000, 1–16
October 2012

Self-Controlled Case Series Method with Smooth Age Effect

Yonas Ghebremichael-Weldeselassie *, Heather J. Whitaker, and C. Paddy Farrington

Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK

**email:* y.g.weldeselassie@open.ac.uk

SUMMARY: The self-controlled case series method, commonly used to investigate potential associations between vaccines and adverse events, requires information on cases only and automatically controls all age-independent multiplicative confounders, while allowing for an age-dependent baseline incidence.

In the parametric version of the method, the age-specific relative incidence is modelled using a piecewise constant function, while in the semiparametric version it is left unspecified. However, mis-specification of age groups in the parametric version can lead to biased estimates of the vaccine effect, and the semiparametric approach runs into computational problems when the number of cases in the study is large. We thus propose to use a penalized likelihood approach where the age effect is modelled using splines.

We use a linear combination of cubic M-splines to approximate the age-specific relative incidence and integrated splines (I-splines) for the cumulative relative incidence. A simulation study was conducted to evaluate the performance of the new approach and its efficiency relative to the semiparametric approach. Results show that the new approach performs better and works well for large data sets. The new spline-based approach is applied to data on febrile convulsions and paediatric vaccines.

KEY WORDS: Case series; M-splines; Penalized likelihood; Poisson; Self controlled; Smoothing parameter.

1. Introduction

The self-controlled case series method (SCCS), or case series method in short, was first developed to estimate the relative incidence of an acute event following transient exposures (Farrington, 1995). It is an alternative to the traditional epidemiologic designs, such as cohort and case control methods. It was derived from a cohort method by conditioning on the total number of events an individual experiences during their observation period. The method requires information only on cases, namely individuals who experience the event of interest.

The self-controlled case series method compares incidence of an event in an exposure period, when an individual is hypothesized to be at a higher or lower risk of developing a disease, with incidence in the control periods, when an individual is not exposed. The incidence in the control period is the baseline incidence. Since the comparison is made within an individual's observation period, the method is self-matched, hence all measured and unmeasured age-independent confounding variables, such as socio-economic status, birth weight, location, severity of underlying disease, gender etc, that act multiplicatively on the baseline incidence, are automatically controlled. Time-varying confounders such as age and season, can be allowed for in the baseline incidence (Farrington and Whitaker, 2006).

Confounding by temporal factors is likely to occur when both the event incidence and the opportunity for exposure vary with age or season. Examples include adverse events and childhood vaccinations; seasonal exposures such as respiratory infections or influenza vaccination; and studies in elderly populations.

Careful control of age effects is particularly important in the study of vaccines and neurological events, such as febrile convulsions. The incidence of such events is highly age-dependent in the first two years of life, which is precisely the age at which many routine vaccinations take place. Partly for this reason, potential associations between vaccination and neurological events have been studied intensively over several decades. These studies have used broad range of methods, including SCCS (Farrington et al., 1995; Barlow et al., 2001; Huang et al., 2010; Miller et al., 1981)

In its original form, the case series model took the multiplicative effect of age on baseline incidence into account by dividing age into selected groups, the age-specific relative incidence function being represented by a piecewise constant step function. The exposure status was also categorical. We refer to this as the parametric version of the case series method. Its limitation is that it is sensitive to mis-specification of age groups which may lead to biased estimates of the association between exposure and event outcome (Farrington and Whitaker, 2006). Hence a semiparametric version in which the age-specific relative incidence function is left unspecified was proposed (Farrington and Whitaker, 2006). However, the number of parameters that must be estimated in the semiparametric model increases with the number of cases in the study, leading to computational problems for large datasets.

Given the limitations of the parametric and semiparametric methods, we propose to extend the method by replacing the piecewise constant age-specific relative incidence function with smooth functions, namely a linear combination of cubic M-splines, and maximize a penalized likelihood. We used M-splines because it is possible to avoid the integral in the case series likelihood by replacing it with integrated M-splines (I-splines). The age-specific cumulative relative incidence function is then approximated by a monotone spline function, a linear combination of I-splines (Ramsay, 1988). The methodology developed here is inspired by Joly et al. (1998), which we adapt for use with the SCCS method, and have programmed in R (R Development Core Team, 2012).

The paper is organized in six sections. Section 2 briefly describes how the likelihood function of an SCCS model is derived, followed by a description of M and I splines and their application to the SCCS model in Section 3. Section 4 describes the performance of the spline-based SCCS and compares it with the parametric and semiparametric versions of

SCCS using a simulation study. In Section 5 we apply the new spline-based SCCS method to a large dataset on febrile convulsions and paediatric vaccines, and in Section 6 we make some final remarks.

2. The case series likelihood

The likelihood function of a case series method may be derived from a cohort likelihood by conditioning on the number of events each individual experiences and on the exposure history over the time period an individual is observed. The assumptions made in deriving the likelihood are: (1) that individuals experience events in a non-homogenous Poisson process; (2) that age-dependent exposures experienced by individuals are exogenous, so exposures are independent of prior events, and (3) that censoring of individuals at the end of the observation period occurs completely at random, i.e the occurrence of the event of interest must not censor or affect the observation period (Farrington, 1995; Whitaker et al., 2006; Weldeselassie et al., 2011).

Let $(a_i,\, b_i]$ be the observation period for individual $i = 1,\, 2,...N$, often determined by a combination of calendar time and age constraints. Let $x_i(t)$ represent the vector of exposures that individual i experiences at age t within the observation period, and x_i^t the history of individual i up to age t. Let $x_i{\equiv}x_i^{b_i}$ be the exposure history of individual i up to the end of their observation period. The unconditional likelihood that individual i experiences n_i events that arise with intensity process $\lambda_i(t|x_i^t)$ at times t_{ij}, $j = 1, 2, ..., n_i$ is

$$L_i^u = \prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i^{t_{ij}}) \exp\left\{ -\int_{a_i}^{b_i} \lambda_i(t|x_i^t) dt \right\}. \tag{1}$$

Conditioning on the total number of events an individual i experiences in the their observation period and on the exposure history x_i, gives the conditional likelihood contribution,

$$L_i^c = \frac{\prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i)}{\left\{\int_{a_i}^{b_i} \lambda_i(t|x_i)dt\right\}^{n_i}}. \tag{2}$$

From this case series likelihood it can be seen that if individual i has no events, $n_i =$

0, then $L_i^c = 0$, implying that only individuals with at least one event in their observation period contribute to the likelihood. Hence, the case series method needs information only on cases. The most convenient way of parameterizing the incidence $\lambda_i(t|x_i)$ is according to the proportional incidence model

$$\begin{aligned}\lambda_i(t|x_i) &= \lambda_0(t)\exp\left\{\gamma_i + x_i(t)^T\beta\right\} \\ &= \varphi\psi(t)\exp\left\{\gamma_i + x_i(t)^T\beta\right\}, \end{aligned} \tag{3}$$

where $\lambda_0(t)$ is the baseline incidence at age t, φ is the underlying incidence at some reference age, $\psi(t)$ is the age-specific relative incidence function, γ_i is a sum of fixed and random individual effects, which might depend on fixed covariates over the period $(a_i, b_i]$, and β is the log-relative incidence that measures the association between exposures and event of interest. The main focus of inference is β. Then, combining (2) and (3), the conditional likelihood is

$$L = \prod_{i=1}^{N}\prod_{j=1}^{n_i} \frac{\psi(t_{ij})\exp\left\{x_i(t_{ij})^T\beta\right\}}{\int_{a_i}^{b_i}\psi(t)\exp\left\{x_i(t)^T\beta\right\}dt}. \tag{4}$$

The terms φ and $\exp(\gamma_i)$ cancel out and hence all fixed covariates that act multiplicatively on the baseline incidence are automatically controlled for.

3. Smooth age effect

To avoid the lack of robustness of parametric SCCS to mis-specification of age groups and the computational problems of semiparametric SCCS in large data sets we propose to represent the age-specific relative incidence $\psi(t)$ in (4) using smooth functions.

Following Joly et al. (1998) we use a linear combination of M-spline basis functions to approximate $\psi(t)$. An M-spline of order q, $q \geqslant 1$, is a combination of polynomial functions of degree $q-1$ connected at knots (de Boor, 1978). First, let $a = \min\{a_i; i = 1, \ldots, N\}$ and $b = \max\{b_i; i = 1, \ldots, N\}$, so the interval $[a, b]$ spans all the observation periods. Let $k_1 < k_2 < \ldots < k_s$, $s \geqslant q+1$, be a non-decreasing sequence of knots in $[a, b]$, where k_1

$= a$ and $k_s = b$. Then add an arbitrary $q-1$ extra knots at the beginning and end of the sequence. Usually the extra knots at the beginning of the sequence are all taken to be equal to k_1 and the ones at the end equal to k_s. Therefore, the new sequence of knots will be: $\tau_1 = \tau_2, ..., = \tau_q < \tau_{q+1} < ... < \tau_{q+s-1} = \tau_{q+s} = ... = \tau_{2q+s-2}$, where $\tau_{q+h} = k_{h+1}$ for $h = 0, 1, ..., s-1$. Then the M-spline basis functions of order q are given recursively by:

$$M_l(t|q) = \begin{cases} \frac{q[(t-\tau_l)M_l(t|q-1)+(\tau_{l+q}-t)M_{l+1}(t|q-1)]}{(q-1)(\tau_{l+q}-\tau_l)}, & \tau_l \leqslant t < \tau_{l+q} \\ 0, & \text{elsewhere,} \end{cases}$$

with

$$M_l(t|1) = \begin{cases} \frac{1}{(\tau_{l+1}-\tau_l)}, & \tau_l \leqslant t < \tau_{l+1} \\ 0, & \text{elsewhere,} \end{cases}$$

for $l = 1, 2, ..., m$ and $m = q + s - 2$. Each $M_l(t|q)$ consists of q polynomial pieces of degree $q-1$ that are joined at $q-1$ inner knots and whose derivatives up to order $q-2$ are continuous at the joining points. Each $M_l(t|q)$ is positive over the interval $\tau_l \leqslant t < \tau_{l+q}$ and zero elsewhere, i.e it is non zero over q intervals in the domain of t, $[a, b]$, and each interval has q positive M-splines. At a given t, q splines are positive.

Our approximation of age-specific relative incidence function is a linear combination of cubic M-splines, namely M-spline basis functions of order 4, and is given by:

$$\psi(t) = \sum_{l=1}^{m} g(\alpha_l)M_l(t) \tag{5}$$

where the coefficients $g(\alpha_l)$ are parameters estimated to determine the shape of the function. M-splines are positive functions and to keep the positivity of $\psi(t)$ the coefficients are constrained to be positive. We use $g(\alpha_l) = \alpha_l^2$, hence $\psi(t) = \sum_{l=1}^{m} \alpha_l^2 M_l(t)$. Combining equations (4) and (5) we obtain the log likelihood function for SCCS as:

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{(\sum_{l=1}^{m} \alpha_l^2 M_l(t_{ij}))\exp\left\{x_i(t_{ij})^T\beta\right\}}{\int_{a_i}^{b_i}(\sum_{l=1}^{m} \alpha_l^2 M_l(t))\exp\left\{x_i(t)^T\beta\right\}dt}\right). \tag{6}$$

The other motivating reason for using M-splines to approximate the age-specific relative incidence function is that the log likelihood function contains integrals. They can be replaced by other spline basis functions known as I-splines. I-splines are piecewise polynomials of degree q obtained by integrating M-splines of degree $q-1$ (Ramsay, 1988) and are thus defined for $\tau_h \leqslant t < \tau_{h+1}$ as $I_l(t|q) = \int_a^t M_l(u|q)du$.

Thus,

$$I_l(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h}(\tau_{m+q+1} - \tau_m)\frac{M_m(t|q+1)}{q+1}, & h-q+1 \leqslant l < h \\ 1 & l < h-q+1. \end{cases}$$

For example, assuming only one exposure period $(c_i, d_i]$ for each individual i and replacing the integrals of M-splines in the log likelihood by I-splines with same coefficients yields

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{(\sum_{l=1}^{m}\alpha_l^2 M_l(t_{ij}))\exp\left\{x(t_{ij})^T\beta\right\}}{I(b_i) - I(a_i) + (\exp(\beta)-1)(I(d_i) - I(c_i))}\right) \tag{7}$$

where

$I(b_i) = \sum_{l=1}^{m}\alpha_l^2 I_l(b_i)$, $I(a_i) = \sum_{l=1}^{m}\alpha_l^2 I_l(a_i)$,
$I(d_i) = \sum_{l=1}^{m}\alpha_l^2 I_l(d_i)$, $I(c_i) = \sum_{l=1}^{m}\alpha_l^2 I_l(c_i)$.

3.1 *Penalized log likelihood*

Approximating functions with non-parametric smooth functions requires determining the number and position of knots in advance. Too few knots leads to under fitting and too many leads to over fitting. Approaches have been proposed in the literature that automatically choose the number and positions of knots, e.g Friedman (1991). On the other hand different approaches have been proposed, e.g Eilers and Marx (1996), that avoid selecting the number of knots by using a relatively large number of knots and introducing a penalty function that controls over-fitting. Following standard practice, we use a penalty function that depends on the second derivative of $\psi(t)$ to control roughness and define our penalized likelihood as

$$
\begin{aligned}
pl &= l - \lambda \int (\sum_{l=1}^{m} g(\alpha_l) M_l^{''}(u))^2 du \\
&= l - \lambda((g(\boldsymbol{\alpha}))^T \mathbf{A} g(\boldsymbol{\alpha})) \qquad (8)
\end{aligned}
$$

where l is the log likelihood in (7), $\mathbf{A}$ is an $m \times m$ matrix with (h, l) element $\int M_h^{"}(u) M_l^{"}(u) du$ and $\lambda \geqslant 0$ is a smoothing parameter that controls the balance between smoothness of the age-specific relative incidence function and the fit to the data. The larger the value of λ, the smoother the age effect.

To obtain the parameter estimates, the penalized log likelihood (8) is maximized for fixed λ. Because $\psi(t)$ represents a relative effect, it is not identifiable without some further constraint (Farrington and Whitaker, 2006). We therefore impose the constraint $\int_a^b \psi(t)dt = 1$. The cumulative relative incidence is represented as the integral of a linear combination of cubic M-spline functions of the form $\int_a^t (\sum_{l=1}^m g(\alpha_l) M_l(u))du$. Since the integral of an M-spline is an I-spline, the cumulative incidence is represented by a linear combination of I-splines of the form $\sum_{l=1}^m g(\alpha_l) I_l(t)$. From the definition of I-splines all the I_l's evaluated at $t = b$ are equal to 1. Hence the required constraint can be achieved by constraining the sum of the coefficients of the linear combination of cubic M-spline functions to be 1. That is, $\sum_l^m g(\alpha_l) = \sum_l^m \alpha_l^2 = 1$.

3.2 *Smoothing parameter selection*

The smoothing parameter λ can be provided by the user or selected using automatic methods. We use a cross-validation method as in Joly et al. (1998), in which an approximate cross-validation score is maximized with β set to zero.

An approximate cross-validation score is maximized to determine the value of λ and is given as follows: Let $\boldsymbol{\alpha}$ be the vector of parameters α_l. Denote the cross-validation score $V(\lambda)$,

$$
V(\lambda) = \sum_i^N l_i(\hat{\boldsymbol{\alpha}}_{-i}) \qquad (9)
$$

where $\hat{\boldsymbol{\alpha}}_{-i} = \hat{\boldsymbol{\alpha}}_{-i}(\lambda)$ is the maximum penalized likelihood estimator of $\boldsymbol{\alpha}$ (with $\beta = 0$) when individual i is removed, and l_i is the log likelihood contribution of individual i. Following O'sullivan (1988), $V(\lambda)$ may be approximated by $\bar{V}(\lambda)$,

$$\bar{V}(\lambda) = l(\hat{\boldsymbol{\alpha}}) - \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H}), \tag{10}$$

where $\hat{H}$ is the likelihood component of the Hessian evaluated at the penalized MLE, $\hat{\boldsymbol{\alpha}}$, and $2\lambda\mathbf{S}$ is the penalized component of the Hessian. The matrix $\mathbf{S}$ depends on $g(\alpha_l)$. In our case $g(\alpha_l) = \alpha_l^2$, therefore, by a similar argument to that presented in Joly et al. (1998), $\mathbf{S} = 4\left(\mathbf{A}\mathrm{o}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\mathrm{diag}(\mathbf{A}\boldsymbol{\alpha}^2))$ (see Appendix) and $\mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H})$ can be interpreted as the model degrees of freeedom. The validity of the approximation was checked in the simulation studies. The penalized log likelihood is maximized for a grid of λ values, and the value of λ that maximizes the approximate cross-validation score is used in a final optimization step with the full model to obtain the relative incidences related to exposure.

3.3 *Fitting a spline-based SCCS*

The information needed to fit a spline-based SCCS is the same as for a standard SCCS. Instead of selecting age groups, a suitable number of knots are chosen. Usually between 8 and 15 are sufficient (Rondeau and Gonzalez, 2005; Joly et al., 1998). The knots will include the values a and b, namely, the minimum age at the start of all observation periods and maximum age at the end of all observation periods.

In a first step, λ is chosen using the approximate cross-validation method ignoring the exposure effect. Then, the parameters are estimated by maximizing the penalized log likelihood function (8) with the chosen value of λ, under the constraint that the sum of the coefficients of the age-specific relative incidence function to be 1.The inverse of the Hessian of the penalized log likelihood is used as a variance estimator of the parameters (Rondeau and Gonzalez, 2005).

Multiple risk periods can readily be incorporated. In addition to incorporating an indicator for the new exposure status in the numerator of the log likelihood function (7) we add $(\exp(\rho) - 1)(\sum_{l=1}^{m} \alpha_l^2 I_l(e_i) - \sum_{l=1}^{m} \alpha_l^2 I_l(s_i))$ in its denominator, where $\exp(\rho)$ is the relative incidence of the new exposure and s_i and e_i are the ages at start and end of the risk period associated with the new exposure for individual i, respectively. Further exposures can be added in the same way; some care is required in handling overlapping risk periods.

The new spline-based SCCS method has been implemented in R 2.15.1 (R Development Core Team, 2012), and the optimization of the constrained log likelihood is done using R function *auglag* from package *alabama*.

4. Simulation study

We conducted a simulation study to investigate the performance of the new method and to compare it with the parametric and semiparametric versions of SCCS. The observation periods for all cases were taken to be from 0 to 500 time units. We assumed only one age-dependent exposure variable and the post exposure risk period was 50 time units.

Three different exposure scenarios were considered: exponentially decreasing, exponentially increasing and uniformly distributed age at exposure. The exposure variable takes value 1 in $(c_i,\, c_i + 50)$ and 0 elsewhere, where c_i is age at first exposure of individual i. Exposure-related relative incidences of 0.5, 1, 2, and 5 were investigated, with sample sizes 50, 100 and 200.

For each combination of age at exposure, exposure relative incidence and sample size, three age-specific relative incidence scenarios were investigated: exponentially decreasing, exponentially increasing and constant. In the simulations it was assumed that each case experiences only one event.

We generated ages at event conditional on the exposure status from multinomial distributions with daily categories. 10000 samples were generated for all combinations of scenarios. The new spline-based SCCS and the semiparametric SCCS were fitted to each of the generated samples. We also fitted an SCCS without age effect to quantify the bias in the exposure effect when age is ignored.

We evaluated the performance of the new method in terms of its fit to the true age specific relative incidence function and in terms of reflecting the true exposure-related relative incidence. For each of the 10000 samples the mean of the estimated integrated square errors (MISE) and their standard deviations were calculated; selected results are presented in Table 1.

The median of exposure-related relative incidence(RI) ($\exp(\hat{\beta})$), log(RI)=$\hat{\beta}$, and the median of standard errors of $\hat{\beta}$'s are shown Table 2. We used the median because there is a non zero probability that all events will occur in the risk period only or in the control period only, so that in finite samples the theoretical bias is undefined. The coverage probability of the 95% confidence interval and the mean square error (MSE) of the $\hat{\beta}$ are also presented. In the simulations, we also checked that the cross-validation score, used in smoothing parameter selection, can be approximated by (10).

From Table 1 we can see that the performance of the new method in approximating the true age-specific relative incidence is better than the semiparametric method as the MISE values are slightly lower for the new method. The results presented in Table 1 are for scenarios when age at exposure decreases exponentially with exponentially decreasing and increasing age-specific relative incidence. Similar results were obtained for the other scenarios.

[Table 1 about here.]

Figure 1 shows the cumulative age-specific relative incidence from the spline-based SCCS, the semiparametric SCCS and the true cumulative incidence for a single sample of simulated data set of 200 cases. From the plots it seems that the two estimation methods give similar results.

[Figure 1 about here.]

For this same sample of simulated data, we evaluated the approximate cross-validation score (10) at a smoothing parameter (λ) value of 100000. We also calculated the exact cross-validation score by leaving out one case at a time and fit spline-based SCCS. The two values were close to each other, 1182.798 and 1183.045 respectively.

Table 2 shows that the bias in estimating the exposure-related relative incidence using the new method is small and when compared to the estimates from semiparametric method the bias is similar or slightly smaller in all scenarios considered. Similarly the mean square errors (MSE) are always slightly lower in spline-based SCCS than the semiparametric method. As expected, the SCCS without age effect produces very biased results.

[Table 2 about here.]

5. Application

The new spline-based and the parametric SCCS (for comparison) were applied to data on febrile convulsions and paediatric vaccines collected in England and Wales. In this application we investigate the association between febrile convulsions and diphtheria/tetanus/pertusis (DTP), and measles/mumps/rubella (MMR) vaccines.

The dataset includes 2389 children aged 29-730 days in the period 1991 to 1994, who had 3826 febrile convulsions. Of the 2389 cases, 2021 cases had an MMR vaccine record. DTP vaccine was given in three doses, DTP1, DTP2 and DTP3. The number of cases vaccinated with DTP1, DTP2 and DTP3 are 1624, 1684 and 1726 respectively. The average ages at DTP1, DTP2, DTP3 and MMR are 74.27, 119.11, 167.71 and 436.998 days respectively.

We estimated relative incidences (RI) for febrile convulsion in risk periods following these vaccines compared to control periods using both spline-based and parametric SCCS methods. For each of the DTP vaccine doses, we used three different risk periods (0-3, 4-7 and 8-14 days after vaccination), and two risk periods after MMR vaccination (6-11 and 15-35 days).

Overlapping risk periods were coded to the latest vaccine. For the parametric method, age was divided into 23 equal intervals of 1 month, apart from the first age group which had 32 days and the last 40 days. In the spline-based SCCS analysis the age-specific relative incidence function was approximated by a linear combination of cubic M-splines with 14 knots.

[Figure 2 about here.]

The fitted age-specific baseline incidence curves obtained from the parametric and the spline-based SCCS are presented in the left panel of Figure 2. The right panel of Figure 2 shows the cumulative age-specific relative incidence curves, where the dashed line is from the spline-based SCCS and the solid line from the parametric SCCS. The smoothing parameter was selected using the approximate cross-validation method; and the model degrees of freedom obtained for the optimum smoothing parameter value was 7.962.

[Table 3 about here.]

Table 3 presents relative incidence estimates from both methods. It shows that the two methods gave similar results for MMR with significant associations between febrile convulsion and MMR vaccines in both risk periods. RI estimates for DTP vaccines in the risk periods 4-7 and 8-14 days were not significantly different from 1 for the two methods. However, there was a difference for the 0-3 days risk period. The RI estimate using the spline-based method was significant whereas with the parametric method it was non-significant. This is due to the very strong age effect in the first year of life, which is inadequately controlled using the parametric model with one month age groups.

[Figure 3 about here.]

The relative incidence estimates were found to be insensitive to the choice of smoothing paramter. Figure 3 shows relative incidence estimates related to DTP and MMR exposures at different values of the smoothing parameter. From the figure it can be seen that the relative incidence estimates remain similar for smoothing parameter values within this range (the optimal value of λ was 1.07×10^9)

6. Final remarks

Modelling the age effect using a linear combination of cubic M-splines avoids the problem of sensitivity to mis-specification of age groups in the parametric version of SCCS method. The performance of the new method is as good as or better than the semiparameteric version of the SCCS method for small and moderate sample sizes. For large samples the semiparametric version is computationally demanding but the new method works well. For example for our convulsions data with 3826 events and 5 risk periods, the spline model took less than 2 minutes to fit on a standard desktop computer.

Acknowledgements

We are grateful to Pierre Joly for sharing his Fortran code. This research was supported by a Royal Society Wolfson research merit award to Paddy Farrington.

References

Barlow, W. E., Davis, R. L., Glasser, J. W., *et al.* (2001). The risk of seizures after receipt of whole-cell pertussis or measles, mumps, and rubella vaccine. *New England Journal of Medicine* **345,** 656–661.

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer-Verlag.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–102.

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. *Journal of the Royal Statistical Society Series C-Applied Statistics* **55,** 553–580.

Farrington, P., Pugh, S., Colville, A., *et al.* (1995). A new method for active surveillance of adverse events from diphteria-tetanus-pertussis and measles mumps rubella vaccines. *Lancet* **345,** 567–569.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics* **19,** 1–67.

Huang, W.-T., Gargiullo, P. M., Broder, K. R., *et al.* (2010). Lack of Association Between Acellular Pertussis Vaccine and Seizures in Early Childhood. *Lancet* **126,** E263–E269.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Miller, D. L., Ross, E. M., Alderslade, R., Bellman, M. H., and Rawson, N. S. B. (1981). Pertussis immunization and serious acute neurological illness in children. *British Medical Journal* **282,** 1595–1599.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *Siam Journal on Scientific and Statistical Computing* **9,** 363–379.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3,** 425–461.

Rondeau, V., and Gonzalez, J. R. (2005). Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine* **80,** 154–164.

Weldeselassie, Y. G., Whitaker, H. J., and Farrington, C. P. (2011) Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection* **139,** 1805–1817.

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.

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.

Appendix

Derivation of the approximate cross-validation score

Using a first order Tailor approximation around $\hat{\boldsymbol{\alpha}}$, the penalized maximum likelihood estimate when all observations are included, we get:

$\mathrm{V}(\lambda) = \sum_{i=1}^{N} l_i(\hat{\boldsymbol{\alpha}}_{-i}) \approx \sum_{i=1}^{N} \left(l_i(\hat{\boldsymbol{\alpha}}) + \frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}})[\hat{\boldsymbol{\alpha}}_{-i} - \hat{\boldsymbol{\alpha}}] \right)$

According to O'Sullivan(1988) $\hat{\boldsymbol{\alpha}}_{-i}$ can approximated as:

$\hat{\boldsymbol{\alpha}}_{-i} \approx \hat{\boldsymbol{\alpha}} - [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{d}_{-i}$, where $\hat{H} = \frac{\partial^2 l}{\partial \boldsymbol{\alpha}^2}(\hat{\boldsymbol{\alpha}})$ is the log likelihood part of the Hessian of the penalized log likelihood evaluated at $\hat{\boldsymbol{\alpha}}$, $\hat{d}_{-i} = -\hat{d}_i = -(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}))$ is a score vector when individual i is removed and $2\lambda\mathbf{S}$ is the penalized part of the Hessian.

Therefore $V(\lambda)$ is approximated by $\bar{V}(\lambda)$, where

$$\begin{aligned} \bar{V}(\lambda) &= \sum_{i=1}^{N} \left(l_i(\hat{\boldsymbol{\alpha}}) + \frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}})[\hat{\boldsymbol{\alpha}} - [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{d}_{-i} - \hat{\boldsymbol{\alpha}}] \right) \\ &= l(\hat{\boldsymbol{\alpha}}) + \mathrm{tr}\left([\hat{H} - 2\lambda\mathbf{S}]^{-1} \sum_{i=1}^{N} \left(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right) \left(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right)^T \right) \end{aligned}$$

Under regularity conditions $\quad E\left(\frac{\partial l_i}{\partial \alpha_l}\frac{\partial l_i}{\partial \alpha_m}\right) = -E\left(\frac{\partial^2 l_i}{\partial \alpha_l \partial \alpha_m}\right)$,

hence

$$\bar{V}(\lambda) \approx l(\hat{\boldsymbol{\alpha}}) - \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H}).$$

Let A $= \int M_h^{"}(u) M_l^{"}(u) du$, $p(\boldsymbol{\alpha}) = (g(\boldsymbol{\alpha}))^T \mathbf{A} g(\boldsymbol{\alpha})$ and $\boldsymbol{\theta} = g(\boldsymbol{\alpha})$. When $g(\alpha_l) = \alpha_l^2$ then, since $\mathbf{A}$ is symmetric, we have

$$\begin{aligned}\mathbf{S} = \frac{1}{2}\frac{\partial^2 p(\alpha)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} &= \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\alpha}}\right)^T \mathbf{A}\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\alpha}}\right) + (\boldsymbol{\theta}^T\mathbf{A}\otimes\mathbf{I})\frac{\partial}{\partial\boldsymbol{\alpha}^T}\left[\mathrm{vec}\left(\frac{\partial\boldsymbol{\theta}^T}{\partial\boldsymbol{\alpha}}\right)\right] \\ &= 4\left(\mathbf{A}\mathrm{o}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\mathrm{diag}(\mathbf{A}\boldsymbol{\alpha}^2)).\end{aligned}$$

LIST OF FIGURES

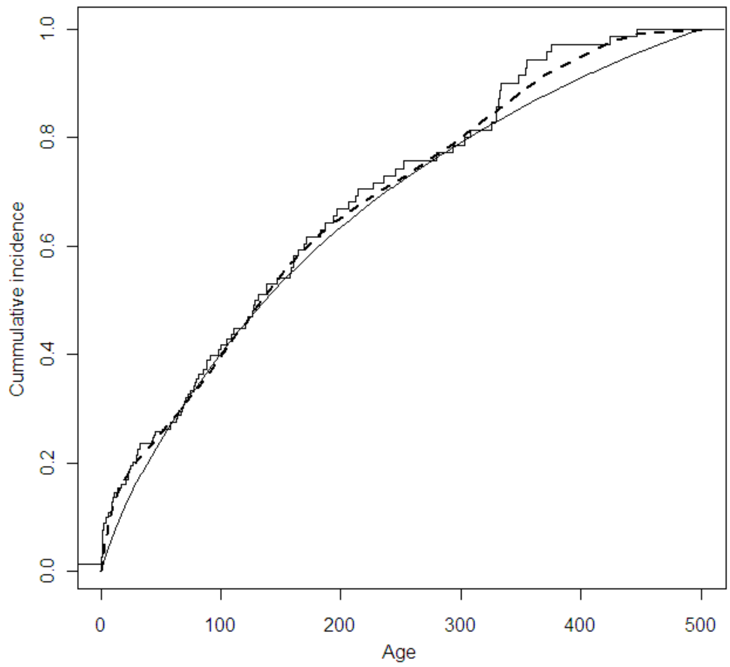

Figure 1. Cumulative age-specific relative incidence for a single sample: True (bold line), estimated from spline-based SCCS (dashed line) and estimated from semiparametric SCCS (step function).

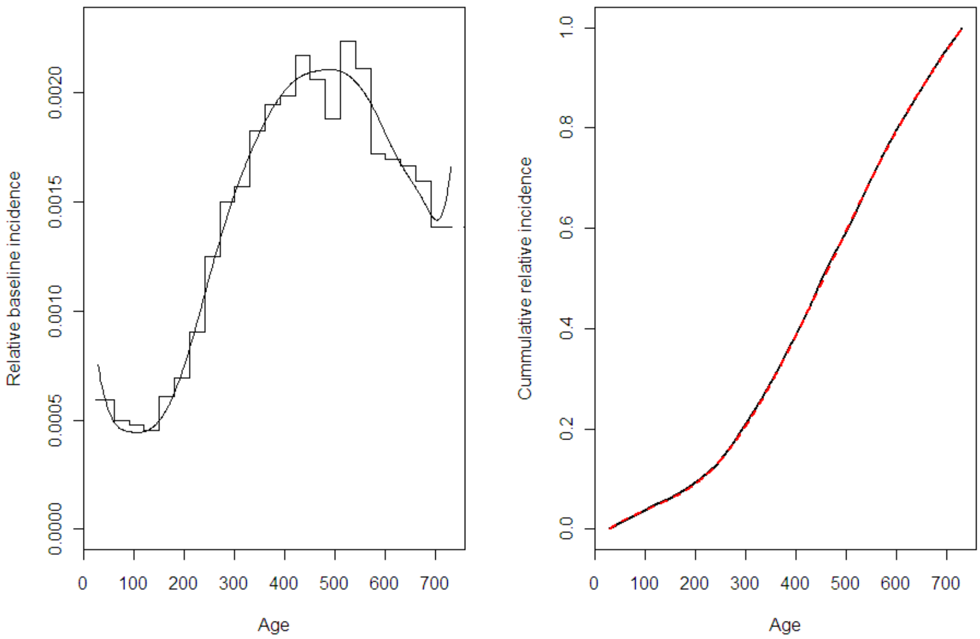

Figure 2. Fitted age-specific relative incidence curves (left) and cumulative age-specific relative incidence curves (right) from parametric and spline-based SCCS methods.

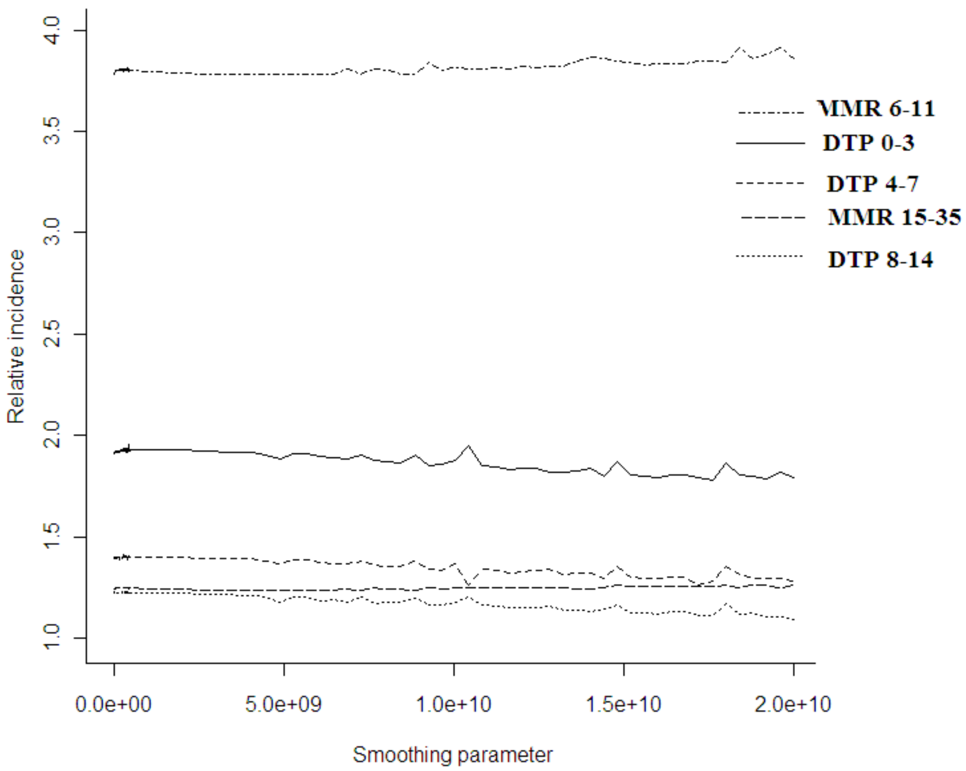

Figure 3. Dependence of exposure-related relative incidence estimates on smoothing parameter.

LIST OF TABLES

Table 1

Mean Integrated Square Error (MISE) and Standard Deviation (SD) from spline-based and semiparamtertic models: simulation with two scenarios of age at exposure (AE), age-specific relative incidence (ASRI) and exposure relative incidence (RI).

	AE decreasing and ASRI increasing, RI=5		AE and ASRI decreasing, RI=2	
# of cases	Spline-based MISE(SD)	Semiparametric MISE(SD)	Spline-based MISE(SD)	Semiparametric MISE(SD)
50	2.004(2.084)	2.208(2.103)	1.911(2.010)	2.102(2.027)
100	0.977(1.012)	1.055(1.012)	0.950(0.962)	1.029(0.965)
200	0.513(0.531)	0.548(0.531)	0.468(0.491)	0.502(0.490)

Table 2
Simulation results when age at exposure and age-specific relative incidence decrease exponentially.

RI	Measure	50 cases			100 cases		
		$\exp(\hat{\beta})(SD)$	$\hat{\beta}(SD)$	SE	$\exp(\hat{\beta})(SD)$	$\hat{\beta}(SD)$	SE
Spline-based SCCS							
5	Median	5.242(4.058)	1.657(0.486)	0.412	5.071(1.636)	1.624(0.287)	0.266
	P95		94.107			94.614	
	MSE		0.238			0.083	
2	Median	2.008(0.923)	0.697(0.394)	0.366	2.008(0.591)	0.697(0.274)	0.261
	P95		94.482			95.358	
	MSE		0.155			0.075	
1	Median	0.995(0.507)	-0.005(0.442)	0.402	0.995(0.528)	-0.005(0.448)	0.405
	P95		94.343			94.098	
	MSE		0.196			0.200	
Semiparametric SCCS							
5	Median	5.244(5.354)	1.657(0.512)	0.471	5.105(1.683)	1.630(0.291)	0.278
	P95		95.600			94.600	
	MSE		0.264			0.085	
2	Median	2.032(0.970)	0.709(0.404)	0.394	2.021(0.599)	0.704(0.276)	0.270
	P95		95.300			94.700	
	MSE		0.163			0.076	
1	Median	1.004(0.524)	0.004(0.453)	0.437	0.996(0.546)	-0.004(0.460)	0.444
	P95		95.080			95.410	
	MSE		0.205			0.212	
SCCS without age effect							
5	Median	8.500(2.979)	2.140(0.305)	0.303	8.375(1.905)	2.125(0.217)	0.214
	P95		56.020			33.540	
	MSE		0.375			0.313	
2	Median	3.259(1.000)	1.181(0.288)	0.287	3.07(0.639)	1.122(0.202)	0.203
	P95		59.480			42.970	
	MSE		0.321			0.225	
1	Median	1.714(0.562)	0.539(0.320)	0.309	1.714(0.556)	0.539(0.314)	0.309
	P95		56.280			54.930	
	MSE		0.393			0.389	

Table 3
Relative incidence estimates for febrile convulsion after DTP and MMR vaccines.

		Spline-based SCCS		Parametric SCCS	
Vaccine	Risk period	RI	95% CI	RI	95% CI
DTP all doses	0-3	1.905	[1.349 , 2.668]	1.420	[0.963 , 2.092]
	4-7	1.391	[0.933 , 2.075]	1.184	[0.774 , 1.812]
	8-14	1.225	[0.899 , 1.670]	0.974	[0.693 , 1.366]
MMR	6-11	3.781	[3.120 , 4.492]	3.451	[2.854 , 4.175]
	15-35	1.241	[1.050 , 1.453]	1.197	[1.013 , 1.414]