Within-individual dependence in self-controlled case series models for recurrent events

C. Paddy Farrington and Mounia N. Hocine
Department of Mathematics and Statistics, The Open University,
Milton Keynes MK7 6AA, UK

June 25, 2009

Abstract

The self-controlled case series model may be used to analyse recurrent events which are conditionally independent given fixed or random individual effects. In order to test the hypothesis of within-individual independence, the model is augmented by an association parameter for diagonal dependence, which provides the focus for a test of independence. Estimation methods are described, and simulations are presented to illustrate the power of the method in relevant scenarios, and to quantify the bias resulting from failure of the independence assumption. The methods are applied to two data sets, relating to a rare bleeding disorder and to myocardial infarction.

Key words: diagonal dependence, Poisson process, recurrent event, self-controlled case series method.

Corresponding Author: Paddy Farrington, Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK.

Email: c.p.farrington@open.ac.uk.

1 Introduction

The self-controlled case series method was originally developed to investigate the association between time-varying exposures, usually of short duration, and rare acute events (Farrington 1995). The method has been used extensively in pharmacoepidemiology, particularly to investigate rare adverse events in relation to childhood vaccination (Whitaker et al 2006). Recently, a semiparametric version of the method has been proposed (Farrington & Whitaker 2006).

The case series model is derived from a Poisson cohort model by conditioning on the total number of events experienced by each individual in the cohort over a pre-determined observation period. Thus, the method applies to recurrent events, arising within individuals in a non-homogeneous Poisson process whose rate parameter might depend on individual factors, fixed or random. In practice, many events of interest are rare and non-recurrent. The method can also be used in this case, with an approximation that becomes ignorable as the underlying event rate tends to zero (Farrington 1995, Farrington & Whitaker 2006).

In some situations, however, recurrences are important, and relatively frequent. Later in the paper, we present two examples where this is the case. The first relates to the rare bleeding disorder idiopathic thrombocytopenic purpura (ITP). While ITP is rare, the incidence in individuals varies greatly, so that some individuals experience frequent recurrences. It is not clear whether such recurrences are independent within individuals. The second example relates to myocardial infarction (MI): occurrence of a first MI increases the risk of further MIs, so recurrences are not independent within individuals in this case.

In analyses with the self-controlled case series method, it is required that recurrent events should be independent within individuals. In particular, occurrence of an event should not affect the rate at which subsequent events may occur. If this assumption is invalid, or its validity is uncertain, a reasonable analysis strategy may be to limit the analysis to first events (Whitaker et al 2006). However, within-individual dependence may itself be of interest, and it would in any case be useful to evaluate the strength of evidence against independence. Furthermore, when studying the effect of an exposure on rare events, it might be desirable to include all such events in order to maximize the efficiency of the exposure effect estimator.

There is a substantial literature on modelling strategies for the various types of dependence that may arise in the study of recurrent events. Cook & Lawless (2007) provide a comprehensive discussion of these methods in survival analysis. However, once we move beyond the non-homogeneous Poisson process, possibly with time-invariant frailties, conditioning on the total number of events to eliminate non-cases no longer factors out multiplicative time-invariant covariates. Thus, the potential for bias due to unmeasured confounders, which the case series approach sought to reduce, is reintroduced. At most, one may hope that a suitable analysis strategy will prove robust to mis-specification of the model. Whatever the analysis strategy, it is important to check the validity or otherwise of the assumption of within-individual independence.

Hocine et al (2005) developed a bivariate version of the case series method, using a copula representation, to study dependence between different types of events, with application to antibiotic resistance (Hocine et al 2007). Hougaard (2000) discusses several measures of dependence in the context of multivariate survival data. However, these measures apply to parallel data, rather than recurrent events for which statistical dependence is present simply by virtue of the order constraints imposed on the sample space (a second event, if it occurs, having necessarily to follow the first). Furthermore, different individuals will experience different numbers of events, so that the dimension of the space is not fixed. These considerations lead us to develop a test specially designed to investigate within-individual dependence of recurrences within the context of the self-controlled case series model.

In Section 2, we describe the case series model and a simple extension of it, using a one-parameter dependence function, to enable testing of the independence assumption. In Section 3 we provide two further interpretations of this extended model. In Section 4 we discuss models for the dependence function. Estimation and tests are discussed in Section 5, including a simple method which can be used for single recurrences within the context of the standard self-controlled case series method. In Section 6 we study the performance of the method using simulations. We apply the proposed methods to the ITP and MI data in Section 7. Finally, analysis strategies are discussed in Section 8.

2 The self-controlled case series model and a simple extension

We consider events occurring within specified boundaries of age and time, which define which events are ascertained. These boundaries determine individual observation periods of the form $(a_i, b_i]$ for individual $i = 1, 2, ..., N$. For simplicity, we shall use age as the primary time line, though other choices (notably calendar time) might be relevant in other applications. Within their observation period, individuals also experience age-dependent exposures. Let $x_i(t)$ denote the vector of exposures experienced by individual i at age t.

2.1 The case series likelihood

Over the period $(a_i, b_i]$, individual i experiences events that arise with intensity process $\lambda_i(t|x_i(t))$. The likelihood that individual i experiences n_i events at times $t_{i1}, ..., t_{in_i}$ (which for the moment we regard as unordered and independent) is

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

Two fundamental assumptions of the case series method is that the exposure process is exogenous and that censoring of observation at b_i occurs completely at random; see Farrington & Whitaker (2006) for further details, and Roy et al (2006) and Farrington et al (2008) for ways round these assumptions. They enable us to condition on the entire exposure history over $(a_i, b_i]$, which we denote x_i. Note that the process of events for individual i is then a non-homogeneous Poisson process with rate $\lambda_i(t|x_i)$. Conditioning on the number of events n_i experienced by individual i in $(a_i, b_i]$ yields the following conditional likelihood:

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

This is the self-controlled case series likelihood. If $n_i = 0$ then $L_i^c = 1$, and consequently individuals who experience zero events contribute trivially and hence need not be sampled: the conditional likelihood can be evaluated from a sample of cases. Hence the name *case series* method. Furthermore, suppose that

$$\lambda_i(t|x_i) = \varphi\psi(t) \exp\left\{\gamma_i + x_i(t)^T \beta\right\} \tag{2}$$

where φ is the baseline incidence at some reference age, $\psi(t)$ is an age effect common to all individuals, γ_i is the sum of random and fixed effects, possibly covariate-dependent, and β is the log relative incidence associated with the exposure. Then

$$L_i^c = \frac{\prod_{j=1}^{n_i} \psi(t_{ij}) \exp\left\{x_i(t_{ij})^T \beta\right\}}{\left\{ \int_{a_i}^{b_i} \psi(t) \exp\left\{x_i(t)^T \beta\right\} dt \right\}^{n_i}}. \tag{3}$$

Thus the term $\varphi \exp(\gamma_i)$ cancels. It follows that the effects of all time-invariant covariates acting multiplicatively on the Poisson rate are necessarily controlled, and cannot confound the analysis: for this reason the method is called the *self-controlled* case series method.

Very often, the exposure variables $x_i(t)$ are age-dependent indicator variables, though continuous exposures may also be used (Whitaker, Hocine & Farrington 2006). The age effect $\psi(t)$ may be represented by a low-dimensional parametric model, such as a step function (Farrington 1995), or modelled non-parametrically (Farrington & Whitaker 2006).

2.2 A simple extension

Note that the conditional likelihood (1) can be written as a product of n_i terms of the form

$$\frac{\lambda_i(t_{ij}|x_i)}{\int_{a_i}^{b_i} \lambda_i(t|x_i)dt}, \quad j = 1, ..., n_i.$$

In other words, the n_i events within individual i are conditionally independent. It follows that representations of clustered recurrences using frailty terms, as often done in the survival literature to accommodate heterogeneity between individuals (Hougaard 2000; Cook & Lawless 2007), will not induce dependence in the case series model, since events remain independent within clusters.

In order to test the hypothesis of within-individual independence, we embed the model in a wider family indexed by an association parameter θ; the value $\theta = 0$ will correspond to independence within individuals, and the test of independence is formulated as a test of $\theta = 0$. Note that the likelihood contribution (1) for an individual i with $n_i = k$ events, ordered as $t_{i1} < ... < t_{ik}$ may be written equivalently

$$L_i^c = \frac{\lambda_i(t_{i1}|x_i) \times ... \times \lambda_i(t_{ik}|x_i)}{\displaystyle\int_{a_i}^{b_i}\int_{t_1}^{b_i} ... \int_{t_{k-1}}^{b_i} \lambda_i(t_k|x_i) dt_k ... dt_2 dt_1}.$$

We extend this expression by introducing a non-negative dependence function $H_k(t_1, ..., t_k; \theta)$, for which $H_k(t_1, ..., t_k; 0) \equiv 1$, so that the likelihood contribution becomes

$$L_i^k = \frac{\lambda_i(t_{i1}|x_i) \times ... \times \lambda_i(t_{ik}|x_i) \times H_k(t_{i1}, ..., t_{ik}; \theta)}{\displaystyle\int_{a_i}^{b_i} \lambda_i(t_1|x_i) \int_{t_1}^{b_i} \lambda_i(t_2|x_i) ... \int_{t_{k-1}}^{b_i} \lambda_i(t_k|x_i) H_k(t_1, ..., t_k; \theta) dt_k ... dt_2 dt_1}. \quad (4)$$

This is a valid likelihood, which reduces to the case series likelihood when $\theta = 0$. Suitable choice of the dependence functions H_k to represent within-individual dependence will be discussed in section 4. Note that if $H(.)$ is symmetric in its arguments, so that for any permutation σ of order k, $H_k(t_1, ..., t_k) = H_k(t_{\sigma(1)}, ..., t_{\sigma(k)})$, then the denominator of (4) is equal to

$$\frac{1}{k!} \int_{a_i}^{b_i} ... \int_{a_i}^{b_i} \lambda_i(t_1|x_i) ... \lambda_i(t_k|x_i) H_k(t_1, ..., t_k; \theta) dt_k ... dt_1.$$

Henceforth, we shall drop explicit reference to θ in $H_k(.; \theta)$.

3 Two alternative derivations of the model

The augmentation of the self-controlled case series model presented above is purely heuristic. In this section we provide two further interpretations of the model, as a multi-dimensional case series model, and as an approximation to a time-varying frailty model.

3.1 A multi-dimensional case series interpretation

Since the derivation of the case series likelihood involves conditioning on the number of events $n_i = k$, say, it is natural to represent the vector $(t_{i1}, ..., t_{ik})$ as a single point in the space $Q_i(k) = \{(t_1, ..., t_k) \in (a_i, b_i]^k : t_1 < ... < t_k\}$. As before, we can model the occurrence of such points as a Poisson process on $Q_i(k)$ with rate $\lambda_i(t_1, ..., t_k|x_i)$ and, conditioning on the occurrence of one such element, obtain the conditional likelihood

$$L_i^k = \frac{\lambda_i(t_{i1}, ..., t_{ik}|x_i)}{\displaystyle\int_{Q_i(k)} \lambda_i(t_1, ..., t_k|x_i) dt_k ... dt_1}. \quad (5)$$

Now set

$$\lambda_i(t_1, ..., t_k|x_i) = \prod_{j=1}^{k} \lambda_i(t_j|x_i) \times H_k(t_1, ..., t_k)$$

where $\lambda_i(t|x_i)$ is the univariate Poisson rate defined in equation (2), and $H_k(t_1, ..., t_k)$ is the dependence function introduced earlier. With this approach, we condition not on the number of events, but on the dimension of the space in which the individual's event history within $(a_i, b_i]$ is inscribed. This shift of perspective enables us to introduce the dependence function H_k rather more naturally within a self-controlled case series model without affecting the key aspects of the method, namely that only cases (now individual event histories of dimension > 0) need be sampled, and that all time-invariant multiplicative effects cancel. Thus, the resulting conditional likelihood remains that of a genuinely self-controlled case series method, albeit one of increased dimension.

The density (5) has support above the diagonal. Other densities have been proposed for ordered data, notably by Jones and Larsen (2004). In our context, the primary focus remains the underlying univariate rate function $\lambda_i(t)$ and, specifically, how it varies with exposure. Typically, a case series dataset will include individuals with 1, 2, 3... events, and the data likelihood will be a product of terms of the form (5) for different values of k, linked by the shared form of $\lambda_i(t)$ and, largely for convenience, a single dependence parameter θ.

3.2 A time-varying frailty interpretation

A standard derivation of multivariate time to event distributions is via copulas, often with interpretations in terms of frailties (Hougaard 2000). As previously explained, a constant frailty will not induce the type of dependence we are interested in, since such frailties describe between-individual, not within-individual heterogeneity. In the present subsection we relate our model to one obtained using time-varying frailties.

Suppose that events for individual i arise according to the non homogeneous Poisson process with rate $U_i(t)\lambda(t|x_i(t))$, where $U_i(t) < \infty$ is a non-negative random variable of mean 1 and constant variance σ^2. The variables $U_i(t)$ are independent between individuals, and represent frailty terms at time t specific to each individual i. We shall assume that the covariance between $U_i(t)$ and $U_i(s)$ depends only on t and s (and hence is common to all individuals). This time-varying frailty model has been discussed by Perperoglou et al (2006).

We write

$$H_2(t,s) = E[U_i(t)U_i(s)]$$

where the expectation is with respect to the joint distribution of $U_i(t)$ and $U_i(s)$. We assume that $U_i(t)$ is integrable, and define

$$I_i = \int_{a_i}^{b_i}\int_{t}^{b_i} U_i(t)\lambda(t|x_i(t))U_i(s)\lambda(s|x_i(s))dsdt.$$

The random variable I_i has mean

$$E(I_i) = \int_{a_i}^{b_i}\int_{t}^{b_i} \lambda(t|x_i(t))\lambda(s|x_i(s))H_2(t,s)dsdt.$$

Suppose now that individual i experiences two events at times t_{i1} and t_{i2}. Given the event process, and conditioning on two events having occurred, the conditional likelihood for individual i is

$$L_i^c(t_{i1}, t_{i2}|U_i(t_{i1}), U_i(t_{i2}), I_i) = \frac{U_i(t_{i1})\lambda(t_{i1}|x_i(t_{i1}))U_i(t_{i2})\lambda(t_{i2}|x_i(t_{i2}))}{\displaystyle\int_{a_i}^{b_i}\int_{t}^{b_i} U_i(t)\lambda(t|x_i(t))U_i(s)\lambda(s|x_i(s))dsdt}.$$

We now eliminate the unobserved frailty terms. Expanding the ratio $U_i(t_{i1})U_i(t_{i2})/I_i$ in a Taylor series and taking expectations yields

$$L_i^c(t_{i1}, t_{i2}) = \frac{\lambda(t_{i1}|x_i(t_{i1}))\lambda(t_{i2}|x_i(t_{i2}))H_2(t_{i1}, t_{i2})}{\int_{a_i}^{b_i}\int_t^{b_i}\lambda(t|x_i(t))\lambda(s|x_i(s))H_2(t,s)dsdt} \times \{1 + O(\eta_i)\}$$

where η_i is the coefficient of variation of I_i and the term in curly brackets represents a power series in η_i. The leading term of this expression is the model we proposed above, with dependence function H_2. We can therefore think of this model as approximating a variable frailty model, the approximation improving as the coefficient of variation η_i of I_i, and hence the frailty variance, get smaller. In this context, the dependence function H_2 is the second moment function about zero. The argument generalises straightforwardly to k events, in which case H_k is the kth moment function.

4 Modelling dependence

In this section we consider possible choices for the dependence function H_k, and its interpretation in terms of long-term and short-term dependence.

4.1 Diagonal dependence

We begin with the bivariate case $k = 2$, which is the most important. An obvious choice for H_2, convenient for modelling purposes, is

$$H_2(t, s) = \begin{cases} \exp(\theta) & \text{if } |t - s| < \delta, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

The dependence parameter θ may be estimated by maximum likelihood. The parameter δ is non-regular (the likelihood is not continuous in δ) and so should be chosen in advance. A reasonable strategy is to explore different values of δ. An alternative, yielding a smooth density, is

$$H_2(t, s) = \exp\left\{\theta \exp -\frac{1}{2}\left(\frac{t-s}{\delta}\right)^2\right\}. \tag{7}$$

For both functions, independence within individuals corresponds to $\theta = 0$. Both functions represent a simple form of dependence, in which more (if $\theta > 0$) or fewer (if $\theta < 0$) events than expected occur close to the diagonal $t = s$. We shall refer to such dependence patterns as *diagonal dependence.*

Generalizing these choices of functions to values $k > 2$ in a sensible manner is not straightforward. It is desirable that the parameter θ should have the same interpretation in all dimensions. The simplest way of achieving this is to build up H_k for $k > 2$ from H_2. There is no unique way to do this. For computational reasons, we shall use the function

$$H_k(t_1, ..., t_k) = \frac{2}{k(k-1)}\sum_{r,s}\{H_2(t_r, t_s) : t_r < t_s\} \tag{8}$$

namely, the average of the dependence functions for all $k(k-1)/2$ distinct pairs of events.

Alternatively, one could introduce further parameters to describe dependence between events m and $m + 2$, between events m and $m + 3$, and so on. We will not pursue this further. Note finally that the functions H_k described here are all symmetric.

4.2 Short and long term dependence

An interesting feature of the likelihood

$$L_i^k = \frac{\prod_{j=1}^{k} \lambda_i(t_j|x_i) \times H_k(t_1, ..., t_k)}{\int_{Q_i(k)} \prod_{j=1}^{k} \lambda_i(t_j|x_i) \times H_k(t_1, ..., t_k) dt_k ... dt_1} \tag{9}$$

is that constant multiplicative terms in H_k cancel out. For example, suppose that $k = 2$ and consider the dependence function

$$H_2(t, s) = \begin{cases} \exp\left(\theta \exp(-\frac{1}{2}(t-s)^2/\delta^2) + \zeta\right) & \text{if } t - s > 0, \\ 1 & \text{otherwise.} \end{cases}$$

This describes a jump in the relative incidence from 1 before to $\exp(\theta + \zeta)$ just after the first event, and declining to $\exp(\zeta)$ long after. Thus $\exp(\zeta)$ can be thought of as long-term dependence on the first event, and $\exp(\theta)$ as an additional short-term factor. However, the terms $\exp(\zeta)$ cancel from the likelihood, so only the short-term effect remains. Note also that $H_2(t, s)$ is not symmetric. Thus it is not the case that $L_i^2 = 2! L_i^c$, and hence the model does not reduce to the independence model if $\theta = 0$.

This has two important implications. First, it is only possible to model short-term dependence in this framework, or at least time-varying dependence, since constant terms cancel out from the likelihood. More precisely, we cannot estimate H_k, only the equivalence class $[H_k] = \{e^\zeta H_k; \zeta \in R\}$ restricted to the subspace $Q_i(k)$. Second, a test of $H_k \equiv 1$ is not sufficient to test for 'full' independence: such a test can only evaluate the evidence against short-term independence since only the equivalence class of H_k is identifiable.

This limitation is fundamental to the fact that only cases are sampled. The information required to determine the value of the constant term $\exp(\zeta)$ is contained primarily in the relative frequencies of singleton cases and pairs (and k-tuples) of cases arising from the underlying cohort of individuals, just as the value of the baseline incidence parameter φ in the standard case series model depends on the relative frequency of events and cannot be estimated from a model involving only cases, unless further information about the case sampling mechanism is available.

4.3 Relation to dependence measures

Consider the special case $k = 2$. Several bivariate local dependence functions have been suggested (Clayton 1978; Holland and Wang 1987). An analytically convenient choice is that of Holland and Wang (1987), namely the cross partial derivative of the log bivariate density, which here is the second derivative of $\log H_2(t, s)$. For the smooth dependence model (7), its sign is that of $\theta\{1 - (t - s)^2/\delta^2\}$. Thus, when $\theta > 0$, dependence is positive for $|t - s| < \delta$, and negative for $|t - s| > \delta$; this also characterises the discontinuous density (6), whence the term *diagonal dependence* introduced earlier to describe both dependence functions.

5 Estimation

In this section we turn to the problem of testing the null hypothesis $\theta = 0$. It will usually also be of interest to study the impact of allowing for dependence on the estimate of β, the parameter for the association with the exposure of primary interest.

5.1 Poisson modelling

Fitting the model (4) involves evaluating multiple integrals. To avoid the resulting complications, we propose a simple Poisson modelling approach based on the conditional Poisson representation (5). We shall use the dependence function (6) and its extension (8) to k dimensions. We take a parametric function for the baseline relative incidence $\psi(t)$, which is assumed to be piecewise constant on pre-defined age groups, represented by a factor with $L+1$ levels, and take the exposure vector to have $M+1$ levels (Farrington 1995).

The likelihood contribution of individuals with a single event (and hence $n_i = 1$) is obtained using the standard case series model (3) (see Farrington 1995, Whitaker et al 2006). Denote the log-likelihood contribution $l_i^1(\alpha,\beta;t_i)$ where α is the L-vector of age parameters and β is the M-vector of exposure parameters.

For an individual with two events at times t_{i1} and $t_{i2} > t_{i1}$ and observation period $(a_i, b_i]$, the likelihood contribution is that of a Poisson model in 2 dimensions. Define an indicator variable taking the value 1 within the diagonal $\{(s,t) : s \leq t < s + \delta\}$ band and 0 outside it. The age groupings, exposures and the diagonal band generate a decomposition of the space $\{(s,t) \in (a_i,b_i]^2 : s < t\}$ into p_i non-overlapping polygons of area $D_{i1},...,D_{ip_i}$, in which the age factors, exposure factors, and the diagonal factor take fixed values. Let N_{ij} denote the count of event pairs (0 or 1) within area D_{ij}, x_{ij}^1 the exposure vector (of length M) corresponding to the first component of area D_{ij}, and x_{ij}^2 that corresponding to its second component, y_{ij}^1, y_{ij}^2 the age vectors (of length L) corresponding to the first and second components, respectively, of D_{ij}, and z_{ij} the value of the $0-1$ diagonal indicator on D_{ij}. The 2-D Poisson likelihood contribution is then

$$\begin{aligned} E(N_{ij}) &= \mu_{ij},\ j = 1,...,p_i, \\ \log(\mu_{ij}) &= \gamma_i + \log(D_{ij}) + \alpha^T(y_{ij}^1 + y_{ij}^2) + \beta^T(x_{ij}^1 + x_{ij}^2) + \theta z_{ij}. \end{aligned}$$

The parameter γ_i is an individual-level factor, a nuisance parameter to obtain the multinomial likelihood from a Poisson model (McCullagh and Nelder, 1989). Let $l_i^2(\alpha,\beta,\theta;t_{i1},t_{i2})$ denote the log-likelihood contribution for an individual with two events. Note that the decomposition into polygons can readily be programmed; an outline of the algorithm is provided in Appendix 1.

An individual i with $k > 2$ events at ages $t_{i1},...,t_{ik}$ has a likelihood contribution comprising weighted likelihoods for the single event times and distinct event pairs, plus a term depending only on θ and the data. It can be shown that, for the dependence function defined in (6) and (8), the log-likelihood contribution $l_i^k(\alpha,\beta,\theta;t_{i1},...,t_{ik})$ is

$$\begin{aligned} l_i^k(\alpha,\beta,\theta;t_{i1},...,t_{ik}) &= \frac{k-2}{k}\sum_{j=1}^{k} l_i^1(\alpha,\beta;t_{ij}) \\ &\quad + \frac{2}{k(k-1)}\sum_{r<s} l_i^2(\alpha,\beta,\theta;t_{ir},t_{is}) + J_k(\theta;t_{i1},...,t_{ik}) \end{aligned} \tag{10}$$

up to a constant term, where

$$J_k(\theta;t_{i1},...,t_{ik}) = \log\left(\frac{2}{k(k-1}\sum_{r<s} H_2(t_r,t_s)\right) - \frac{2}{k(k-1)}\sum_{r<s}\log H_2(t_r,t_s)$$

is the log ratio of the arithmetic mean to the geometric mean of the pairwise dependence functions. The identity (10) is proved in Appendix 2.

The data log-likelihood may thus be written as a weighted likelihood for singletons and pairs of the form $\Sigma w_k l^k(\alpha,\beta,\theta)$, plus a function of θ, $\Sigma J_k(\theta)$. This suggests a simple way of obtaining point and interval estimates for θ: for a given value of θ, maximise $\Sigma w_k l^k(\alpha, \beta, \theta)$ using weighted Poisson regression, and hence obtain the profile log likelihood for θ. The profile log likelihood for β may be obtained implicitly by the same method. A test of the null hypothesis $\theta = 0$ may be based on the maximised log likelihood ratio, or on the profile likelihood confidence interval.

5.2 A simple conditional method for singletons and pairs

The method described above, though readily implemented in any log-linear modelling package, is nevertheless rather cumbersome. If the data involve only singletons and pairs, then a much simpler, conditional method is available, though at the cost of some loss in efficiency.

Individuals with just one event contribute the usual case series likelihood (such cases do not contribute to the estimation of θ, but do contribute to the estimate of age and exposure effects; in a test of $\theta = 0$ they could be left out completely). Individuals with 2 events contribute a conditional likelihood, based on the density of the second event time given the first. From the joint density of t_{i1} and t_{i2} under the augmented model, the conditional density of t_{i2}, given t_{i1} and $n_i = 2$ may be derived as

$$f(t_{i2}|t_{i1}, n_i = 2) = \frac{\lambda_i(t_{i2}|x_i) \times H_2(t_{i1}, t_{i2})}{\displaystyle\int_{t_{i1}}^{b_i} \lambda_i(t|x_i) \times H_2(t_{i1}, t)dt}. \tag{11}$$

Thus the overall likelihood has the form of a standard case series likelihood, except that for individuals with 2 events the observation period $(a_i, b_i]$ is replaced by $(t_{i1}, b_i]$. If the dependence function is that of equation (6), then the standard case series model may be used, with an additional indicator variable to denote proximity to t_{i1}. This approach lends itself particularly easily for use with both a parametric and semiparametric case series model (Farrington and Whitaker 2006).

This simple method involves some loss of information about θ. The asymptotic relative efficiency ARE of the conditional method, relative to the bivariate model, may readily be calculated in the simple scenario where there are no age or exposure effects, and all individuals have the same observation period $(0, 1]$, using the diagonal dependence function (6):

$$ARE = \frac{[e^{\theta}\delta\,(2-\delta) + (1-\delta)^2] \times [1 - \delta - \delta e^{\theta}\log(1 + (1-\delta)/\delta e^{\theta})]}{(1-\delta/2) \times (1-\delta)^2}.$$

As shown in Figure 1, in this scenario the relative efficiency is greater than 0.8 when $\exp(\theta) \geq 1$. Thus, in general, the loss in efficiency might be expected to be moderate. However, the method does not extend readily to individuals with $k > 2$ events, unless dependences between the first $k-1$ event times are ignored.

6 Simulations

In this section we study the performance of the method by simulation. We quantify the bias in β, the association parameter of primary interest, when the assumption of within-individual independence is violated, and study the extent to which estimation from the augmented model reduces this bias. Second, we study the power of the method to identify the presence of intra-individual clustering. We consider three scenarios, both involving only event pairs. In the first, event pairs arise in a planar Poisson process; thus, the simulated data are generated according to the underlying model as described in subsection 3.1. In the second scenario, we assume that occurrence of one event at age t increases the risk of a second event over some period $(t, t+\delta]$. Under this scenario, the data are no longer generated according to the underlying model. Finally, in the third scenario, we assumed that individual event rates were modulated by a time-varying frailty, independently between individuals. The data are no longer generated according to the

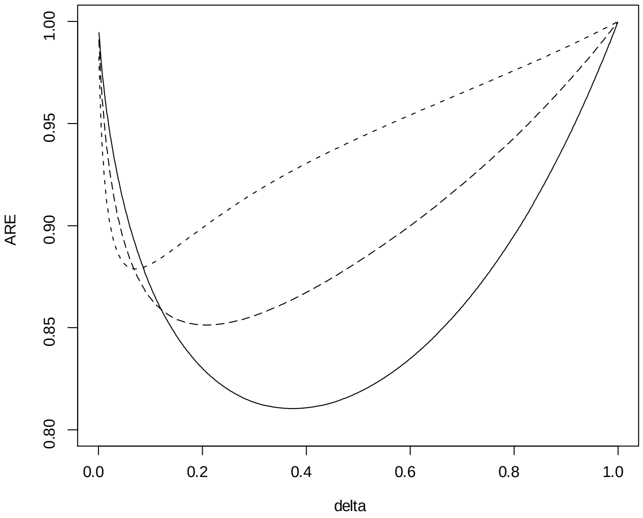

Figure 1: Asymptotic relative efficiency (ARE) of the conditional method, for three values of $\exp(\theta)$: 1 (full line), 2 (long dashes), 5 (short dashes).

underlying model, though it provides an approximation to the time-varying frailty model as described in Subsection 3.2.

In all simulations, we assume that all individuals have the same observation period $(0, 1]$, are all exposed at $v = 0.45$ with a risk period $e = 0.1$, that the marginal event rates do not vary with age and are increased by the factor $\exp(\beta)$ in the risk period. The dependence function is of the form (6).

To study the bias in the exposure effect β, we simulated $N = 10^6$ event pairs (for event-dependent rates, we used $N = 10^5$ to reduce computation time) for each combination of $\delta = 0.15, 0.25$; $\exp(\beta) = 0.5, 1, 1.5, 2, 5, 10$; $\exp(\theta) = 1.5, 2, 3, 5, 10$, and analysed the simulated data using the standard case series model (in which it is assumed that events are independent within individuals) with only the exposure effect, and the augmented model allowing for exposure and diagonal dependence (with the correct δ). For the simulations under the event-dependent and frailty models, we also allowed for age in the augmented model, since model mis-specification can induce spurious age effects. We used the 4 age groups $(0, 0.25], (0.25, 0.5], (0.5, 0.75]$, and $(0.75, 1]$. The bias shown in the tables below is $\widehat{E}(\widehat{\beta}) - \beta$, where $\widehat{E}$ denotes the average of the N simulated values; when small it is roughly equal to the relative bias in $\exp(\beta)$. We also report the impact of different estimation strategies on the standard errors, when $N = 100$ (using 10^3 replicates), in some of the more extreme scenarios, namely $\exp(\theta) = 10$ with $\exp(\beta) = 2$ and 10. We report the empirical standard error of $\widehat{\beta}$, namely the standard deviation of the 10^3 replicates, as well as the mean of the estimated standard errors, and the % difference, namely 100 times the mean of the estimated SE minus the empirical .SE, divided by the empirical SE.

To study power, we simulated 10^3 sets of N event pairs for each combination of $N = 50, 100, 500$; $\delta = 0.15, 0.25$; $\exp(\beta) = 2, 10$; $\exp(\theta) = 1, 1.25, 1.5, 2, 3, 5, 10$. We calculated the proportion of the 10^3 runs in which the null hypothesis of no intra-individual association (i.e. $\theta = 0$) was rejected using the likelihood ratio test.

6.1 Simulations under the assumed model

The subspace $\{(s,t) : 0 < s < t < 1\}$ is partitioned into 11 distinct polygons. Counts of event pairs were generated using multinomial sampling from these 11 categories. The bias in β that results from ignoring dependence between individuals is illustrated in Table 1. The bias increases as $\exp(\beta)$ and $\exp(\theta)$ increase; the relationship with δ is more complex. Under the Poisson model, substantial bias may arise in the estimation of β if the independence assumption is violated. The augmented model completely removes this bias.

Table 1 Poisson model. Bias in $\widehat{\beta}$, without (NC) and with (C) correction for diagonal dependence, for different values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$					$\delta = 0.25$				
		$\exp(\theta)$					$\exp(\theta)$				
$\exp(\beta)$		1.5	2	3	5	10	1.5	2	3	5	10
10	NC	0.155	0.268	0.411	0.568	0.722	0.118	0.190	0.280	0.365	0.438
	C	-0.002	0.002	0.000	0.002	0.000	0.001	-0.001	0.002	0.002	0.002
5	NC	0.099	0.175	0.273	0.384	0.496	0.082	0.136	0.200	0.268	0.321
	C	-0.002	0.000	-0.001	-0.001	-0.002	-0.000	-0.000	-0.001	0.002	-0.002
2	NC	0.043	0.072	0.113	0.165	0.219	0.046	0.076	0.116	0.153	0.191
	C	0.003	0.001	-0.002	-0.000	-0.000	0.001	-0.000	0.000	-0.001	0.001
1.5	NC	0.026	0.045	0.077	0.107	0.148	0.036	0.062	0.092	0.129	0.160
	C	-0.000	-0.002	0.001	-0.003	0.001	-0.001	-0.001	-0.002	0.002	0.002
1	NC	0.011	0.020	0.029	0.047	0.062	0.029	0.045	0.069	0.097	0.126
	C	-0.000	0.000	-0.003	0.001	-0.001	0.001	-0.002	-0.003	0.000	0.005
0.5	NC	-0.001	-0.010	-0.016	-0.028	-0.034	0.014	0.030	0.051	0.065	0.080
	C	0.006	0.002	0.003	-0.000	0.004	-0.003	-0.001	0.005	0.002	0.002

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence and exposure effects.

Table 2 Poisson model. Estimated and empirical standard errors for different values of $\exp(\beta)$ and δ

δ	$\exp(\beta)$		Estimated	Empirical	% difference
0.25	10	NC	0.147	0.168	-12.5
		C	0.140	0.149	-6.0
	2	NC	0.174	0.182	-4.4
		C	0.165	0.163	+1.2
0.15	10	NC	0.155	0.180	-13.9
		C	0.144	0.147	-2.0
	2	NC	0.172	0.193	-10.9
		C	0.154	0.153	+0.7

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence and exposure effects.

Table 2 shows the empirical and average estimated standard errors of $\widehat{\beta}$. In these extreme scenarios, where $\exp(\theta) = 10$, fitting the incorrect model, with exposure effect only, results in standard errors that can be more than 10% lower than the empirical value.

The power is shown in Table 3; for example, it exceeds 80% for $\exp(\theta)$ in excess of 2 when $N = 100$. When $\exp(\theta) > 1$, the power is generally marginally greater for $\delta = 0.25$ and $\exp(\beta) = 2$ than for $\delta = 0.15$ or $\exp(\beta) = 10$.

Table 3 Poisson model. Power to detect dependence for different sample sizes N and values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$			$\delta = 0.25$		
		N			N		
$\exp(\theta)$	$\exp(\beta)$	50	100	500	50	100	500
1	10	0.059	0.049	0.042	0.064	0.047	0.055
	2	0.064	0.053	0.062	0.049	0.047	0.050
1.25	10	0.101	0.161	0.558	0.091	0.160	0.591
	2	0.108	0.167	0.621	0.125	0.199	0.694
1.5	10	0.194	0.371	0.976	0.241	0.426	0.976
	2	0.240	0.442	0.985	0.301	0.536	0.990
2	10	0.541	0.815	1.000	0.541	0.821	1.000
	2	0.663	0.924	1.000	0.652	0.938	1.000
3	10	0.852	0.991		0.873	0.990	
	2	0.962	0.999		0.970	0.999	
.5	10	0.990	1.000		0.987	1.000	
	2	0.999	1.000		1.000	1.000	
10	10	1.000			1.000		
	2	1.000			1.000		

6.2 Simulations under an event-dependent model

Event pairs were simulated as follows. A first event time t_1 was simulated in $(0, 1]$ using the standard case series model (which is just a Poisson model, conditioned on an event occurring). In the interval $(t_1, t_1 + \delta]$ the Poisson rate was then increased by the factor $\exp(\theta)$, and a second event time $t_2 > t_1$ was simulated. If $t_2 \leq 1$ the pair (t_1, t_2) was accepted, otherwise it was discarded and new values t_1 and t_2 were generated. In addition to varying the parameters mentioned in the introduction to this section, we also varied the absolute event rate λ, which its value now affects the results. We took $\lambda = 0.01, 0.1$.

Table 4 Event-dependent model, $\lambda = 0.01$. Bias in $\widehat{\beta}$, without (NC) and with (C) correction for diagonal dependence, for different values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$					$\delta = 0.25$				
		$\exp(\theta)$					$\exp(\theta)$				
$\exp(\beta)$		1.5	2	3	5	10	1.5	2	3	5	10
10	NC	0.156	0.262	0.406	0.570	0.715	0.123	0.188	0.279	0.365	0.438
	C	-0.005	-0.010	-0.004	0.003	0.006	0.033	0.033	0.043	0.056	0.063
5	NC	0.096	0.174	0.275	0.384	0.499	0.082	0.143	0.199	0.266	0.328
	C	-0.005	-0.000	-0.001	0.002	-0.003	0.013	0.034	0.040	0.041	0.059
2	NC	0.047	0.077	0.105	0.165	0.220	0.043	0.079	0.116	0.150	0.189
	C	0.004	0.008	-0.004	-0.000	0.001	0.011	0.026	0.037	0.040	0.048
1.5	NC	0.028	0.045	0.080	0.108	0.148	0.046	0.067	0.099	0.126	0.163
	C	0.007	-0.000	0.007	0.002	0.001	0.023	0.026	0.043	0.042	0.053
1	NC	0.014	0.021	0.029	0.053	0.062	0.033	0.053	0.072	0.103	0.121
	C	0.002	0.002	-0.005	0.001	-0.000	0.021	0.031	0.035	0.051	0.057
0.5	NC	-0.014	-0.020	-0.013	-0.013	-0.039	0.007	0.018	0.048	0.084	0.081
	C	-0.010	-0.012	0.003	0.016	0.003	0.006	0.016	0.041	0.063	0.058

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence, age and exposure effects.

The bias resulting from ignoring dependence is shown in Tables 4 and 5. Substantial bias may arise in the estimation of β if the independence assumption is violated. The augmented model with age and diagonal dependence, as well as exposure, removes much of this bias, though performance is better for smaller values of δ and λ. We also fitted models with just age and exposure (not shown), and in some instances the bias in the estimation of β was a little less than for the model with age, diagonal dependence and exposure.

Table 5 Event-dependent model, $\lambda = 0.1$. Bias in $\widehat{\beta}$, without (NC) and with (C) correction for diagonal dependence, for different values of $\exp(\beta)$, $\exp(\theta)$ and δ, when $\lambda = 0.1$.

		$\delta = 0.15$					$\delta = 0.25$				
		$\exp(\theta)$					$\exp(\theta)$				
$\exp(\beta)$		1.5	2	3	5	10	1.5	2	3	5	10
10	NC	0.168	0.280	0.420	0.546	0.644	0.135	0.215	0.298	0.363	0.417
	C	-0.007	-0.002	-0.000	-0.009	-0.010	0.034	0.059	0.074	0.083	0.110
5	NC	0.117	0.191	0.283	0.368	0.455	0.091	0.144	0.213	0.259	0.312
	C	-0.001	0.001	-0.001	-0.014	-0.015	0.021	0.036	0.050	0.053	0.077
2	NC	0.044	0.069	0.120	0.178	0.209	0.064	0.085	0.118	0.146	0.183
	C	0.006	-0.010	-0.003	0.008	-0.005	0.026	0.027	0.037	0.038	0.048
1.5	NC	0.039	0.059	0.083	0.109	0.140	0.042	0.061	0.094	0.127	0.147
	C	0.004	0.006	-0.001	-0.010	0.001	0.017	0.017	0.032	0.046	0.039
1	NC	0.015	0.037	0.045	0.049	0.065	0.043	0.056	0.081	0.106	0.122
	C	-0.004	0.011	0.009	0.002	0.003	0.026	0.028	0.040	0.053	0.045
0.5	NC	-0.007	-0.009	-0.016	-0.021	-0.041	0.021	0.026	0.055	0.065	0.067
	C	-0.004	-0.000	-0.002	0.007	-0.007	0.018	0.014	0.040	0.048	0.033

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence, age and exposure effects.

Table 6 shows the standard errors in selected scenarios with $\exp(\theta) = 10$. Failure to correct for dependence results in standard errors that are too low, typically by 10% or more, compared to their empirical values. Allowing for age and diagonal dependence reduces substantially the discrepancy between estimated and empirical standard errors.

Table 6 Event-dependent model. Estimated and empirical standard errors for different values of $\exp(\beta)$ and δ

δ	$\exp(\beta)$		Estimated	Empirical	% difference
0.25	10	NC	0.147	0.165	-10.9
		C	0.176	0.185	-4.9
	2	NC	0.174	0.193	-9.8
		C	0.192	0.200	-4.0
0.15	10	NC	0.152	0.178	-14.6
		C	0.180	0.192	-6.3
	2	NC	0.173	0.198	-12.6
		C	0.185	0.191	-3.1

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence, age and exposure effects.

Tables 7 and 8 show the power of the likelihood ratio test of $\theta = 0$, obtained by comparing the log likelihoods of the model with exposure and the model with exposure and diagonal dependence. When $\delta = 0.15$ the power exceeds 80% for $\exp(\theta)$ in excess of 2 when $N = 100$. However, for $\delta = 0.25$ the power is much less than when $\delta = 0.15$. The power is marginally higher for $\lambda = 0.1$ than for $\lambda = 0.01$. If a likelihood ratio test of $\theta = 0$ based on models both including age (as well as exposure), the power is reduced (not shown).

Table 7 Event-dependent model, $\lambda = 0.01$. Power to detect dependence for different sample sizes N and values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$			$\delta = 0.25$		
		N			N		
$\exp(\theta)$	$\exp(\beta)$	50	100	500	50	100	500
1	10	0.042	0.060	0.054	0.061	0.040	0.050
	2	0.049	0.047	0.057	0.051	0.053	0.045
1.25	10	0.111	0.141	0.553	0.052	0.061	0.233
	2	0.118	0.184	0.670	0.069	0.087	0.260
1.5	10	0.234	0.369	0.967	0.073	0.135	0.610
	2	0.281	0.474	0.992	0.101	0.148	0.625
2	10	0.541	0.809	1.000	0.166	0.291	0.955
	2	0.652	0.919	1.000	0.214	0.364	0.967
3	10	0.882	0.994		0.292	0.589	1.000
	2	0.967	0.999		0.380	0.712	1.000
.5	10	0.986	1.000		0.490	0.865	
	2	1.000	1.000		0.645	0.922	
10	10	1.000			0.663	0.975	
	2	1.000			0.801	0.987	

Table 8 Event-dependent model, $\lambda = 0.1$. Power to detect dependence for different sample sizes N and values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$			$\delta = 0.25$		
		N			N		
$\exp(\theta)$	$\exp(\beta)$	50	100	500	50	100	500
1	10	0.050	0.057	0.064	0.045	0.059	0.063
	2	0.044	0.050	0.063	0.050	0.055	0.053
1.25	10	0.124	0.214	0.743	0.060	0.091	0.354
	2	0.155	0.220	0.765	0.057	0.095	0.322
1.5	10	0.303	0.481	0.986	0.107	0.155	0.702
	2	0.324	0.528	0.999	0.116	0.162	0.713
2	10	0.597	0.863	1.000	0.181	0.369	0.988
	2	0.703	0.948	1.000	0.198	0.417	0.989
3	10	0.903	0.994		0.335	0.674	1.000
	2	0.979	1.000		0.435	0.740	1.000
.5	10	0.994	1.000		0.554	0.911	
	2	1.000			0.657	0.936	
10	10	1.000			0.755	0.985	
	2	1.000			0.819	0.992	

6.3 Simulations under a varying frailty model

For each individual, we randomly generated $u \sim U(-\frac{\delta}{2}, 1+\frac{\delta}{2})$ and assumed that the individual's baseline rate was increased by the factor $\exp(\theta)$ on the interval $(u - \frac{1}{2}\delta, u + \frac{1}{2}\delta)$. The resulting within-individual correlation between times s and t declines linearly from 1 when $s = t$ to 0 when $|s - t| \geq \delta$. Event pairs were generated conditionally on the individual frailties using a standard case series model.

The bias in β resulting from analysing the data unconditionally is shown in Table 9. The bias is generally small, except when $\exp(\beta)$ and $\exp(\theta)$ are both large. Fitting the augmented model with age and diagonal dependence has only a moderate ($\delta = 0.25$) or virtually nil ($\delta = 0.15$) effect in reducing the bias.

Table 9 Varying frailty model. Bias in $\widehat{\beta}$, without (NC) and with (C) correction for diagonal dependence, for different values of $\exp(\beta)$, $\exp(\theta)$ and δ.

		$\delta = 0.15$					$\delta = 0.25$				
		$\exp(\theta)$					$\exp(\theta)$				
$\exp(\beta)$		1.5	2	3	5	10	1.5	2	3	5	10
10	NC	-0.009	-0.027	-0.079	-0.183	-0.403	-0.013	-0.038	-0.098	-0.225	-0.450
	C	-0.009	-0.029	-0.087	-0.200	-0.416	-0.010	-0.026	-0.065	-0.154	-0.315
5	NC	-0.006	-0.017	-0.052	-0.131	-0.297	-0.009	-0.023	-0.067	-0.159	-0.319
	C	-0.006	-0.020	-0.061	-0.145	-0.302	-0.006	-0.015	-0.045	-0.110	-0.221
2	NC	-0.004	-0.008	-0.021	-0.054	-0.131	-0.002	-0.012	-0.031	-0.071	-0.154
	C	-0.003	-0.009	-0.022	-0.053	-0.114	0.000	-0.009	-0.023	-0.047	-0.105
1.5	NC	-0.003	-0.005	-0.012	-0.034	-0.082	-0.002	-0.007	-0.024	-0.053	-0.116
	C	-0.003	-0.005	-0.011	-0.028	-0.058	-0.001	-0.003	-0.017	-0.037	-0.083
1	NC	-0.004	-0.001	-0.005	-0.012	-0.027	-0.000	-0.004	-0.013	-0.026	-0.061
	C	-0.002	-0.000	-0.001	-0.002	0.005	-0.000	-0.004	-0.010	-0.017	-0.043
0.5	NC	-0.004	0.000	0.011	0.028	0.055	-0.001	-0.002	-0.001	0.005	-0.002
	C	-0.006	0.001	0.017	0.044	0.099	0.001	-0.001	-0.000	0.004	-0.002

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence, age and exposure effects.

Table 10 shows the standard errors, which present a similar pattern as with the event-dependent model.

Table 10 Estimated and empirical standard errors for different values of $\exp(\beta)$ and δ

δ	$\exp(\beta)$		Estimated	Empirical	% difference
0.25	10	NC	0.144	0.163	-11.7
		C	0.186	0.205	-9.3
	2	NC	0.195	0.208	-6.3
		C	0.216	0.217	-0.5
0.15	10	NC	0.143	0.161	-11.2
		C	0.182	0.196	-7.1
	2	NC	0.193	0.215	-10.2
		C	0.211	0.226	-6.6

NC: standard case series model with exposure effect only.

C: case series model with diagonal dependence, age and exposure effects.

The power, shown in Table 11, is very low even for a sample size of 500 when $\exp(\theta) \leq 2$. It is clear that only time-varying frailties inducing very strong dependence can be detected with moderate sample sizes.

Table 11 Varying frailty model. Power to detect dependence for different sample sizes N and values of $\exp(\beta)$, $\exp(\theta)$ and δ

		$\delta = 0.15$			$\delta = 0.25$		
		N			N		
$\exp(\theta)$	$\exp(\beta)$	50	100	500	50	100	500
1	10	0.054	0.057	0.049	0.060	0.052	0.045
	2	0.062	0.059	0.049	0.057	0.065	0.041
1.25	10	0.066	0.047	0.054	0.052	0.050	0.058
	2	0.056	0.050	0.050	0.049	0.052	0.048
1.5	10	0.049	0.042	0.059	0.062	0.051	0.056
	2	0.057	0.046	0.051	0.053	0.058	0.059
2	10	0.062	0.055	0.052	0.054	0.075	0.099
	2	0.052	0.057	0.075	0.069	0.079	0.183
3	10	0.063	0.082	0.167	0.075	0.103	0.295
	2	0.073	0.130	0.443	0.122	0.198	0.740
5	10	0.129	0.196	0.755	0.181	0.307	0.901
	2	0.320	0.541	0.995	0.412	0.713	1.000
10	10	0.556	0.835	1.000	0.890	0.910	1.000
	2	0.905	0.997	1.000	1.000	0.999	1.000

7 Examples

7.1 ITP and MMR vaccine

Idiopathic thrombocytopenic purpura (ITP) is a rare, potentially recurrent autoimmune disorder in which abnormal bleeding into the skin occurs due to low blood platelet count. Miller et al (2001) studied the association between mumps, measles and rubella vaccine (MMR) and hospital admission for ITP within the South East and North East Thames Regions in the UK. ITP cases arising during the period from October 1991 to September 1994 and aged $12-23$ months were included in the analysis. These time and age boundaries were used to define the observation period for each case. The data set included a total of 44 admissions experienced by 35 children; 5 of these children were admitted twice and 1 was admitted 5 times. It was hypothesised that MMR vaccination may, in rare instances, cause ITP. Risk periods covered the 6 week period after MMR vaccination, and three two-week long risk periods $0-14$, $15-28$ and $29-42$ days after vaccination were used. Of the 35 children, 31 were exposed to MMR vaccine between one and two years of age. Of the 44 events, 13 occurred within 6 weeks after receipt of the MMR vaccine. The intervals preceding the 10 repeat events were (in increasing order) $15, 42, 63, 70, 78, 112, 133, 148, 175, 190$ days.

We analysed these data using a parametric model with 6 age groups, and the dependence function of equation (6) with $\delta = 21, 42$ and 91 days. The estimates of θ with 95% profile confidence intervals given in Table 12 show that the results are very close for different values of δ, particularly for $\delta \geq 42$.

Table 12 Estimates of θ and 95% confidence intervals (CI) for three values of δ

δ (days)	$\widehat{\theta}$	95% profile CI
21	0.47	$(-1.27, 2.50)$
42	0.50	$(-1.30, 2.60)$
91	0.50	$(-1.35, 2.60)$

The estimates of θ are positive, suggestive of positive diagonal dependence. However, they are far from statistically significant, as shown by the wide confidence intervals. The profile log-likelihood for θ

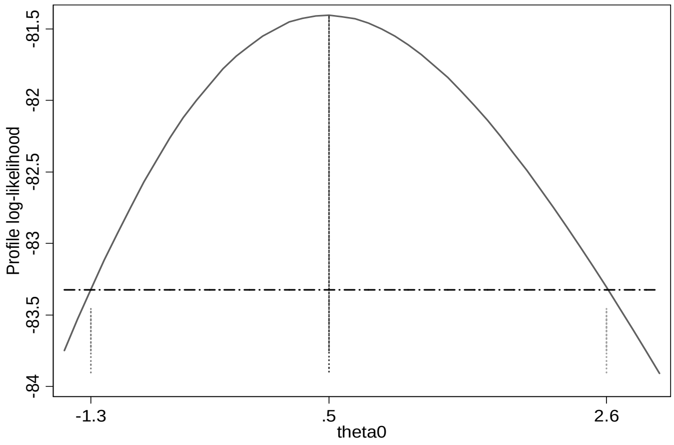

Figure 2: Profile log-likelihood for θ, with $\delta = 42$ days.

with $\delta = 42$ days, shown in Figure 2, is roughly quadratic; similar shapes were obtained for $\delta = 21$ and $\delta = 91$ days.

It is also of interest to examine the variation in the estimated model parameters. Table 13 shows the parameter estimates obtained when $\theta = 0$ and $\theta = 0.5$ (both regarded as fixed); the β are the log relative incidences associated with MMR vaccine in the three post-vaccination risk periods, and the α are the log relative age effects.

Table 13 Parameter estimates (95% confidence intervals) for two values of θ

Parameter	Standard model, $\theta = 0$	Dependence model, $\theta = 0.5$
β_1	$1.31(0.30, 5.73)$	$1.31(0.30, 5.71)$
β_2	$5.95(2.52, 14.07)$	$5.95(2.53, 14.00)$
β_3	$2.60(0.74, 9.07)$	$2.59(0.74, 9.01)$
α_1	$0.66(0.29, 1.46)$	$0.65(0.29, 1.45)$
α_2	$0.21(0.06, 0.74)$	$0.21(0.06, 0.76)$
α_3	$0.29(0.09, 0.90)$	$0.30(0.10, 0.92)$
α_4	$0.39(0.14, 1.13)$	$0.41(0.14, 1.15)$
α_5	$0.40(0.14, 1.15)$	$0.42(0.15, 1.19)$

The parameter estimates are hardly affected by allowing for a positive value of θ. Thus, conclusions based on the standard case series model appear robust to differering dependence assumptions. In particular, there is strong evidence of an association between MMR and ITP, particularly in the $15 - 28$ day period after vaccination. While there is no compelling evidence of dependence between events within individuals, the number of events is too small to rule out such dependence conclusively.

7.2 Myocardial infarction and respiratory infections

This example concerns myocardial infarctions (MI) and their association with respiratory tract infections (RTI), the exposure of interest. The data are a subset of the MI data described in Smeeth et al (2004). Note that MI increases mortality, which formally violates the assumption that observation is censored

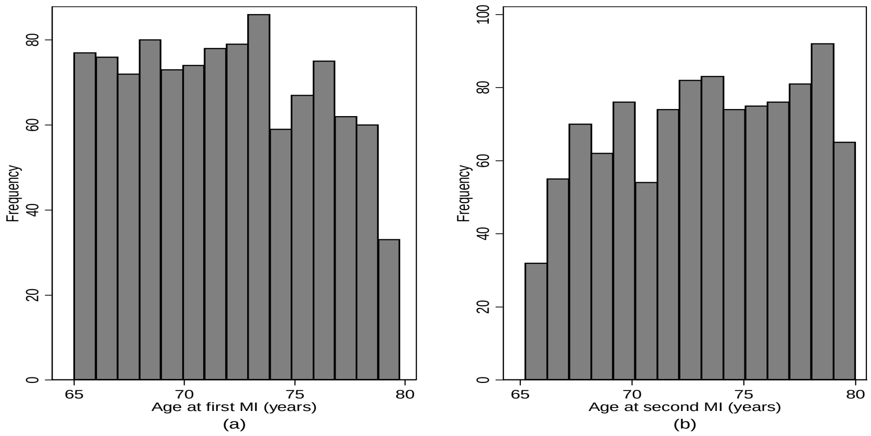

Figure 3: Distribution of age at MI: (a) first episode, (b) second episode.

completely at random. However, Farrington & Whitaker (2006) showed that, for these data, results are robust to failure of this assumption.

For the purposes of the present analysis, we selected individuals who experienced two MI between ages 65 and 80 years. These 1051, individuals experienced up to 8 RTI over this age range. Figure 3 shows the distribution of individuals' age at event date, for first and second episodes of MI, and Figure 4 shows the distribution of the delay between the first and the second MI. The median interval between the first and second MI is 197 days. Figure 5 shows the age distribution of RTIs experienced by the individuals sampled.

First, we analysed these data as if the two MI episodes for each individual were independent occurrences. We applied the standard case series method by fitting a parametric model with five 3-years age groups: $65-68$, $69-71$, $72-74$, $75-77$ and $78-80$ years, and five risk periods: $1-3$, $4-7$, $8-14$, $15-28$ and $29-91$ days after every RTI. The estimates of the relative incidence of MI for each post-RTI risk period, and in the combined risk period $1-91$ days, are given in Table 14. Second, we used the bivariate diagonal dependence model (6) with values of δ from 1 to 24 months in monthly increments. The relative incidence estimates (for the $1-91$ day risk period) and the estimates of $\exp(\theta)$ are shown in Figure 6. Clearly, changing the value of δ, does not affect the estimated exposure effect $\widehat{\beta}$, the parameter of interest, whereas the estimated dependence effect $\widehat{\theta}$ varies strongly with δ. In the first month after the first MI, $\exp(\theta) < 1$, indicating negative dependence. This may be due to intensive therapy, or to requiring that MIs should be at least one month apart to qualify as separate episodes. Thereafter there is a sharp increase of $\exp(\theta)$ to a maximum at 3 months, $\exp\widehat{\theta} = 4.53$, 95% CI $(3.96, 5.18)$, followed by a gradual decline to $\exp\widehat{\theta} = 2.41$, 95% CI $(2.06, 2.83)$ at 2 years. Parameter estimates for two values of δ (91 and 366 days) are given in Table 14. The estimates are very similar to those obtained under independence. Finally, we estimated the dependence using the conditional model (11). The results are shown in Figure 7. The results are similar to those presented in Figure 6, though the parameter estimates are attenuated and the confidence intervals wider.

The overall conclusion from this analysis is that there is strong dependence within individuals, most likely resulting from an increased MI rate following first event. This appears to peak 3 months after the first event, declining to some constant value. The estimates of β do not appear to be overly sensitive to

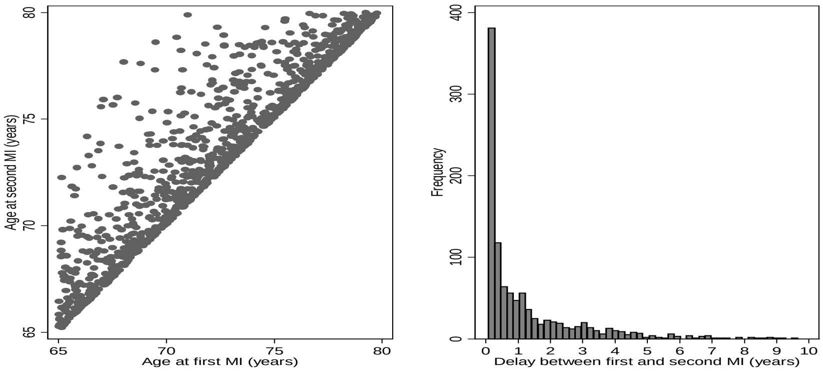

Figure 4: Distribution of the delay between first and second MI episodes.

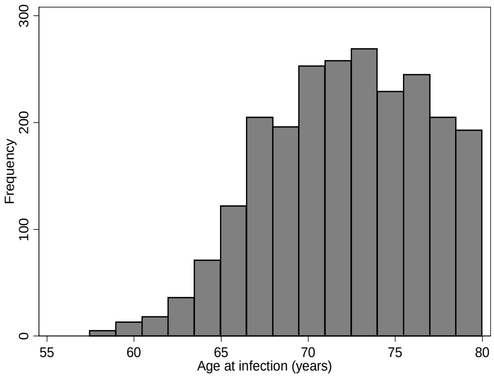

Figure 5: Distribution of age at RTI (all episodes combined)

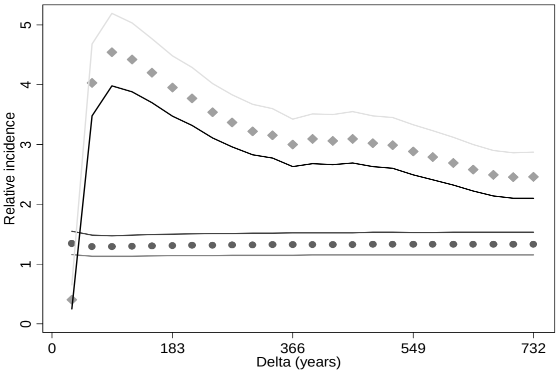

Figure 6: Full bivariate model: relative incidences associated with RTI (dots) and diagonal dependence (squares), with 95% confidence intervals.

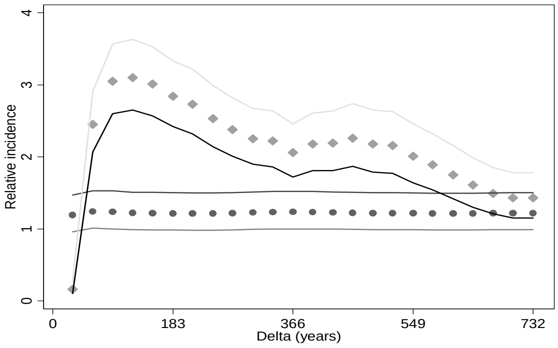

Figure 7: Conditional model: relative incidences associated with RTI (dots) and diagonal dependence (squares), with 95% confidence intervals.

this dependence, though sensitivity to long-term dependence cannot be evaluated. For this reason, it is advisable to analyse first and second MIs separately as well as jointly.

Table 14 Relative incidence of MI by risk period after RTI, for three models

Risk period		Model	
(days)	independent	dependent ($\delta = 91$ days)	dependent ($\delta = 366$ days)
1-3	1.88 (1.11, 3.18)	1.93 (1.12, 3.33)	1.94 (1.12, 3.45)
4-7	2.44 (1.63, 3.65)	2.57 (1.71, 3.88)	2.60 (1.72, 3.92)
8-14	1.45 (0.98, 2.16)	1.28 (0.83, 1.99)	1.29 (0.83, 2.00)
15-28	1.08 (0.78, 1.50)	1.17 (0.84, 1.63)	1.17 (0.84, 1.64)
29-91	1.18 (1.00, 1.39)	1.17 (0.99, 1.38)	1.18 (1.00, 1.40)
1-91	*1.32 (1.15,1.51)*	*1.29 (1.13, 1.47)*	*1.32 (1.15, 1.52)*

8 Discussion

We have extended the self-controlled case series model to allow for a specific departure from within-individual independence between recurrent events. This was achieved by augmenting the model with a diagonal dependence term, which indicates greater or lesser than expected pairwise clustering of events. The augmented model may be represented geometrically as a case series model in multiple dimensions, and was shown to approximate a model in which dependences arise owing to unmeasured time-varying frailties. The method can be applied to any number of recurrences; a very simple conditional method is available when no individual has more than two events. All the models we propose may be explored using standard log-linear modelling techniques.

The primary purpose of the augmented model is to provide a method of testing the within-individual independence assumption required by the case series method. An evaluation by simulation shows that the power available is good in small to moderate samples under the assumed higher-dimensional Poisson model, and when occurence of an event increases the short-term risk of another. In contrast, the power to detect clustering resulting from time-varying frailties is poor in moderate samples. A fundamental limitation of the method, and, arguably, of any method based only on cases, is that long-term dependence cannot be identified: only short-term, or time-varying dependence can be detected.

The estimated exposure effect, which is usually the parameter of primary interest, may be biased if within-individual dependence is present but ignored. Standard errors are also likely to be underestimated. Our simulations have shown that the bias in the point estimates may be substantial when an event increases the short-term risk of subsequent events; in contrast, when clustering is induced by time-varying frailties, the bias is very much less except in the most extreme scenarios. Standard errors are also underestimated, though this effect is relatively small except in extreme scenarios; investigations in a different setting have also shown standard errors to be robust to model mis-specification (Hocine et al, 2006).

A key modelling issue is how to proceed if there is evidence that events are not independent within individuals. The simplest approach is to limit the analysis to first events, provided these are rare, and undertake a separate analysis of recurrences, with observation period starting at the age of the first event, and a term for short-term dependence as used in our conditional model. Alternatively, our simulations suggest that allowing for dependence using the augmented model (with age effects) can substantially reduce bias in the estimated exposure effect (and produce standard errors that are closer to their empirical values). One exception is when events are clustered owing to time-varying frailties (though the bias in such a scenario is usually small). This finding was unexpected, since the frailty model approximates ours, at least when dependence is weak; closer analysis of this scenario shows that the goodness of fit of the model is indeed greatly improved, although the parameter estimates are little affected.

Acknowledgements

We thank Nick Andrews (HPA Centre for Infections) for the ITP data, and Liam Smeeth (London School of Hygiene and Tropical Medicine) for the MI data. We also thank Michel Chavance and Thierry Moreau (INSERM U780, Villejuif, France) for useful discussions. This research was supported by the EPSRC (grant EO2873X/1) and the Fondation pour la Recherche Médicale.

References

Clayton, D. G.(1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141 - 151.

Cook R. J. and Lawless J. F. (2007). *The statistical analysis of recurrent events.* Springer, New York.

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

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

Farrington, C. P., Whitaker, H. J. and Hocine, M. N. (2008) Case series analysis for censored, perturbed or curtailed post-event exposures. *Biostatistics* (in press).

Hocine, M., Guillemot, D., Tubert-Bitter, P. and Moreau, T. (2005). Testing independence between two Poisson-generated multinomial variables in case series and cohort studies. *Statistics in Medicine* **24**, 4035 - 4044.

Hocine, M. Moreau, T and Chavance, M. (2006). Semiparametric analysis of case series data: Contribution to the discussion. *Applied Statistics* **55**, 553 - 594.

Hocine, M. N., Tubert-Bitter, P., Moreau, T., Chavance, M. Varon, E. and Guillemot, D. (2007). Relative-risk ratio was a useful measure of differential association in cohort and case series studies. *Journal of Clinical Epidemiology* **60**, 361 - 365.

Holland, P. W. and Wang, Y. J. (1987). Dependence function for continuous bivariate densities. *Communications in Statistics, Series A* **16**, 863 - 876.

Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer, New York.

Jones, M. C., and Larsen, P. V. (2004). Multivariate distributions with support above the diagonal. *Biometrika* **91**, 975 - 986.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models.* Chapman and Hall, London.

Miller, E., Waight, P., Farrington, P., Stowe, J. and Taylor B. (2001). Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of Disease in Childhood* **84**, 227 - 229.

Perperoglou, A., van Houwelingen, H. C. and Henderson, R. (2006). A relaxation of the gamma frailty (Burr) model. *Statistics in Medicine* **25**, 4253 - 4266.

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

Smeeth, L., Thomas, S. L., Hall, A. J., Hubbard, R., Farrington, P. and Vallance, P..(2004). Risk of myocardial infarction and stroke after acute infection or vaccination. *New England Journal of Medicine* **351**, 2611 - 2618.

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

Appendix 1

Suppose that an individual experiences two events at ages t_1 and t_2 within an observation period $(a, b]$. This can be represented as a single point in the subspace $Q(2) = \{(t, s) \in (a, b]^2;\ s \leq t\}$. The observation period $(a, b]$ is divided into $r + 1$ intervals $A_i = (s_{i-1}, s_i]$ corresponding to the exposure and the age groups where s_i are the ordred exposure and age cutpoints defined as follows:

$$s_i, i = 1, ..., r + 1;\ s_0 = a < s_1 < s_2 < ... < s_r < s_{r+1} = b.$$

We use i to index cutpoints on the horizontal axis and j for the vertical axis. The cutpoints determine a partition of $Q(2)$ into $\frac{1}{2}r(r + 1)$ squares and rectangles and $r + 1$ triangles. We denote the area of the segment with (s_j, s_i) at its top right-hand corner as A_{ij}. In addition, we use the dependence function (6), which determines a further subdivision by the line $t = s + \delta$. This line partitions each area A_{ij} into two, A_{ij0} the area within $t - s > \delta$ and A_{ij1}within $0 < t - s \leq \delta$. The areas A_{ijk} across individuals form the D_{ij} (with different meanings for the indices i and j) referred to in Subsection 5.1.

The values of the A_{ijk} depend on s_i and s_j, and their relation to δ. It turns out that there are 8 possible configurations, as follows.

Configuration 1: if $s_{j-1} - s_i > \delta$ then $A_{ij1} = (s_i - s_{i-1}) \times (s_j - s_{j-1})$ and $A_{ij0} = 0$.

Configuration 2: if $0 \leq s_{j-1} - s_i \leq \delta$ and $s_j - s_i > \delta$ and $s_{j-1} - s_{i-1} > \delta$ then $A_{ij1} = \frac{1}{2}(\delta - s_{j-1} + s_i)^2$ and $A_{ij0} = (s_i - s_{i-1}) \times (s_j - s_{j-1}) - A_{ij1}$.

Configuration 3: if $0 \leq s_{j-1} - s_i \leq \delta$ and $s_j - s_i \leq \delta$ and $s_{j-1} - s_{i-1} > \delta$ then $A_{ij0} = (s_j - s_{j-1}) \times (s_{j-1} - s_{i-1} - \delta) + \frac{1}{2}(s_j - s_{j-1})^2$ and $A_{ij1} = (s_i - s_{i-1}) \times (s_j - s_{j-1}) - A_{ij0}$.

Configuration 4: if $0 \leq s_{j-1} - s_i \leq \delta$ and $s_j - s_i > \delta$ and $s_{j-1} - s_{i-1} \leq \delta$ then $A_{ij0} = (s_i - s_{i-1}) \times (s_j - s_i - \delta) + \frac{1}{2}(s_i - s_{i-1})^2$ and $A_{ij1} = (s_i - s_{i-1}) \times (s_j - s_{j-1}) - A_{ij0}$.

Configuration 5: if $0 \leq s_{j-1} - s_i \leq \delta$ and $s_j - s_i \leq \delta$ and $s_{j-1} - s_{i-1} \leq \delta$ and $s_j - s_{i-1} > \delta$ then $A_{ij0} = \frac{1}{2}(s_j - s_{i-1} - \delta)^2$ and $A_{ij1} = (s_i - s_{i-1}) \times (s_j - s_{j-1}) - A_{ij0}$.

Configuration 6: if $0 \leq s_{j-1} - s_i \leq \delta$ and $s_j - s_{i-1} \leq \delta$ then $A_{ij0} = 0$ and $A_{ij1} = (s_i - s_{i-1}) \times (s_j - s_{j-1})$.

Configuration 7: if $s_i = s_j$ and $s_i - s_{i-1} > \delta$ then $A_{ij0} = \frac{1}{2}(s_i - s_{i-1} - \delta)^2$ and $A_{ij1} = \frac{1}{2}(s_i - s_{i-1})^2 - A_{ij0}$.

Configuration 8: if $s_i = s_j$ and $s_i - s_{i-1} \leq \delta$ then $A_{ij0} = 0$ and $A_{ij1} = \frac{1}{2}(s_i - s_{i-1})^2$.

Appendix 2

In this appendix we prove the identity (10). The log-likelihood contribution of an individual with k events may be written:

$$\begin{aligned} l_i^k(\alpha, \beta, \theta; t_{i1}, ..., t_{ik}) &= \sum_{j=1}^{k} \log \lambda_i(t_{ij}) + \log \frac{2}{k(k-1)} \sum_{r<s} H_2(t_r, t_s) \\ &\quad - \log \int\limits_{Q_i(k)} \lambda_i(t_1)...\lambda_i(t_k) H_k(t_1, ..., t_k) dt_k ... dt_1. \end{aligned} \tag{12}$$

Since $H_k(t_1, ..., t_k)$ has the form given in (8) and is symmetric, the final integral is

$$\begin{aligned} \int\limits_{Q_i(k)} \lambda_i(t_1)...\lambda_i(t_k) H_k(t_1, ..., t_k) dt_k ... dt_1 &= \frac{1}{k!} \int\limits_{(a_i, b_i]^k} \lambda_i(t_1)...\lambda_i(t_k) H_k(t_1, ..., t_k) dt_k ... dt_1 \\ &= \frac{2}{k! k(k-1)} \times \left(\sum_{r<s} \int\limits_{(a_i, b_i]^2} \lambda_i(t_r) \lambda_i(t_s) H_2(t_r, t_s) dt_s dt_r \right) \\ &\qquad \times \left(\int_{a_i}^{b_i} \lambda_i(s) ds \right)^{k-2} \\ &= \frac{2}{k!} \int\limits_{Q_i(2)} \lambda_i(t_r) \lambda_i(t_s) H_2(t_r, t_s) dt_s dt_r \times \left(\int_{a_i}^{b_i} \lambda_i(s) ds \right)^{k-2}. \end{aligned}$$

Hence, the log-likelihood contribution, up to a constant, is:

$$\begin{aligned} l_i^k(\alpha,\beta,\theta;t_{i1},...,t_{ik}) &= \sum_{j=1}^{k}\log\lambda_i(t_{ij}) + \log\sum_{r<s} H_2(t_r,t_s) \\ &\quad - \log\int\limits_{Q_i(2)} \lambda_i(t_r)\lambda_i(t_s)H_2(t_r,t_s)dt_s dt_r - (k-2)\log\int\limits_{a_i}^{b_i}\lambda_i(s)ds. \end{aligned} \tag{13}$$

The right-hand side of the identity (10) may be written:

$$\begin{aligned} &\frac{k-2}{k}\left\{\sum_{j=1}^{k}\log\lambda_i(t_{ij}) - k\log\int\limits_{a_i}^{b_i}\lambda_i(t)dt\right\} \\ &+\frac{2}{k(k-1)}\left\{\sum_{r<s}\{\log\lambda_i(t_{ir})+\log\lambda_i(t_{is})\}\right. \\ &\left.+\sum_{r<s}\log H_2(t_{ir},t_{is}) - \frac{k(k-1)}{2}\log\int\limits_{Q_i(2)}\lambda_i(t)\lambda_i(s)H_2(t,s)dsdt\right\} \\ &+\log\left\{\frac{2}{k(k-1)}\sum_{r<s}H_2(t_{ir},t_{is})\right\} - \frac{2}{k(k-1)}\sum_{r<s}\log H_2(t_r,t_s). \end{aligned} \tag{14}$$

Now,

$$\sum_{r<s}\{\log\lambda_i(t_{ir})+\log\lambda_i(t_{is})\} = (k-1)\sum_{j=1}^{k}\log\lambda_i(t_{ij}),$$

and simplification of (14) results in the same expression as (13), up to an irrelevant constant.