The mean, variance, and correlation for bivariate recurrent event data with a terminal event

Thomas H. Scheike
Frank Eriksson
Siri Tribler

Research Report 18/02

Department of Biostatistics
University of Copenhagen
The mean, variance, and correlation for bivariate recurrent event data with a terminal event

Thomas H. Scheike & Frank Eriksson
Department of Biostatistics, University of Copenhagen
Øster Farimagsgade 5, DK-1014 Copenhagen K, Denmark
email: ts@biostat.ku.dk

Siri Tribler
Department of Medical Gastroenterology, University Hospital of Copenhagen
Blegdamsvej 9, DK-2100, Denmark

Recurrent events in the presence of a terminal event is often encountered in a biomedical setting. The marginal mean of the number of recurrent events in a specified time-period is a useful nonparametric summary of recurrent events data also in the presence of a terminal event. Other useful nonparametric summaries, that are simple to compute, are the distribution function of the number of recurrent events for each point in time and the variance of the number of recurrent events. For bivariate recurrent events, still in the presence of a terminal event, we suggest a simple nonparametric estimator of the covariance or correlation of the marginal number of events for both processes. When there is no terminal event the correlation is useful, but when there is an important terminal event we suggest an adjustment for correlation induced by the terminal event in order to get a measure that reflects the dependence in the recurrent events processes among survivors only. Our estimators can be used for deciding if the two recurrent events are correlated and in what way. We provide large sample properties of our estimators and show their performance in small samples by simulations. The estimators are applied in study of catheter complications among patients receiving home parenteral nutrition.
through a central venous catheter, and show a positive correlation between the number of infections and the number of occlusion defects.

Some key words: Competing risks; Counting process; Correlation; Cumulative incidence; Marginal mean; Recurrent events; Terminal event;

1 Introduction

Patients with chronic intestinal failure receiving home parenteral support through a central venous catheter can experience several complications during the often long-term treatment periods, see Brandt et al. (2016, 2017); Tribler et al. (2017a,b). This work was motivated by a cohort of 715 consecutive patients at the University Hospital of Copenhagen, where the relationship between different types of catheter complications during the treatment period was studied. We focus on the number of catheter-related bloodstream infections and one specific type of mechanical catheter defect. Some of the patients die, due to severe intestinal failure or co-morbidities, and other patients leave the HPS program alive for different reasons during the around 14 years of follow up.

A useful first analysis of recurrent event data is to compute the marginal mean number of events of a specific type, as described in Cook and Lawless (1997), developed further in Ghosh and Lin (2000) and extended to regression models in Ghosh and Lin (2003); Cai and Schaubel (2004). Chen and Cook (2004) considered an extension to deal with multivariate recurrent event processes (in the presence of a terminal event) with focus on making statements about the marginal means for several recurrent event processes jointly.

The possible dependence between multiple recurrent events have been modelled through structured conditional intensity models that describe how the risk of new events depend on the history of the recurrent event processes, see Andersen and Gill (1982); Cook and Lawless (1997). Another approach, utilized in a large number of papers, is to introduce random effects in the intensity functions to induce dependence between the multiple re-
current event processes and the terminal event, see for example Abu-Libdeh et al. (1990); Cook et al. (2010); Rondeau et al. (2007); Ye et al. (2007); Chen et al. (2015); Zeng et al. (2014). In addition, recent work has considered cross hazard ratio models, see Ning et al. (2015). All these approaches are useful and provide insight into relationships formulated via the intensity functions, but they all rely on structural assumptions on the underlying intensities of the event processes.

In this article, we develop a simple nonparametric estimator of the correlation between the number of observed events for two recurrent event processes. Similar to the marginal mean models, our estimator recognizes that the events can only occur for subjects that have not experienced the terminal event yet.

In the presence of a terminal event, interpretation of the simple correlation between the number of observed events is complicated by the terminating event generating dependence even for recurrent events that are otherwise not correlated among survivors. We demonstrate this point by simulations. With this complicated interpretation in mind, we suggest a method for deciding if there is correlation among the survivors, which is typically the question of interest.

We present easily computed summary measures that aim at characterizing the variation in the number recurrent events of a specific type. We provide estimators of the distribution function for each recurrent event, in turn allowing us to estimate the standard deviation of the number of events as a function of time.

An advantage of the nonparametric approach is that it does not rely on structural assumptions such as those required by the random effects and structured conditional regression models. Another advantage of our approach is that, in contrast to the random effects model, it allows us to investigate if the dependence between the recurrent event processes is directional in the sense that events of type 1 are predictive for type 2 events, but not the other way around. The estimators we consider are related to estimators in certain multistate models, see for example Cook and Lawless (2007); Scheike et al. (2014).

The paper is structured as follows. Section 2 presents the model, develops estimators and outlines the asymptotic properties. In Section 3 we consider a different class of esti-
mators based on product-limit techniques. Section 4 formulates some relevant hypothesis testing. Section 5 contains a simulation study. In Section 6 we apply the estimators to a single center population of intestinal failure patients. Finally, Section 7 contains a discussion and some concluding remarks.

2 Model formulation

Let $D$ denote the survival time (the terminal event), and let $N_1^*(t)$ and $N_2^*(t)$ count the number of the two types of recurrent events observed over a time-period $[0,t]$, where $t \leq \tau$. Due to the terminal event, we only observe the recurrent event processes up to $\tau \land D$, where $a \land b = \min(a,b)$. Note that $N_1^*(t) = N_1^*(t \land D)$ and $N_2^*(t) = N_2^*(t \land D)$ because subjects will only have events when still alive. The observation of the two processes may also be censored, thus only making it possible to observe the processes up to the censoring time $C$. Let $\delta = I(D \leq C)$, $T = D \land C$, and let $N_1(t) = N_1^*(t \land T)$, $N_2(t) = N_2^*(t \land T)$ be the observed number of events of the two types and define the at-risk process $Y(t) = I(T \geq t)$. Denote the counting process of the terminal event by $N^D(t)$ and denote its marginal cumulative hazard by $\Lambda^D(t)$. We make the standard assumption that the censoring is independent of $D$ and $(N_1^*(t), N_2^*(t))$. To make the presentation and the technical arguments easier, we assume that both recurrent event counting processes are bounded by a constant $K$. The observations $(N_{1i}(t), N_{2i}(t), T_i, \delta_i) : t \in [0, \tau])$ are assumed to be independent replicates for $i = 1, \ldots, n$.

2.1 Marginal properties

The mean number of events of type 1 is

$$\mu_1(t) = E(N_1^*(t)) = \int_0^t S(s) dR_1(s),$$

(1)
where $S(t) = P(D > t)$ and $dR_1(t) = E(dN_1^*(t) | D > t)$, and can be estimated using the Kaplan-Meier estimator $\hat{S}(t)$ and the Nelson-Aalen estimator

$$\hat{R}_1(t) = \int_0^t \frac{1}{Y_\bullet(s)} dN_{1\bullet}(s)$$

where $Y_\bullet(t) = \sum_{i=1}^n Y_i(t)$ and $N_{1\bullet}(t) = \sum_{i=1}^n N_{i1}(t)$ such that the estimator is

$$\hat{\mu}_1(t) = \int_0^t \hat{S}(s)d\hat{R}_1(s).$$

The normalized estimator converges weakly to a mean-zero Gaussian process with estimable variance, see Ghosh and Lin (2000) for details. The estimator resembles the cumulative incidence estimator and reflects that the number of observed events depends on both survival and the event rate among the survivors.

The mean tells one thing about the recurrent events processes. Another important aspect of the analysis is to describe the heterogeneity. We consider how to estimate the standard deviation of the observed number of events, and more detailed summaries such as the distribution function for the number of events of a specific type at each point in time. In previous work it has been demonstrated that multistate models can be used to describe the probability of observing a specific number of events, see Cook and Lawless (2007) and references therein. We provide estimators that avoid elaborate multistate models, are simple to compute and computationally efficient.

Thanks to the simple counting process structure we observe that $N_1^*(t)^2$ can be written as

$$\sum_{k=0}^{K} \int_0^t I(D > s)I(N_1^*(s-) = k)f(k)dN_1^*(s)$$

with $f(k) = (k + 1)^2 - k^2$, such that its mean can be written as

$$\sum_{k=0}^{K} \int_0^t S(s)f(k)P(N_1^*(s-) = k|D \geq s)E(dN_1^*(s)|N_1^*(s-) = k, D > s)$$
and estimated by

\[ \hat{\mu}_{1,2}(t) = \sum_{k=0}^{K} \int_0^t \tilde{S}(s) f(k) \frac{Y_{1^*}^k(s)}{Y_{1^*}(s)} \frac{1}{Y_{1^*}^k(s)} dN_{1^*}^k(s) = \sum_{i=1}^{n} \int_0^t \tilde{S}(s) f(N_{11}(s-)) \frac{1}{Y_{1^*}(s)} dN_{11}(s), \]

where \( Y_{\bullet}^k(t) = \sum_{i=1}^{n} Y_i(t) I(N_{ij}(t-) = k) \) denotes the number of subjects under risk for the \((k + 1)^{th}\) jump, and \( N_{\bullet}^k(t) = \sum_{i=1}^{n} \int_0^t I(N_{ij}(s-) = k) dN_{ij}(s) \) is the counting process related to the \((k + 1)^{th}\) jump.

The asymptotic expansion of this estimator, for each \( k \), resembles that of \( \hat{\mu}_1(t) \). Using this estimator we can estimate the variance of the observed recurrent events of type 1 by

\[ \hat{\mu}_{1,2}(t) - \hat{\mu}_1(t)^2. \]

We finally note that, for each \( k \), \( P(N_{1^*}(t) \geq k) \) can be estimated by similar simple counting process integrals. To see this, rewrite \( I(N_{1^*}(t) \geq k) \) as

\[ \int_0^t I(D > s) I(N_{1^*}(s-) = k - 1) dN_{1^*}(s), \]

suggesting that its mean can be computed as

\[ \int_0^t S(s) P(N_{1^*}(s-) = k - 1 | D \geq s) E(dN_{1^*}(s)|N_{1^*}(s-) = k - 1, D > s) \]

and estimated by

\[ \tilde{F}_k(t) = \int_0^t \tilde{S}(s) \frac{Y_{1^*}^{k-1}(s)}{Y_{1^*}(s)} \frac{1}{Y_{1^*}^{k-1}(s)} dN_{1^*}^{k-1}(s). \]

These counting process integrals can be written as sums of independent zero mean processes similarly to the marginal mean itself. Note also that \( \hat{\mu}_1(t) = \sum_{k=1}^{K} \tilde{F}_k(t) \).

### 2.2 Covariance estimation

We consider simple nonparametric measures of the correlation or covariance between the number of events of two types. In the situation without a terminal event, \( D = \infty \), a useful first description of the dependence is to compute the correlation between the number of
observed events of the two types over time,

$$\rho(t) = \frac{E(N_1^*(t)N_2^*(t)) - \mu_1(t)\mu_2(t)}{\sqrt{\text{var}(N_1^*(t))} \sqrt{\text{var}(N_2^*(t))}}. \tag{2}$$

In the case with a nonnegligible terminal event we can still estimate this correlation, and it will still describe the correlation between $N_1^*(t)$ and $N_2^*(t)$. However, the interpretation of the correlation will be different, because the terminal event can generate dependence between the observed number of events on top of the possible dependence from the recurrent event processes among survivors.

Consider the covariance between the different types of recurrent events $E(N_1^*(t)N_2^*(t)) - \mu_1(t)\mu_2(t)$. We start by noting that when the two processes do not jump at the same time, then (by integration by parts)

$$E(N_1^*(t)N_2^*(t)) = E(\int_0^t N_1^*(s-)dN_2^*(s)) + E(\int_0^t N_2^*(s-)dN_1^*(s)).$$

Looking at the first term we compute

$$E(\int_0^t N_1^*(s-)dN_2^*(s)) = \sum_{k=1}^{K} \sum_{j=1}^{J} E(\int_0^t kI(N_1^*(s-) = k)I(D > s)dN_2^*(s))$$

$$= \sum_{k=1}^{K} \int_0^t S(s)kP(N_1^*(s-) = k|D \geq s)E(dN_2^*(s)|N_1^*(s-) = k, D > s)$$

and note that this can be estimated by

$$\sum_{k=1}^{K} \int_0^t \hat{S}(s)k\frac{Y_{i,s}^k(s)}{Y_{s,s}^k} \frac{1}{Y_{i,s}^k(s)} d\hat{N}_2^k(s), \tag{3}$$

where $\hat{N}_j^k(t) = \sum_{i=1}^{n} \int_0^t I(N_{ij^o}(s-) = k)dN_{ij}(s)$ is the number of events of type $j$ where the other process $(1^o = 2$ and $2^o = 1)$ is observed with $k$ events. Denote the estimator (3) by $\hat{E}(\int_0^t N_1(s-)dN_2(s))$. 
Thus, the estimator of $E(N_1^*(t)N_2^*(t))$ becomes

$$
\hat{E}(N_1^*(t)N_2^*(t)) = \sum_{k=1}^{K} \int_0^t \hat{S}(s)k \frac{Y_k^1(s)}{Y_k^*(s)} \frac{1}{Y_k^2(s)} d\hat{N}_2^*(s) + \sum_{k=1}^{K} \int_0^t \hat{S}(s)k \frac{Y_k^2(s)}{Y_k^*(s)} \frac{1}{Y_k^1(s)} d\hat{N}_1^*(s).
$$

The correlation compares $E(N_1^*(t)N_2^*(t))$ to $\mu_1(t)\mu_2(t)$, but as pointed out above, this difference reflects both the dependence in the recurrent event processes among the survivors and the terminal event. The terminal event alone will typically generate dependence between the two recurrent events processes even if they are fully independent among survivors. It turns out that it can be constructive to instead compare $E(N_1^*(t)N_2^*(t))$ to its value computed under the assumption that the two processes are not correlated among survivors.

We define not being associated or correlated among survivors via two conditions, namely that $N_1(t-)$ does not contain information about the risk of a jump of type 2, and vice versa, that $N_2(t-)$ does not contain information about the risk of a jump of type 1. Formally we say that $N_2(t-)$ is not predictive about the risk of a jump of type 1 when

$$
E(dN_2^*(t)|N_1^*(t-)|k, D > t) = E(dN_2^*(t)|D > t)
$$

and then the mean can be calculated as

$$
E(\int_0^t N_1^*(s-)dN_2^*(s)) = \int_0^t S(s)E(N_1^*(s-)|D \geq s)E(dN_2^*(s)|D > s)
$$

and estimated by

$$
\int_0^t \hat{S}(s)\left\{ \sum_{k=1}^{K} k \frac{Y_k^1(s)}{Y_k^*(s)} \right\} \frac{1}{Y_k^1(s)} dN_2^*(s),
$$

that we denote $\hat{E}_T(\int_0^t N_2^*(s-)dN_1^*(s))$. If both $N_2(t-)$ is not predictive about $dN_1(t)$ and $N_1(t-)$ is not predictive about $dN_2(t)$, according to the definitions above, we say that the two recurrent event processes are uncorrelated or unassociated among survivors.

Under the assumption that the processes are uncorrelated among survivors, we can
then estimate $E(N_1^*(t)N_2^*(t))$ by

$$
\hat{E}_I(N_1^*(t)N_2^*(t)) = \int_0^t \hat{S}(s) \left\{ \sum_{k=1}^K \frac{Y_k(s)}{Y_\bullet(s)} \right\} \frac{1}{Y_\bullet(s)} dN_2\bullet(s) + \int_0^t \hat{S}(s) \left\{ \sum_{k=1}^K \frac{Y_k^*(s)}{Y_\bullet(s)} \right\} \frac{1}{Y_\bullet(s)} dN_1\bullet(s).
$$

The assumption of no correlation can be investigated by comparing $\hat{E}(N_1^*(t)N_2^*(t))$ to $\hat{E}_I(N_1^*(t)N_2^*(t))$. We emphasize that the purpose for calculating $\hat{E}_I(N_1^*(t)N_2^*(t))$ is for comparison with $\hat{E}(N_1^*(t)N_2^*(t))$, not for inference on its own. More on this in Section 4.

Following similar arguments as used by Ghosh and Lin (2000), both estimators, $\hat{E}(N_1^*(t)N_2^*(t))$ and $\hat{E}_I(N_1^*(t)N_2^*(t))$, can be shown to be asymptotically equivalent to sums of i.i.d. processes. The details are given in the appendix. Using the i.i.d. decomposition we can estimate the standard errors of the estimators and their difference $\hat{E}(N_1^*(t)N_2^*(t)) - \hat{E}_I(N_1^*(t)N_2^*(t))$. Although possible, such a procedure involves a lot of work in setting up the estimators as we would need to set up the i.i.d. decomposition for each $k$, and then add all processes before squaring the residuals. Instead, we implemented a bootstrap that is easy to program and also works reasonably fast. We show later by simulations that this bootstrap works well.

In the worked example later we compare $\hat{E}(N_1^*(t)N_2^*(t))$ and $\hat{E}_I(N_1^*(t)N_2^*(t))$ to conclude that there is positive correlation among survivors for the two recurrent event processes. This is typically more interesting than just concluding that $\hat{E}(N_1^*(t)N_2^*(t))$ is larger than $\hat{E}(N_1^*(t))\hat{E}(N_2^*(t))$, i.e. that there is positive correlation in the observed data.

In Section 5 we demonstrate by simulation that differences between $E(N_1^*(t)N_2^*(t))$ and $E_I(N_1^*(t)N_2^*(t))$ indeed seem to capture what we would call dependence.

Similarly to the univariate situation, we can derive a counting process version of cumulative probability of exceeding $(k_1, k_2)$ events. To see this, note that $I(N_1^*(t) \geq k_1, N_2^*(t) \geq k_2)$ can be written as

\[
\int_0^t I(D > s)I(N_1^*(s-) = k_1 - 1, N_2^*(s-) \geq k_2) dN_1^*(s) + \int_0^t I(D > s)I(N_1^*(s-) \geq k_1, N_2^*(s-) = k_2 - 1) dN_2^*(s)
\]
such that its mean can be computed as

\[
\int_0^t S(s)P(N_1^*(s) = k_1 - 1, N_2^*(s) \geq k_2 | D \geq s) E(dN_1^*(s)|N_1^*(s) = k_1 - 1, N_2^*(s) \geq k_2, D > s)
\]
\[
+ \int_0^t S(s)P(N_1^*(s) \geq k_1, N_2^*(s) = k_2 - 1 | D \geq s) E(dN_1^*(s)|N_1^*(s) = k_1, N_2^*(s) = k_2 - 1, D > s)
\]

and estimated as before using the empirical quantities. A moment estimator of \( E(N_1^*(t)N_2^*(t)) \) may also be based on these estimators.

### 3 Product-limit estimators

In this section we present a different class of estimators that can be implemented with standard software and that have performance comparable to those we have constructed so far. We call these estimators “product-limit estimators” in contrast to the ones considered previously that we call “integral estimators”.

Note that the probability of exceeding \( k \) events as a function of time, \( P(N_1^*(t) \geq k) \), is the cumulative incidence in the competing risk model with the event time \( T_k = \inf \{ t : N_1^*(t) = k \} \) observed subject to the competing risk of a terminal event at time \( D \). Thus \( P(N_1^*(t) \geq k) = P(T_k < t, T_k < D) = F_k(t) \), that can be estimated by the standard product-limit estimator for this competing risks model. We denote this estimator by \( \hat{F}_k(t) \) and give the expression below to make it clear how the different estimators are related.

Note that the marginal mean is \( \mu_1(t) = \sum_{k=1}^K F_k(t) = \sum_{k=0}^K k(F_k(t) - F_{k+1}(t)) \) that can be estimated by \( \hat{\mu}_1(t) = \sum_{k=1}^K \hat{F}_k(t) \).

The two estimators differ in the way the censoring distribution is being estimated as the following expansion demonstrates.

\[
\hat{\mu}_1(t) = \int_0^t \hat{S}(s) \frac{1}{Y_\bullet(s)} dN_\bullet(s) = \sum_{k=1}^K \int_0^t \hat{S}(s) \frac{1}{Y_\bullet(s)} dN^{k-1}_\bullet(s)
\]
\[
= \sum_{k=1}^K \int_0^t \hat{S}_{D,k}(s) \frac{\hat{S}_{C,k}(s)}{\hat{S}_C(s)} \frac{1}{Y^k_\bullet(s)} dN^{k-1}_\bullet(s)
\]  

(6)
where $Y_{k}^{D_{\bullet}}(t) = \sum_{i=1}^{n} I(T_{i} > t, T_{k,i} > t)$, the number under risk in the competing risks model for $T_{k}$, $S_{D,k}(t)$ the Kaplan-Meier estimator of surviving the $k^{th}$ jump and death, $\hat{S}_{C}(t)$ the Kaplan-Meier estimator of the censoring distribution estimated with $T$ and $\delta$, and $\hat{S}_{C,k}(t)$ the Kaplan-Meier estimator of the censoring distribution in the competing risk model with $T_{k}$ and $D$. The last equality in (6) follows from the relations $\hat{S}_{C}(t)\hat{S}_{D}(t) = Y_{\bullet}(t)/n$, and $\hat{S}_{D,k}\hat{S}_{C,k}(t) = Y_{D_{\bullet}}^{k}(t)/n$, see Kaplan and Meier (1958). With this notation $\hat{F}(t) = \int_{0}^{t} \hat{S}_{D,k}(s)(Y_{D_{\bullet}}^{k}(s))^{-1}(s) dN_{D_{\bullet}}^{k-1}(s)$, so the two estimators differ in the ratio of the two estimators of the censoring distribution $\hat{S}_{C,k}(s)/\hat{S}_{C}(s)$.

In particular if there are no censorings the two estimators are equivalent.

Based on simulations it seems that $\hat{\mu}_{1}(t)$ and $\hat{\mu}_{1}(t)$, defined in (1), have different strengths on different parts on the time-scale, see Section 5.2. Note that the standard estimator $\hat{\mu}_{1}(t)$ is easier to compute.

Based on $P(N_{1}^{\bullet}(t) \geq k)$ we can also estimate the second moment, $E((N_{1}^{\bullet}(t))^{2})$, consistently by

$$\hat{\mu}_{1,2}(t) = \sum_{k=0}^{K} k^{2}(\hat{F}(t) - \hat{F}_{k+1}(t)) = \sum_{k=1}^{K} f(k)\hat{F}_{k}(t)$$

with $f(k) = (k + 1)^{2} - k^{2}$. The asymptotic properties follows from the asymptotic properties of the underlying estimators by combining the i.i.d. decompositions for each estimator. We thus have an estimator of the variance of the number of events of type 1, namely $\hat{\mu}_{1,2}(t) - (\hat{\mu}_{1}(t))^{2}$. The integral estimator, $\hat{\mu}_{1,2}(t)$, presented earlier is easier to compute than $\hat{\mu}_{1,2}(t)$ and the difference between estimators lies again in how the censoring distribution is estimated.

Finally, note that the bivariate survival distribution for the recurrent events can also be estimated based on cumulative incidence estimators. First, $P(N_{1}^{\bullet}(t)N_{2}^{\bullet}(t) \geq k)$ is equivalent to the cumulative incidence of the variable $T_{k}^{1,2} = \inf\{ t : N_{1}^{\bullet}(t)N_{2}^{\bullet}(t) \geq k \}$ observed subject to the competing risk of the terminal event, namely $F_{k}^{1,2}(t) = P(T_{k}^{1,2} < t)$. 

11
and can be estimated by the standard cumulative incidence estimator \( \hat{F}_{k}^{1,2}(t) \) for this competing risks model. As a consequence, we can estimate \( E(N_{1}^{*}(t)N_{2}^{*}(t)) \) by

\[
\hat{\mu}_{1,2}(t) = \sum_{k=1}^{K^2} \hat{F}_{k}^{1,2}(t).
\]

In addition, we note that the joint probability \( P(N_{1}^{*}(t) \geq k_{1}, N_{2}^{*}(t) \geq k_{2}) \) can be expressed as the cumulative incidence of the time-variable \( T_{k_{1},k_{2}} = \inf\{t : N_{1}^{*}(t) \geq k_{1}, N_{2}^{*}(t) \geq k_{2}\} \) observed subject to competing risk of the terminal event, namely, \( F_{k_{1},k_{2}}(t) = P(T_{k_{1},k_{2}} < t, T_{k_{1},k_{2}} < D) \).

### 4 Testing

In a situation with a nonnegligible terminal event, one will typically be interested in deciding if the recurrent event processes are correlated among survivors. This can formally be decided by considering the cumulative test statistic

\[
D(t) = \hat{E}(N_{1}^{*}(t)N_{2}^{*}(t)) - \hat{E}_{I}(N_{1}^{*}(t)N_{2}^{*}(t))
\]

for fixed \( t \). The statistic \( D(t) \) is asymptotically normal with variance that can be estimated using a simple bootstrap. If rejecting \( H_0 : D(t) = 0 \), we conclude that there is either positive or negative dependence between the recurrent event processes among survivors. The test statistic is cumulative in the sense that it refers to what is seen up to time \( t \). In a situation where the correlation changes during the study time, the evolution of the correlation can be displayed graphically by plotting \( D(t) \) against time.

Taking full advantage of the way we constructed our estimators \( \hat{E}(N_{1}^{*}(t)N_{2}^{*}(t)) \) and \( \hat{E}_{I}(N_{1}^{*}(t)N_{2}^{*}(t)) \) via one integral that reflects how \( N_{2}(t-) \) affects \( dN_{1}(t) \), and vice versa, we can test specifically if \( N_{2}(t-) \) affects \( dN_{1}(t) \). Consider

\[
D_{N_{2} \rightarrow N_{1}}(t) = \hat{E}(\int_{0}^{t} N_{2}^{*}(s-)dN_{1}^{*}(s)) - \hat{E}_{I}(\int_{0}^{t} N_{2}^{*}(s-)dN_{1}^{*}(s)),
\]
and test $H_0 : \mathcal{D}_{N_2 \rightarrow N_1}(t) = 0$ to conclude that $N_2$ is predictive about $dN_1$. Correspondingly, $H_0 : \mathcal{D}_{N_1 \rightarrow N_2}(t) = 0$ would be a test of if $N_1(t-)$ affects $dN_2(t)$. If both directional correlations are 0, we conclude that $N_2(t)$ and $N_1(t)$ are uncorrelated among survivors.

5 Simulations and data generation

We first considered the covariance estimator under different types of dependence. We used basic intensities, $\lambda^*_j(t)$, that equal the estimated marginals of the recurrent events estimated in the motivating data example, and a basic intensity for the terminal event corresponding to that estimated in the data example. To induce dependence between the counting processes, we generated random effects, either using trivariate log-normally distributed random variables with different types of covariance, or by using (sums of) gamma distributed random variables. Specifically, we simulated observations based on the three intensities

$$\lambda_j(t) = Z_j \lambda^*_j(t)$$

where $j = 1, 2, 3$ refers to the two recurrent events and the terminal event, $Z_j$ are the random effects, and the basic intensities $\lambda^*_j(t)$ are piecewise linear versions of the intensities from the data example estimated by the Nelson-Aalen estimator.

We start by noting that if $Z_1 = \tilde{Z}_1, Z_2 = \tilde{Z}_2, Z_3 = \tilde{Z}_1 + \tilde{Z}_2$, where $\tilde{Z}_k$ are independent positive random variables, then we get recurrent event processes that are uncorrelated among survivors, even though both recurrent event processes are correlated with the terminal event. In Appendix B, we formally show that $E(N^*_1(t)N^*_2(t)) = E_{t}(N^*_1(t)N^*_2(t))$, reflecting our notion of being uncorrelated among survivors.

We considered four scenarios; the two first being Scenario I, with $Z_1 = \tilde{Z}_1, Z_2 = \tilde{Z}_2, Z_3 = \tilde{Z}_3$, and Scenario II, with $Z_1 = \tilde{Z}_1, Z_2 = \tilde{Z}_2, Z_3 = (\tilde{Z}_1 + \tilde{Z}_2)/2$, where the random variables $\tilde{Z}_k$ are independent and Gamma distributed with mean 1 and variance 1. In Scenario I all three random effects are independent and in Scenario II we have independence among the recurrent event processes for the survivors.
In Scenarios III and IV, we used a trivariate log-normally distributed random effect such that all three components had variance 0.5 on the log-hazard scale. To parameterize the covariance between the random effects on the log-hazard scale, we let $\rho_{12}$ denote the correlation between the two recurrent events, $\rho_{j3}$ the correlation between the $j$th recurrent and the terminal event for $j = 1, 2$, and organized these into a vector $\rho = (\rho_{12}, \rho_{13}, \rho_{23})$. Specifically, in Scenario III we considered positive dependence $\rho = (0.5, 0.5, 0.5)$, and in Scenario IV, negative dependence among the recurrent event processes $\rho = (-0.4, 0.5, 0.5)$. We generated censoring times from a uniform distribution over the observation period $[0, 5000]$, leading to 20% of censoring.

To see how the estimated correlations captured the dependencies, we generated 3000 subjects according to the four scenarios. Figure 1 shows the general estimator $\hat{E}(N_1^*(t)N_2^*(t))$ (solid line), the estimator assuming no correlation among survivors $\hat{E}_I(N_1^*(t)N_2^*(t))$ (dotted line), and finally $\hat{E}(N_1^*(t))E(N_2^*(t))$ (broken line). Panels (a) and (b) show Scenario I and II, showing that the observed processes are highly correlated but when the dependence due to the terminal event is taken out, the two processes are indeed uncorrelated among survivors. We note that Scenarios I and II look very similar, even though Scenario I is simulated under independence, and in Scenario II, the random effects of the recurrent events both depend on the random effect of the terminal event. Similarly, the additional positive/negative correlation is captured by the difference between $\hat{E}(N_1^*(t)N_2^*(t))$ and $E_I(N_1^*(t)N_2^*(t))$ in panels (c) and (d).

Figure 1 about here

5.1 Performance of estimators

We demonstrate that our estimators of $\hat{E}(N_1^*(t)N_2^*(t))$ and $\hat{E}_I(N_1^*(t)N_2^*(t))$ are unbiased by comparing them to the empirical means of $n = 100000$ simulated recurrent event processes without censoring. With this sample size, the empirical quantities were completely stable. We also show that the standard errors are reasonably well approximated by a simple block bootstrap.

We generated data from 200, 400, 800 and 1600 subjects along the lines of the four
different scenarios, but we only report the results of scenario III that is closest to the motivating data example. In each setting we performed 5000 replicated analyses, each using a bootstrap with only 100 replications.

We stress that the simulated data is rather wild with a large number of events of type 1 as illustrated in Figure 4. In fact, we observe up to 54 observations per subject when simulating 100000 subjects. About 7% have more than 10 events and around 2% have more than 16 events. In contrast, the largest number of events of type 2 per subject is only 12 (over the 100000 subjects). Only around 1% have more than 4 events.

Table 1 around here

In all scenarios, the estimators worked well, see Table 1. In particular, we observe that the estimators are both essentially unbiased even for small sample sizes. The general estimator $\hat{E}(N_1(t)N_2(t))$ seems, as expected, more difficult to estimate and we observed a small relative bias. Standard errors and coverage worked well for $\hat{E}_I(N_1^*(t)N_2^*(t))$ already at the smallest sample size 200, and even though the coverage was slightly too small for $n = 200$ this improved as the sample size increased. In contrast, it was more difficult to estimate the standard error of $\hat{E}(N_1^*(t)N_2^*(t))$, and we noticed that the standard errors were estimated a bit too small for $n = 200$ and remained slightly too small for the larger sample sizes, but this improved and lead to reasonable coverage for the largest sample size $n = 1600$.

5.2 Product-limit or integral estimators

In Section 2 and 3 we showed that the mean, the variance and the probability of exceeding $k$ events can be estimated by a product-limit estimator or a computationally simpler integral estimator that uses one single estimate of the censoring distribution across all $k$, $\hat{F}_k(t)$ and $\tilde{F}_k(t)$, respectively. For the mean we used either $\tilde{\mu}(t)$ (4) or $\hat{\mu}(t)$ (3). We call $\tilde{\mu}(t)$ and the corresponding variance estimator product-limit estimators when based on underlying product-limit estimators $\hat{F}_k(t)$.

We simulated data from a model with independence, positive dependence (gamma random effects), and negative dependence between the recurrent events and death. We
only report the results for the situation with negative dependence, since results were quite
similar in the other cases. We used the intensities from the previous section together with
random effects drawn from a trivariate normal distribution with \( \rho = (0.5, -0.4, -0.4) \).
We did 5000 repetitions with \( n = 200 \) to compare the different estimators.

Figure 2 about here

Figure 2 panel a shows the empirical means of the marginal means for two recurrent
events based on the product-limit estimator (pl, broken line), or integral estimators (idN,
dotted line), and true mean (solid line). The dotted line is hidden by the solid line,
indicating essentially no bias for this estimator. The three upper curves are for \( N_1(t) \),
and the three lower curves for \( N_2(t) \). We note that the integral estimator is slightly better
in terms of bias, but this bias disappeared when \( n \) was increased to 400 (not shown). In
panel b we show the standard deviation of the product-limit based estimators relative to
the standard deviation of the integral estimators for \( N_1(t) \) (solid line) and \( N_2(t) \) (broken
line). We note that the product-limit estimator has smaller variance for early times,
and that this is reversed for later times. In panel c, we show the mean of the estimated
variance of two recurrent events based on product-limit estimators (pl, broken line),
integral estimators (idN, dotted line), and true mean (solid line). The three top curves
are for \( N_1 \), and the three lower curves for \( N_2(t) \). Panel d gives the standard deviation
of the product-limit based estimators relative to the standard deviation of the integral
estimators for \( N_1(t) \) (solid line) and \( N_2(t) \) (broken line). The picture is quite similar to
that for the estimators of the mean functions. Finally, panel e shows the mean estimated
probability of exceeding 1, 3, 5, and 10 events, respectively, for the product-limit estimator
(pl, broken line), the integral estimator (idN dotted line), and true mean (solid line) for
\( N_1(t) \). The standard deviation of the product-limit estimators relative to the standard
deviation of the integral estimators exceeding 1 event (solid line), 3 events (broken line),
5 events (dotted line) and 10 events (broken-dotted line), respectively, are given in panel
f. The probability of exceeding 1 event is generally better estimated by the product-limit
estimator, but for the other number of events, again the probability is well estimated by
both estimators. The standard error is smaller early on for the product-limit estimator,
and this is reversed for later times.

It seems that the smaller standard error of the product-limit estimator of exceeding $k$ events early on is inherited in the derived estimators of the mean and variance. Another observation is that the mean is more biased when estimated via the product-limit estimators. Note again, that the integral estimators are easier to compute.

6 Example

We consider 715 patients with intestinal failure receiving home parenteral support (HPS) through a central venous catheter (CVC). These patients may experience several types of events while part of the HPS program. In particular, they may develop catheter-related bloodstream infections (CRBSI) and catheter occlusions. The latter is defined as problems with infusion or aspiration on the CVC, and can occur due to thrombotic product or precipitates of the content in the parenteral nutrition in the lumen of the CVC. This material may prone to microorganism adhesion to the CVC, thereby promoting biofilm-formation, which is a precursor for developing subsequent CRBSIs, thus possibly generating more infections in patients with occlusions. On the contrary we did not expect infections to lead to an increased level of occlusions. We observed a total of 1400 CRBSIs and 58 occlusions in the 715 patients during the 14 years of follow up.

In this section we used the integral estimators. Figure 3 shows the marginal means and standard deviations of the two events, the probability of staying in the HPS program, and the estimated standard deviation of the two recurrent event counting processes. The standard deviation of the number of CRBSIs prior to the terminal event is quite large compared to its mean, showing a rather large heterogeneity among patients. Similarly, the standard deviation of the number of occlusions is also large compared to its mean. Figure 3 below provides further insight into this.

The large heterogeneity among patients is illustrated in Figure 4. Panel a shows the probability of having more than $k$ CRBSIs and panel c shows the probability of more
that \( k \) occlusions, while panel b and d, respectively, show the proportion of patients with \( k \) CRBSIs or \( k \) occlusions among those under risk.

Figure 4 about here

The estimated mean \( E(N_1^*(t)N_2^*(t)) \) is shown in Figure 5, panel c. For comparison, the plot also includes the mean estimated under the assumption that the processes are independent among survivors. Panels a, and b, show the estimators of the two integrals \( E(\int_0^t N_2^*(s-)dN_1^*(s)) \) and \( E(\int_0^t N_1^*(s-)dN_2^*(s)) \), respectively, both estimated with or without the assumption of independence of the two processes among survivors. Panel a demonstrates that \( N_2(t-) \) is predictive for \( dN_1(t) \), i.e. the risk of CRBSI is dependent on the level of occlusions, whereas in contrast, panel b provides no evidence of CRBSI being predictive for occlusions. Finally, panel d shows the difference \( E(N_1^*(t)N_2^*(t)) - E(N_1^*(t)N_2^*(t)) \) (solid line) with 95 % confidence intervals based on a bootstrap. Panel d suggests that there is borderline positive correlation between the processes among the survivors, and comparing panels a and b, we conclude that this correlation is driven by the dependence of \( N_1(s) \) given \( N_2(s-) \), and not the other way around. That is, the presence of occlusions seems to suggest an increased level of infections as indeed hypothesized prior to the analyses.

Figure 5 about here

Formally, testing if \( H_0 : \mathcal{D}(\tau) = 0 \) and \( H_0 : \mathcal{D}_{N_2\to N_1}(\tau) = 0 \) (occlusions being predictive for crbsi) we get the p-values 0.087 and 0.055, respectively.

7 Discussion

We have provided a simple nonparametric estimator of the covariance or correlation of the mean number of events for two recurrent event processes, possibly in the presence of a terminal event. In the case of a terminal event, we suggest a correction for the dependence due the terminal event by estimating the covariance also under nondependence between survivors in both directions. This seems to capture different types of dependence well.

In the worked example we used the fact that our estimator is decomposed into two
terms, each reflecting how one process is driven by the other process. This revealed that the dependence in CRBSI and occlusions in our intestinal failure patients were indeed driven by the increased risk of infections due to the presence of occlusions.

An alternative to the estimators suggested here are the inverse probability of censoring weighting estimators that are easy to use and give similar estimates. One difficulty with these estimators is that they are harder to correct for dependence driven by a terminal event.

We have given some easily computed additional summaries of the heterogeneity in the recurrent event processes. We presented two types of estimators; one class similar to the marginal mean estimator, the integral estimators, and another class of estimators that we called product-limit estimators. The properties of these estimators need to be investigated and compared further. In our simulation study we saw that it is not clear which estimator to prefer.

The suggested methods may be extended to regression settings by considering covariate dependent models for the different components of our estimators. Indeed, one could model how the recurrent events are related to covariates by for example modelling the risk of CRBSI given the history of the processes and other covariates, see for example Cook and Lawless (2007). In some settings this may also be needed to make the censoring distribution (conditionally) independent of the processes of interest. However, the aim of our work was to demonstrate how simple nonparametric estimators can be constructed.

All figures and curves can be reproduced using the development version of the mets R package on github.

Acknowledgement

We thank the Department of Medical Gastroenterology at the University Hospital Rigshospitalet for letting us use their data.
Appendix A

Asymptotic expansion

Define, for $j = 1, 2$ and $k = 1, \ldots, K$, $m_j(t) = E(N_j^0(t) | D > t)$, $\hat{m}_j(t) = \sum_{k=1}^{K} k Y_j^k(t)/\hat{Y}_j(t)$, $\pi(t) = P(Y(t) = 1)$, $\hat{\pi}(t) = Y_j(t)/n$, $\pi_j^k(t) = P(N_j^0(t-\cdot) = k | Y(t) = 1)$, $\hat{\pi}_j^k(t) = Y_j^k(t)/\hat{Y}_j(t)$, $dR_j^k(t) = E(dN_j^0(t) | D > t, N_j^0(t-\cdot) = k)$, $\hat{R}_j^k(t) = \int_{0}^{t} \frac{1}{Y_j^k(s)} d\hat{N}_j^k(s)$ recall that $1^o = 2, 2^o = 1$, and that $\hat{N}_j^k(t) = \sum_{i=1}^{n} \int_{0}^{t} I(N_{ij}^0(s) = k) dN_{ij}(s)$. Further let $\nu_j(t) = \int_{0}^{t} S(s) \pi_j(s) dR_j(s)$ and $\nu_j^k(t) = \int_{0}^{t} S(s) k \pi_j^k(s) dR_j^k(s)$.

Under independence among survivors we estimate $E(\int_{0}^{t} N_1^0(s-\cdot) dN_2^0(s))$ by

$$
\int_{0}^{t} \hat{S}(s) \{\sum_{k=1}^{K} k \frac{Y_j^k(s)}{\hat{Y}_j(s)}\} \frac{1}{\hat{Y}_j(s)} dN_2^0(s) = \int_{0}^{t} \hat{S}(s) \hat{m}_1(s) d\hat{R}_2(s),
$$

see (5). Consider

$$
n^{1/2} \left( \int_{0}^{t} \hat{S}(s) \hat{m}_1(s) d\hat{R}_2(s) - \int_{0}^{t} S(s) m_1(s) dR_2(s) \right)
= n^{1/2} \int_{0}^{t} (\hat{S}(s) - S(s)) m_1(s) dR_2(s) + n^{1/2} \int_{0}^{t} S(s) m_1(s) d(\hat{R}_2(s) - R_2(s))
+ n^{1/2} \int_{0}^{t} S(s)(\hat{m}_1(s) - m_1(s)) dR_2(s) + \mathcal{O}_p(1).
$$

Following Ghosh and Lin (2000), the first and second term above can be expanded and written as a sum i.i.d. zero mean processes

$$
n^{-1/2} \sum_{i=1}^{n} \left( \int_{0}^{t} \frac{S(s) m_1(s)}{\pi(s)} dM_{i2}(s) - \nu_2(t) \int_{0}^{t} \frac{1}{\pi(s)} dM_i^D(s) + \int_{0}^{t} \frac{\nu_2(s)}{\pi(s)} dM_i^D(s) \right) + \mathcal{O}_p(1),
$$

where $M_i^D(t) = N_i^D(t) - \int_{0}^{t} Y_i(s) dA^D(s)$, and $M_{i2}(t) = N_{i2}(t) - \int_{0}^{t} Y_i(s) dR_2(s)$. Finally, the last term has the expansion

$$
n^{1/2} \int_{0}^{t} S(s)(\hat{m}_1(s) - m_1(s)) dR_2(s)
= n^{-1/2} \sum_{i=1}^{n} \int_{0}^{t} S(s) \frac{Y_i(s)}{\pi(s)} (N_{i1}(s-\cdot) - m_1(s)) dR_2(s) + \mathcal{O}_p(1),
$$

20
using dominated convergence.

We now turn to the general estimator of $E(\int_0^t N_1^*(s-)dN_2^*(s))$ given in (3) and look at the term for each $k$ separately,

$$\int_0^t \hat{S}(s)k \frac{Y_{k2}(s)}{Y_{k1}(s)}d\hat{N}_{k2}(s) = \int_0^t \hat{S}(s)k\hat{\pi}_1^k(s)d\hat{R}_{k2}(s)$$

Similarly to in independence among survivor setting above,

$$n^{1/2} \left( \int_0^t \hat{S}(s)k\hat{\pi}_1^k(s)d\hat{R}_{k2}(s) - \int_0^t S(s)k\pi_1^k(s)dR_{k2}(s) \right),$$

can be written as the sum of

$$n^{-1/2} \sum_{i=1}^n \left( \int_0^t \frac{S(s)\pi_1^k(s)}{\pi_1^t(s)}dM_{k2}^i(s) - \nu_2^k(t) \int_0^t \frac{1}{\pi_1^t(s)}dM_{k1}^i(s) + \int_0^t \frac{\nu_2^k(s)}{\pi_1^t(s)}dM_{k1}^i(s) \right) + o_P(1)$$

where $M_{k2}^i(t) = N_{k2}^i(t) - \int_0^t Y_i(s)I(N_{i1}(s-) = k)dR_{k2}^i(s)$, and

$$n^{1/2} \int_0^t S(s)k(\pi_1^k(s) - \pi_1^t(s))dR_{k2}^t(s)$$

$$= n^{-1/2} \sum_{i=1}^n \int_0^t S(s)k\frac{Y_i(s)}{\pi_1^t(s)}(I(N_{i1}(s-) = k) - \pi_1^t(s))dR_{k2}^i(s) + o_P(1)$$

For finite $K$ we sum all these terms to get the i.i.d. expansion of $n^{1/2}(\hat{E}(\int_0^t N_1^*(s-)dN_2^*(s)) - E(\int_0^t N_1^*(s-)dN_2^*(s)))$. We obtain similar terms for $E(\int_0^t N_2^*(s-)dN_1^*(s))$. Both estimators for $E(N_1^*(t)N_2^*(t))$, (3) and (3), and their difference, $\hat{E}(N_1^*(t)N_2^*(t)) - \hat{E}(N_1^*(t)N_2^*(t))$ can thus be written as sums of i.i.d. processes.
Appendix B

Assume that conditional on the random effect $V = (V_1, V_2, V_D) \sim f_V, N_k(t), k = 1, 2$ are nonstationary Poisson processes with intensities $v_k d\Lambda_k(t)$ and the hazard of the terminal event is $v_D d\Lambda^P(t)$. Also assume that $(N_1, N_2, D)$ are mutually independent given $V$.

To compute $E(dN_1(t)|D > t, N_2(t-) = k)$ and $E(dN_1(t)|D > t)$ we need to compute $E(V_1|D > t, N_2(t-) = k)$ and $E(V_1|D > t)$.

\[
E[V_1|N_2(t-), D > t] = \frac{\int v_1 f(N_2(t-), D > t|V)f_V(V)dv}{\int f(N_2(t-), D > t|V)f_V(V)dv} \\
= \frac{\int v_1 f(N_2(t-)|V)f(D > t|V)f_V(V)dv}{\int f(N_2(t-)|V)f(D > t|V)f_V(V)dv} \\
= \frac{\int v_1 e^{-v_2\Lambda_2(t)} (v_2\Lambda_2(t))^{N_2(t-)} / N_2(t-)!e^{-v_D\Lambda^P(t)} f_V(V)dv}{\int v_2 N_2(t-) e^{-v_2\Lambda_2(t)-v_D\Lambda^P(t)} f_V(V)dv} \\
= \frac{\int v_1 v_2 N_2(t-) e^{-v_2\Lambda_2(t)-v_D\Lambda^P(t)} f_V(V)dv}{\int v_2 N_2(t-) e^{-v_2\Lambda_2(t)-v_D\Lambda^P(t)} f_V(V)dv} \tag{7}
\]

\[
E[V_1|D > t] = \frac{\int v_1 e^{-v_D\Lambda^P(t)} f_V(V)dv}{\int e^{-v_D\Lambda^P(t)} f_V(V)dv} \tag{8}
\]

Uncorrelated processes among survivors

Consider a frailty structure where the random effects $V_1$ and $V_2$ are independent. Specifically, the $V$’s are related to the independent random variables $W_{1D} \sim f_{W_{1D}}$ and $W_{2D} \sim f_{W_{2D}}$ through $V_1 = W_{1D}, V_2 = W_{2D}$ and $V_D = \phi_1 W_{1D} + \phi_2 W_{2D}$.

Using (7) and (8),

\[
E[V_1|N_2(t-), D > t] = \frac{\int w_{1D} e^{-w_{1D}\phi_1\Lambda^P(t)} f_{w_{1D}}(w_{1D})dw_{1D} \int w_{2D}^{N_2(t-)} e^{-w_2D(\Lambda_2(t)+\phi_2\Lambda^P(t))} f_{w_2D}(w_2D)dw_2D}{\int e^{-w_{1D}\phi_1\Lambda^P(t)} f_{w_{1D}}(w_{1D})dw_{1D} \int w_{2D}^{N_2(t-)} e^{-w_2D(\Lambda_2(t)+\phi_2\Lambda^P(t))} f_{w_2D}(w_2D)dw_2D} \\
= \frac{\int w_{1D} e^{-w_{1D}\phi_1\Lambda^P(t)} f(w_{1D})dw}{\int e^{-w_{1D}\phi_1\Lambda^P(t)} f(w_{1D})dw} \\
E[V_1|D > t] = \frac{\int w_{1D} e^{-w_{1D}\phi_1\Lambda^P(t)} f(w_{1D})dw_{1D}}{\int e^{-w_{1D}\phi_1\Lambda^P(t)} f(w_{1D})dw_{1D}}
\]

Thus, we have $E[V_1|N_2(t-), D > t] = E[V_1|D > t]$ and the processes are uncorrelated among survivors.

In contrast, a multiplicative random effects structure, as for example generated from
We now consider random effects, correlation among survivors. Consider the random effects structure \( V_1 = \exp(b_1) \), \( V_2 = \exp(b_2) \) and \( V_D = \exp(\phi_1 b_1 + \phi_2 b_2) \) where \( b_1 \sim f_{b_1} \) and \( b_2 \sim f_{b_2} \) are independent,

\[
E[V_1 | N_2(t-), D > t] = \int \frac{e^{b_1} e^{N_2(t-)-b_2} \Lambda \exp(t) f_{b_1}(b_1) f_{b_2}(b_2) dB_1 dB_2}{\int e^{N_2(t-)-b_2} \Lambda \exp(t) f_{b_1}(b_1) f_{b_2}(b_2) dB_1 dB_2} \\
\quad \quad \quad \quad \quad \quad \quad \quad = \int \frac{e^{b_1} e^{\phi_1 b_1 + \phi_2 b_2} \Lambda \exp(t) f_{b_1}(b_1) f_{b_2}(b_2) dB_1 dB_2}{\int e^{\phi_1 b_1 + \phi_2 b_2} \Lambda \exp(t) f_{b_1}(b_1) f_{b_2}(b_2) dB_1 dB_2} \\
\quad \quad \quad \quad \quad \quad \quad \quad = \int \frac{w_1 e^{-w_1 \Lambda \exp(t)} f_v(v) dv}{\int e^{-w_1 \Lambda \exp(t)} f_v(v) dv} \\
\quad \quad \quad \quad \quad \quad \quad \quad = E[V_1 | D > t].
\]

**Correlation among survivors**

We now consider random effects \( V_1 \) and \( V_2 \) that are dependent. Let \( V_1 = W_{12} + W_{1D} \), \( V_2 = W_{12} + W_{2D} \) and \( V_D = W_{1D} + W_{2D} \), where the \( W \)’s are independent. In this case we show that \( E(V_1 | D > t, N_2(t-)) \neq E(V_1 | D > t) \).

Consider the denominator in the right-hand side of (7),

\[
\int (w_{12} + w_{2D})^{N_2(t-)} e^{-(w_{12}+w_{2D})\Lambda_2(t)-(w_{1D}+w_{2D})\Lambda(t)} f_v dv \\
\quad \quad = \int \sum_{l=0}^{N_2(t-)} \binom{N_2(t-)}{l} w_{12}^{N_2(t-)-l} w_{2D}^l \\
\quad \quad \quad \quad \times e^{-w_{12}\Lambda_2(t)} e^{-w_{2D}(\Lambda_2(t)+\Lambda(t))} e^{-w_{1D}t} f_v(v) dv \\
\quad \quad = \int e^{-w_{1D}t} f_{w_1D}(w_{1D}) dw_{1D} \\
\quad \quad \quad \quad \times \sum_{l=0}^{N_2(t-)} \binom{N_2(t-)}{l} \int w_{12}^{N_2(t-)-l} e^{-w_{12}\Lambda_2(t)} f_{w_{12}}(w_{12}) dw_{12} \\
\quad \quad \quad \quad \times \int w_{2D}^{l} e^{-w_{2D}\Lambda_2(t)} f_{w_{2D}}(w_{2D}) dw_{2D}.
\]

Similarly, with this random effects structure, the numerator on the right-hand side of (7)
\[
\int w_{1D}(w_{12} + w_{2D})N_2(t) e^{-(w_{12}+w_{2D})\Lambda_2(t)-(w_{1D}+w_{2D})\Lambda_D(t)} f_{\nu}(\mathbf{v}) d\mathbf{v} \\
+ \int w_{12}(w_{12} + w_{2D})N_2(t) e^{-(w_{12}+w_{2D})\Lambda_2(t)-(w_{1D}+w_{2D})\Lambda_D(t)} f_{\nu}(\mathbf{v}) d\mathbf{v} \\
= \int w_{1D}e^{-w_{1D}\Lambda_D(t)} f_{w_{1D}}(w_{1D}) dw_{1D} \\
\times \sum_{l=0}^{N_2(t)} \left( \binom{N_2(t)}{l} \right) \int w_{12}^{N_2(t)-l} e^{-w_{12}\Lambda_2(t)} f_{w_{12}}(w_{12}) dw_{12} \\
\times \int w_{2D}^{l} e^{-w_{2D}(\Lambda_2(t)+\Lambda_D(t))} f_{w_{2D}}(w_{2D}) dw_{2D} \\
+ \int e^{-w_{1D}\Lambda_D(t)} f_{w_{1D}}(w_{1D}) dw_{1D} \\
\times \sum_{l=0}^{N_2(t)} \left( \binom{N_2(t)}{l} \right) \int w_{12}^{N_2(t)-l+1} e^{-w_{12}\Lambda_2(t)} f_{w_{12}}(w_{12}) dw_{12} \\
\times \int w_{2D}^{l} e^{-w_{2D}(\Lambda_2(t)+\Lambda_D(t))} f_{w_{2D}}(w_{2D}) dw_{2D}
\]

Combining the two displays above,

\[
E [V_1 | N_2(t-), D > t] \\
= E[W_{1D} | D > t] \\
+ \frac{\sum_{l=0}^{N_2(t-)} \left( \binom{N_2(t-)}{l} \right) \int w_{12}^{N_2(t-)-l} e^{-w_{12}\Lambda_2(t)} f_{w_{12}}(w_{12}) dw_{12} \int w_{2D}^{l} e^{-w_{2D}(\Lambda_2(t)+\Lambda_D(t))} f_{w_{2D}}(w_{2D}) dw_{2D}}{\sum_{l=0}^{N_2(t-)} \left( \binom{N_2(t-)}{l} \right) \int w_{12}^{N_2(t-)-l} e^{-w_{12}\Lambda_2(t)} f_{w_{12}}(w_{12}) dw_{12} \int w_{2D}^{l} e^{-w_{2D}(\Lambda_2(t)+\Lambda_D(t))} f_{w_{2D}}(w_{2D}) dw_{2D}}
\]

On the other hand, (8) yields

\[
E [V_1 | D > t] = \frac{\int (w_{1D} + w_{12}) e^{-(w_{1D}+w_{2D})\Lambda_D(t)} f_{\nu}(\mathbf{v}) d\mathbf{v}}{\int e^{-(w_{1D}+w_{2D})\Lambda_D(t)} f_{\nu}(\mathbf{v}) d\mathbf{v}} = E[W_{1D} | D > t] + E[W_{12}]
\]

That is, the processes \( N_1(t) \) and \( N_2(t) \) are correlated among survivors as \( E [dN_1(t) | N_2(t-), D > t] \neq E [dN_1(t) | D > t] \).
References


Table 1: Bias, standard errors and coverage for $\hat{E}(N_1^*(t)N_2^*(t))$ and $\hat{E}_I(N_1^*(t)N_2^*(t))$. Relative bias, the mean of the estimated standard errors (mean se), and the standard deviation of the estimators (sd), and coverage of 95% confidence intervals (cov) based on 5000 repetitions. Simulation mimics the motivating data.
Figure 1: Correlation for different types of dependence. Simulated data $N = 3000$, estimators $\hat{E}(N_1(t)N_2(t))$ (solid line), $\hat{E}_I(N_1(t)N_2(t))$ (dotted line), and $\hat{E}(N_1(t))E(N_2(t))$ (broken line). Panel a (full independence) scenario I, panel b (independence among survivors) scenario II, panel c (positive dependence among survivors) scenario III, panel d (negative dependence among survivors) scenario IV. Scenarios described in text.
Figure 2: Performance of product-limit and integral estimators see text for details.
Figure 3: Marginal means of CRBSI, occlusions (with pointwise 95 % confidence intervals), the standard deviations of the number of CRBSI and occlusions as well as the probability of surviving terminal event.
Figure 4: Probability distribution CRBSI over time (a) and probability of number of previous events of CRBSI for those under risk (b). Probability distribution of occlusions over time (c) and probability of number of previous events of occlusions for those under risk (d).
Figure 5: Dependence between CRBSI and occlusions over time. See text for details.
15/01 Mansourvar, Z., Martinussen, T. & Scheike, T. Semiparametric regression for restricted mean residual life under right censoring.

15/02 Mansourvar, Z., Martinussen, T. & Scheike, T. An additive-multiplicative restricted mean residual life model.

15/03 Mansourvar, Z. & Martinussen, T. Estimation of average causal effect using the restricted mean residual lifetime as effect measure.


15/06 Müller, M. & Kreiner, S. Item Fit Statistics in Common Software for Rasch Analysis.

16/01 Andersen, P.K. Life years lost among patients with a given disease.

16/02 Strohmaier, S., Haase, N., Weterslev, J. & Lange, T. A simple to implement algorithm for natural direct and indirect effects in survival studies with a repeatedly measured mediator.

16/03 Andersen, P.K., Syriopoulou, E. & Parner, E. Causal inference in survival analysis using pseudo-observations.

16/04 Paul Blanche, Michael W. Kattan & Thomas A. Gerds. The c-index is not proper for the evaluation of $t$-year predicted risks.

17/01 Liis Starkopf, Mikkel Porsborg Andersen, Thomas Alexander Gerds, Christian Torp-Pedersen & Theis Lange. Comparison of five software solutions to mediation analysis.

17/02 Klemen Pavlič, Torben Martinussen & Per Kragh Andersen. Goodness of fit tests for estimating equations based on pseudo-observations.
Julie K. Furberg, Christian B. Pipper & Thomas Scheike. Flexible tests for treatment equivalence in a survival setting.

Luise Cederkvist, Klaus K. Holst, Klaus K. Andersen & Thomas H. Scheike. Modelling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing.

Cristina Boschini, Klaus K. Andersen & Thomas H. Scheike. Excess risk estimation for matched cohort survival data.

Paul Blanche. Confidence intervals for the cumulative incidence function via constrained NPMLE.

Per Kragh Andersen, Jules Angst & Henrik Ravn. Modeling marginal features in studies of recurrent events in the presence of a terminal event.

Thomas H. Scheike, Frank Eriksson & Siri Tribler. The mean, variance, and correlation for bivariate recurrent event data with a terminal event.