Existence and uniqueness of relative incidence estimates in case-series analysis

This item was submitted to Loughborough University’s Institutional Repository by the/an author.


Additional Information:

- This article was published in the journal, Computational Statistics and Data Analysis [© Elsevier]. The definitive version is available from: http://www.sciencedirect.com/science/article/pii/S0167947305000319

Metadata Record: https://dspace.lboro.ac.uk/2134/9158

Version: Accepted for publication

Publisher: © Elsevier

Please cite the published version.
This item was submitted to Loughborough’s Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
Existence and uniqueness of relative incidence estimates in case-series analysis

Baibing Li \textsuperscript{a,*}, Xiangning Huang \textsuperscript{b}

\textsuperscript{a}Business School, Loughborough University, Loughborough, LE11 3TU, UK
\textsuperscript{b}Pfizer Global Research and Development, Sandwich, CT13 9NJ, UK

Abstract

Case-series analysis is used to estimate relative incidences of clinical events in defined time intervals after vaccination compared to a control period. It has advantages, in terms of both a reduction in data collection effort, because it uses only data on cases, and a reduction in the resultant variances of estimates, due to individuals being self-controlled. The existence and uniqueness of relative incidence estimates in case-series analysis are investigated. For the relative incidence of a clinical event, a simple condition for existence and uniqueness of the estimate of the parameter vector in a case-series model is established. An algorithm is developed to examine the established condition, which provides a clue for remedy when the condition for existence and uniqueness is not satisfied.

Keywords: Case-series analysis; Existence and uniqueness; Maximum likelihood estimate; Nonlinear programming; Relative incidence

1. Introduction

Evaluation of vaccine safety is an important aspect of vaccine programme surveillance. For example, several research projects have recently investigated the possible relationship between the MMR vaccine and autism (see, e.g. \cite{1,2}).

Most of the existing methods for evaluation are data-intensive, involving large cohorts or the careful selection and matching of controls \cite{3,4}. Farrington \cite{4} developed a case-series model that solely uses data on cases, thus substantially reducing the data collection effort. It combined aspects of the case control and cohort methods, using retrospectively ascertained vaccination histories in cases to estimate the relative incidence of an event on different time intervals after vaccination relative to a control period. Consequently, the resultant variances of estimates reduce due to individuals being self-controlled and individual effects and covariates being canceled out through conditioning. In contrast, in case control studies, individual or strata effects have to be taken into account during modeling and inference; see for instance, Kim and Cohen \cite{5} for a semiparametric/nonparametric modeling approach for effect modification in matched studies. The case-series model has, for example, been applied to assess vaccine safety in a study of intussusception among infants given an oral rotavirus vaccine \cite{6}, and in a study of the association between diphtheria/tetanus/pertussis (DTP) vaccination and febrile convulsion \cite{7}.
The unknown parameter vector of the case-series model developed in [4] consists of a relative incidence part and an association part correcting age effects. After obtaining an estimate that maximizes the log-likelihood kernel of the case-series model, the estimate of log relative incidence can be used for the evaluation of vaccine safety.

The question arises as to whether a relative incidence estimate exists, and if so, whether it is unique. When it does not exist or is not unique, the solution obtained through an optimization algorithm may give misleading indications for vaccine safety evaluation. In general, one does not assume infinite parameter values in underlying populations, and thus infinite parameter estimates are not normally acceptable in practice [8, 9]. The problems of infinite parameter values typically occur with small to medium-sized data sets. One possible solution is to reparameterize the relevant parameters during numerical calculations. However, interpretation of the results in terms of the original parameters, if they are of primary interest, may still be problematic. A well-known example is the separation problem in logistic regression analysis. The separation problem may result in at least one parameter estimate diverging to infinity, and lead to Wald confidence intervals of infinite width [8, 9].

The issue of existence and uniqueness has been given much attention in the statistical literature. Silvapulle and Burridge [10] presented a general condition for the existence of maximum likelihood estimates (MLE) in regression models. Specifically, the MLE of a parameter vector, say \( \mathbf{\theta} \), in a regression model exists if and only if there does not exist a non-zero vector \( \mathbf{u} \) having the same dimension of \( \mathbf{\theta} \) such that \( Z \mathbf{u} \leq 0 \), where \( Z \) is a matrix calculated from observation data. Mathematically this condition is simple and can be verified by linear programming methods [10]. Unfortunately, when an MLE does not exist, this condition does not provide any information for diagnosis and remedy. Recently, for the logistic regression analysis, Heinze and Schemper [9] developed a procedure to remedy the separation problem in logistic regression analysis.

For a case-series model, the issue of diagnosis and remedy is particularly important and useful because the definition of age intervals in a case-series model can be adjusted to a considerable extent, and thus even if the MLE of a parameter vector does not exist under one partition of age intervals, it may exist under another. A diagnosis of the problem may thus provide information for re-defining age intervals. In this paper, we establish a simple condition for the existence and uniqueness of relative incidence estimates in a case-series model. The condition is explicitly associated with observation data and can be checked through a simple algorithm. This algorithm results in some certain diagnostic information and thus provides a clue of how to re-define age intervals for remedy when an MLE does not exist.

2. The case series model of relative incidences

In this section, the case-series model of relative incidences is briefly summarized; see [4] for further details. Suppose that a random sample of size \( N \) from a population is observed in a defined calendar period and over a defined age interval, \( I \). For the \( i \)th individual with a vector of fixed covariates \( \mathbf{x}_i \), let \( c_i \in I \) and \( d_i \in I \) be the ages of the \( i \)th individual at the start and end of observation respectively, and let \( v_i \) be the age of the \( i \)th individual at vaccination if vaccinated prior to age \( d_i \), and \( v_i = +\infty \) if unvaccinated by age \( d_i \). The \( r \)th risk interval for the \( i \)th individual is denoted as \( R_i(v_i) = (v_i + a_r,v_i + b_r) \), where \( a_r \geq 0 \), \( b_r \geq 0 \), \( b_r > a_r \).
(1 ≤ r ≤ s ; s ≥ 1) and \( b_i < a_{r+1} \) (1 ≤ r ≤ s – 1). The control period \( R_0(v_i) \) is defined as \( \bigcup_{r=1}^{s} R_r(v_i) \).

Let \( \lambda_{00}(t) \) denote the baseline of an incidence function. Although in general it is a continuous function of time \( t \), it is approximated by a step function in [4] in order to handle the baseline \( \lambda_{00}(t) \) more easily. Specifically, it is approximately assumed that \( \lambda_{00}(t) \) is a piecewise constant on disjoint age intervals \( T_j = (t_j, t_{j+1}] \) ( \( t_j < t_{j+1} \)) ( \( j = 0, \ldots, m ; m \geq 1 \)) which cover the age interval \( I \):

\[
\lambda_{00}(t) = \exp(\varphi_i + \rho_j) \quad \text{if} \quad t \in T_j,
\]

where \( \varphi_i \) represents individual effects and \( \rho_j \) represents age effects with \( \rho_0 = 0 \). Throughout this paper, we assume that \( T_0 \subseteq R_0(v_i) \) for all \( i \). This holds when, for instance, \( t_i < v_i \) for all \( i \).

Let the set of cutpoints for the \( i \)th individual’s observation period be \( \{v_j ; v_i + a_i ; v_i + b_i ; c_i ; d_i \} \), \( j = 0, \ldots, m ; r = 1, \ldots, s ; t_i, v_i + a_i, v_i + b_i \in [c_i, d_i] \), and \( \Omega_i = \{ \tau_{ik} | k = 0, \ldots, h(i) \} \) be the set of ordered distinct cutpoints in the \( i \)th individual’s observation period, where \( h(i) + 1 \) is the number of elements in \( \Omega_i \), \( \tau_{i0} = c_i \) and \( \tau_{i(h(i))} = d_i \). Let \( y_{ik} = [y_{ik1}, \ldots, y_{ikm}]^T \), where \( y_{ikj} = 1 \) if \( \tau_{ik} \in T_j \) and 0 otherwise ( \( j = 0, \ldots, m \)); \( y_{ik} = 0 \) if \( \tau_{ik} \in T_0 \); \( z_{ik} = [z_{ik1}, \ldots, z_{ikh(i)}]^T \), where \( z_{ikr} = 1 \) if \( \tau_{ik} \in R_r(v_i) \) and 0 otherwise ( \( r = 1, \ldots, s \)); \( z_{ik} = 0 \) if \( \tau_{ik} \in R_0(v_i) \).

Due to equation (1), the incidence in a case-series model for individual \( i \) is assumed to be constant on \( (\tau_{ik-1}, \tau_{ik}] \):

\[
\lambda_{ik} = \exp(\varphi_i) \exp(\gamma^T x_i + \beta^T y_{ik} + \rho^T z_{ik}),
\]

where \( \beta = [\beta_1, \ldots, \beta_s]^T \) is a vector of log relative incidences and \( \rho = [\rho_1, \ldots, \rho_m]^T \) of age effects.

Let \( n_{ik} \) be the number of events the \( i \)th individual experiences on \( (\tau_{ik-1}, \tau_{ik}] \). The likelihood kernel is given by [4]:

\[
L(\beta, \rho) = \prod_{i=1}^{N} \prod_{h(i)} \left[ e_{ik} \exp(\rho^T y_{ik} + \beta^T z_{ik}) / \sum_{j=1}^{h(i)} e_{ij} \exp(\rho^T y_{ij} + \beta^T z_{ij}) \right]^{n_{ik}},
\]

where \( e_{ik} = \tau_{ik} - \tau_{ik-1} > 0 \) is the observation time of the \( i \)th individual on \( (\tau_{ik-1}, \tau_{ik}] \). Note that both individual effects \( \varphi_i \) and covariates \( x_i \) of individuals are canceled out through conditioning.

The notations above can be further simplified as follows. Denote \( u_{ik} = [y_{ik1}, z_{ik1}]^T \) and \( a = [\rho^T, \beta^T]^T \). Let \( P_{ik}(a) = e_{ik} \exp(a^T u_{ik}) / \sum_{j=1}^{h(i)} e_{ij} \exp(a^T u_{ij}) \). Then for any \( i \) (1 ≤ i ≤ N), we have

\[
\sum_{k=1}^{h(i)} P_{ik}(a) = 1 \quad \text{and} \quad P_{ik}(a) \geq 0 \quad \text{for} \quad k = 1, \ldots, h(i).
\]
Equation (2) can thus be rewritten as \( L(\alpha) = \prod_{n=1}^{N} \prod_{k=1}^{h(i)} [P_{ik}(\alpha)]^{n_k} \) and the log-likelihood kernel is \( \log[L(\alpha)] = \sum_{i=1}^{N} \sum_{k=0}^{h(i)} n_{ik} \log[P_{ik}(\alpha)]. \) The estimate of the unknown parameter vector of the above model, \( \hat{\alpha} \), is a solution to the following optimization problem:

\[
\min_{\alpha} -\log[L(\alpha)].
\] (4)

3. Existence and uniqueness

In this section we investigate the existence and uniqueness of relative incidence estimates in case-series analysis, i.e. the existence and uniqueness of solutions to problem (4). First we state a lemma.

**Lemma 1.** The function \( -\log[L(\alpha)] \) is convex.

**Proof.** From the chain rule, for any \( k = 1, \ldots, h(i) \) and any \( i = 1, \ldots, N \), we have

\[
\partial^2 \{-\log(P_{ ik}(\alpha))\} / \partial \alpha^2 = \sum_{l=1}^{h(i)} u_{ il} P'_{ il}(\alpha) - \sum_{l=1}^{h(i)} u_{ il} P_{ il}(\alpha) \sum_{l=1}^{h(i)} u_{ il}^T P_{ il}(\alpha).
\]

For any \((m + s) \times 1\) vector \( \pi \), let \( \xi_{ il} = u_{ il}^T \pi \). Then for two sequences \( \{\xi_{ il} P_{ il}^{1/2}(\alpha)\} \) and \( \{P_{ il}^{1/2}(\alpha)\} \) \((l = 1, \ldots, h(i))\), from the Cauchy’s inequality (see, e.g. [11, p30]), we have

\[
\sum_{l=1}^{h(i)} \xi_{ il}^2 P_{ il}(\alpha) \sum_{l=1}^{h(i)} P_{ il}(\alpha) \geq \left| \sum_{l=1}^{h(i)} \xi_{ il} P_{ il}(\alpha) \right|^2. \]

Hence, combining this inequality with equation (3), we obtain

\[
\pi^T \partial^2 \{-\log(P_{ ik}(\alpha))\} / \partial \alpha^2 \pi = \sum_{l=1}^{h(i)} \xi_{ il}^2 P_{ il}(\alpha) - \left| \sum_{l=1}^{h(i)} \xi_{ il} P_{ il}(\alpha) \right|^2 \geq 0. \] (5)

This implies that function \( -\log(P_{ ik}(\alpha)) \) is convex, and thus \( -\log(L(\alpha)) = \sum_{i=1}^{N} \sum_{k=1}^{h(i)} -\log[P_{ ik}(\alpha)] \) is also a convex function. This completes the proof.

From Lemma 1, to ensure that solutions to problem (4) are finite, we need to find an appropriate condition for boundedness; and to show the uniqueness of the solution, we need to show that function \( -\log[L(\alpha)] \) is strictly convex under such a condition. These will be established in Theorem 1 and Theorem 2 respectively later in this section.

Let \( L_0 = L(\hat{\alpha}) \) which is positive since \( e_{ ik} = \tau_{ ik} - \tau_{ ik-1} > 0 \). Hence, an optimal solution to problem (4), \( \hat{\alpha} \), satisfies \( L(\hat{\alpha}) \geq L_0 > 0 \). Define \( S_0 = \{\alpha | L(\alpha) \geq L_0 / 2\} \). Clearly \( S_0 \neq \emptyset \) since \( 0 \in S_0 \), where \( \emptyset \) denotes an empty set. Note that any \( \alpha \in R^{m+1} \setminus S_0 \) satisfies \( L(\alpha) < L_0 / 2 \), thus not being an optimal solution to (4). Consequently, it is sufficient to search for the optimal solutions only in set \( S_0 \). Problem (4) is then equivalent to

\[
\min_{\alpha \in S_0} -\log[L(\alpha)].
\] (6)

Next, define a set

\[ S = \{\alpha | P_{ ik}(\alpha) \geq (L_0 / 2) \text{ for } n_{ ik} > 0, k = 1, \ldots, h(i); i = 1, \ldots, N\}. \]

Clearly, we have \( S_0 \subseteq S \).
For $\alpha \in S$ and any given $i$ and $q$ satisfying $n_{iq} > 0 \ (1 \leq q \leq h(i) ; 1 \leq i \leq N)$, if existing $p \ (1 \leq p \leq h(i) ; \ p \neq q)$ such that $\alpha^T (u_{iq} - u_{pq}) \to +\infty$, we have

$$P_{iq}(\alpha) = 1/[1 + (e_p / e_q) \exp{\{\alpha^T (u_p - u_q)\}}] + \sum_{l=p,l \neq q} (e_l / e_q) \exp{\{\alpha^T (u_p - u_q)\}} \to 0,$$

which contradicts the assumption that $\alpha \in S$ , i.e. $P_{iq}(\alpha) \geq (L_0 / 2) > 0$ for $n_{iq} > 0$. Hence, we have the following important property for set $S$:

**Lemma 2.** For any given $i$ and $q$ satisfying $n_{iq} > 0 \ (1 \leq q \leq h(i) ; 1 \leq i \leq N)$, $\alpha \in S$ implies

$$\alpha^T (u_{iq} - u_{pq}) < +\infty \quad \text{for all} \ p \ (1 \leq p \leq h(i) ; \ p \neq q).$$

### 3.1. A condition of existence and uniqueness

A simple example of non-uniqueness of the solutions to problem (6) is that there exists a risk interval that is always coincident with a certain age interval, resulting in the corresponding parameters of the relative incidence and of the age effect being unestimatable. In this subsection, we consider a condition for existence and uniqueness.

Intuitively, if any of the intervals which is associated with an unknown parameter to be estimated does not have any observed individual, this unknown parameter is not estimable. The question is, if each of such intervals has at least one observation, is this sufficient to guarantee the existence and uniqueness of the solution to problem (6)? Such a condition has to be presented precisely and the existence and uniqueness under this condition has to be established mathematically.

Roughly speaking, the following condition requires that for each of the age intervals there is at least one observed individual experiencing an event in the intersection of the control period and that age interval, and there is at least one observed individual experiencing an event on each of the risk intervals. In addition, since the idea of case-series analysis is to estimate relative incidences on risk intervals compared to a control period, the observation interval of the individual should cover at least part of the control period.

Now we state the condition below.

**Condition.**

(i) For any age interval $T_j \ (j > 0)$, there exists an observed individual $i \ (1 \leq i \leq N)$ experiencing an event on $T_j \cap R_0(v_i)$ and the observation interval $[c_i,d_i]$ covers part of $T_0$;

(ii) There exists an observed individual $i \ (1 \leq i \leq N)$ experiencing an event on $T_0$ and any risk interval of the individual does not completely cover any age interval, i.e. $[c_i,d_i] \cap T_j \cap R_0(v_i) \neq \emptyset$ for all $j > 0$.

(iii) For any given $r \ (1 \leq r \leq s)$, there exists an observed individual, $i \ (1 \leq i \leq N)$, experiencing an event on $R_r(v_i)$ and the observation interval of the individual covers at least part of the control period;

(iv) There exists an observed individual, $i$, associated with an observation interval $[c_i,d_i]$ which overlaps all risk intervals $R_r(v_i)$ $(r = 1,\ldots,s)$, experiencing an event on $R_0(v_i)$.

It is clear that part (i) of the condition excludes the situation where there exists a risk interval that is always coincident with a certain age interval.

### 3.2. Existence and uniqueness of estimates

In this subsection, we first investigate the boundedness of solutions, which is summarized in Theorem 1. Then, we will show that function $-\log[L(\alpha)]$ is strictly convex under the
condition given above, which is important to ensure the uniqueness of the solution. Finally, the existence and uniqueness of the solution will be established in Theorem 3.

**Lemma 3.** (i) For any $\alpha \in S$, if parts (i) and (ii) of the condition hold, then $|\alpha_j| = |\beta_j| < +\infty$ for $j$ satisfying $0 < j \leq m$.

(ii) For any $\alpha \in S$, if parts (i)-(iv) of the condition hold, then $|\alpha_{m+r}| = |\beta_r| < \infty$ for $r$ satisfying $1 \leq r \leq s$.

The proof of Lemma 3 is given in the Appendix. From the above lemma, we obtain the following theorem immediately:

**Theorem 1.** If the condition holds, then set $S$ is bounded.

Next, for any given $i$ and $q$ satisfying $n_q > 0$, let $\Delta_i$ be the matrix with its rows consisting of $(u_{pq} - u_{iq})^T$ $(p = 1, ..., h(i); \; p \neq q)$, and let $\Lambda = [\Lambda_i^T, ..., \Lambda_N^T]^T$. Note that for recurrent events, there may exist multiple indexes $q$ satisfying $n_q > 0$ for the same individual $i$.

**Lemma 4.** If the condition holds, then matrix $\Delta$ is of full column-rank.

See the Appendix for proof.

**Theorem 2.** If the condition holds, then

(i) function $-\log[L(\alpha)]$ is strictly convex in set $S$;

(ii) set $S_0$ is a convex set.

**Proof.** (i) From the proof of Lemma 1, we have

$$\sum_{k=1}^{h(i)} \xi_{ik}^2 P_{ik}^T (\alpha) \sum_{k=1}^{h(i)} P_{ik} (\alpha) \geq \left[ \sum_{k=1}^{h(i)} \xi_{ik} P_{ik} (\alpha) \right]^2$$

with equality holding if and only if the sequences $\{\xi_{ik} P_{ik}^{1/2} (\alpha)\}$ and $\{P_{ik}^{1/2} (\alpha)\}$ $(k = 1, ..., h(i))$ are proportional ([11, p30]), i.e. $\xi_{ik} = u_{ik}^T \pi = C_i$ $(k=1, ..., h(i))$, for a constant $C_i$. Hence, we obtain

$$(u_{ip} - u_{iq})^T \pi = 0$$

for $p = 1, ..., h(i)$ and $p \neq q$, where $n_q > 0$ or equivalently, $\Delta \pi = 0$.

From Lemma 4, we conclude that $\pi = 0$. Hence the equality in equation (5) holds if and only if $\pi = 0$ in the proof of Lemma 1. This implies that $\partial^2 [-\log[L(\alpha)]] / \partial \alpha^2$ is positive definite and function $-\log[L(\alpha)]$ is strictly convex.

(ii) Since $-\log[L(\alpha)]$ is a strictly convex function in $S_0 \subseteq S$, set $S_0$ is a convex set (see, e.g. [13, p88]). This completes the proof.

Combining Theorems 1 and 2, we obtain the following main result:

**Theorem 3.** If the condition holds, then the solution to problem (6) exists and is unique.

**Proof.** It is clear that set $S_0$ is closed. The set $S_0$ is also bounded since $S_0 \subseteq S$ and $S$ is bounded from Theorem 1. Hence, $S_0$ is a compact set ([12], p172). The continuous function, $-\log[L(\alpha)]$, then attains a minimum in the compact set $S_0$ ([12], p170) at, say, $\hat{\alpha} \in S_0$. Clearly, $\hat{\alpha}$ is finite and is not a boundary point of $S_0$. Finally, from Theorem 2, the uniqueness follows by noting that problem (6) is a convex programming problem [13].

4. An algorithm

In the previous section, we defined $\Lambda = [\Lambda_1^T, ..., \Lambda_N^T]^T$, where $\Lambda_i$ is a matrix with its rows consisting of $(u_{ip} - u_{iq})^T$ $(p = 1, ..., h(i); \; p \neq q)$ for any given $q$ satisfying $n_q > 0$. Let $\Lambda_{ik}$ be
the $k$th row of the matrix $\Delta_i$, and $\delta_{ik}$ be the last $s$ entries of $\Delta_i$. Let $\varepsilon_{ik}$ denote an $l \times 1$ vector having 1 at the $k$th entry and 0 otherwise.

From the proof of Lemma 3 in the Appendix, it can be seen that if parts (i) and (ii) of the condition are satisfied for an age interval $j$, then there exists a row vector,

$$(u_{jp} - u_{jq})^T = [-\varepsilon_{mj}, 0]^T = -\varepsilon_{(m+s)j}^T,$$

for some $i$, $p$, and $q$; and there exists a row vector,

$$(u_{jp'} - u_{jq'})^T = [\varepsilon_{mj}, 0]^T = \varepsilon_{(m+s)j}^T,$$

for some $i'$, $p'$, and $q'$. From the proof of Lemma 3, these two row vectors guarantee the boundedness of the corresponding parameter $\rho_j$. Hence, if we can find out such row vectors in matrix $\Delta$, this indicates that parts (i) and (ii) of the condition hold for $j$. We can similarly check if other parts of the condition are satisfied.

An algorithm is given below to detect whether the condition holds. The vectors $A_1$, $A_2$, $A_3$ and $A_4$ are used to indicate on which of these intervals each part of the condition, (i), (ii), (iii) and (iv), fails to hold. For example, if the third element of $A_1$ is zero, it indicates that part (i) of the condition does not hold for $j=3$.

**Given Data:** $\Delta = [\Delta_1^T, \Delta_2^T, \Delta_3^T, \Delta_4^T]^T$.

**LET** $A_1 = [a_1(1), \ldots, a_m(1)]^T = 0_{m \times 1}, A_2 = [a_1(2), \ldots, a_m(2)]^T = 0_{m \times 1}, A_3 = [a_1(3), \ldots, a_m(3)]^T = 0_{m \times 1}, A_4 = [a_1(4), \ldots, a_m(4)]^T = 0_{m \times 1}$.

**A = \prod_{j=1}^{m} a_j(1) a_j(2) \prod_{j=1}^{i} a_j(3) a_j(4).**

**WHILE** {$i \leq N$ AND $A = 0$}

**WHILE** {$k \leq \text{number of rows of } \Delta_i$ AND $A = 0$}

IF $\Delta_{ik} = -\varepsilon_{(m+s)q}^T$ for some $q$ ($1 \leq q \leq m$), then LET $a_q(1) = 1$;

ELSEIF $\Delta_{ik} = \varepsilon_{(m+s)q}^T$ for some $q$ ($1 \leq q \leq m$), then LET $a_q(2) = 1$.

ENDIF

IF $\delta_{ik} = -\varepsilon_{sq}^T$ for some $q$ ($1 \leq q \leq s$), then LET $a_q(3) = 1$;

ELSEIF $\delta_{ik} = \varepsilon_{sq}^T$ for some $q$ ($1 \leq q \leq s$), then LET $a_q(4) = 1$.

ENDIF

LET $k = k + 1$; UPDATE $A$.

END LOOP $k$.

LET $i = i + 1$; $k = 1$; UPDATE $A$.

END LOOP $i$.

When the algorithm is terminated and $A = 0$, it implies that some parts of the condition fail to hold. Next, we discuss the issue of remedy in such a case.

Recall that in equation (1) the continuous function $\hat{\lambda}_0(t)$ of the baseline incidence is approximated by a step function on disjoint age intervals $T_j = (t_j, t_{j+1})$ ($t_j < t_{j+1}$) ($j = 0, \ldots, m; m \geq 1$) which cover the age interval $I$:

$$\hat{\lambda}_0(t) = \exp(\phi_i + \rho_j) \quad \text{if} \quad t \in T_j.$$ 

Since an approximation of the baseline incidence function as above is solely for the purpose of numerical computation, it allows us considerable flexibility in the choice of different partitions of age intervals. In particular, when some parts of the condition fail to hold, one simple solution is to re-define age intervals by enlarging a “problematic” age interval so as to ensure that some necessary observations fall within it. The vectors $A_1$, $A_2$, $A_3$, $A_4$ may be
used to indicate on which intervals the condition fails to hold and thus which intervals need to be re-defined.

5. An example

In this section, we use a simple hypothetical example to illustrate the condition and the related issues of diagnosis and remedy when the condition fails to hold.

Consider a sample of twelve children as displayed in Table 1, where the first row is the ID numbers of individuals, the second and third rows give the time points at which individuals experienced an event and got vaccination respectively, and the fourth row provides risk intervals for each of individuals. It is assumed that the period 11-20 days after vaccination was defined as a risk interval, i.e. $a_1 = 11$ and $b_1 = 20$, and $R_i(v_i) = (v_i + 11, v_i + 20]$. All remaining time periods were included in the control period $R_0(v_i)$. The age intervals were defined by $t_0 = 365$ (days), $t_1 = 400$, and $t_2 = 430$. For simplicity, suppose that the observation interval $[c_i, d_i]$ of individual $i$ was the same as $[t_0, t_2]$.

For this example, it is easy to verify that part (i) of the condition does not hold since for the age interval $T_i = (t_1, t_2)$, there does not exist an observed individual $i$ (1 ≤ $i$ ≤ 12) experiencing an event on $T_i \cap R_i(v_i)$. Intuitively, this may cause the log relative incidence $\beta_i$ and the age effect $\rho_i$ to be unestimable. In fact, the likelihood kernel attains the maximum as the estimates satisfy $\hat{\beta}_i = -\hat{\rho}_i$ and $\hat{\rho}_i \to -\infty$. Hence, the maximum likelihood estimates do not exist.

Table 1.
Events and Vaccination (days)

<table>
<thead>
<tr>
<th>Individual ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced event</td>
<td>375</td>
<td>380</td>
<td>384</td>
<td>390</td>
<td>397</td>
<td>398</td>
</tr>
<tr>
<td>Vaccination</td>
<td>391</td>
<td>391</td>
<td>395</td>
<td>398</td>
<td>400</td>
<td>407</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual ID</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced event</td>
<td>402</td>
<td>406</td>
<td>407</td>
<td>415</td>
<td>416</td>
<td>420</td>
</tr>
<tr>
<td>Vaccination</td>
<td>390</td>
<td>390</td>
<td>395</td>
<td>402</td>
<td>400</td>
<td>402</td>
</tr>
<tr>
<td>Risk interval</td>
<td>(401, 410]</td>
<td>(401, 410]</td>
<td>(406, 415]</td>
<td>(413, 422]</td>
<td>(411, 420]</td>
<td>(413, 422]</td>
</tr>
</tbody>
</table>

Fig. 1. Configuration of age intervals

Note that for this example, $h(i) = 4$, $u_{i1} = [0, 0]^T$, $u_{i2} = [1, 0]^T$, $u_{i3} = [1, 1]^T$ and $u_{i4} = [1, 0]^T$ for 1 ≤ $i$ ≤ 12. In addition, $n_{i1} = 1$ for 1 ≤ $i$ ≤ 6, whilst $n_{i3} = 1$ for 7 ≤ $i$ ≤ 12. After applying the algorithm in section 4 to check this data set, we obtain $a_1(11) = 0$, indicating that part (i) of the condition does not hold for the age interval $T_i = (t_1, t_2)$. Based on this diagnostic
information, a simple remedy may be decided to shift \( t_1 \) towards \( t_0 \) such that \( T_i \cap R_0(v_i) \neq \emptyset \) for some individual \( i \), for instance, by letting \( t_1 = \tau^n = 396 \).

References


Appendix

Proof of Lemma 3. Let \( \varepsilon_\hat{k} \) be an \( l \times 1 \) vector having 1 at the \( k \)th entry and 0 otherwise. We first prove part (i) of Lemma 3. Suppose that part (i) of the condition holds, and an individual, \( i \), experienced an event in \( (\tau_{iq-1}, \tau_{iq}] \subseteq T_j \cap R_0(v_j) \). Then \( n_{iq} > 0 \), and \( \tau_{iq} \) is the right endpoint of either \( T_j \) or \( R_0(v_j) \), i.e. \( \tau_{iq} \) is either \( t_{j+1} \) or the left endpoint of a risk interval. Hence, \( u_{iq} = [e_{mj}^T, 0^T]^T \).

In addition, since the observation interval \([c_i, d_j]\) covers part of \( T_0 \), let \( \tau_{iq} \) be the right endpoint of \( T_0 \), i.e. \( \tau_{iq} = t_i \). This implies that \( u_{iq} = 0_{(m+p+1)} \). Hence by noting \( n_{iq} > 0 \) and \( u_{iq} - u_{iq} = [e_{mj}^T, 0^T]^T \), from Lemma 2 we obtain \( \alpha^T(u_{iq} - u_{iq}) = -\alpha_j = -\rho_j < +\infty \).

When part (ii) of the condition holds, we can similarly prove that there exist \( p \) and \( q \) (with \( n_{iq} > 0 \)) such that \( u_{iq} - u_{iq} = [e_{mj}^T, 0^T]^T \) for age interval \( j \). Then from Lemma 2 we have \( \alpha^T(u_{iq} - u_{iq}) = \alpha_j = \rho_j < +\infty \). This completes the proof of (i).
Next, we prove part (ii) of Lemma 3. Suppose that part (iii) of the condition holds, and an individual, \(i\), experienced an event during \((\tau_{iq-1}, \tau_{iq}] \subseteq R_q(v_i)\) for some \(q\). We suppose that \(\tau_{iq}\) lies in, say, the \(j\)th age interval \(T_j\). Then \(n_{iq} > 0\), and \(\tau_{iq}\) is either \(b_i + v_i\) or \(t_{j+1}\). This implies that \(u_{iq} = [e_{iq}^T, e_{iq}^\tau]^T\).

In addition, from part (iii) of the condition, the observation interval of the individual covers at least part of the control period, i.e. \(R_q(v_i) \cap [c_i, d_i] \neq \emptyset\). Let \(\tau_{iq}\) denote a cutpoint belonging to \(R_q(v_i) \cap [c_i, d_i]\) which lies in, say, the \(k\)th age interval, \(\tau_{iq} \in T_k\). Then \(u_{iq} = [e_{ik}^T, 0]^T\). Hence, noting \(n_{iq} > 0\) and \(u_{iq} - u_{iq} = [e_{ik}^T - e_{iq}^T, e_{iq}^\tau]^T\), from Lemma 2 we have \(a^T (u_{iq} - u_{iq}) = \rho_k - \rho_j - \beta_r < +\infty\). Since from part (i) of Lemma 3 we have that \(\rho_i\) are bounded for all \(i\), we conclude that \(-\beta_r < +\infty\).

Finally, suppose that part (iv) of the condition holds, we can similarly prove that there exists \(q\) with \(n_{iq} > 0\) such that \(u_{iq} = [e_{iq}^T, 0]^T\) for some age interval \(T_k\), and there exists \(p\) such that \(u_{ip} = [e_{ip}^T, e_{ip}^\tau]^T\) for some age interval \(T_j\). Since \(u_{ip} - u_{iq} = [e_{iq}^T - e_{ip}^T, e_{iq}^\tau]^T\), from Lemma 2 we obtain \(a^T (u_{ip} - u_{iq}) = \rho_j - \rho_k + \beta_r < +\infty\), which implies \(\beta_r < +\infty\) since \(\rho_i\) are bounded for all \(i\). This completes the proof.

**Proof of lemma 4.** For any \((m+s)\times1\) vector \(\pi\), we will show that \(\pi = 0\) if \(\Delta \pi = 0\). From the proof of Lemma 3, for any \((1 \leq j \leq m)\), there exists an individual \(i\) with two cutpoints, \(\tau_{iq}\) and \(\tau_{iq}\) (\(n_{iq} > 0\)), such that \(u_{iq} = [e_{iq}^T, 0]^T\) and \(u_{iq} = 0\). Hence, from \(\Delta \pi = 0\) we have \((u_{ip} - u_{iq})^T \pi = \pi_j = 0\) (\(1 \leq j \leq m\)). Similarly, we can prove that \(\pi_{m+r} = 0\) for \(1 \leq r \leq s\). Hence, \(\Delta\) is of full column-rank if the condition holds.