Published in final edited form as:
Biometrics. 2010 March ; 66(1): 39–49. doi:10.1111/j.1541-0420.2009.01266.x.

Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring

C.-Y. Huang[1,*], **J. Qin**[1], and **M.-C. Wang**[2]

[1] Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

[2] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

Summary

Recurrent event data analyses are usually conducted under the assumption that the censoring time is independent of the recurrent event process. In many applications the censoring time can be informative about the underlying recurrent event process, especially in situations where a correlated failure event could potentially terminate the observation of recurrent events. In this paper, we consider a semiparametric model of recurrent event data that allows correlations between censoring times and recurrent event process via frailty. This flexible framework incorporates both time-dependent and time-independent covariates in the formulation, while leaving the distributions of frailty and censoring times unspecified. We propose a novel semiparametric inference procedure that depends on neither the frailty nor the censoring time distribution. Large sample properties of the regression parameter estimates and the estimated baseline cumulative intensity functions are studied. Numerical studies demonstrate that the proposed methodology performs well for realistic sample sizes. An analysis of hospitalization data for patients in an AIDS cohort study is presented to illustrate the proposed method.

Keywords

Comparable recurrence times; Frailty; Pairwise pseudolikelihood; Proportional rate model

1. Introduction

Recurrent event data arise in a wide variety of settings, including public health, biomedical research, reliability studies, and social sciences. In these studies each subject is at risk of experiencing repeated events, and the observation of recurrent events is terminated at or before the end of the study. For example, an HIV patient may experience multiple opportunistic infections during follow-up, and a schizophrenic patient may be repeatedly admitted to a psychiatric hospital. The recording of recurrent events could stop early if the patient withdraws or dies before the study period ends. In the examples, the analysis of recurrent event data is more relevant than time to first infection or hospital admission, because recurrent events are more informative about a patient's underlying health condition and can be used to assess whether there is evidence for deterioration over the long-term course of the disease.

[*] email: huangchi@niaid.nih.gov.

A key feature of the recurrent event data structure is that the event recurrence times within a subject are stochastically ordered and typically correlated. Hence recurrent event data can be viewed as clustered “survival” time data. However, methods for clustered survival time data cannot be applied directly because the cluster size, i.e. the number of events of a subject, is correlated with the underlying distribution of recurrent events. Various methodologies have been proposed to study the risk of event occurrence over time. Extending the methodologies of survival analysis, Prentice, Williams and Peterson (1981), Andersen and Gill (1982), and Chang and Wang (1999) considered conditional models based on intensity and hazard functions. Others, such as Pepe and Cai (1993), Lawless and Nadeau (1995), and Lin et al. (2000), proposed marginal multiplicative rate models based on the number of recurrent events. Schaubel, Zeng, and Cai (2006) considered semiparametric additive rate models, where the effect of covariates acts additively on the marginal rate of recurrent event. As a useful alternative, Ghosh (2004) and Sun and Su (2008) proposed accelerated rate models where the covariate effect transforms the time scale of the baseline rate function. Finally, Zeng and Lin (2006) proposed a class of semiparametric transformation models to allow for more flexible modeling of the recurrent event process.

The analysis of recurrent event data is usually conducted with the assumption of independent censoring; hence, at any time point, the collection of subjects under observation is a random sample of the study population defined at the time origin. In many instances, however, censoring can be informative about the recurrent event process, especially when a failure event serves as a part of the censoring mechanism and precludes further observation of recurrent events. The afforementioned problem in dealing with recurrent event data under informative censoring complicates analyses for the ALIVE (AIDS Link to Intraveneous Experience) study (Vlahov et al. 1991), where nearly 3000 injection drug users (IDU) from Baltimore Maryland were recuited through extensive community outreach efforts. During the study period substantial information was collected including repeated hospitalizations, HIV test results, race, gender, and death. Hospitalization is an important integrated measure of an IDU's health status, as it reflects the negative health consequences from a variety of causes, e.g. violence, opportunistic infections, liver-related complications, or complications of injection drug use, such as infections of the skin and soft tissue (abscess, cellulitis, and necrotizing fasciitis), bacteremia, endocarditis, and osteomyelitis. Both time-dependent covariates, e.g. HIV status, and time-independent covariates, e.g. race and gender, might influence the rehospitalization risk. The potential informative censoring makes it difficult to properly assess the effect of these risk factors. The censoring among HIV-negative users is likely to be caused by informative dropouts from healthier IDUs. Moreover, HIV positive IDUs tend to be hospitalized more often and also have a higher mortality rate. A naive analysis of the hospitalization rate is likely to yield a biased estimate of the effect of HIV status on overall hospitalization.

To account for the dependence of recurrent events and the censoring event, researchers have proposed two types of approaches, marginal models and frailty models. Ghosh and Lin (2003) studied a correlated marginal model for the joint distribution of recurrent events and the failure time. By artificially censoring some observed event times, the authors developed a semiparametric estimation procedure for estimating the effect of time-independent covariates without modeling the correlation between the recurrent event process and the failure time. Lancaster and Intrator (1998) considered a joint fully parametric model of the recurrent event process and the failure time, where the dependency between the two outcomes is induced by sharing an unobserved frailty in the intensity of recurrent event process and the hazard of the failure event. Extending the work of Lancaster and Intrator (1998), Liu, Wolfe, and Huang (2004), Ye, Kalbfleisch, and Schaubel (2007), Huang and Liu (2007) and Rondeau et al. (2007) studied joint semiparametric models with a specified frailty distribution and estimated the model parameters by maximizing the semiparametric likelihoods. On the other hand, Wang, Qin, and Chiang (2001), Huang and Wang (2004), Huang, Wang, and Zhang (2006), Sun, Tong, and He (2007) proposed nonparametric and semiparametric estimation procedures where, in addition to the baseline intensity function and the baseline hazard function, the distribution of the unobserved frailty was also left unspecified. The estimation procedures proposed by these authors, however, cannot deal with time-dependent covariates.

We consider a semiparametric model for recurrent event data that accounts for both time-dependent and time-independent covariates in modeling the risk of recurrent events. This model allows the censoring time to be correlated with the recurrent event process via an unobserved frailty, relaxing the independent censoring assumption required by the usual risk set methods. Current methods that deal with time-dependent covariates require either an independent censoring assumption (Lin et al. 2000) or a parametric specification of the frailty distribution (Lancaster and Intrator 1998, Liu et al. 2004, and Ye et al. 2007). In practice, however, these methods may suffer from assumption violations due to informative censoring or lack of model checking techniques for the distribution of unobserved frailty. In this paper we present a novel semiparametric estimation procedure that depends on neither the distributions of frailty variables nor the failure times. We propose to estimate the regression coeffcients of time-dependent covariates by applying the conditional likelihood method to comparable pairs of event times, so the the nuisance parameters are eliminated from the conditional likelihood function. We then construct a weighted truncation product-limit estimator, adjusting for sampling bias in the risk sets, to estimate the shape function of the recurrent event process. Finally, we solve estimating equations formulated based on expected number of observed events to estimate the effects of time-independent covariates. We apply this approach to the ALIVE study and show that HIV-positive status is associated with an increased risk in repeated hospitalization.

2. Model Setup

Suppose that $[0, \tau]$ is the time period of interest or the study time interval, where the recurrent events could potentially be observed up to τ. Let subscript i be the index for a subject, $i = 1, 2, \ldots, n$. For subject i, let $N_i(t)$ represent the number of events that occur over the interval $[0, t]$, $\boldsymbol{X}_i(\cdot)$ be a bounded p-dimensional time-dependent covariate process evolving in the time interval $[0, \tau]$, and $\boldsymbol{W}_i$ be a $q \times 1$ vector of time-independent covariates. Denote by $\boldsymbol{X}_i(t) = \{\boldsymbol{X}_i(u) : 0 \leqslant u \leqslant t\}$ the covariate history of $\boldsymbol{X}_i$ up to time t, and assume $\boldsymbol{X}_i$ is a left-continuous covariate process. Let Y_i be the censoring time at which the observation of recurrent events is terminated on $[0,\tau]$. Note that Y_i can be the time of a composite censoring event, $Y_i=\min\left(Y_{i0}^*, Y_{i1}^*\right)$, where Y_{i0}^* represents a noninformative censoring time, such as the end of the study, that is independent of $N_i(\cdot)$, and Y_{i1}^* represents an informative censoring time, such as the time of death, that is correlated with $N_i(\cdot)$.

We introduce a nonnegative valued frailty variable Z_i, with $E(Z^2) < \infty$, to account for the dependence between the underlying recurrent event process $N_i(\cdot)$ and censoring time Y_i. For identifiability reason we assume $E[Z_i\boldsymbol{W}_i,\boldsymbol{X}_i(\tau)] = 1$ throughout the paper. Conditioning on Z_i, $N_i(\cdot)$ is a nonhomogeneous Poisson process with intensity function given by

$$\lambda_i(t) = Z_i \lambda_0(t) \exp\left\{\boldsymbol{X}_i(t)'\beta + \boldsymbol{W}_i'\gamma\right\}, \quad t \in [0, \tau], \tag{1}$$

where β and γ are $p \times 1$ and $q \times 1$ vectors of parameters, and the baseline intensity function $\lambda_0(t)$ is an arbitrary continuous function with $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$. We further assume that, conditional on $(Z_i,\boldsymbol{X}_i,\boldsymbol{W}_i)$, Y_i is independent of $N_i(\cdot)$; thus the censoring time can be correlated with $N_i(\cdot)$ through the connection with unobserved frailty Z_i and covariates $(\boldsymbol{X}_i,\boldsymbol{W}_i)$.

Note that two different types of informative censoring are often encountered in real application: informative dropouts and terminal events that could preclude further occurrence of recurrent events. For the case of informative dropouts the proposed model can be interpreted straightforwardly. For the latter case, recurrent events after the terminal event can be considered latent and modelled as if they could have occurred, analogous to latent failure times in the setting of competing risks (see discussions in Ghosh and Lin 2003). The validity of the proposed estimation procedure in Section 3 only relies on the population structure for recurrent events occurring before the terminal event.

Assumption (1) implies that the occurrence of recurrent events follows a proportional intensity model, where the unobserved frailty Z_i inflates/deflates the intensity. Because of the memoryless property of a Poisson process, conditional on Z_i, the rate function equals the intensity function of the recurrent event process. Under (1) and $E[Z_i \mid \boldsymbol{W}_i,\boldsymbol{X}_i(\tau)] = 1$, the rate function of event occurrence at time t in a random population is given by $\lambda_0(t)\exp\left\{\boldsymbol{X}_i(t)'\beta+\boldsymbol{W}_i'\gamma\right\}$. Thus (1) implies the proportional rate model for recurrent event data studied by Lin et al. (2000) and many others. The proposed model also reduces to the semiparametric model studied in Wang et al. (2001) in the absence of time-dependent covariate $\boldsymbol{X}_i(t)$, and is in line with the model for case series data studied in Farrington and Whitaker (2006) in the absence of time-independent covariate $\boldsymbol{W}_i$.

3. Estimation Procedure

3.1 Estimation of β

We denote by m_i the number of recurrent events that occurred before time Y_i and $t_{i1}, \ldots, t_{im_i}$ the observed event times for subject i. For ease of notation, we use m_i and t_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m_i$, to denote either random variables or realized values. Let D_i denote the observed data of the ith subject, that is, $D_i = \{(Y_i,\boldsymbol{W}_i,\boldsymbol{X}_i(Y_i),m_i, (t_{i1}, \ldots, t_{im_i})\}$. Assume that $\{(Z_i,\boldsymbol{X}_i(\cdot),\boldsymbol{W}_i, Y_i,N_i(\cdot)); i = 1, \ldots, n\}$ are independent and identically distributed (iid), so that the $\{D_1, \ldots, D_n\}$ are also iid.

For estimation of the regression parameter β, an initial attempt might use the conditional technique used in Wang et al. (2001) to eliminate nuisance parameters from the likelihood. Under (1) the event times $(t_{i1}, \ldots, t_{im_i})$of the ith subject conditional on $(Z_i, Y_i,m_i,\boldsymbol{W}_i,\boldsymbol{X}(Y_i))$ are order statistics of a set of iid random variables with the density function

$$\frac{Z_i\lambda_0(t)\ \exp\left\{\boldsymbol{X}_i(t)'\beta+\boldsymbol{W}_i'\gamma\right\}}{\int_0^{Y_i} Z_i\lambda_0(u)\ \exp\left\{\boldsymbol{X}_i(t)'\beta+\boldsymbol{W}_i'\gamma\right\}du},\qquad 0\leqslant t\leqslant Y_i.$$

Note that both the unobserved frailty Z_i and the time-independent covariates $\boldsymbol{W}_i$ are eliminated from the conditional density function. The conditional likelihood based on all subjects is proportional to

$$\prod_{i=1}^{n}\prod_{j=1}^{m_i}\frac{\lambda_0(t)\ \exp\left\{\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\}}{\int_0^{Y_i}\lambda_0(u)\ \exp\left\{\boldsymbol{X}_i(u)'\beta\right\}du}=\prod_{i=1}^{n}\prod_{j=1}^{m_i}\frac{d\pi_i\left(t_{ij}\right)}{\pi_i(Y_i)},\qquad (2)$$

where $d\pi_i(t)$ is the shape function of the recurrent event process given by

$$d\pi_i(t)=\frac{\lambda_0(t)\ \exp\left\{\boldsymbol{X}_i(t)'\beta\right\}}{\int_0^{\tau}\lambda_0(u)\ \exp\left\{\boldsymbol{X}_i(u)'\beta\right\}du}I(0\leqslant t\leqslant\tau),$$

and $\pi_i(t)=\int_0^t d\pi_i(u)$. Note that π_i defines a proper distribution function with $\pi_i(\tau) = 1$.

When the recurrent event model only includes time-independent covariates, the conditional density function further reduces to $\lambda_0(t)/\Lambda_0(Y_i)$, which is in the form of a truncated density function. It is easy to see that for this special case the conditional likelihood depends only on λ_0 and is computationally equivalent to the nonparametric likelihood of independently right-truncated data. Hence the reduced conditional likelihood is maximized by the product-limit estimator for independently right-truncated data (Wang, Jewell, and Tsai 1986). In the presence of time-varying covariates, however, the conditional likelihood (2) involves both the parametric component $\boldsymbol{X}_i(t)'\beta$ and the nonparametric component λ_0. Maximizing the conditional likelihood function is challenge because the integral in the denominator of the conditional likelihood does not have a closed form with λ_0 unspecified. Motivated by Liang and Qin (2000) and Kalbfleisch (1978), we propose an alternative estimation procedure for $\boldsymbol{\beta}$ that does not involve the nonparametric component λ_0, and hence has the advantage of computational convenience.

Because (2) is computationally equivalent to the semiparametric likelihood of a set of independently right-truncated random variables, we can reformulate the problem as estimating the regression parameter $\boldsymbol{\beta}$ using the data $\{t_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i\}$, where t_{ij} is an observed event time with the distribution function π_i and is subject to independent right truncation Y_i. The pairwise pseudolikelihood method considered by Liang and Qin (2000), however, can not be applied directly to truncated data: event times are not necessary comparable because they are subject to different truncation times. The observation of t_{ij} is subject to the constraint $t_{ij} \leqslant Y_i$, hence any two event times t_{ij} and t_{kl}, $i \neq k$, are comparable if t_{ij} belongs to the observation interval of t_{kl} and t_{kl} belongs to the observation interval of t_{ij}. These constraints amount to $t_{ij} \leqslant Y_i \wedge Y_k$ and $t_{kl} \leqslant Y_i \wedge Y_k$, where $\wedge$ denotes minimum.

For any two event times t_{ij} and t_{kl}, let $\delta_{ijkl} = 1$ if (t_{ij}, t_{kl}) is a comparable pair, and 0 otherwise. We condition on having observed the values $\{t_{ij}, t_{kl}\}$ for a given pair, but without knowing the order. We refer to this as conditioning on the order statistics of (t_{ij}, t_{kl}). The conditional distribution is degenerate at the observed values if (t_{ij}, t_{kl}) are not comparable. By conditioning on the order statistics of (t_{ij}, t_{kl}) and $\delta_{ijkl} = 1$, the pairwise pseudolikelihood of (t_{ij}, t_{kl}), $i < k$, is given by

$$\begin{aligned}&\frac{d\pi_i(t_{ij})d\pi_k(t_{kl})}{d\pi_i(t_{ij})d\pi_k(t_{kl})+d\pi_i(t_{kl})d\pi_k(t_{ij})}\\ &\quad=\frac{\exp\left[\left\{\boldsymbol{X}_i(t_{ij})+\boldsymbol{X}_k(t_{kl})\right\}'\beta\right]}{\exp\Big[\left\{\boldsymbol{X}_i(t_{ij})+\boldsymbol{X}_k(t_{kl})\right\}'\beta+\exp\Big[\left\{\boldsymbol{X}_i(t_{kl})+\boldsymbol{X}_k(t_{ij})\right\}'\beta\Big]}=\frac{1}{1+\exp\left\{\rho_{ik}(t_{ij},t_{kl})'\beta\right\}},\end{aligned}\tag{3}$$

where $\boldsymbol{\rho}_{ik}(t, u) = \boldsymbol{X}_i(u)+\boldsymbol{X}_k(t)-\boldsymbol{X}_i(t)-\boldsymbol{X}_k(u)$. Interestingly, the pairwise pseudolikelihood depends on the regression parameter $\boldsymbol{\beta}$ but not the nonparametric component λ_0. Hence $\boldsymbol{\beta}$ can be estimated by maximizing the pairwise pseudolikelihood

$$\prod_{i<k}\prod_{j=1}^{m_i}\prod_{l=1}^{m_k}\left[\frac{1}{1+\exp\left\{\rho_{ik}\left(t_{ij},t_{kl}\right)'\beta\right\}}\right]^{\delta_{ijkl}}.$$

The score function derived from the logged pairwise pseudolikelihood can be expressed as

$$\sum_{i<k}\int\int I\left(t\leqslant Y_{ik},u\leqslant Y_{ik}\right)\left[\frac{-\exp\left\{\rho_{ik}(t,u)'\beta\right\}}{1+\exp\left\{\rho_{ik}(t,u)'\beta\right\}}\rho_{ik}\left(t,u\right)\right]dN_i\left(t\right)dN_k\left(u\right),$$

where $Y_{ik} = Y_i \wedge Y_k$. Recall that D_i and D_k denote the observed data of the ith and the kth subject. Define the function

$$\boldsymbol{h}\left(d_i,D_k;\beta\right)=\int\int I\left(t\leqslant Y_{ik},u\leqslant Y_{ik}\right)\left[\frac{-\exp\left\{\rho_{ik}(t,u)'\beta\right\}}{1+\exp\left\{\rho_{ij}(t,u)'\beta\right\}}\rho_{ik}\left(t,u\right)\right]dN_i\left(t\right)dN_k\left(u\right),$$

and denote the score function by

$$\boldsymbol{S}_p\left(\beta\right)=\frac{1}{\binom{n}{2}}\sum_{i<k}\boldsymbol{h}\left(D_i,D_k;\beta\right).$$

It is easy to see that $\boldsymbol{h}$ is permutation symmetric in its arguments and $\boldsymbol{S}_p$ is a U-statistic with the kernel $\boldsymbol{h}(\cdot,\cdot)$. If $\boldsymbol{\beta}$ is the true parameter value, it can be shown that the score function $\boldsymbol{S}_p(\boldsymbol{\beta}) = \mathbf{0}$. Applying the projection method developed by Hoeffding (1948) and under the assumption that $\boldsymbol{X}(t)$ is bounded by M, we can show that score function $\sqrt{n}\boldsymbol{S}_p\left(\beta\right)$ converges to a normal distribution with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{V}_1 = 4E\{\boldsymbol{h}(D_1, D_2; \boldsymbol{\beta})\,\boldsymbol{h}(D_1, D_3; \boldsymbol{\beta})'\}$. We then study the asymptotic properties of $\hat{\boldsymbol{\beta}}$ using delta method. The large sample properties of $\hat{\boldsymbol{\beta}}$ are stated in Theorem 1, with proofs given in the appendix.

Theorem 1—*Assume that* $\boldsymbol{X}(t)$ *is bounded by* M *and* $E[N(\tau)] < \infty$. *Let* $\hat{\boldsymbol{\beta}}$ *be the solution of* $\boldsymbol{S}_p(\boldsymbol{\beta}) = 0$. Then $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Furthermore, $\sqrt{n}\left(\widehat{\beta}-\beta\right)\rightarrow N\left(0,\boldsymbol{V}\right)$, *where* $\boldsymbol{V}=\boldsymbol{V}_2^{-1}\boldsymbol{V}_1\boldsymbol{V}_2^{-1}$ *with* $\boldsymbol{V}_1 = 4E\{\boldsymbol{h}(D_1, D_2; \boldsymbol{\beta})\boldsymbol{h}(D_1, D_3; \boldsymbol{\beta})'\}$ *and* $\boldsymbol{V}_2 = -E\{\partial\boldsymbol{h}(D_1, D_2; \boldsymbol{\beta})/\partial\boldsymbol{\beta}\}$.

Note that the variance covariance matrix $\boldsymbol{V}$ can be estimated by $\widehat{\boldsymbol{V}}_2^{-1}\widehat{\boldsymbol{V}}_1\widehat{\boldsymbol{V}}_2^{-1}$, where

$$\widehat{\boldsymbol{V}}_1=\frac{4}{n}\sum_{i=1}^{n}\frac{1}{\binom{n-1}{2}}\sum_{i<j<k}\boldsymbol{h}\left(D_i,D_j;\widehat{\beta}\right)\boldsymbol{h}\left(D_i,D_k;\widehat{\beta}\right)',$$

and

$$\widehat{\boldsymbol{V}}_2 = -\frac{1}{\binom{n}{2}} \sum_{i<k} \frac{\partial \boldsymbol{h}\left(D_i, D_k;\widehat{\beta}\right)}{\partial \beta}.$$

Also note that based on the score function $\boldsymbol{S}_p$, a test statistic for testing the hypothesis $\boldsymbol{\beta} = \mathbf{0}$ can be formulated based on

$$\frac{1}{\binom{n}{2}} \sum_{j<k} N_i(Y_{ik}) N_k(Y_{ik}) \int_0^{Y_{ik}} \{\boldsymbol{X}_i(u) - \boldsymbol{X}_k(u)\} \left\{\frac{dN_i(u)}{N_i(u)} - \frac{dN_k(u)}{N_k(u)}\right\}.$$

Interestingly, this test statistic for the effects of time-dependent covariates does not require information about time-independent covariates, and hence is useful for checking proportionality assumption in the usual proportional rate model.

3.2 Estimation of Λ_0 and γ

By definition $\pi_i(t)$ can be considered as the distribution function of a biased sample from the distribution $F(t) = \Lambda_0(t)/\Lambda_0(\tau)$, where the observations are sampled with a probability proportional to $\exp\{\boldsymbol{X}_i(t)'\boldsymbol{\beta}\}$. It is easy to see that, under the assumption (1), the conditional likelihood (2) is computationally equivalent to the likelihood of a set of independent random variables, where the data are a biased sample from distribution function $F(t)$ with sampling weight proportional to $\exp\{\boldsymbol{X}_i(i)'\boldsymbol{\beta}\}$ and are right truncated by Y_i. Thus event times in the risk set are observed with different sampling probabilities, where the probabilities are proportional to $\exp\{\boldsymbol{X}_i(t)'\boldsymbol{\beta}\}$. If $\boldsymbol{\beta}$ is known, the probability structure of the risk set can be recovered by using the inverse probability weighting technique. Following the spirit of the Breslow estimator of the cumulative hazard in the Cox model, we modify the truncation product-limit estimator (Wang et al. 1986) as follows to estimate F by assigning each event in the risk set a weight that is proportional to the inverse of the sampling weight function:

$$\prod_{s_l>t} \left\{1 - \frac{d_l(\beta)}{R_1(\beta)}\right\},$$

where $d_l(\beta) = \frac{1}{n}\Sigma_{i=1}^{n}\Sigma_{j=1}^{m_i} I\left(t_{ij}=s_l\right) \exp\left\{-\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\}$ and $R_l(\beta) = \frac{1}{n}\Sigma_{i=1}^{n}\Sigma_{j=1}^{m_i} I\left(t_{ij} \leqslant s_l \leqslant Y_i\right) \exp\left\{-\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\}$. Note that $\boldsymbol{\beta} = 0$ implies that the assigned weight is a unit weight and the proposed estimator reduces to the product-limit estimator that maximizes the nonparametric likelihood function for truncated data. By replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, we estimate F with

$$\widehat{F}\left(t;\widehat{\beta}\right) = \prod_{s_l>t} \left\{1 - \frac{d_1\left(\widehat{\beta}\right)}{R_1\left(\widehat{\beta}\right)}\right\}.$$

The large sample properties of $\hat{F}(t)$ are stated in Lemma 1, with proofs given in Appendix.

Lemma 1—*Assume that (a)* $\Lambda_0(\tau) > 0$, *(b)* $Pr(Y > \tau, Z > 0) > 0$, *and (c)* $G(u) = E\{ZI(Y \geqslant u)\exp(\boldsymbol{W}'\boldsymbol{\gamma})\}$ *is a continuous function for* $u \in [0, \tau]$. *For* $\inf\{y : \Lambda_0(y) > 0\} < t \leqslant \tau$, $\sqrt{n}\left\{\widehat{F}(t) - F(t)\right\}$ converges weakly to a normal distribution with mean 0 *and variance* $4F(t)^2E\{\kappa(D_1, D_2; t, \boldsymbol{\beta})\kappa(D_1, D_3; t, \boldsymbol{\beta})\}$, *where* κ *is defined in Appendix.*

To estimate the regression parameters of time-independent covariates, we note that, conditioning on $\{Z_i, Y_i, \boldsymbol{W}_i, \boldsymbol{X}_i(Y_i)\}$, the expected value of m_i is given by

$$E\{m_i|Z_i, Y_i, \boldsymbol{W}_i, \boldsymbol{x}_i(Y_i)\} = Z_i\Lambda_0(\tau)\exp\left\{\boldsymbol{W}_i'\gamma\right\}\int I(Y_i \geqslant u)\exp\left\{\boldsymbol{X}_i(u)'\beta\right\}dF(u).$$

Thus, following the assumption $E[Z_i \mid \boldsymbol{W}_i, \boldsymbol{X}_i(\tau)] = 1$ and by double expectation, we have

$$E\left\{\frac{m_i}{\int_0^{Y_i}\exp\left\{\boldsymbol{X}_i(u)'\beta\right\}dF(u)}|\boldsymbol{W}_i\right\} = \Lambda_0(\tau)\exp\boldsymbol{W}_i'\gamma).$$

We propose to estimate $\boldsymbol{\gamma}$ and $\Lambda_0(\tau)$ by solving the following estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i^*\left[\frac{m_i}{\int_0^{Y_i}\exp\left\{\boldsymbol{X}_i(u)'\widehat{\beta}\right\}d\widehat{F}(u)} - \exp\left\{\left(\boldsymbol{W}_i^*\right)'\eta\right\}\right] = 0, \tag{4}$$

where $\boldsymbol{W}_i^* = \left(1, \boldsymbol{W}_i'\right)'$ and $\boldsymbol{\eta} = (\ln\Lambda_0(\tau), \boldsymbol{\gamma}')'$. Let $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1\ \hat{\boldsymbol{\gamma}}')'$ be the root of the estimating equations. Then $\Lambda_0(t)$ can be estimated by $\hat{\Lambda}_0(t) = \hat{F}(t) \times \exp(\hat{\eta}_1)$.

Following Theorem 1 and Lemma 1, we can establish the asymptotic properties for $\boldsymbol{\gamma}$ and $\Lambda_0(t)$ stated in Theorem 2, with proofs given in the appendix.

Theorem 2—Assume that the conditions specified in Theorem 1 and Lemma 1 hold, $\sqrt{n}(\widehat{\gamma} - \gamma)$ converges weakly to a multivariate normal distribution with mean 0 and *variance-covariance matrix* $E(-\partial\boldsymbol{\xi}/\partial\boldsymbol{\gamma})^{-1}\Sigma\{E(-\partial\boldsymbol{\xi}/\partial\boldsymbol{\gamma})'\}^{-1}$, *provided* $E(-\partial\boldsymbol{\xi}/\partial\boldsymbol{\gamma})^{-1}$ *exits, where* $\boldsymbol{\xi}$ *and* Σ *are defined in Appendix. Moreover, when* $n \to \infty$ *and* $\inf\{y : \Lambda_0(y) > 0\} < t \leqslant \tau$, $\sqrt{n}\left\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\right\}$ converges weakly to a normal distribution with mean 0 *and variance*

$$4F(t)^2\exp(2\eta_1)E\left[\{f(D_1, D_2) + \kappa(D_1, D_2; t, \beta)\}\{f(D_1, D_3) + \kappa(D_1, D_3; t, \beta)\}\right],$$

where $f(D_i, D_k)$ *is the first entry of the vector function* $E(-\partial\boldsymbol{\xi}/\partial\boldsymbol{\gamma})^{-1}\xi(D_i, D_k)$.

4. Simulation Studies

We conduct Monte Carlo simulation studies to evaluate the finite-sample properties of the proposed estimator. For all simulations, we generate 1,000 simulated datasets, each with n = 100 and n = 400 independent subjects. For subject i, the frailty Z_i is generated from a gamma distribution with unit mean and variance 4, and the time-independent covariate W_i, which corresponds to treatment assignment, is generated taking values 0 or 1 with probability 0.5. We assume that the time-dependent covariate $X_i(t)$ takes the form $X_i\log(t)$, where X_i has a uniform [0, 1] distribution. We generate recurrent event times from model (1), where the subject's underlying recurrent event process is a nonhomogeneous Poisson process with intensity function $Z_i\lambda_0(t)\ \exp\left\{X_i(t)'\beta+W_i'\gamma\right\}$. To examine the performance of proposed estimators under different choices of (β, γ) and λ_0, we consider combinations corresponding to (β, γ) = (0, 0), (0.3, 0), (0, 0.3) and (0.3, 0.3), and $\lambda_0(t)$ = 1/2 and $\sqrt{t}/4$. Under each scenario we generate a failure event time Y_{i1}^* from an exponential distribution with mean 10 for subjects in the treatment arm ($W_i = 1$), and from exponential distributions with mean $6Z_i$ +4 and $10X_i + 5$ for subjects with $X_i > 0.5$ and $X_i <= 0.5$, respectively, in the control arm (W_i = 0). Let $Y_{i0}^*=10$ be the time of end of the study, and $Y_i=\min\left(Y_{i0}^*,Y_{i1}^*\right)$ be the censoring time. Note that the censoring time is conditionally independent of $N_i(\cdot)$ given Z_i, $\boldsymbol{X}_i(\tau)$, and W_i, but not unconditionally.

We compare the results from the proposed estimation procedure to that from the Lin et al. (2000), termed as LWYY, method. The empirical bias, empirical standard error and the mean square error of the estimates based on 1,000 samples are shown in Table 1. Figures 1 and 2 show the estimates and the pointwise 95% confidence intervals of the baseline cumulative intensity function. As summarized in the table, the average length of follow-up period is approximately 5.8, and the average number of observed recurrent events ranges from 3.3 to 4.7 in these simulation studies. The estimators of β and γ from the proposed method perform well in that the biases in the estimates of β and γ are small, and the averages of $\Lambda_0(t)$ are almost indistinguishable from the true curve. On the other hand, using the LWYY method, which requires independent censoring to draw valid inference, results in biased estimation as well as greater mean squared errors of the covariate effects. Under the specified conditions in our simulations, simulated control subjects ($W_i = 0$) with higher frailty values tend to have a longer observation period. Thus risk sets are more likely to consist of sicker subjects at later time points. As a result, the LWYY method that compares subjects within risk sets underestimates the difference between treatment group and control group in the intensity of recurrent events. Note that, compared to the LWYY method, the relative efficiency of the proposed estimator depends on the degree of association between the censoring mechanism and the recurrent event process. Under the independent censoring assumption, the proposed method is expected to be less efficient than the LWYY method since the estimation steps involves procedures to handle the informative censoring which results in loss of estimation efficiency.

4.1 Analysis of ALIVE Study

We use our method to analyze data from the AIDS Link to Intravenous Experience (ALIVE) study (Vlahov et al. 1991). During the period February 1988 through March 1989, a total of 2,946 active injection drug users in the city of Baltimore, Maryland were recruited into the ALIVE study through extensive community outreach efforts. Participants underwent a screening interview in which data on sociodemographic factors, history of drug use and sexual practice over the previous 10 years were obtained. Information on HIV testing was collected at follow-up visits scheduled at 6-month intervals. The date of seroconversion was estimated as the midpoint between the dates of the first positive HIV test and the date of the last negative HIV test, if available. We analyze hospital admissions recorded between July 16, 1993 and December 31, 1997 from 1,896 intravenous drug users who had at least one follow-up visit at the study clinic during the same period of time. Among these participants, there were 1,412 (74%) males and 1,781 (95%) Africa Americans. The number of hospital admissions ranges from 0 to 19, averaging 3.5 per subject, whereas 1,026 (54%) subjects had no hospitalization record. A total of 244 (13%) participants died during the four-and-a-half study period, among them 200 were male (82%) and 233 (95%) were black.

We apply the proposed method to study patient's risk of hospitalization since time 0 (July 16, 1993). The covariates of interest in our analysis include patient's HIV status ($X_i(t) = 1$ if

HIV-positive at time t, 0 if HIV-negative), gender (W_{1i} = 1 if male, 0 if female), and race (W_{2i} = 1 if African American, 0 if non-African American). Censoring time Y_i is defined as the time of the last visit at the study clinic or date of death recorded by the study before December 31, 1997, and τ denotes the maximum time of Y_i's. Let β be the regression coefficient for HIV status ($X_i(t)$), and γ_1 and γ_2 be the coefficients for gender (W_{i1}) and race (W_{i2}). To estimate the 95% confidence intervals of $\hat{\beta}$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\Lambda}_0(t)$, we adopt a nonparametric bootstrap method for clustered data by repeatedly sampling 1,781 subjects with replacement, using subject as the sampling unit, from the AIDS cohort data. We repeat the resampling procedure 1,000 times and use the 2.5th and 97.5th percentiles of the empirical distribution based on these 1,000 estimates as the 95% bootstrap confidence interval. The nonparametric bootstrap method is adopted because the variance estimates of the proposed estimation procedure are quite complicated. Also, moment estimators (asymptotic variance estimator is one of those) are less robust when outliers are present.

For model (1), the estimated $\hat{\beta}$ is 0.16 with 95% bootstrap confidence interval (−0.62, 0.99), and the estimated $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are −0.26 and −0.06, with corresponding 95% confidence intervals (−0.45, −0.08) and (−0.06, 0.19). The HIV-positive status is associated with 17% increase in the risk of hospitalization, but the difference is not statistically significant. African Americans have a insignificantly lower rate of hospitalization compared to non-African Americans, while males have a significantly lower risk (23% lower) than females. Figure 3 shows the estimated $\hat{\Lambda}_0(t)$ and the pointwise 95% bootstrap confidence intervals. The estimated baseline cumulative rate function is close to a straight line, suggesting that the risk of hospitalization is approximately constant over time. For comparison, we also apply the LWYY method to analyze the ALIVE data. The estimated $\hat{\beta}$ is 0.50 with 95% bootstrap confidence interval (−0.36, 0.67), and the estimated $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are −0.27 and −0.18, with corresponding 95% confidence intervals (−0.44, −0.10) and (−0.48, 0.19). The direction of covariate effects estimated in the LWYY model are consistent with the estimates under the proposed model.

5. Discussion

In this paper, we develop a flexible estimation procedure that does not require parametric assumptions about the distributions of the frailty variable and the censoring time. In particular, we apply the modified pairwise pseudolikelihood method (Liang and Qin 2000) based on all comparable pairs of event times to estimate $\boldsymbol{\beta}$. The proposed pseudolikelihood-based estimation procedure involves neither frailty nor the time-independent covariates, and hence is very useful for checking proportionality assumption in the widely used proportional rate models.

Naturally we can consider applying the pseudolikelihood method to three or more comparable event times, and expect a potential gain in efficiency. Due to comparability constraints, however, the number of comparable triples may not be much more than the number of comparable pairs. Hence the efficiency gain can be very limited, while the computation time may increase substantially. As an alternative to increase efficiency, we consider the disjoint time intervals $[0, Y_{(1)}]$, $(Y_{(1)}, Y_{(2)}]$, …, $(Y_{(n-1)}, Y_{(n)}]$, where $(Y_{(1)}, \ldots, Y_{(n)})$ are order statistics of the censoring times $\{Y_1, \ldots, Y_n\}$. Note that any event time is comparable with all other event times within the same time interval. Thus, we can formulate a pseudolikelihood for each disjoint interval by conditioning on the order statistics of all the event times within that interval, and the denominator of the pseudolikelihood can be approximated by drawing a subset of all possible permutations of event times within the interval. We then estimate $\boldsymbol{\beta}$ by combining pseudolikelihoods of disjoint time intervals.

As discussed in Section 3.2, the conditional likelihood (2) is computationally equivalent to the likelihood of a set of independent observations that are subject to two sources of bias: right truncation and biased sampling. In the literature, the product-limit estimator is used for inference under random truncation (Wang et al. 1986), while inverse probability weighting estimators are often used to correct for sampling bias (Horvitz and Thompson 1952). In our setting, however, both sources of bias are present simultaneously. To correct the bias in estimating the shape function of the recurrent event process, we propose an estimator that properly combines the inverse probability weighting technique with the product limit estimator by assigning each event time in the risk set a weight proportional to the inverse of its sampling probability. Interestingly, the estimation of the shape function does not require information about time-independent covariates and the unobserved frailty, and is reduced to the usual product limit estimator when $\boldsymbol{\beta} = 0$.

Note that the censoring time in our model formulation is treated as a nuisance. In many applications, especially when a failure event precludes the observation of recurrent events, study interests are placed on the joint inference of the recurrent event process and the failure times. Shared frailty models such as the fully parametric approach by Lancaster and Intrator (1998) and semiparametric approaches by Liu et al. (2004) and Huang and Wang (2004) have been used to study recurrent events and failure time data jointly. The first two models require a parametric assumption for the unobserved frailty distribution, and thus suffer from lack of model checking techniques; moreover, censoring mechanism other than the failure event of interest is required to be random for validating their inferential results. On the other hand, the Huang and Wang (2004) model is more robust in that the frailty distribution is left unspecified and censoring mechanism other than the failure event is allowed to be correlated with the recurrent event process and the failure times. Their estimation procedure, however, inherits the properties of the estimator studied by Wang et al. (2001), and hence can not handle time-dependent covariates. By properly combining the new methodology proposed in this paper and the “borrow-strength estimation procedure” studied in Huang and Wang (2004), we can easily extend their joint inference procedure to handel both time-dependent and time-independent covariates. The properties of the new estimation procedure will be studied elsewhere.

This article does not intend to develop formal model checking methods. Under informative censoring, model checking is expected to be a difficult task in general. we simply suggest a possible approach for validating model assumptions. A rigorous study will be done elsewhere. To check on the proportional intensity model assumption imposed by (1), we could replace Z_i with $\widehat{Z_i}=m_i/\left\{\int_0^{Y_i}\lambda_0(u)\ \exp\left\{\boldsymbol{X}_i(t)'\widehat{\beta}+\boldsymbol{W}_i'\widehat{\gamma}\right\}du\right\}$ to derive the Schoenfeld residuals (Schoenfeld 1982). If the assumption of proportional intensity model holds, the derived residuals are expected to randomly scattered around 0.

Acknowledgments

Mei-Cheng Wang's work was supported by National Institutes of Health grant R01 AI078835. The authors thank Drs. David Vlahov and Steffanie Strathdee for providing the data from the ALIVE study, which was supported by NIDA grants DA 04334 and DA 08009. Comments from Drs. Mike Proschan and Dean Follmann have improved the presentation of this paper.

Appendix

Proof of Theorem 1

Because $-\log[1 + \exp\{\boldsymbol{\rho}_{ik}(t_{ij}, t_{kl})'\boldsymbol{\beta}\}]$ is the log-likelihood of t_{ij} and t_{kl} conditional on $\{Z_i, m_i, Y_i, \boldsymbol{X}_i(Y_i), \boldsymbol{W}_i\}$, $\{Z_k, m_k, Y_k, \boldsymbol{X}_k(Y_k), \boldsymbol{W}_k\}$ and the order statistics of $\{t_{ij}, t_{ik}\}$, the pairwise pseudolikelihood (4) achieves its maximum at the true parameter value. By the conditional Kullback-Leibler information inequality (Andersen 1970), the maximum pairwise pseudolikelihood estimator $\hat{\boldsymbol{\beta}}$ is consistent.

For convenience, we denote $a^2 = aa'$ for any vector a. Applying Taylor expansion to $S_p(\boldsymbol{\beta})$, we have $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \{-\partial S_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}^{-1} S_p(\boldsymbol{\beta})+o_p(n^{-1/2})$. By noting that $E\{\partial\boldsymbol{h}(D_i, D_k; \boldsymbol{\beta})/\partial\boldsymbol{\beta}\}^2 \leqslant (4M)^4E\{N_i(\tau)^2N_k(\tau)^2\} < \infty$, it follows from the strong law of large numbers for U-statistics that $-\partial S_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ converges almost surely to $E\{-\partial S_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\} = E\{-\partial\boldsymbol{h}(D_1, D_2; \boldsymbol{\beta})/\partial\boldsymbol{\beta}\} = \boldsymbol{V}_2$.

Hence $\widehat{\beta}-\beta=V_2^{-1}\frac{1}{\binom{n}{2}}\Sigma_{i<k}\boldsymbol{h}\left(D_i, D_k;\beta\right)+o_p\left(n^{-1/2}\right)$. By the central limit theorem for the U-statistics (Serfling 1980, Chap 5), $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ converges weakly to the normal distribution with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{V}=\boldsymbol{V}_2^{-1}\boldsymbol{V}_1\boldsymbol{V}_2^{-1}$.

Proof of Lemma 2

Define the functions $Q(u)=\int_0^u G(v)\,d\Lambda_0(v)$ and $R(u) = G(u)\Lambda_0(u)$. It can be shown that n the empirical averages $\tilde{Q}(u;\beta)=\frac{1}{n}\Sigma_{i=1}^n\Sigma_{j=1}^{m_i}I\left(t_{ij}\leqslant u\right)\exp\left(\left\{-\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\}\right.$ and $\tilde{R}(u;\beta)=\frac{1}{n}\Sigma_{i=1}^n\Sigma_{j=1}^{m_i}I\left(t_{ij}\leqslant u\leqslant Y_i\right)\exp\left\{-\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\}$ are unbiased estimators of $Q(u)$and $R(u)$ if $\boldsymbol{\beta}$ is the true parameter value. Let $\hat{Q}(u) = \tilde{Q}(u; \hat{\boldsymbol{\beta}})$ and $\hat{R}(u) = \tilde{R}(u; \hat{\boldsymbol{\beta}})$. A Taylor series expansion of $\hat{Q}(u) = \tilde{Q}(u; \hat{\boldsymbol{\beta}}$ about $\boldsymbol{\beta}$ yields $\hat{Q}(u)-\tilde{Q}(u; \boldsymbol{\beta}) = V_{\tilde{Q}}(u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+o_p(n^{-1/2})$, where $V_{\tilde{Q}}(u) = E\{\partial\tilde{Q}(u; \boldsymbol{\beta})/\partial\boldsymbol{\beta}\}'$. Similarly, we have $\hat{R}(u) - \tilde{R}(u; \boldsymbol{\beta}) = V_{\tilde{R}}(u)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(n^{-1/2})$, where $V_{\tilde{R}}(u) = E[\partial\tilde{R}(u; \boldsymbol{\beta})/\partial\boldsymbol{\beta}]'$. The weak convergence of $\sqrt{n}\left\{\widehat{Q}(u)-Q(u)\right\}$ and $\sqrt{n}\left\{\widehat{R}(u)-R(u)\right\}$ follows directly from Slutsky's theorem and Theorem 1.

Further, it is easy to see that $\int_t^\tau R(u)^{-1}dQ(u)=-\ln F(t)$. Note that assumptions (a) and (b) implies that $R(\tau) > 0$. Thus for any constant $c^* > \inf\{y : \Lambda_0(y) > 0\}$, one has $R(u) > 0$ for $c^* \leqslant u < \tau$. As $n \to \infty$, $\hat{Q}(u)$ and $\hat{R}(u)$ converge almost surely to $Q(u)$ and $R(u)$ uniformly in $u \in [c^*, \tau]$. By approximation techniques for product-limit estimators and the inequality $0 < -\ln(1 - u^{-1}) - u^{-1} < u^{-1}(u - 1)^{-1}$, for $u > 1$, we can show that $-\ln\widehat{F}(t)-\int_t^\tau d\widehat{Q}(u)/\widehat{R}(u)\to 0$ almost surely for each $t \in [c^*, \tau]$, with the usual convention $0/0 = 0$. Hence,

$$\widehat{F}(t)=\exp\left\{-\int_t^\tau\frac{d\widehat{Q}(u)}{\widehat{R}(u)}\right\}+O_p\left(n^{-1/2}\right).$$

Because the mapping of $\int_t^\tau d\widehat{Q}(u)/\widehat{R}(u)$ from the two empirical processes, under mild regularity conditions, is compactly differentiable with respect to the supremum norm and the two processes converge weakly to their limits (see example 2.11.16 of van der Vaart and Wellner 1996), we apply the functional delta method to $\int_t^\tau d\widehat{Q}(u)/\widehat{R}(u)$ and establish its asymptotic representation

$$\begin{aligned}\int_t^\tau \frac{d\widehat{Q}(u)}{\widehat{R}(u)} - \int_t^\tau \frac{dQ(u)}{R(u)} &= \int_t^\tau \frac{d\left\{\tilde{Q}\left(u;\widehat{\beta}\right)-\tilde{Q}(u;\beta)\right\}}{R(u)} - \int_t^\tau \frac{\left\{\tilde{R}\left(u;\widehat{\beta}\right)-\tilde{R}(u;\beta)\right\}dQ(u)}{R(u)^2}\\ &\quad + \int_t^\tau \frac{d\tilde{Q}(u;\beta)}{R(u)} - \int_t^\tau \frac{\tilde{R}(u;\beta)}{R(u)^2} dQ(u) + o_p\left(n^{-1/2}\right)\\ &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i<k} \phi(D_i, D_k; t, \beta) + \frac{1}{n}\sum_{i=1}^{n} \psi(D_i; t, \beta) + o_p\left(n^{-1/2}\right),\end{aligned}$$

where $\phi(D_i, D_k; t, \beta) = \int_t^\tau \left\{R(u)^{-1} d\boldsymbol{V}_{\tilde{Q}}(u) - \boldsymbol{V}_{\tilde{R}}(u) R(u)^{-2} dQ(u)\right\} \boldsymbol{V}_2^{-1} \boldsymbol{h}(D_i, D_k; \beta)$ and $\psi(D_i; t, \beta) = \Sigma_{j=1}^{m_i} I\left(t<t_{ij}\leqslant\tau\right) R\left(t_{ij}\right)^{-1} - \int_t^\tau I\left(t_{ij}\leqslant u\leqslant Y_i\right) \exp\left\{-\boldsymbol{X}_i\left(t_{ij}\right)'\beta\right\} R(u)^{-2} dQ(u)$. Let $\kappa(D_i, D_k; t, \boldsymbol{\beta}) = \boldsymbol{\phi}(D_i, D_k; t, \boldsymbol{\beta}) + \{\psi(D_i; t, \boldsymbol{\beta}) + \psi(D_k; t, \boldsymbol{\beta})\}/2$. It can be verified that $\begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \Sigma_{i<k} \kappa(D_i, D_k; t, \beta)$ is a U-statistic with $E\{\kappa(D_1, D_2; t, \boldsymbol{\beta})^2\} < \infty$. Hence, for fixed t, $\sqrt{n}\left\{\int_t^\tau d\widehat{Q}(u)/\widehat{R}(u) - \int_t^\tau dQ(u)/R(u)\right\}$ converges weakly to a normal distribution with zero mean and variance $4E\{\kappa(D_1, D_2; t, \boldsymbol{\beta})\kappa(D_1, D_3; t, \boldsymbol{\beta})\}$ by the central limit theorem for U-statistics. Further, we have $\widehat{F}(t) - F(t) = \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \Sigma_{i<k} \kappa(D_i, D_k; t, \beta) F(t) + o_p\left(n^{-1/2}\right)$, and, for fixed t, $\sqrt{n}\left\{\widehat{F}(t) - F(t)\right\}$ converges weakly to a normal distribution with mean 0 and variance $4F(t)^2 E\{\kappa(D_1, D_2; t, \boldsymbol{\beta})\kappa(D_1, D_3; t, \boldsymbol{\beta})\}$.

Proof of Theorem 2

Let H bet the joint probability measure of $(\boldsymbol{W}^*, m, \boldsymbol{X}, Y)$. Arguing as in the proof of Theorem 1 in Wang et al. (2001), we show that the left-hand side of (4) can be expressed as $\begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \Sigma_{i<k} \xi(D_i, D_k) + o_p(1)$ where

$$\begin{aligned}\xi(D_i, D_k) &= \int \frac{-\boldsymbol{w}^* m}{\left[\int_0^y \exp\left\{\boldsymbol{x}(u)'\beta\right\} dF(u)\right]^2} \left[\int_0^y \boldsymbol{x}(u)' \boldsymbol{V}_2^{-1} \boldsymbol{h}(D_1, D_k; \beta) \exp\left\{\boldsymbol{x}(u)'\beta\right\} dF(u)\right.\\ &\qquad\qquad \left. + \int_0^y \exp\left\{\boldsymbol{x}(u)'\beta\right\} d\{\kappa(D_i, D_k; u, \beta) F(u)\}\right] dH(\boldsymbol{w}^*, m, \boldsymbol{x}, y)\\ &\quad + \frac{1}{2}\boldsymbol{W}_i^* \left\{\frac{m_i}{\int_0^{Y_i} \exp\left\{\boldsymbol{X}_i(u)'\beta\right\} dF(u)} - \exp\left\{ \left(\boldsymbol{W}_i^*\right)'\eta\right\}\right\}\\ &\quad + \frac{1}{2}\boldsymbol{W}_k^* \left\{\frac{m_k}{\int_0^{Y_k} \exp\left\{\boldsymbol{X}_k(u)'\beta\right\} dF(u)} - \exp\left\{\left(\boldsymbol{W}_k^*\right)'\eta\right\}\right\}.\end{aligned}$$

It can be verified that $\begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \Sigma_{i<k} \kappa(D_i, D_k; t, \beta)$ is a U-statistic with $E\{\xi(D_i, D_k)\} = 0$ and $E\{\boldsymbol{\xi}(D_i, D_k)^2\} < \infty$. Applying Taylor expansion to the estimating equation (4) gives $\sqrt{n}\left(\widehat{\gamma} - \gamma\right) = \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} E(-\partial\xi/\partial\gamma)^{-1} \Sigma_{i<k} \xi(D_i, D_k) + o_p(1)$. Hence it follows the central limit theorem for U-statistics that $\sqrt{n}\left(\widehat{\gamma} - \gamma\right)$ converges weakly to a multivariate normal distribution with mean 0 and variance-covariance matrix $E(-\partial\xi/\partial\gamma)^{-1}\Sigma\{E(-\partial\xi/\partial\gamma)'\}^{-1}$, where $\Sigma = 4E\{\xi(D_1, D_2)\xi(D_1, D_3)'\}$.

To study the asymptotic normality of $\sqrt{n}\left\{\widehat{\Lambda}_0(t)-\Lambda_0(t)\right\}$, we write

$$\begin{aligned}\sqrt{n}\left\{\widehat{\Lambda}_0(t)-\Lambda_0(t)\right\} &= \sqrt{n}F(t)\,e^{\eta_1}\left(\widehat{\eta}_1-\eta_1\right)+\sqrt{ne}^{\eta_1}\left\{\widehat{F}(t)-F(t)\right\}+o_p(1)\\ &= \sqrt{ne}^{\eta_1}F(t)\,\frac{1}{\binom{n}{2}}\sum_{i<k}\left\{f(D_i,D_k)+\kappa(D_i,D_k;t,\beta)\right\}+o_p(1),\end{aligned}$$

where $f(D_i, D_k)$ is the first entry of the vector function $E(-\partial\xi/\partial\gamma)^{-1}\xi(D_i, D_k)$. Hence Theorem 2 follows from the central limit theorem for U-statistics.

References

Andersen EB. Asymptotic Properties of Conditional Maximum-likelihood Estimators. Journal of the Royal Statistical Society, Series B. 1970; 32:283–301.

Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Annals of Statistics. 1982; 10:1100–1120.

Chang S-H, Wang M-C. Conditional regression analysis for recurrence time data. Journal of the American Statistical Association. 1999; 94:1221–1230.

Farrington CP, Whitaker HJ. Semiparametric analysis of case series data. Journal of the Royal Statistical Society, Series C. 2006; 55:553–594.

Ghosh D. Accelerated rates regression models for recurrent failure data. Lifetime Data Analysis. 2004; 10:247–261. [PubMed: 15456106]

Ghosh D, Lin DY. Semiparametric analysis of recurrent events in the presence of dependent censoring. Biometrics. 2003; 59:877–885. [PubMed: 14969466]

Huang C-Y, Wang M-C. Joint modeling and estimation for recurrent event processes and failure time data. Journal of the American Statistical Association. 2004; 99:1153–1165.

Huang C-Y, Wang M-C, Zhang Y. Analysing panel count data with informative observation times. Biometrika. 2006; 93:763–775.

Huang X, Liu L. A joint frailty model for survival time and gap times between recurrent events. Biometrics. 2007; 63:389–397. [PubMed: 17688491]

Hoeffding W. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statististics. 1948; 19:293–325.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.

Kalbfleisch JD. Likelihood methods and nonparametric tests. Journal of the American Statistical Association. 1978; 73:167–170.

Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV-positive patients. Journal of the American Statistical Association. 1998; 93:46–53.

Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. Technometrics. 1995; 37:158–168.

Liang K-Y, Qin J. Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. Journal of the Royal Statistical Society, Series B. 2000; 62:773–786.

Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society, Series B. 2000; 62:711–730.

Liu L, Wolfe RA, Huang XL. Shared frailty models for recurrent events and a terminal event. Biometrics. 2004; 60:747–756. [PubMed: 15339298]

Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. Journal of the American Statistical Association. 1993; 88:811–820.

Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981; 68:373–379.

Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. Biostatistics. 2007; 8:708–721. [PubMed: 17267392]

Schaubel DE, Zeng D, Cai J. A semiparametric additive rates model for recurrent event data. Lifetime Data Analysis. 2006; 12

Serfling, RJ. Approximation Theorems of Mathematical Statistics. Wiley; New York: 1980.

Sun L, Su B. A class of accelerated means regression models for recurrent event data. Lifetime Data Analysis. 2008; 14:357–375. [PubMed: 18516715]

Sun J, Tong X, He Xin. Regression analysis of panel count data with dependent observation times. Biometrics. 2007; 63:1053–1059. [PubMed: 18078478]

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. Springer-Verlag; New York: 1996.

Vlahov D, Anthony JC, Muñoz A, Margolick J, Nelson KE, Celentano DD, Solomon L, Polk BF. The ALIVE study, a longitudinal study of HIV-1 infection in intravenous drug users: description of methods and characteristics of participants. The Journal of Drug Issues. 1991; 21:759–776.

Wang M-C, Jewell NP, Tsai W-Y. Asymptotic properties of the product limit estimate under random truncation. Annals of Statistics. 1986; 14:1597–1650.

Wang M-C, Qin J, Chiang C-T. Analyzing recurrent event data with informative censoring. Journal of the American Statistical Association. 2001; 96:1057–1065.

Ye Y, Kalbfleisch JD, Schaubel ES. Semiparametric analysis of correlated recurrent and terminal events. Biometrics. 2007; 63:78–87. [PubMed: 17447932]

Zeng D, Lin DY. Efficient estimation of semiparametric transformation models for counting processes. Biometrika. 2006; 93:627–640.

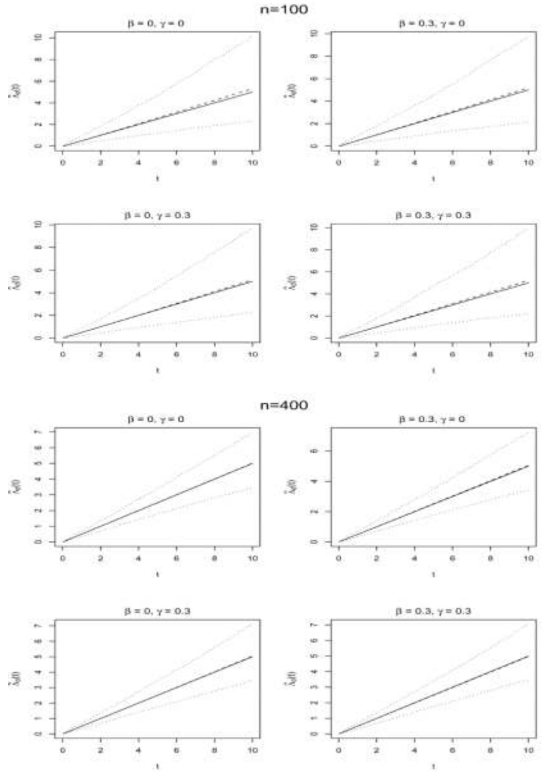

Figure 1.

Plots of estimated $\hat{\Lambda}_0(t)$ with pointwise 95% confidence intervals for Scenario I: $\Lambda_0(t) = t/2$ (—: true curve, - - - : empirical average, ⋯: pointwise 95% confidence interval).

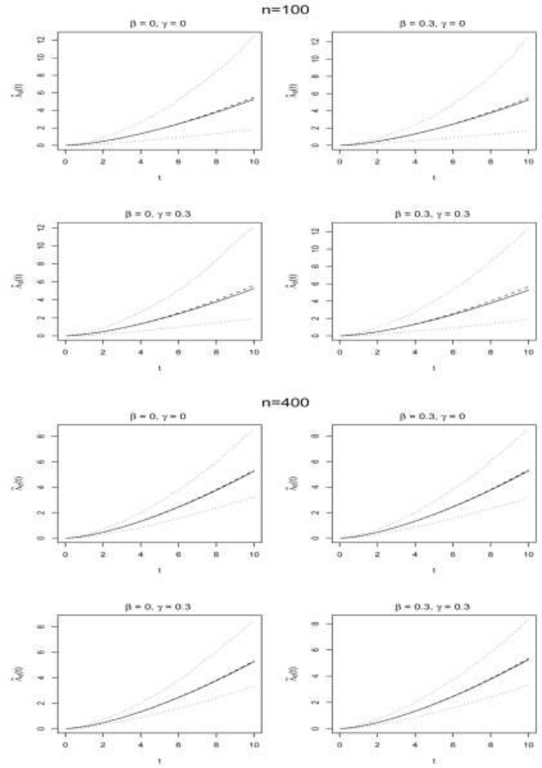

Figure 2.

Plots of estimated $\hat{\Lambda}_0(t)$ with pointwise 95% confidence intervals for Scenario II: $\Lambda_0(t) = t^{3/2}/6$ (—: truecurve, --- : empirical average, ⋯: pointwise 95% confidence interval).

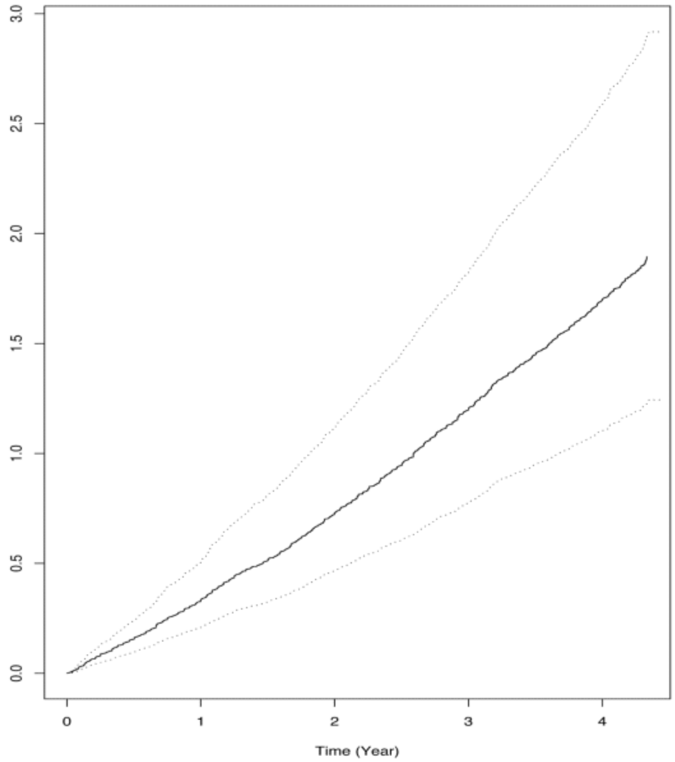

Figure 3.
Plot of $\hat{\Lambda}_0(t)$ for the ALIVE Cohort Data, With Pointwise 95% Bootstrap Confidence Intervals (—: proposed estimate, ···: pointwise 95% confidence interval).

Table 1

Summary of the simulation study

			Proposed Method						LWYY Method					
n	***(β, γ)***	***E*[*m*$_i$]**	***B*$_β$**	***ESE*$_β$**	***MSE*$_β$**	***B*$_γ$**	***ESE*$_γ$**	***MSE*$_γ$**	***B*$_β$**	***ESE*$_β$**	***MSE*$_β$**	***B*$_γ$**	***ESE*$_γ$**	***MSE*$_γ$**
Scenario I: $\lambda_0(t) = 1/2$														
100	(0, 0)	3.3	−0.011	0.238	0.057	−0.016	0.475	0.226	0.267	0.452	0.275	−0.258	0.487	0.304
	(0.3, 0)	4.1	0.021	0.259	0.068	−0.026	0.500	0.250	0.240	0.479	0.287	−0.299	0.512	0.352
	(0, 0.3)	3.9	0.005	0.224	0.050	0.005	0.466	0.217	0.220	0.441	0.243	−0.224	0.478	0.279
	(0.3, 0.3)	4.6	0.000	0.234	0.055	−0.003	0.451	0.204	0.216	0.473	0.270	−0.278	0.474	0.302
400	(0, 0)	3.3	0.005	0.107	0.011	0.007	0.232	0.054	0.244	0.219	0.107	−0.235	0.224	0.105
	(0.3, 0)	4.1	0.005	0.111	0.012	0.004	0.233	0.054	0.256	0.241	0.124	−0.292	0.236	0.141
	(0, 0.3)	3.9	−0.003	0.098	0.010	0.003	0.230	0.053	0.193	0.226	0.088	−0.238	0.237	0.113
	(0.3, 0.3)	4.6	0.006	0.105	0.011	0.013	0.235	0.055	0.222	0.243	0.108	−0.28	0.247	0.139
Scenario II: $\lambda_0(t) = \sqrt{t} / 4$														
100	(0, 0)	3.1	0.018	0.378	0.143	0.016	0.546	0.298	0.243	0.527	0.337	−0.294	0.528	0.365
	(0.3, 0)	3.8	0.017	0.370	0.137	0.014	0.506	0.256	0.270	0.518	0.341	−0.366	0.510	0.394
	(0, 0.3)	3.7	−0.004	0.344	0.118	−0.005	0.527	0.278	0.241	0.525	0.333	−0.291	0.516	0.351
	(0.3, 0.3)	4.7	−0.002	0.348	0.121	−0.028	0.534	0.286	0.233	0.541	0.347	−0.366	0.537	0.423
400	(0, 0)	3.1	0.008	0.161	0.026	−0.013	0.285	0.081	0.268	0.245	0.132	−0.306	0.248	0.155
	(0.3, 0)	3.8	0.008	0.164	0.027	−0.009	0.287	0.082	0.271	0.261	0.142	−0.354	0.263	0.195
	(0, 0.3)	3.7	0.002	0.149	0.022	0.002	0.282	0.079	0.230	0.264	0.122	−0.302	0.248	0.152
	(0.3, 0.3)	4.7	0.005	0.145	0.021	−0.002	0.270	0.073	0.224	0.260	0.118	−0.359	0.246	0.190

NOTE: $B\beta$ and $B\gamma$ are the empirical biases of $\hat{\beta}$ and $\hat{\gamma}$; $ESE\beta$ and $ESE\gamma$ are the empirical standard errors of $\hat{\beta}$ and $\hat{\gamma}$;$MSE\beta$ and $MSE\gamma$ are the mean squared errors of $\hat{\beta}$ and $\hat{\gamma}$.