(2015), **0**, 0, *pp.* 1–25
doi://

Spline-based self-controlled case series method

YONAS GHEBREMICHAEL-WELDESELASSIE*, HEATHER J. WHITAKER, C. PADDY FARRINGTON

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

yonas.weldeselassie@open.ac.uk

Summary

The self-controlled case series (SCCS) method is an alternative to study designs such as cohort and case control methods and is used to investigate potential associations between the timing of vaccine or other drug exposures and adverse events. It requires information only on cases, individuals who have experienced the adverse event at least once, and automatically controls all fixed confounding variables that could modify the true association between exposure and adverse event. Time-varying confounders such as age, on the other hand, are not automatically controlled and must be allowed for explicitly. The original SCCS method used step functions to represent risk periods (windows of exposed time) and age groups. The SCCS method has been extended by modelling only the age effect or only the time-varying exposure effect using splines while representing the other by a piecewise constant step function. In these two extensions, there is a need to pre-specify exposure risk periods or age groups a priori, but a poor choice of group boundaries may lead to biased estimates. In this paper, we propose a non-parametric SCCS

*To whom correspondence should be addressed.

method in which both age and exposure effects are represented as linear combinations of cubic M-splines at the same time. To avoid a numerical integration of the product of these two spline functions in the likelihood function of the SCCS method we defined the first, second and third integrals of I-splines based on the definition of integrals of M-splines. Simulation studies showed that the new method performs well. This new method is applied to data on paediatric vaccines.

Key words: Integral of I-splines; M-splines; Self-controlled case series; Smooth risk functions; Splines.

1. Introduction

The self-controlled case series method is used to investigate potential associations between time varying exposures to vaccines or other drugs and adverse health events (Farrington, 1995). It yields estimates of relative incidence, that is, the incidence in exposure risk periods relative to all other time over which cases are observed. The method uses information only from cases, individuals who have experienced the event of interest at least once, and implicitly controls all measured and unmeasured confounding variables that act multiplicatively on the hazard. However, time varying confounders such as age are not automatically controlled and hence should be included in the model. The standard case series method uses piecewise constant step functions to represent both age and exposure effects. Poor choice of the a priori chosen age groups or exposure risk periods in the standard method may result in biased estimates of exposure-related relative incidences. Usually the choice of exposure risk periods is motivated by reference to previous studies or hypotheses, by biologically plausible mechanisms or by expert opinion, but it is not uncommon to face a situation in which there is little knowledge of the precise timing defining true exposure risk periods. Recently Ghebremichael-Weldeselassie *et al* (2014) extended the SCCS method by modeling the age effect using splines to avoid the limitations of the standard SCCS model. In addition Ghebremichael-Weldeselassie *et al* (2015) used linear combinations of cubic M-splines (piecewise polynomials of degree three) to represent the exposure risk effect. However, in these two extensions either age or exposure risk effects are represented by piecewise constant functions requiring a priori choice of cut points. Therefore, in this paper we extend the SCCS method further by modelling both age and exposure effects using splines to have a fully non-parametric SCCS method. This extension is non-trivial as it involves integrals of spline products.

The paper is organized as follows; after some initial remarks in Section 2, the likelihood function of the spline-based SCCS method is derived in Section 3. In this section, we also describe and define derivatives and integrals of M and I splines, and the integral of a product of two spline functions. Section 4 presents the penalised log-likelihood function of the spline-based SCCS method and discusses the selection of smoothing parameters. In Section 5, we evaluate the performance of the new method using simulations. We apply the spline-based SCCS method to data on febrile convulsion and MMR vaccine in Section 6 and follow this with final remarks in Section 7.

2. Modelling Age and Exposure Effects Using Splines

The use of regression splines in the context of the self-controlled case series method has shown an improved performance compared to the use of step functions as presented in Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015). Among the motivations for using regression splines based on M-splines in these papers were that the spline functions give flexible and plausible shapes of age and exposure-related relative incidence functions and avoid numerical integration of the integral in the denominator of the SCCS likelihood function. This numerical integration is avoided because the integral of an M-spline is an I-spline, therefore the integral of a linear combination of M-splines can be expressed as a linear combination of I-splines. Based on similar arguments, both age and exposure effects can be represented as linear combinations of M-spline basis functions. In this paper, since age and exposure are to be represented by linear combinations of M-splines at the same time, the denominator of the SCCS likelihood function will involve the integral of a product of two spline functions. This cannot be represented by a linear combination of I-splines only, so the integration cannot be avoided in the same way used by Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015). Therefore, based on the definition of the integral of an M-spline developed by Ramsay (1988), we define first, second and third integrals of an I-spline to avoid numerical integration of the product of two spline functions. In the following section we derive the likelihood function of the SCCS method when both age and exposure effects are approximated by linear combinations of M-spline basis functions.

3. Likelihood Function

To derive the likelihood function of the spline-based SCCS method, we begin with the general SCCS likelihood function given in Farrington and Whitaker (2006). The likelihood is specified over an observation period defined by age or calendar time boundaries $(a_i, b_i]$ within which an event has been observed and the full exposure history is known. Note that in this paper we take the underlying time line as age, while in practice this can be replaced with calendar time. For one exposure risk period, the likelihood is given as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} \psi(t) \exp\{x_i(t)\beta\}\, dt} \tag{3.1}$$

where a_i and b_i are the start and end of the observation period for individual i, $i = 1, 2, \ldots, N$, n_i is number of events experienced by individual i within the observation period $(a_i, b_i]$, t_{ij} is age at the j^{th} event of individual i, $x_i(t)$ is the exposure status of individual i at t, $\exp(\beta)$ is the exposure-related relative incidence, and $\psi(t)$ is the age-related relative incidence function.

Equation 3.1 can be generalized as follows by using time since start of exposure as an argument for the exposure effect:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij})\omega(t_{ij} - c_i)}{\int_{a_i}^{b_i} \psi(t)\omega(t - c_i)dt}, \tag{3.2}$$

where $\omega(t-c)$ is the exposure-related relative incidence function which takes the value one if the age at event is not between age at start of exposure (c_i) and age at end of exposure (d_i). In the standard SCCS method, $\psi(t)$ and $\omega(t-c)$ are represented by step functions; in the semi-parametric version of SCCS (Farrington and Whitaker, 2006), $\psi(t)$ is left unspecified and $\omega(t-c)$ is fitted as a step function; in Ghebremichael-Weldeselassie *et al* (2014), $\psi(t)$ is approximated by spline functions and $\omega(t-c)$ by a step function, and in Ghebremichael-Weldeselassie *et al* (2015), $\psi(t)$ is represented as a step function and $\omega(t-c)$ as a linear combination of M-spline functions.

In this paper, we approximate both $\psi(t)$ and $\omega(t-c)$ as linear combinations of cubic M-spline basis functions. M-splines, which are variants of B-splines, are piecewise polynomials connected at points known as knots. Given a knot sequence $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$, an M-spline of order q is defined as

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

with

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

The integrals of M-splines were defined by Ramsay (1988) as I-splines. I-splines are piecewise polynomials of order $q+1$ obtained by integrating M-splines of order q and are thus defined for $k_h \leqslant t < k_{h+1}$ as $I_l(t|q) = \int_o^t M_l(u|q)du$, where the lower limit of the integral is the minimum interior knot denoted by o.

Thus for the same sequence of interior knots used in defining M-splines, I-splines are defined as

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

$\psi(t)$ is defined between $a = \min\{a_i; i = 1, \ldots, N\}$ and $b = \max\{b_i; i = 1, \ldots, N\}$, where N is the total number of cases in the study. Since $\psi(t)$ is a relative effect it has to be a positive function and to obtain such a function based on M-splines we constrain the coefficients to be non-negative to give the following expression for $\psi(t)$:

$$\psi(t) = \sum_{l=1}^{m_1} g(\alpha_l) M_{1l}(t) = \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t). \tag{3.3}$$

The $g(\alpha_l)$ are parameters used to determine the shape of $\psi(t)$ and are constrained to be non-negative by taking $g(\alpha_l) = \alpha_l^2$. $M_{1l}(t)$ is the l^{th} M-spline basis function related to age, m_1 is the number of parameters or the number of M-spline basis functions which is equal to the sum of the number of interior knots and the order of the basis functions.

Similarly, the exposure-related relative incidence function with non-negative coefficients is defined between 0 and $\max\{(d_i - c_i); i = 1, \ldots, N\}$, where c_i and d_i are the start and end of age at exposure respectively for individual i. When the exposure is a point exposure, e.g. a vaccine, a nominal maximum risk period is defined which can be unbounded to the right. The nominal risk period is a period within the observation period where the exposure-related relative incidence can be different from 1; outside this period the exposure-related relative incidence function takes the value 1. Therefore, it is defined as:

$$\omega(t-c) = \begin{cases} \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c), & c < t \leqslant d \\ 1, & \text{otherwise,} \end{cases} \tag{3.4}$$

where m_2 is the number of M-spline basis functions used to define the exposure-related relative incidence function $\omega(t-c)$ and $M_{2l}(t-c)$ is the l^{th} basis function related to exposure. The knots which are used to define the M-splines related to the age effect and the exposure effect are chosen to be equidistant including the arbitrary knots added below and above the minimum and maximum values of the variable.

Now replacing $\psi(t)$ and $\omega(t-c)$, in Equation (3.2), by the spline functions in Equations (3.3) and (3.4) respectively gives the likelihood function for the spline-based SCCS as

$$l = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t_{ij})\right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t_{ij} - c_i)\right)^{I(c_i < t_{ij} \leqslant d_i)}}{\int_{a_i}^{b_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t - c_i)\right)^{I(c_i < t \leqslant d_i)} dt}$$

and the log-likelihood function is

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\left(\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t_{ij})\right)\left(\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t_{ij}-c_i)\right)^{I(c_i<t_{ij}\leqslant d_i)}}{\int_{a_i}^{b_i}\left(\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t)\right)\left(\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t-c_i)\right)^{I(c_i<t\leqslant d_i)} dt}\right). \qquad (3.5)$$

To further simplify the denominator of the log-likelihood function (3.5), to avoid numerical integration, we will use integration by parts. This will involve derivatives and integrals of linear combinations of M-spline functions and integrals of their integrals. Therefore, before we proceed with simplifying the log-likelihood function, we describe derivatives of M-splines and define integrals of I-splines in the following subsections.

3.1 *Derivatives of M-splines*

The derivative of an M-spline of order q is given by (de Boor, 1978)

$$\frac{dM_l(t|q)}{dt} = \frac{q}{k_{l+q}-k_l}\left(M_l(t|(q-1)) - M_{l+1}(t|(q-1))\right).$$

In general, the j^{th} derivative of an M-spline function of order q, $M_l(t|q)$, is

$$\frac{d^j M_l(t|q)}{dt^j} = \frac{q}{k_{l+q}-k_l}\left(\frac{d^{j-1}M_l(t|(q-1))}{dt^{j-1}} - \frac{d^{j-1}M_{l+1}(t|(q-1))}{dt^{j-1}}\right),$$

and the j^{th} derivative of a function which is a linear combination of M-spline basis functions, $f(t) = \sum_{l=1}^{m}\alpha_l M_l(t|q)$, can be given as

$$\frac{d^j f(t)}{dt^j} = \sum_{l=1}^{m}\alpha_l \frac{d^j M_l(t|q)}{dt^j}.$$

3.2 *Integrals of I-splines*

Based on the definition for the integral of an M-spline given by Ramsay (1988) and shown above we define the integral of an I-spline. Let the integral of $I_l(t|q)$ be denoted by $I_l^1(t|q) = \int_o^t I_l(u|q)du$. Using the same sequence of interior knots employed to define the M-splines, for $k_h \leqslant t < k_{h+1}$ the integral of an I-spline, $I_l^1(t|q)$, has three different expressions depending on the value of l. For

$l > h$ the value of an I-spline is zero so its indefinite integral will be a constant, and hence

$$I_l^1(t|q) = \int_o^t I_l(u|q)du = 0.$$

For $h - q + 1 \leqslant l \leqslant h$ an I-spline, $I_l(t|q)$, is given by

$$I_l(t|q) = \sum_{m=l}^{h} (k_{m+q+1} - k_m) \frac{M_m(t|q+1)}{q+1}$$

therefore its integral will be

$$\begin{aligned} I_l^1(t|q) &= \int_o^t \sum_{m=l}^{h} (k_{m+q+1} - k_m) \frac{M_m(u|q+1)}{q+1} du \\ &= \sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \int_o^t M_m(u|q+1)du. \end{aligned}$$

$\int_o^t M_m(u|q+1)du$ in the above expression is the integral of an M-spline of order $q+1$ that gives another I-spline, $I_m(t|(q+1)) = \sum_{n=m}^{h} (k_{n+q+2} - k_n) \frac{M_n(t|q+2)}{q+2}$ for $h - q \leqslant m \leqslant h$, so

$$I_l^1(t|q) = \sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \sum_{n=m}^{h} (k_{n+q+2} - k_n) \frac{M_n(t|q+2)}{q+2}.$$

For $l < h - q + 1$, that is for any value of $t > k_{l+q}$ the value of $I_l(t|q) = 1$. This is because $M_l(t|q) = 0$ for all values of $t > k_{l+q}$. Now the integral of $I_l(t|q)$ has two parts for $t > k_{l+q}$, the integral of the function up to k_{l+q} and from k_{l+q} to t. That is,

$$\int_o^{k_{l+q}} I_l(u|q)du + \int_{k_{l+q}}^t I_l(u|q)du = \left(\sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \int_o^{k_{l+q}} M_m(u|q+1)du \right) + (t - k_{l+q}).$$

Therefore, in summary the integral of an I-spline is given by

$$I_l^1(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \sum_{n=m}^{h} (k_{n+q+2} - k_n) \frac{M_n(t|q+2)}{q+2}, & h - q + 1 \leqslant l \leqslant h \\ t - k_{l+q} + \sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \sum_{n=m}^{h} (k_{n+q+2} - k_n) \frac{M_n(k_{l+q}|q+2)}{q+2}, & l < h - q + 1. \end{cases}$$

The second integral of an I-spline $I_l^2(t|q) = \int_o^t I_l^1(u|q)du$ and the third integral $I_l^3(t|q) = \int_o^t I_l^2(u|q)du$ can be obtained in a similar way (see supplementary material).

3.3 *Integrating the Product of Two Spline Functions*

Now going back to the log-likelihood function (3.5), since the exposure-related relative incidence function, $\omega(t-c)$, takes the value 1 in the control periods, $(a_i, c_i]$ and $(d_i, b_i]$, within the observation period, the denominator of the log-likelihood function can be rewritten as

$$\int_{a_i}^{c_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt + \int_{c_i}^{d_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right)\left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c_i)\right)dt + \int_{d_i}^{b_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt$$

Furthermore, the first and the last terms are integrals of only one function, the age-specific relative incidence $\psi(t)$, whereas the second term is the integral of a product of two spline functions. Since the integral of an M-spline of order q is an I-spline of order $q+1$, hence the integral of a linear combination of M-splines can be expressed as a linear combination of I-splines. Therefore, we replace the integrals in the first and third terms by linear combinations of I-spline basis functions which leads to a denominator with the expression

$$\sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\Big|_{a_i}^{c_i} + \int_{c_i}^{d_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right)\left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c_i)\right)dt + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\Big|_{d_i}^{b_i}.$$

The $I_{1l}(t)$ are I-splines related to the age effect and $I_{2l}(t)$ will be used to denote I-splines related to the exposure effect. The remaining part in the denominator of the log-likelihood function of the spline-based SCCS is in the risk period $(c_i, d_i]$ where the exposure-related relative incidence can take a value different from 1. This part contains an integral of the product of the two spline functions, $\psi(t)$ and $\omega(t-c)$.

To evaluate this integral we use integration by parts as follows:

$$\int \psi(t)\omega(t-c)dt = \psi(t)\int \omega(t-c)dt - \int \left(\psi'(t)\int \omega(t-c)dt\right)dt \tag{3.6}$$

where $\psi'(t)$ is the first derivative of $\psi(t)$. Since $\psi(t)$ and $\omega(t-c)$ are linear combinations of M-spline basis functions, $\int \omega(t-c)dt$ can be expressed as a linear combination of I-splines denoted by $I_E(t-c)$

$$I_E(t-c)=\int_c^t \omega(u-c)du=\int_c^t \sum_{l=1}^{m_2}\beta_l^2 M_{2l}(u-c)du=\sum_{l=1}^{m_2}\beta_l^2 I_{2l}(t-c).$$

Letting the integral of the linear combination of I-splines $I_E(t-c)$ be denoted by $I_E^1(t-c)$, the integral of $I_E^1(t-c)$ by $I_E^2(t-c)$ and the integral of $I_E^2(t-c)$ by $I_E^3(t-c)$ that is,

$$I_E^1(t-c)=\int I_E(t-c)dt, \quad I_E^2(t-c)=\int I_E^1(t-c)dt \quad \text{and} \quad I_E^3(t-c)=\int I_E^2(t-c)dt,$$

the expression in Equation (3.6) becomes

$$\int \psi(t)\omega(t-c)dt=\psi(t)I_E(t-c)-\int\left(\psi'(t)I_E(t-c)\right)dt.$$

The last term of this equation is again an integral of a product of two non-constant functions. We therefore apply integration by parts repeatedly until none of the terms is an integral of two non-constant functions and get:

$$\begin{aligned}\int \psi(t)\omega(t-c)dt &= \psi(t)I_E(t-c)-\int\left(\psi'(t)I_E(t-c)\right)dt\\ &= \psi(t)I_E(t-c)-\psi'(t)I_E^1(t-c)+\psi''(t)I_E^2(t-c)-\psi'''(t)I_E^3(t-c)\end{aligned}$$

where $\psi'(t)$, $\psi''(t)$ and $\psi'''(t)$ are the first, second and third derivatives of $\psi(t)$ respectively. $\psi'''(t)$ is a constant function because $\psi(t)$ is a piecewise cubic function.

Then the log-likelihood function of the spline-based SCCS method, obtained by replacing the appropriate expressions for the terms $\int_{a_i}^{c_i}\psi(t)dt$, $\int_{c_i}^{d_i}\psi(t)\omega(t-c)dt$ and $\int_{d_i}^{b_i}\psi(t)dt$ in the denominator, is

$$l=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left(\frac{\left(\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t_{ij})\right)\left(\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t_{ij}-c_i)\right)^{I(c_i<t_{ij}\leqslant d_i)}}{B}\right) \tag{3.7}$$

where

$$
\begin{aligned}
B = & \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\big|_{a_i}^{c_i} + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\big|_{d_i}^{b_i} \\
& + \left(\psi(t) I_E(t-c_i) - \psi'(t) I_E^1(t-c_i) + \psi''(t) I_E^2(t-c_i) - \psi'''(t) I_E^3(t-c_i)\right)\big|_{c_i}^{d_i}
\end{aligned}
$$

and

$$
\begin{aligned}
I_E^1(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^1(t-c), \\
I_E^2(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^2(t-c), \\
I_E^3(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^3(t-c)
\end{aligned}
$$

where $I_{2l}^1(t-c)$, $I_{2l}^3(t-c)$ and $I_{2l}^3(t-c)$ are the first, second and third integrals of the l^{th} I-spline $(I_{2l}(t-c))$ related to exposure, respectively.

So far, the methodology developed in this paper has considered only one exposure period. However, it can be applied to multiple exposures provided that the exposure risk periods do not overlap. If multiple exposure risk periods, within an observation, do not overlap then the relative incidence at each interval will be represented by the age-specific relative incidence function in the control periods and the product of the age-specific relative incidence function and the relative incidence function related to only one of the exposures. For example, if we have a second non-overlapping exposure risk period $(e_i, f_i]$, the numerator of the log-likelihood function (3.7) will be multiplied by a spline function related to the new exposure, $\left(\sum_{l=1}^{m_3} \gamma_l^2 M_{3l}(t_{ij} - e_i)\right)^{I(e_i < t_{ij} \leqslant f_i)}$. The denominator then becomes

$$
\begin{aligned}
B = &\sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\big|_{a_i}^{c_i} + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\big|_{d_i}^{e_i} + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\big|_{f_i}^{b_i} \\
&+ \big(\psi(t) I_E(t-c_i) - \psi'(t) I_E^1(t-c_i) + \psi''(t) I_E^2(t-c_i) - \psi'''(t) I_E^3(t-c_i)\big)\,\big|_{c_i}^{d_i} \\
&+ \big(\psi(t) I_{E2}(t-e_i) - \psi'(t) I_{E2}^1(t-e_i) + \psi''(t) I_{E2}^2(t-e_i) - \psi'''(t) I_{E2}^3(t-e_i)\big)\,\big|_{e_i}^{f_i}
\end{aligned}
$$

where $I_{E2}(t-e_i)$, $I_{E2}^1(t-e_i)$, $I_{E2}^2(t-e_i)$ and $I_{E2}^3(t-e_i)$ are I-splines and their integrals all related to the second exposure. Further exposures can be incorporated in a similar way.

4. Penalised Log-Likelihood

The numbers of knots, which determine the numbers of M-spline basis functions that make up the age-specific and exposure-related relative incidence functions are chosen a priori. Maximising the log-likelihood function (3.7) after choosing too large a number of knots over-fits the true curves, while selecting too small a number of knots leads to under-fitting overly smoothed curves. Therefore, to control the smoothness of the estimated functions we fix the numbers of knots at higher values than are believed to be enough to represent the functions and introduce roughness penalty terms to the log-likelihood function (3.7). Following Joly and Commenges (1999), we choose a roughness measure to be the sum of the square norms of the second derivatives of the age and exposure effect functions. This leads to the penalised log-likelihood function

$$
\begin{aligned}
pl &= l(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \lambda_1 \int \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}^{''}(u) \right)^2 du - \lambda_2 \int \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}^{''}(u) \right)^2 du \\
&= l(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \lambda_1((\boldsymbol{\alpha^2})^T \mathbf{A}_1 \boldsymbol{\alpha^2}) - \lambda_2((\boldsymbol{\beta^2})^T \mathbf{A}_2 \boldsymbol{\beta^2})
\end{aligned}
\tag{4.8}
$$

where $\boldsymbol{\alpha}$ is a vector of parameters $\alpha_1, \ldots, \alpha_{m_1}$, that define the age-specific relative incidence function and $\boldsymbol{\alpha^2} = \alpha_1^2, \ldots, \alpha_{m_1}^2$, $\boldsymbol{\beta}^2 = \beta_1^2, \ldots, \beta_{m_2}^2$ are parameters related to the exposure effect, $\mathbf{A}_1$ is an $m_1 \times m_1$ matrix with (r, l) element $\int M_{1r}^{''}(u) M_{1l}^{''}(u) du$, $\mathbf{A}_2$ is an $m_2 \times m_2$ matrix with (r, l) element $\int M_{2r}^{''}(u) M_{2l}^{''}(u) du$, $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the log-likelihood function (3.7). λ_1 and λ_2 are non-negative smoothing parameters that control the trade-off between the model fit and the smoothness of the functions. The penalised log-likelihood function (4.8) is maximised, for fixed λ_1 and λ_2 values, to estimate the parameters related to age and exposure effects.

4.1 *Selection of Smoothing Parameters*

We choose the smoothing parameters by maximizing approximate cross-validation scores, as proposed by O'Sullivan (1988). λ_1 is first chosen by ignoring the exposure effect then λ_2 by ignoring the age effect.

Denote the cross-validation scores by $V_1(\lambda_1)$ and $V_2(\lambda_2)$,

$$V_1(\lambda_1) = \sum_{i}^{N} l_i(\hat{\boldsymbol{\alpha}}_{-i}) \tag{4.9}$$

$$V_2(\lambda_2) = \sum_{i}^{N} l_i(\hat{\boldsymbol{\beta}}_{-i}) \tag{4.10}$$

where $\hat{\boldsymbol{\alpha}}_{-i} = \hat{\boldsymbol{\alpha}}_{-i}(\lambda_1)$ is the maximum penalized likelihood estimator of $\boldsymbol{\alpha}$ (with the exposure effect excluded from the model) when individual i is removed, and l_i is the log likelihood contribution of individual i. Following O'Sullivan (1988), $V_1(\lambda_1)$ may be approximated by $\bar{V}_1(\lambda_1)$,

$$\bar{V}_1(\lambda_1) = l(\hat{\boldsymbol{\alpha}}) - \text{tr}([\hat{H}_1 - 2\lambda_1 \mathbf{S}_1]^{-1}\hat{H}_1), \tag{4.11}$$

where $\text{tr}(X)$ is the trace of a matrix X, $l(\hat{\boldsymbol{\alpha}})$ is the log-likelihood function in Equation (3.7) where no exposure effect is included and evaluated at the maximum penalized likelihood estimates $(\hat{\boldsymbol{\alpha}})$. $\hat{H}_1 = \frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}(\hat{\boldsymbol{\alpha}})$ is the log-likelihood part of the Hessian of the penalized log-likelihood evaluated at the penalized maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$. The matrix $\mathbf{S}_1$ depends on the expression for $g(\alpha_l)$. If $g(\alpha_l) = \alpha_l$ then $\mathbf{S}_1 = \mathbf{A}_1$, however here we take $g(\alpha_l) = \alpha_l^2$. Therefore, $\mathbf{S}_1 = 4\left(\mathbf{A}_1 \circ (\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\text{diag}(\mathbf{A}_1\boldsymbol{\alpha}^2))$ (Ghebremichael-Weldeselassie *et al*, 2014), where $\circ$ is the

Hadamard product of matrices. Similarly, to choose the smoothing parameter of the exposure-related relative incidence function, $V_2(\lambda_2)$ can be approximated as

$$\bar{V}_2(\lambda_2) = l(\hat{\boldsymbol{\beta}}) - \mathrm{tr}([\hat{H}_2 - 2\lambda_2 \mathbf{S}_2]^{-1}\hat{H}_2), \tag{4.12}$$

where $l(\hat{\boldsymbol{\beta}})$ is the log-likelihood (3.7) taking no age effect into consideration, $\hat{H}_2 = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}})$ is the Hessian when no age effect is included and $\mathbf{S}_2 = 4\left(\mathbf{A}_2 \circ (\boldsymbol{\beta}\boldsymbol{\beta}^T)\right) + 2(\mathrm{diag}(\mathbf{A}_2\boldsymbol{\beta}^2))$.

Then after choosing the smoothing parameters the log-likelihood function (4.8) is maximised for fixed λ_1 and λ_2 values.

5. Simulation Study

To evaluate the performance of the new spline-based SCCS method and to compare it with the extensions made to the standard SCCS method by Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015), we conducted a simulation study. The methods developed by Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015) showed that the use of splines has a better performance in terms of efficiency than the standard SCCS methods.

5.1 *Design of the Simulation Study*

The number of cases used in this simulation was 1000, each with ages at the start and end of the observation period of 0 and 730 days respectively. For each case, the risk period between the age at start of exposure c_i and age at end of exposure d_i was taken as 49 days. The baseline incidence is generated from a sine function, defined as $\lambda_0(t) \propto 8(\sin(0.01 \times t)) + 9$ at age t. The true age-related relative incidence function is presented in Panel a of Figure 1. Ages at start of exposure c_i, for $i : 1, \ldots, 1000$, were sampled within (0,730] from an exponential density with rate 0.003. The histogram of c_i is shown in Panel b of Figure 1.

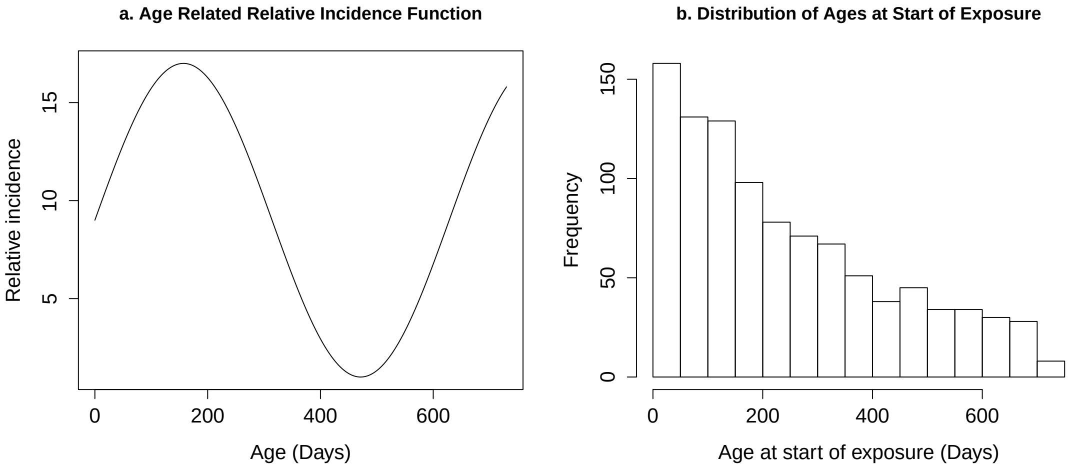

Fig. 1. *True age-related relative incidence function in Panel* (*a*) *and distribution of ages at start of exposure in Panel* (*b*)*, which were used to simulate data sets.*

For the given age-related relative incidence function and distribution of age at exposure, we investigated four scenarios for the exposure-related relative incidence function, $\omega(t-c)$. These functions take the value one outside the risk period $(c_i, d_i]$, that is when time since start of exposure $t-c \leqslant 0$ or $t-c > 49$. Without loss of generality we consider each case to have experienced only one event. The daily incidence rates within the observation period are evaluated as the product of the age-related relative incidence and the exposure-related relative incidence. An event day for each individual was generated from a multinomial distribution. The probability of an event at a given day within the observation period was computed as the incidence rate for that day divided by the sum of the rates for all the days within the observation period. For each scenario 100 data sets were simulated.

The data sets generated were analyzed in three ways:

1. SCCS with smooth age effect and parametric exposure effect (step function) (Ghebremichael-Weldeselassie *et al*, 2014),

2. SCCS with parametric age effect (step function) and spline-based exposure effect (Ghebremichael-

Weldeselassie *et al*, 2015), and

3. the spline-based SCCS, the new method proposed in this paper.

For the first method, seven exposure groups of length seven days between 0 and 49 were chosen to represent the exposure effect by a step function. For methods (1) and (3), to represent the age effect with a spline function 9 interior knots between the minimum of ages at the start of observation (zero) and the maximum of the ages at the end of observation periods (730) were chosen. For the age effect, since exposure-related parameters are not duly sensitive to changes in the smoothing parameter related to the age effect (Ghebremichael-Weldeselassie *et al*, 2014), we chose a smoothing parameter for the first sample in a given scenario by the cross validation method and used the same value for the remaining samples.

For the second method, where age is represented with a piecewise constant function, six age groups with cut points at 0, 120, 240, 360, 480, 600 and 730 days were pre-specified. To represent the exposure effect with a spline function in methods (2) and (3), a nominal risk period of 49 days was chosen. 12 interior knots between zero and 49 were selected. The smoothing parameter of the exposure was chosen by the cross validation method for all the samples in the two methods. In addition, we fitted method (2), but with only three age groups with cut points at 0, 240, 480 and 730 days, to see how a change in age groups affects the results.

To compare the performance of the three methods in terms of estimating the age-specific relative incidence and the exposure related relative incidence we used the mean of the integrated squared errors (MISE) and the standard deviation of the integrated squared errors (SD). For the age effect we constrained the cumulative relative incidence function to be one at the maximum age to make the three methods comparable. The integrated squared error (ISE) for each sample is defined as

$$\int_0^{730} (\Psi(t) - \hat{\Psi}(t))^2 dt,$$

where $\Psi(t)$ is the true age-specific cumulative relative incidence constrained to be one at age 730 days and $\hat{\Psi}(t)$ is the estimated cumulative relative incidence. After fitting the models for each sample we estimated the cumulative relative incidence at each day of age from 0 - 730 and approximated the ISE values as:

$$\sum_{t=0}^{730} (\Psi(t) - \hat{\Psi}(t))^2.$$

We then evaluated the MISE values as the mean of the ISE values of the 100 samples in each scenario and the SD as the standard deviation of the ISE values. Similarly the ISE value for the exposure-related relative incidence is defined as

$$\int_c^{d} (\omega(t-c) - \hat{\omega}(t-c))^2 dt,$$

where $\omega(t-c)$ is the true exposure-related relative incidence function, and $\hat{\omega}(t-c)$ is the estimated relative incidence function. The true exposure-related relative incidence functions used in the simulations are presented in Figure 2.

5.2 *Results of the Simulation Study*

Table 1 presents the MISE and SD for estimating the age and exposure effects using the three methods. The method proposed by Ghebremichael-Weldeselassie *et al* (2015) was fitted twice for each generated data set using 6 and 3 age groups.

The results in Table 1 suggest that the new method performs well. In estimating the age-specific relative incidence function the spline-based method has equivalent performance to method (1) with smooth age effect and has better performance as compared to method (2).

In estimating the exposure-related relative incidence function, the fully spline-based method showed the highest performance as compared to both methods (1) and (2). For method (2),

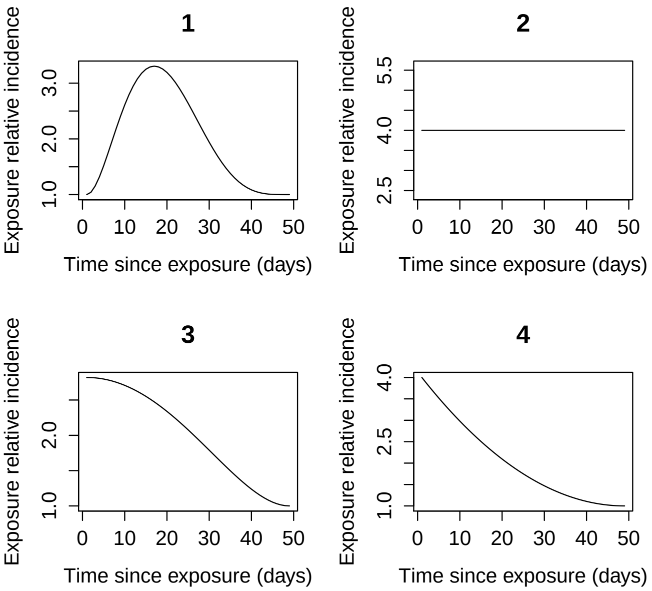

Fig. 2. *True exposure-related relative incidence functions.*

Table 1. *Mean integrated squared errors (MISE) and standard deviation (SD) obtained from the three spline-based SCCS methods: SCCS with smooth age effect, SCCS with smooth exposure effect (twice with 6 and 3 age groups) and SCCS with both age and exposure effects represented by splines. Each simulated data set was fitted by the three methods using a nominal risk period of* 49 *days. The true age-specific relative incidence function was generated from a sine function.*

		Method 1	Method 2 6 age groups	Method 2 3 age groups	Method 3
Scenario	Effects	MISE (SD)	MISE (SD)	MISE (SD)	MISE (SD)
	Exposure	13.182 (6.581)	7.318 (4.792)	7.393 (4.835)	7.220 (4.433)
1	Age	0.110 (0.103)	0.181 (0.086)	1.466 (0.102)	0.110 (0.106)
	Exposure	22.959 (10.249)	10.849 (12.996)	10.507 (12.678)	9.298 (7.188)
2	Age	0.117 (0.105)	0.202 (0.107)	1.483 (0.102)	0.123 (0.106)
	Exposure	9.856 (5.597)	5.438 (6.466)	5.552 (6.597)	4.393 (4.372)
3	Age	0.107 (0.089)	0.187 (0.093)	1.476 (0.111)	0.109 (0.090)
	Exposure	10.007 (4.882)	6.388 (8.451)	6.424 (8.207)	4.890 (6.328)
4	Age	0.126 (0.108)	0.204 (0.103)	1.490 (0.121)	0.129 (0.107)

when the age groups used in modelling the age effect are reduced to three, the performance of the method reduces, which indicates that mis-specification of age groups may lead to a reduced performance of this method. However, for scenario 2 surprisingly the performance increased when

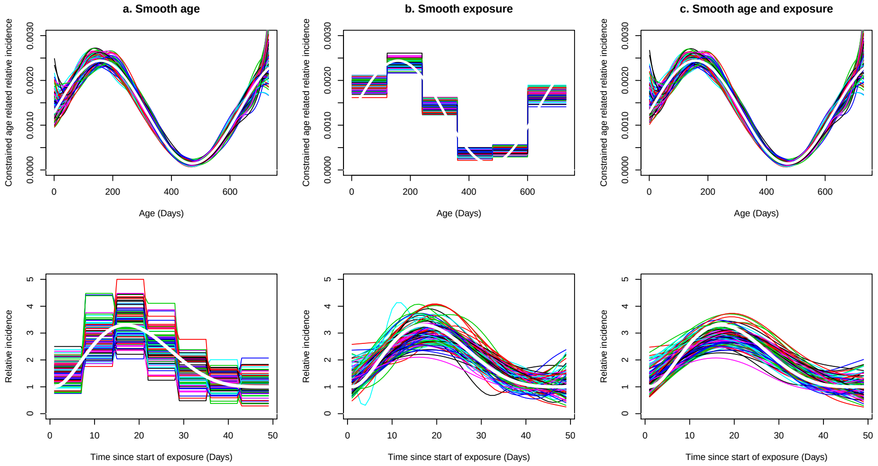

Fig. 3. *Estimated relative incidence curves for scenario* 1*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

the number of age groups is reduced. The spline-based method developed in this paper does not have a limitation related to mis-specification of age and exposure groups.

The estimated age-related and exposure-related relative incidence functions along with their true curves are presented in Figures 3, 4, 5 and 6 for scenarios 1, 2, 3 and 4 respectively (the model with three age groups is not presented). The curves related to the age effect are plotted by constraining the cumulative relative incidence at the maximum of the ages at the end of observation period to be one.

The figures suggest that the spline-based method performs well in estimating both the age and exposure-related relative incidence curves. In all cases the true functions are within the range of the estimated curves and the estimated curves equally follow the trend of the true functions. However there are some estimated exposure-related curves that over-fitted the true curve for scenario 2, (Figure 4), where the true function is a constant.

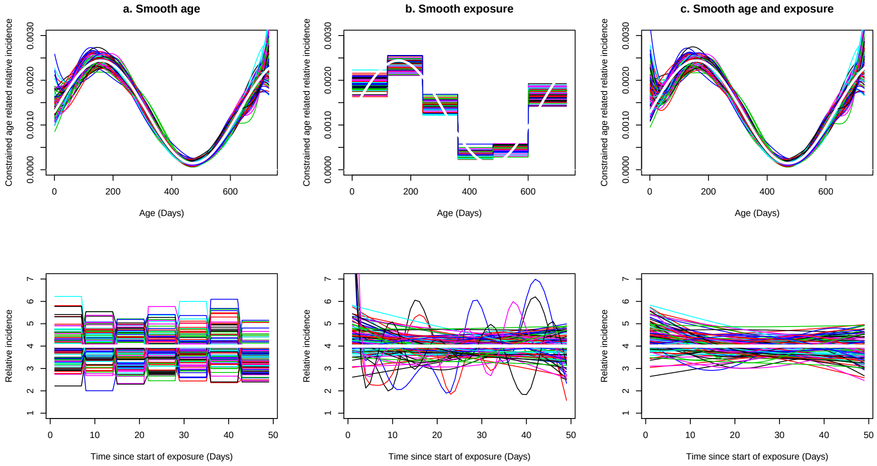

Fig. 4. *Estimated relative incidence curves for scenario* 2*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

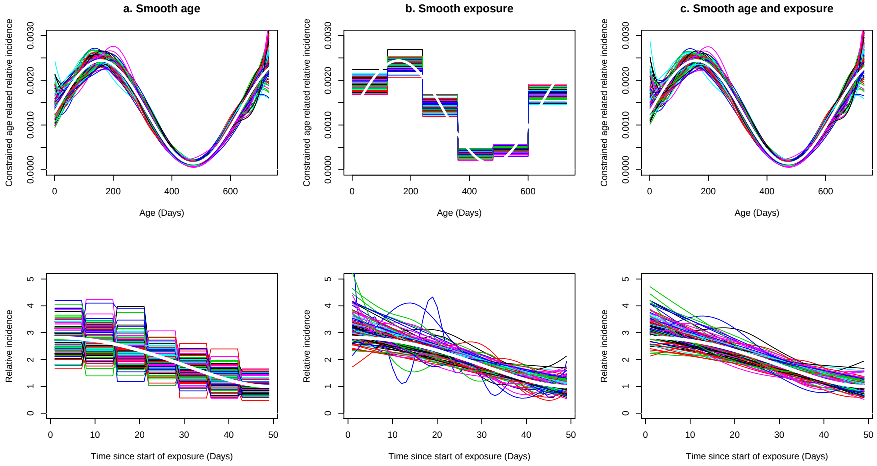

Fig. 5. *Estimated relative incidence curves for scenario* 3*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

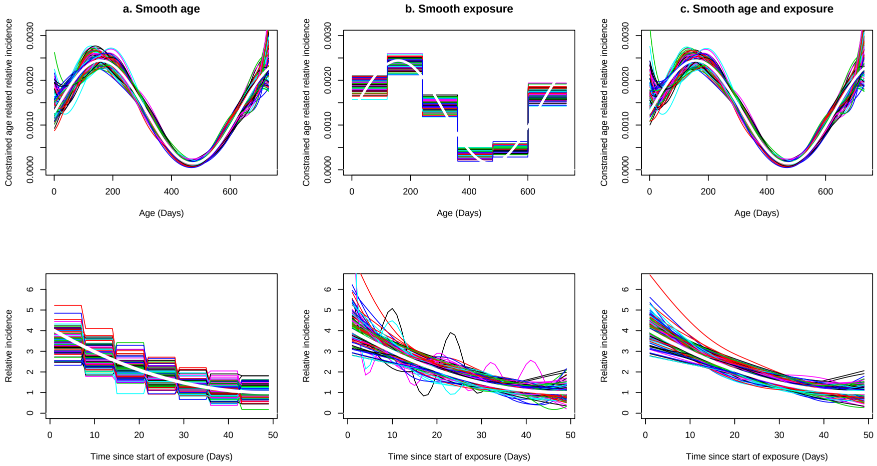

Fig. 6. *Estimated relative incidence curves for scenario 4; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

6. Application

We illustrate the spline-based self-controlled case series method by applying it to data on measles, mumps and rubella (MMR) vaccines and febrile convulsions. The data set includes $2,389$ cases aged between 29 and 730 days with $3,826$ events. The data were collected in England and Wales in the period 1991-1994. The objective was to investigate a potential association between febrile convulsion and exposure to MMR vaccine. We used the spline-based SCCS method developed in this paper where linear combinations of cubic M-splines are used to represent the age and exposure effects. For the MMR vaccine related relative incidence function we chose a nominal risk period of 50 days. We used 12 equally spaced interior knots between 0 and 50. The smoothing parameter λ_2 for the exposure effect was chosen by the cross validation method and was found to be 0.031. For the age-related relative incidence, we used 12 interior knots between 29 and 730 and chose the smoothing parameter using the cross validation method. The value selected was 1.07×10^9. Then for the given values of the smoothing parameters, we maximised the spline-based SCCS

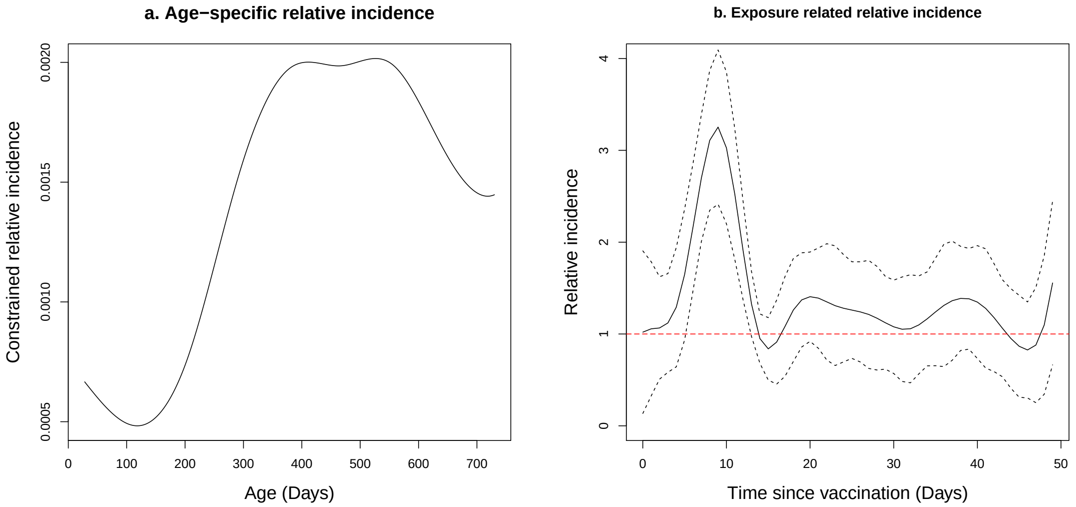

Fig. 7. *Relative incidence curves estimated by fitting spline-based SCCS. Panel (a) shows the estimated constrained age-related relative incidence function Panel (b) represents estimated exposure-related relative incidence curve (solid line) along with 95% confidence bands denoted by the dashed lines.*

penalised log-likelihood function (4.8). The estimated age and exposure-related relative incidence curves are presented in Figure 7.

Panel (a) of Figure 7 shows the estimated age-related relative incidence function, where the cumulative age effect is constrained to have the value one at the maximum end of observation period. Panel (b) of the figure shows the relative incidence curve post MMR vaccine. From the figure, it can be seen that there is a significant increase in the risk of febrile convulsion from six to 12 days after exposure to MMR vaccine. Five and 13 days after vaccination have a borderline non-significant risk of febrile convulsion. There is no increased risk in other periods.

The dotted lines in Panel (b) of Figure 7 are approximate variability bands obtained by using the Bayesian-like technique proposed by O'Sullivan (1988) to generate confidence bands. The 95% coverage probabilities of these confidence bands in the context of the SCCS method were studied in a simulation study by Ghebremichael-Weldeselassie *et al* (2015).

7. Final remarks

The method developed here combines the extensions to the standard SCCS method developed by Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015). In Ghebremichael-Weldeselassie *et al* (2014), only the age effect was approximated by a linear combination of M-spline basis functions and the exposure effect was represented by a piecewise constant function. In Ghebremichael-Weldeselassie *et al* (2015), splines were used only to estimate the exposure-related relative incidence function and age was taken into account based on step functions. In this paper, the effects of both age and exposure in the SCCS model are represented by linear combinations of M-spline basis functions simultaneously. The new method avoids the mis-specification bias that may occur due to poor choice of age and\or exposure groups in the standard and the previous spline based methods due to the use of step functions.

The denominator of the log-likelihood function of the new method includes the integral of a product of two spline functions, namely the age-related and the exposure-related relative incidence functions. Rather than using numerical integration techniques, we evaluated this integral analytically using integration by parts. This required evaluation of the first, second and third integrals of an I-spline function, based on the definition of the integral of an M-spline given by Ramsay (1988).

A simulation study was conducted to evaluate the performance of the new method, spline-based SCCS. It was found that the new method has better performance as compared to the extensions presented in Ghebremichael-Weldeselassie *et al* (2014) and Ghebremichael-Weldeselassie *et al* (2015).

In other versions of the SCCS method either the age effect or the exposure risk period, or both, are represented by step functions that can yield biased estimates if the a priori selected groups are poorly chosen or mis-specified. The new method avoids this. When the exposure risk period is represented by step functions greater accuracy can be achieved by defining several contiguous risk periods, likewise when the age effects are represented by step functions bias is minimised by defining more age groups. The new approach offers greater efficiency in comparison when fewer parameters need to be estimated overall. In addition it offers the advantage of representing the exposure risk with biologically plausible shapes, graphical displays of which may be of particular interest.

8. Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Funding

This work was supported by a Royal Society Wolfson research merit award to Paddy Farrington and Medical Research Council [grant number HGXF A4468].

References

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

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-Applied Statistics* **55**, 553–580.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Self controlled case series method with smooth age effect. *Statistics in Medicine* **33(4)**, 639 – 649.

GHEBREMICHAEL-WELDESELASSIE, Y., WHITAKER, H. J., AND FARRINGTON, C. P. (2015). Flexible modelling of vaccine effect in self-controlled case series models. *Submitted*

JOLY, P., AND COMMENGES, D. (1999). A Penalized Likelihood Approach for a Progressive Three-State Model with Censored and Truncated Data: Application to AIDS. *Biometrics* **55(3)**, 887–890.

JOLY, P., COMMENGES, D., HELMER C., AND LETENNEUR L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3(3)**, 433–443.

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

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.