STATISTICS IN MEDICINE
Statist. Med. **19**, 13–33 (2000)

SURVIVAL ANALYSIS FOR RECURRENT EVENT DATA: AN APPLICATION TO CHILDHOOD INFECTIOUS DISEASES

PATRICK J. KELLY* AND LYNETTE L-Y. LIM

Centre for Clinical Epidemiology and Biostatistics, The University of Newcastle, Level 3, David Maddison Building, Royal Newcastle Hospital, Newcastle, NSW, 2300, Australia

SUMMARY

Many extensions of survival models based on the Cox proportional hazards approach have been proposed to handle clustered or multiple event data. Of particular note are five Cox-based models for recurrent event data: Andersen and Gill (AG); Wei, Lin and Weissfeld (WLW); Prentice, Williams and Peterson, total time (PWP-CP) and gap time (PWP-GT); and Lee, Wei and Amato (LWA). Some authors have compared these models by observing differences that arise from fitting the models to real and simulated data. However, no attempt has been made to systematically identify the components of the models that are appropriate for recurrent event data. We propose a systematic way of characterizing such Cox-based models using four key components: risk intervals; baseline hazard; risk set, and correlation adjustment. From the definitions of risk interval and risk set there are conceptually seven such Cox-based models that are permissible, five of which are those previously identified. The two new variant models are termed the 'total time – restricted' (TT-R) and 'gap time – unrestricted' (GT-UR) models. The aim of the paper is to determine which models are appropriate for recurrent event data using the key components. The models are fitted to simulated data sets and to a data set of childhood recurrent infectious diseases. The LWA model is not appropriate for recurrent event data because it allows a subject to be at risk several times for the same event. The WLW model overestimates treatment effect and is not recommended. We conclude that PWP-GT and TT-R are useful models for analysing recurrent event data, providing answers to slightly different research questions. Further, applying a robust variance to any of these models does not adequately account for within-subject correlation. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Event times may be correlated due to either clustered or multiple events. Clustered events occur when each subject experiences one event but the event times are correlated due to the subjects being in groups or clusters, for example, siblings and schools. Multiple event data occurs when each subject can have more than one event, so that the events from the same subject are potentially correlated. Multiple event data can be further divided into two categories: recurrent or multiple-type events. Recurrent event data is where the subject experiences repeated occurrences of the same type of event, for example repeated asthma attacks. Multiple-type event data is where the subject experiences events of entirely different natures, for example the occurrence of tumours at different sites in the body.

* Correspondence to: Patrick J. Kelly, Centre for Clinical Epidemiology and Biostatistics, The University of Newcastle, Level 3, David Maddison Building, Royal Newcastle Hospital, Newcastle, NSW. 2300, Australia. E-mail: pkelly@mail.newcastle.edu.au

CCC 0277–6715/2000/010013–21$17.50

Received December 1997
Accepted January 1999

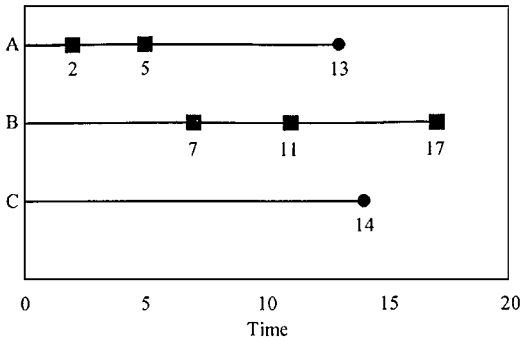

Figure 1. A hypothetical example of three subjects with recurrent events. Observations start at the same time; ■ is the occurrence of an event and ● is censoring. Subject A has two events before being censored; subject B has three events, ending the period of observation with an event; and subject C has no events before being censored

Many survival models based on the Cox proportional hazards have been proposed that handle clustered and multiple event data. In particular, there are five prominent Cox-based models for recurrent event data: Andersen and Gill[1] (AG); Prentice Williams and Peterson[2] gap time (PWP-GT) and total time (PWP-CP); Lee, Wei and Amato[3] (LWA); and Wei, Lin and Weissfeld[4] (WLW). Some of these models have been compared using real and simulated data, and it is well known that these models give different results.[5–9] However, it remains unclear which models are suitable for recurrent event data, as the differences between these models are subtle and no attempt has been made to systematically identify how the models differ from each other.

We propose a systematic way of characterizing such Cox-based models using four key components: risk intervals; baseline hazard; risk set, and correlation adjustment. There are two aims of this paper. First, to determine which models are suitable for recurrent event data, using the key components. Secondly, to ascertain whether applying a robust variance is adequate when there is within-subject correlation.

The four key components of a model are described in Section 2. In Section 3 the models and their associated partial likelihoods are constructed using the key components. In Section 4 the models are applied to different simulations that examine recurrent event data when treatment is constant, when treatment is effective for the first event only and when there is within-subject correlation. Section 5 applies the models to a data set of recurrent respiratory infections in children. Section 6 discusses previous work and the software available. Section 7 concludes with the recommended models for recurrent event data.

2. KEY MODEL COMPONENTS

There are four components to a Cox-based recurrent event model: definition of the risk interval; definition of the risk set; choice of a common versus event-specific baseline hazards, and handling of within-subject correlation. These concepts are illustrated with a hypothetical example (Figure 1) of the experiences of three subjects with recurrent events.

2.1. Risk intervals

Risk intervals define when a subject is at risk of having an event along a given time scale. Three formulations are available: gap time; total time, and counting process formulation. Figure 2

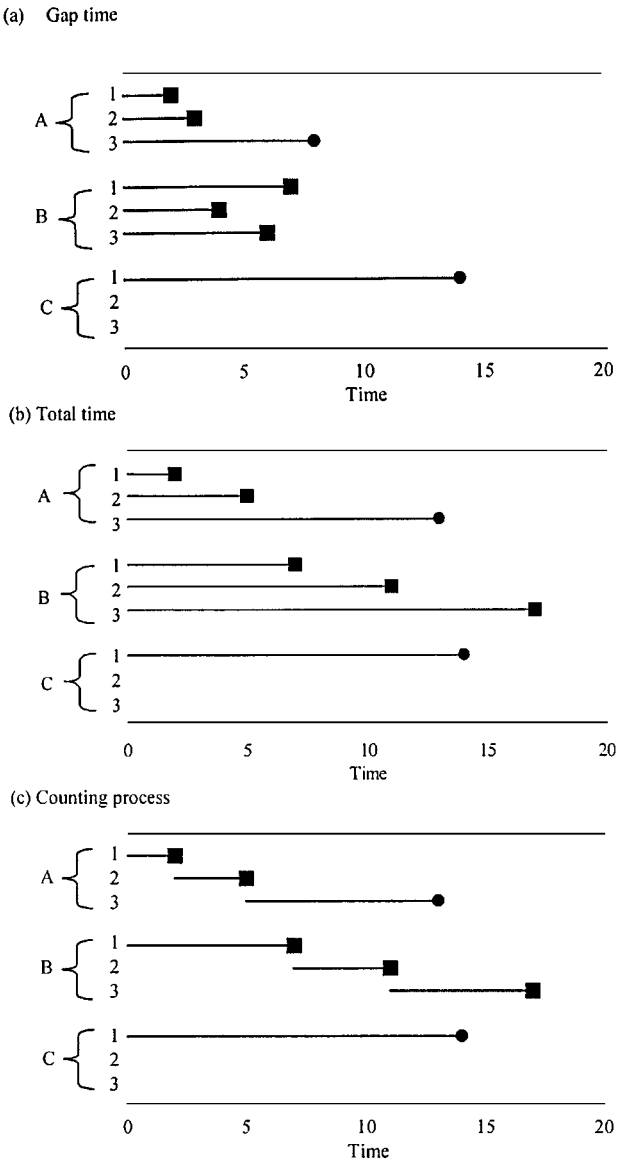

Figure 2. Illustrations of the risk interval formulations: (*a*) gap time; (*b*) total time; (*c*) counting process, using the hypothetical data from Figure 1, where ■ is an event and ● is censoring. Each time to an event or censoring is a separate risk interval, hence subjects A and B have three separate intervals

displays the three types of risk intervals for the subjects from Figure 1. Gap time is the time from the prior event, that is, the clock restarts after each event. For example, with gap time, subject A (Figure 2(*a*)) is at risk of the first event during (0, 2], and of the second and third events during (0, 3] and (0, 8], respectively. Total time is the time from a chosen point, usually the time from the start of treatment. With total time, subject A (Figure 2(*b*)) is at risk for the first, second and third

events during (0, 2], (0, 5] and (0, 13], respectively. Counting process[10] uses the same time scale as total time, but recognizes that a subject may have a delayed entry or censored period before the subject becomes at risk for the event; with recurrent events a subject is not considered to be at risk for the kth event until after the $(k-1)$th event. In the counting process formulation, subject A (Figure 2(*c*)) is at risk for the first event during (0, 2], and the second and third events during (2, 5] and (5, 13], respectively. In both gap time and counting process formulations the subject is at risk for the same length of time. The risk interval for the first event is the same for all three risk interval definitions.

The risk interval determines whether a model is either marginal or conditional. Gap time and counting process are conditional since a subject cannot be at risk until the end of the previous event, that is, the subject is at risk conditioned on previous events. Total time is marginal since the subject is a risk from the start of treatment and does not depend on any previous events.

2.2 Baseline Hazard

There are two choices for the baseline hazard function for recurrent event models: common and event-specific. A model with a common baseline hazard has the same underlying hazard for all events. An event-specific baseline hazard is a stratified baseline hazard that allows the baseline hazard to be different for each kth event. Stratifying by event is essentially fitting a separate model for each kth event.

2.3. Risk Set

The kth risk set contains the individuals who are at risk for the kth event. There are three possible risk sets: unrestricted; restricted, or semi-restricted. The risk set definition incorporates the choice of baseline hazard. The risk set at a given point in time depends on the individuals included in the set and when those individuals are at risk, that is, the risk interval.

When the risk set is unrestricted, all the subjects' risk intervals may contribute to the risk set for any given event, regardless of the number of events experienced by each subject. Therefore, a subject's second event time may contribute to the risk set corresponding to another subject's first event. For example, consider the second event of B. In the gap time formulation, the risk set for the second event of subject B at time 4 (Figure 2(*a*)) includes information from the third event of A, the first, second and third event of B, and the first event of C. In the total time formulation, contributions to the risk set corresponding to the second event of subject B at time 11 (Figure 2(*b*)) includes information from the third event time of A, the second and third event times of B and the single event time of C. In the counting process formulation, the second event time of subject B (Figure 2(*c*)) the risk set includes the contribution of the third event of subject A, the second event of B and the first event of C. An unrestricted risk set has a common baseline hazard function for all events.

With a restricted risk set, contributions to the kth risk set is restricted to only include the kth event risk intervals of those subjects who have experienced $(k-1)$ events. For example, only subjects who have already had three events will be considered to be at risk for the fourth event. In both the gap time and total time formulation, the risk set for the second event of subject A at time 3 (Figure 2(*a*) and 2(*b*)) includes information from the second event of A and B. In the counting process formulation, contributions to the risk set corresponding to the second event of subject A at time 5 (Figure 2(*c*)) only includes the second event time of A. A restricted risk set has event-specific baseline hazards.

Table I. Definition of risk interval and risk set for each model

Model characteristics Risk interval	Risk set/baseline hazard		
	Unrestricted/ common	Semi-restricted/ event-specific	Restricted/ event-specific
Gap time	Possible (GT-UR)	Not possible	PWP-GT
Total time	LWA	WLW	Possible (TT-R)
Counting process	AG	Possible	PWP-CP

Semi-restricted risk sets have event-specific baseline hazards but allow subjects who have less than $(k-1)$ events to be at risk for the kth event through the creation of 'dummy' risk intervals.[4] Thus a subject who has had none or one event can be considered at risk of a fourth event. However, a semi-restricted risk set does not allow information from the kth event risk interval to contribute to the risk set for an earlier event. This risk set only applies to the total time and counting process formulation with event-specific baseline hazards. For example, in total time the second event of subject B and the 'dummy' risk interval of subject C contributes to the risk set for the second event of subject A (Figure 2(*b*)). However, the first risk interval for subject B cannot contribute to the risk set of this event due to the event-specific baseline hazard.

2.4. Within-subject correlation

Three approaches have been proposed for accounting for the within-subject correlation between events: conditional; marginal, and random effects. The conditional approach assumes that the current event is unaffected by earlier events that occurred to the subject. This assumption can be relaxed by introducing time-dependent covariates in the model, such as the number of prior recurrences, which may capture the dependence structure among the recurrence times. The marginal approach assumes that the events within a subject are independent and estimates a robust variance using a 'sandwich' estimator.[11–13] The unadjusted variance estimates are called naive estimates. The random effects approach, also called frailty models, introduces a random covariate into the model that induces dependence among the recurrent event times.[14,15]

3. MODELS

Of the four key components, two are pivotal for constructing a model – risk interval and risk set. These also determine the interpretation of the resulting model. Baseline hazard is linked to risk set definition. That is, an unrestricted risk set is necessarily associated with a common baseline hazard and the semi-restricted and restricted risk sets are necessarily associated with event-specific hazards.

Table I shows the possible models cross-classified by choice of risk set and risk interval, with the five well-known Cox-based models indicated: AG; PWP-GT; PWP-CP; LWA, and WLW. The model comprising a semi-restricted risk set and the gap time risk interval is not permissible because it depends on the time between events. The remaining cells define permissible models that have not been explicitly identified in the literature. These are gap time with an unrestricted risk set (gap time – unrestricted model, GT-UR) and total time with a restricted risk set (total time

– restricted model, TT-R). It is possible to have a semi-restricted counting process model, but it does not make any sense to use this model for recurrent event data since the counting process intervals are used to prevent a subject from being at risk for an event until after the previous event, which is violated when using a semi-restricted risk set. In practical terms a robust variance can be calculated for any of the permissible models.[5,16]

The features of the five well-known Cox-based models in relation to the key components, with their associated partial likelihoods, are explained in the following sections.

3.1. Notation

Let T_{ik} be the true total time of the kth event for the ith subject, C_{ik} is the censoring time of the kth event in the ith subject, and X_{ik} is the corresponding observation time, $X_{ik} = \min(T_{ik}, C_{ik})$. Let $I(\cdot)$ be the indicator variable where $I(E) = 1$ if E is true and $I(E) = 0$ otherwise. The censoring variable is $\delta_{ik} = I(T_{ik} \leqslant C_{ik})$. Let $G_{ik} = X_{ik} - X_{i,k-1}$ be the gap time with $X_{i0} = 0$. Let $\lambda_{ik}(t)$ denote the hazard function for the kth event of the ith subject at time t, $\lambda_0(t)$ represents a common baseline hazard for all events, and $\lambda_{0k}(t)$ is an event-specific baseline hazard for the kth event. Let $Z_{ik} = (Z_{1ik}, \ldots, Z_{pik})'$ denote the covariate vector for the ith subject with respect to the kth event, $Z_i = (Z'_{i1}, \ldots, Z'_{iK})$ is the covariate vector for the ith subject, where K is the maximum number of events within a subject, and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters.

3.2. Partial Likelihood

The partial likelihood (L) is defined as

$$L(\beta) = \prod_{j=1}^{d} \frac{\lambda(t_{(j)})}{\sum_{k \in R(t_{(j)})} \lambda(t_k)}$$

where d is the observed number of events.[17] Let the ordered event times be $t_{(1)} < t_{(2)} \cdots < t_{(d)}$, where $t_{(j)}$ is the jth ordered event time. The corresponding risk set, denoted $R(t_{(i)})$, is the set of individuals at risk at time $t_{(i)}$.

Table II displays the hazard function and partial likelihood equation for each model, and illustrates the formulation of the partial likelihoods using the hypothetical example from Figure 1. Table II is divided into two parts: (*a*) displays the three unstratified models, AG, LWA and GT-UR; (*b*) displays the four stratified models, PWP-GT, PWP-CP, WLW and TT-R. Examination of the example partial likelihoods is helpful for understanding how parameter estimates of the various models differ. The example partial likelihood for the seven models (Table II) is the product of five terms. The first two correspond to the events experienced by subject A (at times 2 and 5), and the last three terms to those experienced by subject B (at times 7, 11 and 17). In the example, i = A, B, C; $k = 1, 2, 3$. Let A1 denote the first event of subject A, and so on. For example, the hazard for the second event of subject B at time 11 is denoted, $\lambda_{B2}(11)$.

3.3. Conditional Models

The AG model defines the risk intervals using counting process, with an unrestricted risk set and hence assumes a common baseline hazard for all events. The example AG partial likelihood (Table II(*a*)) is constructed using Figure 2(*c*). Note that with counting process a subject can only make one contribution to the risk set (denominator) for a given event.

Even though the PWP-CP has been called a 'total time' model,[2] it actually uses counting process formulation. The PWP-CP model is a stratified AG model; it has event-specific baseline hazards and risk set is restricted. To determine the example partial likelihood in Table II(*b*) use Figure 2(*c*) again, but only include in the risk set those who have had the same number of previous events. The PWP-GT model differs from PWP-CP by using gap time instead of counting process. Hence the example partial likelihood in Table II(*b*) employs Figure 2(*a*).

Since the PWP models have event-specific baseline hazards, we can have either an overall estimate or event-specific estimates for each covariate. The overall estimate, $\hat{\beta}$, is obtained by fitting the single covariate vector Z_i to the model. The event-specific estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_k$, are obtained by fitting event-specific covariates to the model, such that $Z_i = (Z_{i1}, 0, \ldots, 0)'$, $Z_i = (0, Z_{i2}, \ldots, 0)', \ldots,$ and $Z_i = (0, \ldots, 0, Z_{iK})'$, for $k = 1, 2, \ldots, K$, respectively. In practice the data may need to be limited to a maximum number of events if the risk set becomes very small for later strata and the event-specific estimates become too unreliable.

The AG and PWP models as defined by their orginators make no adjustment to account for the correlation within subjects, except by including covariates in the regression equation. However, the robust variance 'sandwich' estimator has been applied by other researchers to these conditional models in a further attempt to adjust for correlation.[1,16]

3.4. Marginal models

The LWA model employs total time and assumes a common baseline hazard with an unrestricted risk set. The example partial likelihood (Table II(*a*)) uses Figure 2(*b*), including all risk intervals at time t. The LWA model allows a subject to be at risk for several events simultaneously. For example, the risk set for the first event of subject A includes subjects A and B three times. That is, a subject with n risk intervals may contribute to a risk set n times; none of the other models allows this. The LWA model accounts for the within-subject correlation by adjusting the variance via the 'sandwich' estimator.

The WLW is an event-specific LWA model with a semi-restricted risk set. To calculate the partial likelihood (Table II(*b*)), Figure 2(*b*) is used. Since there is a maximum of three events, each subject is potentially at risk for three events, due to the semi-restricted risk set. Therefore, 'dummy' records are created so that there are three records for each subject. For example, the risk interval for subject C is (0, 17] for the first event, (0, 17] for the second, and (0, 17] for the third.

The estimates of WLW, like PWP, can be either event-specific or overall. However, the overall estimate proposed by WLW is defined differently; it is the weighted average of the event-specific estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_K$, such that the corresponding weighted average of the robust variance is the smallest possible.[4] The WLW model also requires limitation of the data to a maximum number of events if event-specific estimates become unreliable.

4. SIMULATIONS

Simulations were conducted to investigate any biases and evaluate the inferences of the permissible models for recurrent event data with and without within-subject correlation. For each subject the time to the next event, t_{ik} , was generated using an exponential distribution such that

$$\log t_{ik} = \beta_0^* + \beta_k^{*'} Z_{ik} + v_i^* \tag{1}$$

Table II. The equations of the hazard functions and partial likelihoods for the unstratified (unrestricted) and stratified models. The partial likelihoods are illustrated with the hypothetical data of Figure 1

Partial likelihood	Model	Hazard	Example
(a) *Unstratified (unrestricted) models*			
$L(\beta)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n}\sum_{l=1}^{K} Y_{jl}(X_{ik})e^{\beta' Z_{jl}(X_{ik})}}\right)^{\delta_{ik}}$	AG	$\lambda_{ik}(t;Z_{ik})=\lambda_0(t)e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)}\times\frac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{B1}(5)+\lambda_{C1}(5)}$
		$Y_{ik}(t)=I(X_{i,k-1}<t\leqslant X_{ik})$	$\times\frac{\lambda_{B1}(7)}{\lambda_{A3}(7)+\lambda_{B1}(7)+\lambda_{C1}(7)}\times\frac{\lambda_{B2}(11)}{\lambda_{A3}(11)+\lambda_{B2}(11)+\lambda_{C1}(11)}\times\frac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$
	LWA	$\lambda_{ik}(t;Z_{ik})=\lambda_0(t)e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{A2}(2)+\lambda_{A3}(2)+\lambda_{B1}(2)+\lambda_{B2}(2)+\lambda_{B3}(2)+\lambda_{C1}(2)}$
		$Y_{ik}(t)=I(X_{ik}\geqslant t)$	$\times\frac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{A3}(5)+\lambda_{B1}(5)+\lambda_{B2}(5)+\lambda_{B3}(5)+\lambda_{C1}(5)}$
			$\times\frac{\lambda_{B1}(7)}{\lambda_{A3}(7)+\lambda_{B1}(7)+\lambda_{B2}(7)+\lambda_{B3}(7)+\lambda_{C1}(7)}$
			$\times\frac{\lambda_{B2}(11)}{\lambda_{A3}(11)+\lambda\lambda_{B2}(11)+\lambda_{B3}(11)+\lambda_{C1}(11)}\times\frac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$
	GT-UR	$\lambda_{ik}(t;Z_{ik})=\lambda_{0k}(t-t_{k-1})e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{A2}(2)+\lambda_{A3}(2)+\lambda_{B1}(2)+\lambda_{B2}(2)+\lambda_{B3}(2)+\lambda_{C1}(2)}$
		$Z_{jk}(X_{ik})$ replaced by $Z_{ik}(X_{i,k-1}+G_{ik})$	$\frac{\lambda_{A2}(3)}{\lambda_{A2}(3)+\lambda_{A3}(3)+\lambda_{B1}(3)+\lambda_{B2}(3)+\lambda_{B3}(3)+\lambda_{C1}(3)}$
		$Y_{ik}(t)=I(G_{ik}>t)$	$\times\frac{\lambda_{B1}(7)}{\lambda_{A3}(7)+\lambda_{B1}(7)+\lambda_{C1}(7)}$
			$\times\frac{\lambda_{B2}(4)}{\lambda_{A3}(4)+\lambda_{B1}(4)+\lambda_{B2}(4)+\lambda_{B3}(4)+\lambda_{C1}(4)}$
			$\times\frac{\lambda_{B3}(6)}{\lambda_{A3}(6)+\lambda_{B1}(6)+\lambda_{B3}(6)+\lambda_{C1}(6)}$

(*b*) *Stratified models*

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) e^{\beta' Z_{ik}(X_{ik})}} \right)^{\delta_{ik}}$$

PWP-CP	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}$	$\dfrac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \dfrac{\lambda_{A2}(5)}{\lambda_{A2}(5)}$
	$Y_{ik}(t) = I(X_{i,k-1} < t \leqslant X_{ik})$	$\times \dfrac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \dfrac{\lambda_{B2}(11)}{\lambda_{B2}(11)} \times \dfrac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$
PWP-GT	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t-t_{k-1})e^{\beta' Z_{ik}(t)}$	$\dfrac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \dfrac{\lambda_{A2}(3)}{\lambda_{A2}(3)+\lambda_{B2}(3)}$
	$Z_{jk}(X_{ik})$ replaced by $Z_{ik}(X_{i,k-1}+G_{ik})$	$\times \dfrac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \dfrac{\lambda_{B2}(4)}{\lambda_{B2}(4)} \times \dfrac{\lambda_{B3}(6)}{\lambda_{B3}(6)}$
	$Y_{ik}(t) = I(G_{ik} > t)$	
WLW*	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}$	$\dfrac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \dfrac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{B2}(5)+\lambda_{C1}(5)}$
	$Y_{ik}(t) = I(X_{ik} \geqslant t)$	$\times \dfrac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \dfrac{\lambda_{B2}(11)}{\lambda_{B2}(11)+\lambda_{C1}(11)} \times \dfrac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$
TT-R*	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}$	$\dfrac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \dfrac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{B2}(5)}$
	$Y_{ik}(t) = I(X_{ik} \geqslant t)$	$\times \dfrac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \dfrac{\lambda_{B2}(11)}{\lambda_{B2}(11)} \times \dfrac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$

*WLW and TT-R share the same equation, but their partial likelihoods differ due to the 'dummy' risk intervals employed by WLW

which is equivalent to

$$\log \lambda_{ik}(t) = \beta_0 + \beta_k' Z_{ik} + v_i$$

or

$$\lambda_{ik}(t) = \lambda_0(t) e^{(\beta_k' Z_{ik} + v_i)}$$

where $\beta^* = -\beta$, $v_i^* = -v_i$ and $\lambda_0(t) = e^{\beta_0}$. For our examples, $\beta_0 = 3$, $i = 500$ subjects with 250 receiving treatment, and $k = 1, 2, 3, 4$ events. That is, we consider subjects to have a maximum of four events. Events are censored if the additive total time since the start of the study is greater than 120 days. We consider two different treatment scenarios:

(i) treatment is constantly effective where $\beta_1^* = \beta_2^* = \beta_3^* = \beta_4^* = 1{\cdot}0$;
(ii) treatment is effective for only the first event, where $\beta_1^* = 1{\cdot}0$, $\beta_2^* = \beta_3^* = \beta_4^* = 0$.

For the gap time models, given how the data are simulated, it is clear that $\beta_k = -1{\cdot}0$ for all k when treatment is constant and $\beta_1 = -1{\cdot}0, \beta_2 = \beta_3 = \beta_4 = 0$ when treatment is effective for the first event only. However, for the total time models, since time is defined from the start of the study, β_k may be different compared to the gap time models. By rearranging equation (1) it can be shown that the expected value of β_k^* for total time is the log(expected time for treated subjects since start of study) minus the log(expected time for placebo subjects since start of study). Using this equation and $\beta_k = -\beta_k^*$, the β_k values for total time are $-1{\cdot}0$ for all k when treatment is constant, and $-1{\cdot}0$, $-0{\cdot}62$, $-0{\cdot}45$ and $-0{\cdot}36$, for $k = 1, 2, 3, 4$, respectively, when treatment is effective for the first event only. Hence, when treatment is effective for the first event the total time model displays a 'carry-over' effect where treatment effect diminishes with each consecutive event. This is observed since, if treatment is effective for one event, the total time to an event for a treated subject will always be longer compared to a placebo subject. However, as the time since treatment increases, the relative difference between treated and placebo subjects will get smaller, and so a gradual decrease in treatment effect is observed.

To induce within-subject correlation we include a random effect covariate v_i^*, which is normally distributed with a mean of 0 and variance σ^2. v_i^* is a constant value for events within a subject, but different between subjects. We examine three different values of σ^2, 0, 0·1 and 0·4. Large values of σ reflect greater heterogeneity between subjects and a stronger association among within-subject events. Sequential events are independent when v_i is zero.

A 100 independent data sets are generated for both treatment scenarios. Table III shows the results for constant treatment effect when: (a) $\sigma^2 = 0$; (b) $\sigma^2 = 0{\cdot}1$, and (c) $\sigma^2 = 0{\cdot}4$. Table IV shows the results when treatment is effective for the first event only with: (a) $\sigma^2 = 0$; (b) $\sigma^2 = 0{\cdot}1$, and (c) $\sigma^2 = 0{\cdot}4$. Each table displays: the true treatment effect; the mean, standard deviation and bias of the estimated treatment; the mean naive and robust standard errors; and the empirical coverage of these estimates using the naive and robust standard error at the nominal 95 per cent confidence level.

4.1. Constant treatment effect

4.1.1. Independent events

First consider when treatment is constant and there is no induced within-subject correlation (Table III(*a*)):

(i) The event-specific treatment estimates for the PWP-GT and PWP-CP models have very similar results with negligible bias. The total time models overestimate the treatment effect

Table III. The event-specific, weighted average* and common† treatment estimates when treatment effect is constant with: (*a*) $\sigma^2 = 0$; (*b*) $\sigma^2 = 0{\cdot}1$ and (*c*) $\sigma^2 = 0{\cdot}4$. NSE and RSE are the naive and robust standard errors, respectively, with corresponding empirical coverage of 95 per cent confidence intervals

Model		β	Mean($\hat{\beta}$)	SD($\hat{\beta}$)	Bias	Mean estimated NSE($\hat{\beta}$)	Mean estimated RSE($\hat{\beta}$)	Coverage naive 95 per cent CI	Coverage robust 95 per cent CI
(*a*) $\sigma^2 = 0$									
PWP-GT	Event 1	−1·0	−1·013	0·100	−0·013	0·100	0·099	0·97	0·95
	Event 2	−1·0	−1·014	0·113	−0·014	0·106	0·106	0·96	0·93
	Event 3	−1·0	−0·987	0·137	0·013	0·125	0·124	0·94	0·93
	Event 4	−1·0	−1·008	0·170	−0·008	0·168	0·166	0·95	0·95
	Weighted average	−1·0	−1·004	0·058	−0·004		0·058		0·97
	Common	−1·0	−1·006	0·059	−0·006	0·059	0·059	0·97	0·97
PWP-CP	Event 1	−1·0	−0·013	0·100	−1·013	0·100	0·099	0·97	0·95
	Event 2	−1·0	−1·010	0·113	−0·010	0·112	0·111	0·97	0·97
	Event 3	−1·0	−0·985	0·140	0·015	0·133	0·132	0·92	0·92
	Event 4	−1·0	−0·997	0·179	0·003	0·174	0·172	0·94	0·94
	Weighted average	−1·0	−1·002	0·061	−0·002		0·060		0·98
	Common	−1·0	−1·004	0·061	−0·004	0·061	0·061	0·98	0·98
TT-R	Event 1	−1·0	−1·013	0·100	−0·013	0·100	0·099	0·97	0·95
	Event 2	−1·0	−1·303	0·112	−0·303	0·109	0·110	0·21	0·22
	Event 3	−1·0	−1·369	0·139	−0·369	0·127	0·125	0·18	0·17
	Event 4	−1·0	−1·366	0·154	−0·366	0·169	0·162	0·38	0·35
	Weighted average	−1·0	−1·190	0·084	−0·190		0·081		0·36
	Common	−1·0	−1·224	0·082	−0·224	0·059	0·080	0·07	0·24
WLW	Event 1	−1·0	−1·013	0·100	−0·013	0·100	0·099	0·97	0·95
	Event 2	−1·0	−1·459	0·112	−0·459	0·109	0·110	0·02	0·01
	Event 3	−1·0	−1·845	0·137	−0·845	0·128	0·126	0·00	0·00
	Event 4	−1·0	−2·296	0·151	−1·296	0·169	0·165	0·00	0·00
	Weighted average	−1·0	−1·263	0·108	−0·263		0·092		0·25
	Common	−1·0	−1·521	0·097	−0·521	0·058	0·095	0·00	0·00
GT-UR	Common	−1·0	−1·004	0·056	−0·004	0·057	0·057	0·94	0·96
AG	Common	−1·0	−1·000	0·054	−0·000	0·056	0·056	0·96	0·96
LWA	Common	−1·0	−0·914	0·055	−0·086	0·056	0·057	0·68	0·72
(*b*) $\sigma^2 = 0{\cdot}1$									
PWP-GT	Event 1	−1·0	−0·938	0·090	0·090	0·098	0·099	0·94	0·94
	Event 2	−1·0	−0·922	0·098	0·078	0·105	0·106	0·91	0·91
	Event 3	−1·0	−0·881	0·121	0·119	0·123	0·123	0·87	0·85
	Event 4	−1·0	−0·878	0·159	0·122	0·157	0·156	0·85	0·85
	Weighted average	−1·0	−0·910	0·054	0·090		0·059		0·70
	Common	−1·0	−0·912	0·054	0·088	0·058	0·060	0·70	0·73

Table III. Continued

Model		β	Mean($\hat{\beta}$)	SD($\hat{\beta}$)	Bias	Mean estimated NSE($\hat{\beta}$)	Mean estimated RSE($\hat{\beta}$)	Coverage naive 95 per cent CI	Coverage robust 95 per cent CI
PWP-CP	Event 1	− 1·0	− 0·938	0·090	0·062	0·098	0·099	0·94	0·94
	Event 2	− 1·0	− 0·884	0·101	0·116	0·110	0·110	0·83	0·83
	Event 3	− 1·0	− 0·829	0·127	0·171	0·128	0·128	0·75	0·72
	Event 4	− 1·0	− 0·818	0·167	0·182	0·162	0·160	0·78	0·80
	Weighted average	− 1·0	− 0·880	0·053	0·120		0·059		0·45
	Common	− 1·0	− 0·883	0·054	0·117	0·060	0·060	0·51	0·50
TT-R	Event 1	− 1·0	− 0·938	0·090	0·062	0·098	0·099	0·94	0·94
	Event 2	− 1·0	− 1·149	0·103	− 0·149	0·107	0·108	0·70	0·72
	Event 3	− 1·0	− 1·171	0·117	− 0·200	0·124	0·122	0·73	0·71
	Event 4	− 1·0	− 1·153	0·149	− 0·153	0·158	0·150	0·87	0·83
	Weighted average	− 1·0	− 1·063	0·075	− 0·063		0·079		0·92
	Common	− 1·0	− 1·081	0·074	− 0·081	0·058	0·079	0·64	0·82
WLW	Event 1	− 1·0	− 0·938	0·090	0·062	0·098	0·099	0·94	0·94
	Event 2	− 1·0	− 1·315	0·099	− 0·315	0·107	0·108	0·17	0·16
	Event 3	− 1·0	− 1·634	0·115	− 0·634	0·125	0·123	0·00	0·00
	Event 4	− 1·0	− 1·988	0·140	− 0·988	0·158	0·154	0·00	0·00
	Weighted average	− 1·0	− 1·152	0·094	− 0·152		0·093		0·62
	Common	− 1·0	− 1·365	0·086	− 0·365	0·057	0·095	0·00	0·03
GT-UR	Common	− 1·0	− 0·922	0·053	0·078	0·057	0·059	0·74	0·78
AG	Common	− 1·0	− 0·937	0·053	0·063	0·056	0·059	0·82	0·86
LWA	Common	− 1·0	− 0·838	0·056	0·162	0·056	0·059	0·20	0·24
(*c*) $\sigma^2 = 0{\cdot}4$									
PWP-GT	Event 1	− 1·0	− 0·808	0·108	0·192	0·097	0·099	0·50	0·51
	Event 2	− 1·0	− 0·736	0·114	0·264	0·106	0·106	0·32	0·32
	Event 3	− 1·0	− 0·668	0·123	0·332	0·121	0·120	0·25	0·24
	Event 4	− 1·0	− 0·627	0·138	0·373	0·143	0·140	0·21	0·21
	Weighted average	− 1·0	− 0·725	0·066	0·275		0·061		0·01
	Common	− 1·0	− 0·729	0·067	0·271	0·057	0·062	0·01	0·01
PWP-CP	Event 1	− 1·0	− 0·808	0·108	0·192	0·097	0·099	0·50	0·51
	Event 2	− 1·0	− 0·649	0·118	0·351	0·108	0·108	0·11	0·11
	Event 3	− 1·0	− 0·557	0·124	0·443	0·125	0·124	0·05	0·06
	Event 4	− 1·0	− 0·499	0·142	0·501	0·148	0·143	0·06	0·06
	Weighted average	− 1·0	− 0·658	0·063	0·342		0·057		0·00
	Common	− 1·0	− 0·664	0·062	0·336	0·058	0·059	0·00	0·00
TT-R	Event 1	− 1·0	− 0·808	0·108	0·192	0·097	0·099	0·50	0·51
	Event 2	− 1·0	− 0·886	0·113	0·114	0·106	0·106	0·78	0·78
	Event 3	− 1·0	− 0·860	0·127	0·140	0·121	0·118	0·72	0·71
	Event 4	− 1·0	− 0·816	0·132	0·184	0·143	0·138	0·79	0·77
	Weighted average	− 1·0	− 0·836	0·085	0·164		0·076		0·41
	Common	− 1·0	− 0·842	0·085	0·158	0·057	0·077	0·27	0·44

Table III. Continued

Model		β	Mean($\hat{\beta}$)	SD($\hat{\beta}$)	Bias	Mean estimated NSE($\hat{\beta}$)	Mean estimated RSE($\hat{\beta}$)	Coverage naive 95 per cent CI	Coverage robust 95 per cent CI
WLW	Event 1	− 1·0	− 0·808	0·108	0·192	0·097	0·099	0·50	0·51
	Event 2	− 1·0	− 1·058	0·111	− 0·058	0·106	0·106	0·90	0·90
	Event 3	− 1·0	− 1·266	0·126	− 0·266	0·121	0·119	0·48	0·48
	Event 4	− 1·0	− 1·479	0·147	− 0·479	0·143	0·140	0·10	0·09
	Weighted average	− 1·0	− 0·943	0·106	0·057		0·094		0·83
	Common	− 1·0	− 1·091	0·105	− 0·091	0·056	0·097	0·56	0·83
GT-UR	Common	− 1·0	− 0·754	0·066	0·246	0·056	0·063	0·03	0·04
AG	Common	− 1·0	− 0·795	0·068	0·205	0·055	0·066	0·08	0·13
LWA	Common	− 1·0	− 0·698	0·068	0·302	0·056	0·063	0·01	0·02

* Weighted average of the event-specific estimates such that the robust standard error is the smallest, as defined by WLW.
† Model fitted with one parameter

after the first event, with the overestimation becoming larger with each consecutive event in the WLW model. WLW concludes that for the 4th event treatment is over twice as effective than it is in reality. For all event-specific estimates the robust standard errors are similar to the naive standard errors. Event-specific standard errors between models are also similar. Hence, the empirical coverage probabilities of the confidence intervals are close to the nominal 95 per cent level for all event-specific estimates for PWP-GT and PWP-CP model but is lower for the total time models, especially WLW.

(ii) The common and weighted average estimates in the event-specific models are in general very similar. These overall treatment estimates reflect the 'average' of the event-specific estimates. Only the WLW model had a statistically significant difference between the two overall treatment effects, with the weighted average being less biased and having a better empirical coverage of the nominal confidence interval level. The robust standard errors are similar to the naive for the common treatment estimate in the PWP-GT and PWP-CP models, but are inflated for the total time models.

(iii) For the unrestricted models it can be seen that the common treatment estimate for GT-UR and AG are similar to their corresponding restricted models, PWP-GT and PWP-CP, and have minimal bias. LWA, on the other hand, underestimates the treatment effect which is opposite to the other total time models, TT-R and WLW. The robust standard errors are the same as the naive in all unrestricted models.

4.1.2. Within-subject correlation

Tables III(*b*) and III(*c*) show that inducing within-subject correlation decreases the treatment estimates as compared to the corresponding model estimates when the events are independent (Table III(*a*)). As the within-subject correlation increases, the smaller the treatment estimates become. All models are similarly affected by within-subject correlation, however, since total time models overestimate treatment effect for independent events it may appear that these models

Table IV. The event-specific, weighted average* and common† treatment estimates when treatment effect is effective for the first event only with: (*a*) $\sigma^2 = 0$; (*b*) $\sigma^2 = 0{\cdot}1$ and (*c*) $\sigma^2 = 0{\cdot}4$. NSE and RSE are the naive and robust standard errors, respectively, with corresponding empirical coverage of 95 per cent confidence intervals

Model		β	Mean($\hat{\beta}$)	SD($\hat{\beta}$)	Bias	Mean estimated NSE($\hat{\beta}$)	Mean estimated RSE($\hat{\beta}$)	Coverage naive 95 percent CI	Coverage robust 95 percent CI
(*a*) $\sigma^2 = 0$									
PWP-GT	Event 1	−1·00	−1·012	0·106	−0·012	0·100	0·099	0·94	0·94
	Event 2	0·00	0·001	0·097	0·001	0·095	0·095	0·94	0·94
	Event 3	0·00	0·005	0·095	0·005	0·099	0·099	0·94	0·94
	Event 4	0·00	−0·002	0·099	−0·002	0·107	0·106	0·96	0·95
	Weighted average	−0·25	−0·254	0·468	−0·004		0·050		0·97
	Common	−0·25	−0·257	0·047	−0·007	0·051	0·049	0·97	0·97
PWP-CP	Event 1	−1·00	−1·012	0·106	−0·012	0·100	0·099	0·94	0·94
	Event 2	0·00	0·006	0·104	0·006	0·099	0·098	0·94	0·93
	Event 3	0·00	0·005	0·096	0·005	0·101	0·101	0·96	0·96
	Event 4	0·00	−0·002	0·104	−0·002	0·108	0·107	0·94	0·94
	Weighted average	−0·25	−0·259	0·049	−0·009		0·050		0·95
	Common	−0·25	−0·265	0·050	−0·015	0·052	0·052	0·94	0·95
TT-R	Event 1	−1·00	−1·012	0·106	−0·012	0·100	0·099	0·94	0·94
	Event 2	−0·62	−0·620	0·107	−0·000	0·097	0·099	0·95	0·95
	Event 3	−0·45	−0·445	0·096	0·008	0·099	0·100	0·95	0·95
	Event 4	−0·36	−0·329	0·090	0·028	−0·106	0·106	0·97	0·97
	Weighted average	−0·61	−0·625	0·068	−0·017		0·078		0·98
	Common	−0·61	−0·613	0·078	−0·006	0·051	0·078	0·80	0·95
WLW	Event 1	−1·00	−1·012	0·106	−0·012	0·100	0·099	0·94	0·94
	Event 2	−0·62	−0·843	0·097	−0·222	0·098	0·098	0·39	0·40
	Event 3	−0·45	−0·742	0·102	−0·289	0·100	0·100	0·20	0·20
	Event 4	−0·36	−0·680	0·098	−0·322	0·107	0·107	0·11	0·11
	Weighted average	−0·61	−0·862	0·087	−0·255		0·090		0·21
	Common	−0·61	−0·825	0·090	−0·217	0·051	0·090	0·10	0·35
GT-UR	Common	−0·25	−0·355	0·048	−0·105	0·050	0·050	0·47	0·46
AG	Common	−0·25	−0·421	0·560	−0·171	0·049	0·058	0·07	0·16
LWA	Common	−0·61	−0·483	0·061	0·125	0·049	0·064	0·29	0·49
(*b*) $\sigma^2 = 0{\cdot}1$									
PWP-GT	Event 1	−1·00	0·920	0·087	0·080	0·098	0·098	0·93	0·93
	Event 2	0·00	0·015	0·099	0·015	0·096	0·095	0·94	0·94
	Event 3	0·00	0·035	0·092	0·035	0·100	0·100	0·97	0·97
	Event 4	0·00	0·029	0·115	0·029	0·108	0·108	0·90	0·90
	Weighted average	−0·25	−0·223	0·056	0·027		0·052		0·87
	Common	−0·25	−0·225	0·051	0·025	0·511	0·519	0·94	0·94
PWP-CP	Event 1	−1·00	−0·920	0·087	0·080	0·098	0·098	0·93	0·93
	Event 2	0·00	0·057	0·106	0·057	0·098	0·099	0·88	0·88
	Event 3	0·00	0·071	0·095	0·071	0·102	0·102	0·93	0·92
	Event 4	0·00	0·053	0·113	0·053	0·109	0·109	0·88	0·90
	Weighted average	−0·25	−0·206	0·058	0·044		0·051		0·79
	Common	−0·25	−0·206	0·052	0·044	0·052	0·052	0·87	0·87

Table IV. Continued

Model		β	Mean($\hat{\beta}$)	SD($\hat{\beta}$)	Bias	Mean estimated NSE($\hat{\beta}$)	Mean estimated RSE($\hat{\beta}$)	Coverage naive 95 per cent CI	Coverage robust 95 per cent CI
TT-R	Event 1	−1·00	−0·920	0·087	0·080	0·098	0·098	0·93	0·93
	Event 2	−0·62	−0·513	0·090	0·107	0·097	0·098	0·78	0·78
	Event 3	−0·45	−0·353	0·103	0·100	0·100	0·100	0·85	0·84
	Event 4	−0·36	−0·243	0·115	0·115	0·108	0·107	0·77	0·77
	Weighted average	−0·61	−0·539	0·070	0·068		0·077		0·84
	Common	−0·61	−0·524	0·075	0·084	0·051	0·077	0·57	0·81
WLW	Event 1	−1·00	−0·920	0·087	0·080	0·098	0·098	0·93	0·93
	Event 2	−0·62	−0·747	0·084	−0·127	0·098	0·098	0·78	0·78
	Event 3	−0·45	−0·644	0·091	0·191	0·101	0·101	0·54	0·54
	Event 4	−0·36	−0·574	0·101	−0·217	0·108	0·108	0·49	0·49
	Weighted average	−0·61	−0·774	0·078	−0·167		0·090		0·60
	Common	−0·61	−0·730	0·079	−0·122	0·051	0·091	0·43	0·74
GT-UR	Common	−0·25	−0·317	0·050	−0·067	0·050	0·053	0·71	0·73
AG	Common	−0·25	−0·384	0·054	−0·134	0·049	0·061	0·31	0·44
LWA	Common	−0·61	−0·425	0·064	0·183	0·049	0·065	0·11	0·23
(*c*) $\sigma^2 = 0{\cdot}4$									
PWP-GT	Event 1	−1·00	−0·809	0·105	0·191	0·097	0·098	0·46	0·48
	Event 2	0·00	0·079	0·103	0·079	0·098	0·098	0·89	0·89
	Event 3	0·00	0·089	0·111	0·089	0·104	0·104	0·85	0·84
	Event 4	0·00	0·081	0·113	0·081	0·111	0·111	0·90	0·91
	Weighted average	−0·25	−0·171	0·060	0·079		0·057		0·70
	Common	−0·25	−0·168	0·062	0·082	0·052	0·057	0·59	0·66
PWP-CP	Event 1	−1·00	−0·809	0·105	0·191	0·097	0·098	0·46	0·48
	Event 2	0·00	0·198	0·113	0·198	0·100	0·104	0·45	0·50
	Event 3	0·00	0·174	0·116	0·174	0·105	0·107	0·61	0·63
	Event 4	0·00	0·146	0·112	0·146	0·122	0·113	0·75	0·75
	Weighted average	−0·25	−0·117	0·060	0·133		0·052		0·27
	Common	−0·25	−0·105	0·059	0·145	0·053	0·053	0·25	0·26
TT-R	Event 1	−1·00	−0·809	0·105	0·191	0·097	0·098	0·46	0·48
	Event 2	−0·62	−0·343	0·113	0·277	0·098	0·099	0·23	0·23
	Event 3	−0·45	−0·227	0·112	0·226	0·104	0·103	0·42	0·42
	Event 4	−0·36	−0·154	0·117	0·204	0·111	0·110	0·54	0·54
	Weighted average	−0·61	−0·435	0·079	0·172		0·076		0·38
	Common	−0·61	−0·406	0·085	0·202	0·052	0·076	0·13	0·24
WLW	Event 1	−1·00	−0·809	0·105	0·191	0·097	0·098	0·46	0·48
	Event 2	−0·62	0·606	0·111	0·014	0·099	0·099	0·89	0·89
	Event 3	−0·45	−0·504	0·114	−0·051	0·104	0·104	0·91	0·91
	Event 4	−0·36	−0·442	0·125	−0·084	0·111	0·111	0·85	0·85
	Weighted average	−0·61	−0·670	0·098	−0·062		0·093		0·87
	Common	−0·61	−0·604	0·104	0·004	0·051	0·094	0·67	0·93
GT-UR	Common	−0·25	−0·257	0·063	−0·007	0·051	0·059	0·88	0·96
AG	Common	−0·25	−0·335	0·074	−0·085	0·051	0·067	0·56	0·75
LWA	Common	−0·61	−0·353	0·075	0·254	0·051	0·668	0·03	0·05

* Weighted average of the event-specific estimates such that the robust standard error is the smallest, as defined by WLW
† Model fitted with one parameter

perform better than the gap time and counting process models. The robust standard errors are similar to the naive for the event-specific estimates – if anything the robust standard errors are slightly smaller for later events. The robust standard errors for the common and weighted average estimates in the total time models are inflated, but this is same behaviour as displayed when events are independent. There is a small increase in the robust standard error for the common and weighted average estimates for PWP-GT and the unrestricted models, AG, GT-UR and LWA, but not for PWP-CP. However, the inflated standard error are insufficient to compensate for the bias in the estimated treatment effect.

4.2. Treatment effective for first event only

4.2.1. Independent events

Consider when events are independent (Table IV(*a*)). PWP-GT and PWP-CP show that treatment is effective for the first event only, again with minimal bias and empirical coverage close to the nominal level. The TT-R model displays the expected 'carry-over' effect, with treatment effect slowly diminishing with each consecutive event, showing minimal bias with empirical coverage close to the nominal level. The WLW again overestimates the treatment effect after the first event and hence empirical coverage is lower than the nominal level.

When treatment effect is not constant the unrestricted models make little sense and do not reflect the expected 'average' treatment effect as estimated by the weighted average and common estimates of the restricted models. Even for the event-specific models using an overall estimate of treatment effect when treatment effect is not constant between events probably has little value and should be interpreted cautiously. If it is used it should be stated clearly that treatment effect is not constant.

4.2.2. Within-subject correlation

Inducing within-subject correlation (Table IV(*b*) and (*c*)) has the same effect on estimates as before when treatment is constant; it underestimates the treatment effect compared to the corresponding model when events are independent.

4.3. Comments on the results

Recurrent event data have two closely linked features: the subject can only be at risk for one event at a time; and the events are ordered, for example, the first event occurs before the second event, where the hazards may change after each event. The choice of the model characteristics should be determined by this data structure and research question of interest. How the model characteristics, risk interval and risk set are defined determines the results and their interpretation; these are discussed below.

4.3.1. Risk Interval Definition

Total times have a 'carry-over' effect. Total times within a subject tend to be highly correlated even when gap times are not.[19] For instance, the total time of the second risk interval includes the first interval; the third risk interval contains the first and the second intervals, and so on. An analysis based on total times which estimates a large treatment effect for the first event will carry over this effect for subsequent events as seen with the simulated data where treatment is effective

for the first event only, even when an analysis of gap times suggests that the treatment is no longer effective. When treatment effect is constant, it is expected that the event-specific estimates in a total time model are constant, however, the simulated results show an increase in treatment after each event. This observation could be the same carry-over effect occurring.

The results show that counting process estimates are extremely close to gap time estimates. The subjects for these models are at risk for the same length of time but on a different time scale. The counting process estimates are not close to the total time estimates, which share the same time scale but differ in the length of time at risk. Hence, the length of time at risk may influence the results much more than the time scale. However, using gap time has a different hazard ratio interpretation compared with total time and counting process estimates because of the different time scale – for gap time the hazards are since the last event, and for total time or counting process the hazards are since treatment.

4.3.2. Risk Set Definition

Semi-restricted risk sets have a carry-over effect. This is clear from comparing the results of TT-R and WLW. For example, if treatment is effective then treated subjects will have fewer events in the same period. A semi-restricted risk set includes all subjects in each event-stratum, and so with each consecutive event the number of treated subjects with a censored observation increases. These censored observations are compared to those untreated subjects who are experiencing an event, and so exaggerate the treatment effect in the later strata.

A restricted risk set is essential for recurrent event data if the hazards are expected to change after each event. For our results, it was important to stratify when the treatment effect was not constant, otherwise the model estimate did not reflect the 'average' treatment effect. Event-specific models also have the advantage of observing how effects change with each subsequent event.

An unrestricted risk set is suitable when the baseline hazard does not change with each event. However, this is only appropriate for the GT-UR and AG models and not for LWA. In the results, the LWA was the only model to underestimate the treatment effect because it was the only model that allowed a subject to be at risk several times for the same event.

4.3.3. Robust Variance

A deflated robust standard error means there is more variation within subjects than between, while an inflated robust standard error means there is less variation within subjects than between. From the simulated results the robust standard errors of the weighted average and common estimates in the total time models are inflated regardless whether events are independent within subjects or not. For all the other models, except PWP-CP, the robust standard errors of the weighted average and common estimates are the same as the naive when within-subject events are independent, and become inflated when events are not independent. The naive and robust standard errors for the weighted average and common estimates for PWP-CP appear to be the same regardless of the within-subject correlation.

5. DATA: MORVITA TRIAL

The permissible models are applied to data from a trial that investigated the effect of childhood MORbidity from supplementation of VITamin A – the MORVITA trial.[20] This was a randomized double-blinded placebo-controlled trial with 1405 subjects aged 6–47 months (11 deleted due

Table V. The event-specific, weighted average* and common† treatment estimates for the MORVITA trial

Model	Estimates	$\hat{\beta}$	Estimated naive SE ($\hat{\beta}$)	Estimated robust SE ($\hat{\beta}$)
PWP-GT	Event 1	0·152	0·0577	0·0573
	Event 2	−0·011	0·0649	0·0643
	Event 3	0·001	0·0786	0·0779
	Event 4	0·105	0·1050	0·1030
	Event 5	0·171	0·1577	0·1573
	Weighted average	0·071		0·0349
	Common	0·072	0·0347	0·0349
PWP-CP	Event 1	0·152	0·0577	0·0573
	Event 2	0·007	0·0650	0·0598
	Event 3	0·030	0·0787	0·0716
	Event 4	0·139	0·1053	0·0883
	Event 5	0·212	0·1587	0·1396
	Weighted average	0·086		0·0347
	Common	0·088	0·0347	0·0347
TT-R	Event 1	0·152	0·0577	0·0573
	Event 2	0·063	0·0650	0·0645
	Event 3	0·129	0·0786	0·0782
	Event 4	0·252	0·1046	0·1040
	Event 5	0·327	0·1572	0·1573
	Weighted average	0·144		0·0442
	Common	0·142	0·0347	0·0444
WLW	Event 1	0·152	0·0577	0·0573
	Event 2	0·098	0·0649	0·0647
	Event 3	0·117	0·0785	0·0783
	Event 4	0·247	0·1044	0·1041
	Event 5	0·500	0·1579	0·1571
	Weighted average	0·140		0·0557
	Common	0·158	0·0346	0·0602
GT-UR	Common	0·089	0·0346	0·0378
AG	Common	0·087	0·0346	0·0371
LWA	Common	0·090	0·0346	0·0279

* Weighted average of the event-specific estimates such that the robust standard error is the smallest as, defined by WLW

† Model fitted with on parameter

to missing values). Once a child was randomized (to vitamin A or placebo) they received the same treatment throughout the study. The treatment was a single dose given every four months. The event outcome was acute respiratory illness (ARI).

For this paper the analyses includes data from the time since the first treatment dose until the end of four months, but before the second dose. The only covariate considered is treatment, which is 0 if the subject received placebo or 1 if the subject received vitamin A. Only 170 (12·2 per cent)

subjects had more than five events and the event-specific estimates were unreliable for greater than five events. To allow direct comparisons between all models the data were truncated after five events.

The risk set size between definitions varies greatly, especially in the later strata for the event-specific risk sets, semi-restricted and restricted. The size of the unrestricted risk set is 1394 and the size of the semi-restricted risk set for each kth event is always 1394. The restricted risk set size is 1394, 1203, 948, 648 and 369, for the first to the fifth event, respectively. The restricted and semi-restricted risk sets are only the same for the first event.

The results of the fitted models are shown in Table V. The pattern of the results for the models are remarkably similar to the simulated examples presented earlier. The results show that treatment has a significant adverse effect on the first event ($\hat{\beta}_1 = 0{\cdot}152$, 95 per cent CI is 0·039 to 0·264). That is, the hazard of experiencing the first ARI event since dosage is 1·16 times higher for those who received treatment compared to those who receive the placebo. PWP-GT and PWP-CP models show that treatment effect is not statistically significant for any other events, although the point estimates are positive again for the fourth and fifth events. TT-R and WLW display the characteristic 'carry-over' treatment effect as those observed with the simulated results. TT-R and WLW conclude that there is no treatment effect for the second and third event, but there is an adverse effect for the fourth and fifth event since treatment. The unrestricted models are of no use since treatment effect is not constant. The adverse treatment effect may have been underestimated, as suggested by the simulation results when there exists within-subject correlation.

6. DISCUSSION

6.1. Related Work

The results from previous work that have made comparison between any of the above models follow the same patterns that we observe with our results, especially the 'carry-over' effect for the WLW.[5,8,17,21] There has been some debate in the literature regarding whether WLW is suitable for recurrent event data. It is commonly used and has been explicitly recommended for analysing recurrent event data.[5,8,9,21] However, those against state it is not sensible due to the semi-restricted risk set – a subject should not be at risk of the 4th event if the subject has had only one event.[22] We also do not recommend using WLW since using an unrestricted risk set has a 'carry-over' effect that overestimates the true treatment effect.

6.2. Software

The models discussed can be implemented in various software packages. For example, SAS version 6·10 or later, and STATA version 5.0. The *SAS* 6.10 *STAT Enhancements and Changes* manual[23] clearly explains how to use PROC PHREG to fit the AG, WLW and PWP models. The variance 'sandwich' estimator is a little clumsy, requiring PROC IML. In STATA the survival commands include STSET, which tells STATA what are the survival times, censoring and identification variables, and uses STCOX to fit Cox proportional hazard models. The STATA 5.0 manual[24] is not explicit on how to fit the various models, but it does have a simple robust variance command. Time-varying covariates may be difficult to handle in STATA 5.0.

6.3. Robust variance models versus frailty models

It is important to adjust for any within-subject correlation. The simulation results show that using a robust variance was not adequate to account for the within-subject correlation. This is a surprising and interesting result, but it is unclear why this is so. One possibility is due to the choice of the underlying distribution for the subject random effect, which was Gaussian. The robust variance may be adequate for other distributions, such as the positive stable or gamma. Clearly, further work is required to determine why the robust variance failed to adjust for the within-subject correlation.

When there exists within-subject correlation, perhaps other methods should be used instead of heavily relying on the robust variance estimate to adjust for the misspecification of the model. For example, correlation can be modelled using a random effect (frailty model), such as a gamma distribution.[25] However, it is more complicated than applying a robust variance. Two statistical software packages that will handle this approach for survival analysis are BUGS[26] and MLn[27].

7. CONCLUSION

We recommend using the PWP-GT model and the TT-R model to analyse recurrent event data when within-subject events are independent. Each model provides the answer to slightly different research questions. The PWP-GT model would be used to determine whether the treatment is effective for the *k*th event since the time from the previous event. For example, it is only effective for delaying the first infection but afterwards there is no difference? The TT-R model would be used to assess whether treatment is effective for the *k*th event since the start of treatment. However, be cautious when using TT-R since it may be biased due to the 'carry-over' effect from using total time. The AG and the GT-UR models are adequate if a common baseline hazard can be assumed, but these lack the detail and versatility of the event-specific models.

Applying a robust variance may not be adequate when there is within-subject correlation. Further work is required to investigate when and why the robust variance estimate fails to adjust for within-subject correlation. Perhaps other methods, such as frailty models, should be used instead.

The WLW and LWA are not appropriate for recurrent event data. WLW has a semi-restricted risk set that allows subjects to be at risk of the 4th event even if the subjects have only had one event, which leads to overestimation of the treatment effect. LWA allows subjects to be at risk several times for the same event. However, these models are suitable in other situations. The WLW is most appropriate to data where there are different types of events from the same person, where the baseline hazard is potentially different for each type, such as multi-type event data, for example, tumours at several different sites in the body. The LWA model is suitable for clustered data such as siblings or pairs of eyes, where it can be assumed that the baseline hazard is the same, and the beginning of risk of the event is the same within the cluster.

ACKNOWLEDGEMENTS

We wish to thank the MORVITA Trial research team, especially Dr. M. J. Dibley, Dr. T. Sadjimin and Dr. C. L. Kjolhede, for the use of data. We would also like to thank the reviewers for their constructive and helpful comments.

REFERENCES

1. Andersen, P. K. and Gill, R. D. 'Cox's regression model for counting processes: a large sample study', *Annals of Statistics*, **10**, 1100–1120 (1982).
2. Prentice, R. L., Williams, B. J. and Peterson, A. V. 'On the regression analysis of multivariate failure time data', *Biometrika*, **68**, 373–379 (1981).
3. Lee, E. W., Wei, L. J. and Amato, D. A. 'Cox-type regression analysis for large numbers of small groups of correlated failure time observations', *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kluwer Academic Publisher, Dordrecht, 1992, pp. 237–247.
4. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065–1073 (1989).
5. Therneau, T. M. and Hamilton, S. A. 'RhDNase as an example of recurrent event analysis', *Statistics in Medicine*, **12**, 2029–2047 (1997).
6. Gao, S. and Zhou, X-H. 'An empirical comparison of two semi-parametric approaches for the estimation of covariate effects from multivariate failure time data', *Statistics in Medicine*, **16**, 2049–2062 (1997).
7. Clayton, D. 'Some approaches to the analysis of recurrent event data', *Statistical Methods in Medical Research*, **3**, 244–262 (1994).
8. Lin, D. Y. 'Cox regression analysis of multivariate failure time data: the marginal approach', *Statistics in Medicine*, **13**, 2233–2247 (1994).
9. Wei, L. J. and Glidden, D. V. 'An overview of statistical methods for multiple failure time data in clinical trials', *Statistics in Medicine*, **16**, 833–839 (1997).
10. Andersen, P. K., Borgan, O., Gill, R. D. and Keilding, N. *Statistical Models Based on Counting Process*, Springer-Verlag, New York, 1993.
11. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1982).
12. Royall, R. M. and Cumberland, W. G. 'Variance estimation in finite population sampling', *Journal of the American Statistical Association*, **73**, 351–358 (1978).
13. Lin, D. Y. and Wei, L. J. 'The robust inference for the Cox proportional hazard model', *Journal of the American Statistical Association*, **84**, 1074–1078 (1989).
14. Klein, J. P. 'Semiparametric Estimation of random effects using the Cox model based on the EM algorithm', *Biometrics*, **48**, 795–806 (1992).
15. Pickles, A. and Crouchley, R. 'A comparison of frailty models for multivariate survival data', *Statistics in Medicine*, **14**, 1447–1461 (1995).
16. Finkelstein, D. M., Schoenfeld, D. A. and Stamenovic, E. 'Analysis of multiple failure time data from an AIDS clinical trial', *Statistics in Medicine*, **16**, 951–961 (1997).
17. Cox, D. R. 'Regression analysis and life table (with discussion)', *Journal of the Royal Statistical Society*, *Series B*, **34**, 187–222 (1972).
18. Kish, L. *Survey Sampling*, Wiley, New York, 1965.
19. Lipschutz, K. H. and Snapinn, S. M. 'Discussion of paper by Wei and Glidden', *Statistics in Medicine*, **16**, 846–848 (1997).
20. Dibley, M. J., Sadjimin, T., Kjolhede, C. L. and Moulton, L. H. 'Vitamin A supplementation fails to reduce incidence of acute respiratory illness and diarrhea in preschool-age Indonesian children', *Journal of Nutrition*, **126**, 434–442 (1996).
21. Barai, U. and Teoh, N. 'Multiple statistics for multiple events, with application to repeated infections in the growth factor studies', *Statistics in Medicine*, **16**, 941–949 (1997).
22. Cook, R. J. and Lawless, J. F. 'Discussion of paper by Wei and Glidden', *Statistics in Medicine*, **16**, 841–843 (1997).
23. SAS Institute Inc. *SAT/STAT Software*: *Changes and Enhancements*, *Release* 6.10, SAS Institute Inc., Cary, NC, 1994, pp. 81–113.
24. StataCorp. *Stata Statistical Software*: *Release* 5.0, Stata Corporation, College Station, TX 1997.
25. Oakes, D. A. 'Frailty models for multiple event times', *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kluwer Academic Publisher, Dordrecht, 1992, pp. 371–379.
26. Spiegelhalter, D. J., Thomas, A., Best, N. and Gilks, W. R. '*BUGS*: *Bayesian Inference using Gibbs Sampling*, *Version* 0.50', MRC Biostatistics Unit, Cambridge, UK, 1995.
27. Goldstein, H. *Multilevel Statistical Models*, Edward Arnold, London, Wiley, New York, 1995.

