censoring process does not depend on antipsychotic use. This does not appear unreasonable. However, no case series model - nor indeed any other model - is immune from unobserved time-varying confounders, through which exposure is correlated with a time-varying but unmeasured factor associated with stroke. If such a factor is responsible for the association between exposure to antipsychotics and stroke, it would act differentially in patients with and without dementia. It is also conceivable that a time-varying unmeasured counfounder may be associated with both exposure to antipsychotics and censoring. To impact on results, such a confounder would need to act differentially in patients with and without dementia. While the mean of the exponential component did vary slightly between patients with (0.091 years) and without (0.11 years) dementia (likelihood ratio test, $p = 0.077$ for model EWA), the associations with stroke were not sensitive to this difference. We conclude that it is unlikely that such effects could explain the marked contrast between associations with antipsychotic use observed in patients with and without dementia.

6 DISCUSSION

The standard self-controlled case series model requires that observation periods be censored completely at random (CCAR). In this paper, we have extended the method to allow for event-dependent censoring, or censoring at random (CAR) while remaining within a case series framework in which only individuals who have experienced the event of interest are sampled. The method works by conditioning explicitly on the age at censoring in cases, and involves weighting cases by the density of the time from event to censoring (or the corresponding survival probability in cases which are followed to the nominal end of observation). The approach is similar in spirit to that first proposed by Robins *et al* (1995) and applied in conditional log-linear models by Roy *et al* (2006), but with an additional trick to enable us to remain within the case series framework. Under this model, the exposure effect parameters β retain the same interpretation as under the standard case series model. The age effect, however, takes on a different interpretation: it now represents the combined effect of age-specific relative incidence and censoring. In most applications, the focus of the investigation is the exposure effect, and the age-specific relative incidence is a nuisance parameter. If censoring of observation periods is CAR but not CCAR, the age-specific relative incidence cannot in general be estimated from case series data alone, but requires information from non-cases. One important exception is when censoring for reasons unrelated to the event is rare; in this case the age-specific relative incidence retains its original interpretation.

The method is applicable provided that conditioning on censoring time does not render the event time determinate, as would occur if the event were death. In this situation, the method developed by Farrington *et al* (2009) can be used instead, provided the risk period is not indefinite. More generally, we might expect the efficiency of the method to decline as the variance of age at event, conditional on

censoring time, is reduced.

The method may be implemented in standard loglinear modelling software, with estimated offsets $\log\{W_{ij}(c_i)\}$ estimated in a separate modelling process. The standard errors obtained using this simple method are conservative, though our simulations suggest that such losses are likely to be small, at least as far as the exposure effect parameters β are concerned. As noted by Rathouz (2004), further loss of efficiency may also be expected owing to the fact that the censoring process is ancillary for β: the density (4) does not involve β, yet the conditional likelihood (3) depends on the censoring times c_i. In a different but related context, Rathouz (2004) found that a small increase in efficiency could be obtained by removing the β-ancillary information in the censoring process. Overall, however, the gains in efficiency over the simple log-linear model obtained by these more elaborate methods are likely to be small, especially for estimating the exposure effect.

This extension expands the field of possible applications of the self-controlled case series method, particularly in pharmacoepidemiology. One of the most important biases in observational drug safety studies - confounding by indication - is rendered largely irrelevant by the self-controlled aspect of the method. However, many serious potential adverse drug effects requiring evaluation increase the risk of short/medium term death, and therefore violate the standard case series assumptions. The present extension allows the case series method to be used reliably in such circumstances, provided other assumptions are not also violated. As in the standard case series model, confounders that act multiplicatively on the event rate $\lambda_i(t|x_i(t))$ factor out and hence may be ignored (at least as main effects). However, this is not true for confounders acting on the censoring process $\mu_i(t|h(t))$, as they can influence the estimation of the weights $w_i(c_i; t_i)$. Because of this limitation, it remains of interest to establish conditions under which the standard case series model remains robust

to event-dependent censoring, and to study the robustness of the present extension to mis-specification of the censoring process.

We have illustrated the method with an application to an important dataset concerning the effect of antipsychotic drugs on the risk of stroke. We have shown that ignoring event-dependent censoring of observation periods can produce substantial bias, though the main finding of the original paper still stands, namely that taking antipsychotics is associated with a substantially increased risk of stroke in patients with dementia. In this application, the time from event to study end or censoring was modelled by a two-component mixture in which the first component represented the effect of stroke and the second the underlying, stroke-independent, censoring process. This formulation provided a novel opportunity for internal model validation, in which the second component is checked by dropping the first. This mixture modelling approach and validation technique are likely to be applicable more widely.

7 SUPPLEMENTAL MATERIAL

Simulations: Details of the simulations briefly summarized in subsection 5.2, including simulation scheme, full results, and discussion (pdf).

8 REFERENCES

Douglas, I. J. and Smeeth, L. (2008), "Exposure to antipsychotics and risk of stroke: self controlled case series study," *British Medical Journal* , 337:a1227, doi:101136/bmj.a1227.

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)," *Journal of the Royal Statistical Society, Series C*, 55, 553-594.

Farrington, C. P., Whitaker, H. J. and Hocine, M. N. (2009), "Case series analysis for censored, perturbed or curtailed post-event exposures," *Biostatistics*, 10, 3 - 16.

Rathouz, P. J. (2004), "Fixed effects models for longitudinal binary data with drop-outs missing at random," *Statistica Sinica*, 14, 969 - 988.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90, 106 - 121.

Roy J., Alderson D., Hogan J. W. and Tachima, K. T. (2006) "Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women," *Journal of the American Statistical Association*, 101, 424 - 434.

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: 1768-1798.

Whitaker, H. J., Hocine, M. N. and Farrington, C. P (2009), "The methodology of self-controlled case series studies," *Statistical Methods in Medical Research*, 18, 7 - 26.

Tables

Table 1 Model fit

Model	EWA	EWI	EGA	EGI
Loglik	-7849.55	-7855.09	-7855.95	-7853.07
AIC	15751.10	15762.18	15763.90	15758.14

Table 2 Relative incidence of stroke by risk period, with 95% confidence interval

	Model				
Risk period	EWA	EWI	EGA	EGI	Standard
Exposed	1.29	1.28	1.28	1.28	1.54
	(1.19,1.40)	(1.18,1.39)	(1.19,1.39)	(1.19,1.39)	(1.42,1.67)
Washout 1	0.93	0.95	0.94	0.95	1.53
	(0.83,1.03)	(0.85,1.05)	(0.85,1.04)	(0.85,1.05)	(1.38,1.70)
Washout 2	0.89	0.92	0.91	0.91	1.48
	(0.77,1.03)	(0.79,1.06)	(0.78,1.05)	(0.79,1.05)	(1.28,1.70)
Washout 3	0.94	0.96	0.95	0.96	1.44
	(0.79,1.11)	(0.81,1.14)	(0.80,1.13)	(0.81,1.14)	(1.22,1.70)
Washout 4	0.75	0.76	0.76	0.76	1.11
	(0.60,0.92)	(0.62,0.94)	(0.61,0.93)	(0.62,0.94)	(0.90,1.36)
Washout 5	0.70	0.71	0.71	0.71	0.96
	(0.55,0.88)	(0.56,0.90)	(0.56,0.89)	(0.56,0.90)	(0.76,1.20)
Loglik	-15791.22	-15770.31	-15763.35	-15768.67	-16895.21

Note: Each model has 225 parameters.

Table 3 Relative incidence of stroke in risk period by dementia status and antipsychotic type, with 95% confidence interval

Anti-	Prior	Model				
psychotic	dementia	EWA	EWI	EGA	EGI	Standard
Typical	No	1.05	1.06	1.05	1.06	1.31
		(0.95,1.16)	(0.96,1.16)	(0.96,1.16)	(0.96,1.17)	(1.18,1.44)
Atypical	No	0.94	0.95	0.95	0.95	1.45
		(0.67,1.33)	(0.67,1.34)	(0.67,1.34)	(0.67,1.34)	(1.02,2.05)
Typical	Yes	3.10	2.84	2.87	2.87	2.91
		(2.57,3.73)	(2.36,3.41)	(2.39,3.44)	(2.39,3.45)	(2.42,3.50)
Atypical	Yes	2.80	2.74	2.76	2.75	3.40
		(1.44,5.41)	(1.44,5.25)	(1.45,5.29)	(1.43,5.26)	(1.79,6.46)
Loglik		-15711.79	-15697.64	-15689.16	-15695.41	-16840.45

Note: Each model has 255 parameters.

Table 4 Relative incidences (and 95% CI) without event-dependent censoring

Anti-	Prior	Model		
psychotic	dementia	EWA, $\pi \equiv 0$	EGA, $\pi \equiv 0$	Standard
Typical	No	1.29	1.30	1.31
		(1.17,1.43)	(1.17,1.43)	(1.18,1.44)
Atypical	No	1.42	1.42	1.45
		(1.00,2.01)	(1.00,2.01)	(1.02,2.05)
Typical	Yes	2.83	2.83	2.91
		(2.35,3.40)	(2.35,3.40)	(2.42,3.50)
Atypical	Yes	3.18	3.19	3.40
		(1.67,6.06)	(1.68,6.07)	(1.79,6.46)
Loglik		-16840.10	-16841.74	-16840.45

Figure captions

Figure 1: Intervals from stroke to end of observation (years), and censoring values (rug).

Figure 2: Goodness-of-fit of model EWA. Left: logarithm of the empirical survival function of the fitted cumulative hazards (dots) and a unit exponential (line). Right: residuals *versus* expected values (open circles are censored observations) of intervals from event to end of observation.

Figure 3: Features of the fitted EWA model: functional dependence of parameters $\pi(t, x)$, $\mu(a, x)$, $\nu(a, x)$ on age at stroke t, age at start of the GPRD record a, gender x_1 and presence of dementia x_2. Full lines: males without dementia; tight dashes: females without dementia; dots and dashes: males with dementia; loose dashes: females with dementia.

Figure 4: Logarithms of the age effects (5-year age groups) estimated using two case series models, relative to age 60 - 65 years. Full dots and lines: standard model; circles and dashes: model EWA.

Figure 1

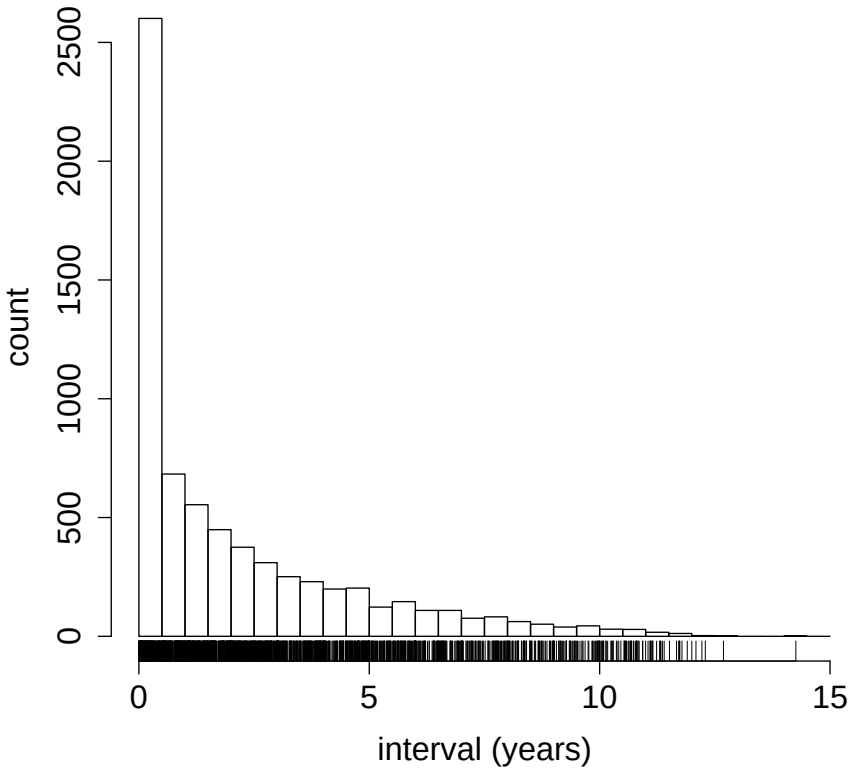


Figure 2

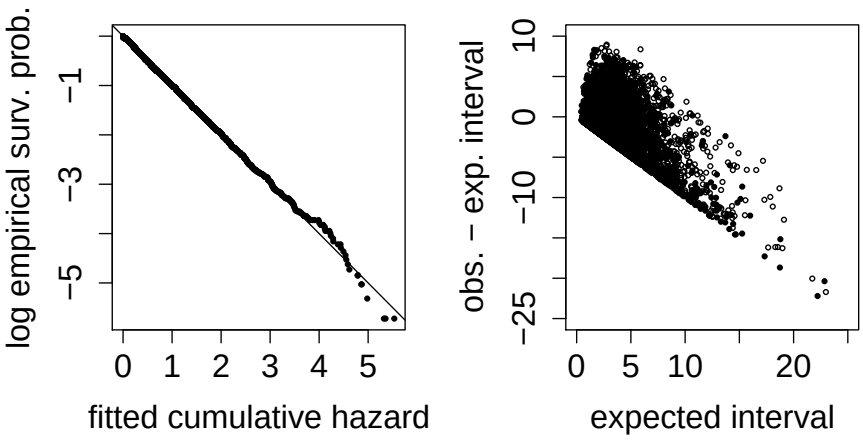

log empirical surv. prob.
−5
−3
−1
0 1 2 3 4 5
fitted cumulative hazard
obs. − exp. interval
−25
−10
0
10
0 5 10 20
expected interval

Figure 3

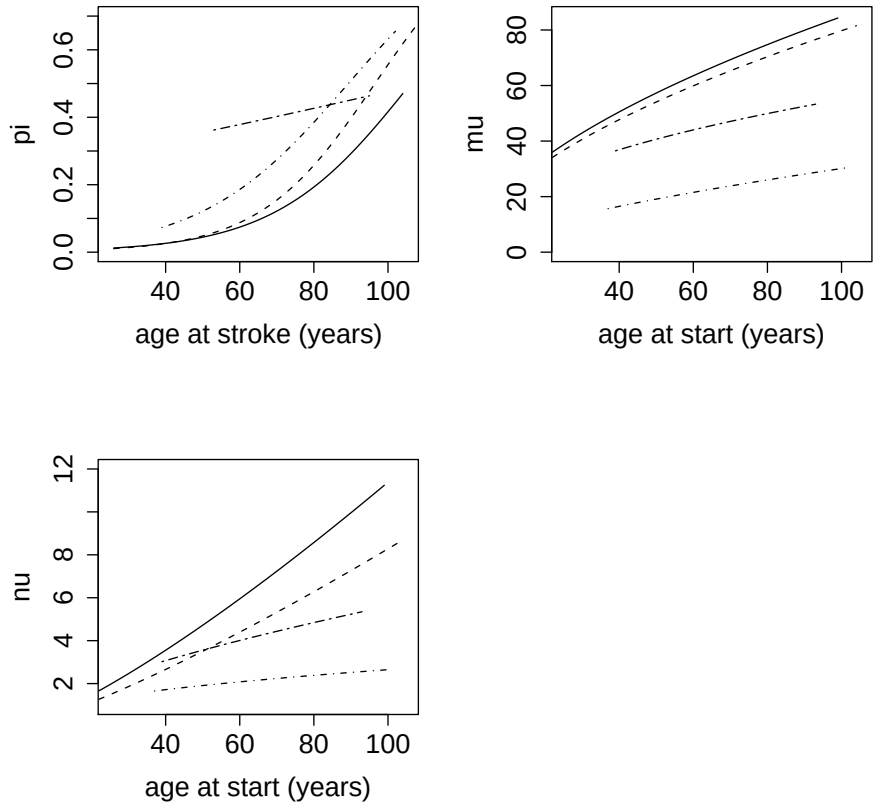
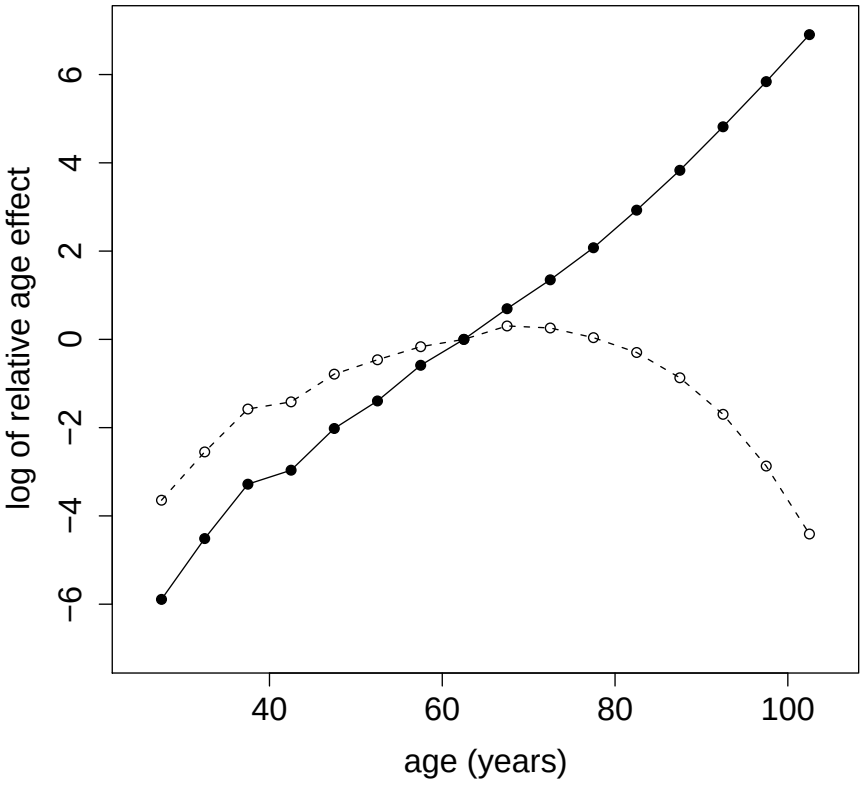

pi
0.0
0.2
0.4
0.6
40
60
80
100
age at stroke (years)
mu
0
20
40
60
80
40
60
80
100
age at start (years)
nu
2
4
6
8
12
40
60
80
100
age at start (years)

Figure 4



Supplemental Material

This Supplemental Material describes the simulations summarized in section 5.2 of the paper. It includes details of the simulation scheme used, full results, and discussion.

A The simulation scheme

The simulations were undertaken to evaluate the likely impact of event-dependent observation periods on the estimated association between antipsychotics and stroke in the dataset under consideration, and to evaluate the adjustment method described in the paper.

Accordingly, we used the observed exposure histories $x_i(t)$ and observation periods $(a_i, c_i^*]$ (where $c_i^* = b_i \wedge c_i$]) for each case i. For the purpose of the simulations, we assumed that time from stroke to termination of the GPRD record was exponentially distributed with mean τ^{-1}, for various values of τ. We obtained the $W_{ij}(c_i)$ from equation (8) of the paper. For an interval $(a_{ij}, a_{ij+1}]$,

$$\log(W_{ij}(c_i)) = \log(\tau) I(c_i \geq b_i) + \log\{e^{-(c_i^* - a_{ij+1})/\tau} - e^{-(c_i^* - a_{ij})/\tau}\}.$$

Note that since the term $\log(\tau) I(c_i \geq b_i)$ is constant for individual i, it cancels out in equation (7) of the paper, and hence it makes no difference in this case whether $c_i < b_i$ or $c_i \geq b_i$.

We chose a range of values for the relative incidence e^{β} associated with exposure, and fixed the relative incidences associated with each 5-year age group to values similar to those estimated from the data using the standard case series model. These values indicate an exponentially increasing risk with age, and are plotted in Figure 4

of the paper. We used the same risk periods as in the data, but no washout periods: the relative incidence in these was assumed to be 1.

We simulated event times t_i for each case, conditionally on the observed exposures and observation periods, and thus on the c_i^*. This conditioning is required because estimation with the case series method is conditional on these quantities. The simulation of the t_i was done in two steps. First, for each case, the age-exposure interval containing t_i was selected by multinomial sampling based on equation (7) of the paper. Second, the event time within that interval, $(a, b]$ say, was selected using the density $\tau^{-1}\exp(y/\tau)/\{\exp(b/\tau) - \exp(a/\tau)\}$. The resulting intervals $c_i^* - t_i$ are then approximately exponentially distributed with mean τ^{-1}. The approximation is valid when τ is small in relation to the observation period; the validity of the approximation was checked. Note that, in this sampling scheme, all observation periods are censored by the end of the GPRD record.

For each run, the data were analysed by three methods: (a) the standard case series method; (b) the new method with known τ; and (c) the new method with estimated $\widehat{\tau} = \Sigma_{i=1}^{6790}(c_i^* - t_i)/6790$.

One thousand runs were done for each combination of $\tau = 50, 100, 150, 200$ days and $\exp(\beta) = 1, 2, 5, 10$.

B Results

Tables A1 to A4 give the results for $\tau = 50, 100, 150, 200$ days, respectively. The summary statistics presented include the means of the 1000 values of $\hat{\beta}$ estimated using the three methods described in section A; the mean of the estimated standard errors of $\hat{\beta}$ obtained using the three methods; and the standard deviation of the 1000 values of $\hat{\beta}$.

C Discussion of the simulations

The simulations show that the standard case series method can produce substantially biased results when observation periods are subject to event-dependent censoring. For example, when the true value of the relative incidence is 1 then the bias is positive and the estimated relative incidence is typically about 1.3 for the values of τ considered. Both the size of the bias and its direction vary with τ and β. For a given value of τ, the bias becomes negative for sufficiently large values of β. For example, with $\tau = 50$ days, the estimated relative incidence is typically about 7.4 for a true value of 10.

More generally, the magnitude and direction of bias will depend on the dataset, as both are determined by the relative lengths of the risk and control periods curtailed by censoring. The value $\tau = 50$ days best mimics the stroke data. Our simulations with this τ suggest that relative incidences close to 1 will be biased upwards in a standard analysis, whereas relative incidences slightly lower than 5 will be roughly unbiased.

The adjusted case series method eliminates the bias effectively for all values of β and τ. Estimating τ does not introduce any further bias.

There is good agreement between the average standard errors and the standard deviations of the simulated $\hat{\beta}$, especially for the correctly specified models. Unsurprisingly, the standard errors are larger for the adjusted models than for the standard model, though the mean squared errors (not shown) are lower for the adjusted models in these data.

Theoretical considerations described in subsection 3.3 of the paper suggest that efficiency may be improved by estimating τ even when it is known. One might therefore expect to find the empirical standard deviation of $\hat{\beta}$ in the adjusted analysis with τ estimated to be less than mean standard error in the adjusted analysis with

τ fixed at its known value. There is some evidence of this, though the differences are small and not practically important. The mean standard errors estimated using the loglinear model with τ estimated and τ known differ only in the fourth decimal place. We conclude that the loglinear modelling framework is suitable for this application.

Table A1: Simulation results for $\tau = 50$ days

	$\beta : \log(RI)$			
Analysis	$0 : \log(1)$	$0.6931 : \log(2)$	$1.6094 : \log(5)$	$2.3026 : \log(10)$
Unadjusted				
mean $\hat{\beta}$	0.2579	0.8459	1.5433	1.9959
mean SE	0.0407	0.0367	0.0360	0.0361
std dev	0.0358	0.0331	0.0298	0.0276
Adjusted: true τ				
mean $\hat{\beta}$	0.0008	0.6935	1.6093	2.3061
mean SE	0.0407	0.0406	0.0431	0.0466
std dev	0.0410	0.0405	0.0426	0.0459
Adjusted: estimated τ				
mean $\hat{\beta}$	0.0036	0.6925	1.5838	2.2346
mean SE	0.0408	0.0406	0.0427	0.0457
std dev	0.0410	0.0402	0.0419	0.0430

RI: Relative incidence.

Table A2: Simulation results for $\tau = 100$ days

	$\beta : \log(RI)$			
Analysis	$0 : \log(1)$	$0.6931 : \log(2)$	$1.6094 : \log(5)$	$2.3026 : \log(10)$
Unadjusted				
mean $\hat{\beta}$	0.3287	0.9748	1.7635	2.2918
mean SE	0.0376	0.0362	0.0356	0.0359
std dev	0.0361	0.0348	0.0325	0.0296
Adjusted: true τ				
mean $\hat{\beta}$	0.0012	0.6939	1.6103	2.3025
mean SE	0.0385	0.0380	0.0396	0.0423
std dev	0.0378	0.0384	0.0402	0.0412
Adjusted: estimated τ				
mean $\hat{\beta}$	0.0013	0.6960	1.6126	2.2962
mean SE	0.0385	0.0381	0.0396	0.0422
std dev	0.0379	0.0384	0.0402	0.0408

RI: Relative incidence.

Table A3: Simulation results for $\tau = 150$ days

	$\beta : \log(RI)$			
Analysis	$0 : \log(1)$	$0.6931 : \log(2)$	$1.6094 : \log(5)$	$2.3026 : \log(10)$
Unadjusted				
mean $\hat{\beta}$	0.3158	0.9844	1.8169	2.3932
mean SE	0.0374	0.0359	0.0352	0.0356
std dev	0.0369	0.0363	0.0324	0.0317
Adjusted: true τ				
mean $\hat{\beta}$	0.0002	0.6937	1.6102	2.3019
mean SE	0.0378	0.0370	0.0378	0.0398
std dev	0.0375	0.0385	0.0373	0.0397
Adjusted: estimated τ				
mean $\hat{\beta}$	-0.0034	0.6927	1.6153	2.3108
mean SE	0.0378	0.0371	0.0380	0.0400
std dev	0.0376	0.0387	0.0376	0.0399

RI: Relative incidence.

Table A4: Simulation results for $\tau = 200$ days

	$\beta : \log(RI)$			
Analysis	$0 : \log(1)$	$0.6931 : \log(2)$	$1.6094 : \log(5)$	$2.3026 : \log(10)$
Unadjusted				
mean $\hat{\beta}$	0.2904	0.9712	1.8303	2.4412
mean SE	0.0374	0.0358	0.0349	0.0353
std dev	0.0380	0.0345	0.0321	0.0320
Adjusted: true τ				
mean $\hat{\beta}$	-0.0004	0.6932	1.6099	2.3048
mean SE	0.0374	0.0364	0.0367	0.0383
std dev	0.0381	0.0358	0.0353	0.0375
Adjusted: estimated τ				
mean $\hat{\beta}$	-0.0083	0.6878	1.6128	2.3174
mean SE	0.0375	0.0365	0.0369	0.0386
std dev	0.0382	0.0360	0.0358	0.0382

RI: Relative incidence.