Research Article

Statistics in Medicine

Received 9 February 2010, Accepted 2 May 2011 Published online 12 July 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4297

A joint model of recurrent events and a terminal event with a nonparametric covariate function

Zhangsheng Yu[a*†] and Lei Liu[b]

We extend the shared frailty model of recurrent events and a dependent terminal event to allow for a nonparametric covariate function. We include a Gaussian random effect (frailty) in the intensity functions of both the recurrent and terminal events to capture correlation between the two processes. We employ the penalized cubic spline method to describe the nonparametric covariate function in the recurrent events model. We use Laplace approximation to evaluate the marginal penalized partial likelihood without a closed form. We also propose the variance estimates for regression coefficients. Numerical analysis results show that the proposed estimates perform well for both the nonparametric and parametric components. We apply this method to analyze the hospitalization rate of patients with heart failure in the presence of death.

Keywords: proportional hazards model; counting process; informative censoring; smoothing parameter; survival analysis

1. Introduction

Recurrent events are often encountered in longitudinal studies [1]. In many situations, we observe a terminal event accompanying the recurrent events. For example, patients may experience recurrent hospitalizations that are terminated by death. Furthermore, these two processes are likely correlated: patients with poorer health tend to visit the hospital more frequently; they are also at a higher risk of death. This dependence entails the joint analysis of recurrent and terminal events.

It is of interest to many researchers to account for the dependent terminal event in the analysis of recurrent events data. Lancaster and Intrator [2] proposed a shared frailty model in the joint analysis of recurrent and terminal events. Cook and Lawless [1, Sections 6.6 and 6.8] provided a comprehensive review. More recently related works include [3–7]. However, all previously mentioned analyses assume linear regression coefficients in log intensity functions. In practice, some of the covariate effects could be nonlinear. It is thus of interest to capture such nonlinear patterns to describe the covariate effects more accurately. We will use spline methods for the nonparametric function estimation. For independent survival data, many researchers estimated the nonparametric covariate functions by the smoothing spline method, for example, [8–12]. However, the smoothing spline method uses all the observed covariate values as the spline knots, which requires an enormous computation burden. In contrast, Ruppert *et al.* [13] showed that penalized spline performs well in various setting. Compared with smoothing spline, it requires much less computational resource [8, 9]. Therefore, we will use the penalized spline method.

Our motivating example is a study of the recurrent hospitalization rate for heart failure patients in the clinical data repository (CDR) from the University of Virginia Health System. Among a total of 1172 heart failure patients, 295 patients died during the follow-up period. The recurrent events

[a] *Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA*
[b] *Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, 22908-0717, USA*
**Correspondence to: Zhangsheng Yu, Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA.*
[†] *E-mail: yuz@iupui.edu*

(hospitalizations) processes were thus terminated by death. Previous study [14] showed that the recurrent hospitalizations and death are correlated for these patients. Failure to accommodate such a correlation may lead to bias in the estimation of recurrent events model [14, 15]. Therefore, it is desirable to use a joint frailty model for the recurrent events and the terminal event to capture their correlation and obtain unbiased estimates. Furthermore, our preliminary analysis reveals that although the age effect is linear on log death hazard in a Cox proportional hazards model, age has a significant nonlinear effect on the log intensity of hospitalizations. We, therefore, extend the joint model of recurrent events and a terminal event [15] to accommodate both nonlinear and linear covariate effects. It is worth noting that recurrent events data with an independent terminal event can be treated as a special case of our model and fitted by the proposed estimation procedure.

For our joint model, we incorporate a Gaussian frailty in both recurrent and death event components to introduce the correlation between these two processes. We employ the Laplace approximation to approximate the marginal likelihood in a similar fashion to generalized linear mixed models. For the nonlinear covariate function in the recurrent events component, we will use the penalized spline method [13]. We will estimate the variance component and smoothing parameter using a likelihood method.

We arrange the remainder of this work as follows. We introduce the joint model of recurrent events and terminal event in Section 2. We present the estimation for regression parameters, variance components, and smoothing parameter in Section 3, using penalized spline and Laplace approximation. We evaluate the performance of the estimation procedure using simulation studies in Section 4. In Section 5, we apply the proposed method to the heart failure data and conclude the article with a discussion in Section 6.

2. Models

For the ith subject ($i = 1, 2, ..., n$), we observe the follow-up time $T_i = \min(C_i, D_i)$, where C_i is the (independent) censoring time and D_i is the death time. Let $\Delta_i = I(D_i \leqslant C_i)$ be the indicator for death. Let $0 < R_{i1} < R_{i2} < \cdots < R_{ir_i} < T_i$ be the r_i recurrent event times observed before T_i. Let $Y_i(t) = I(T_i \geqslant t)$ be the at-risk process. Denote the actual and observed death processes by $N_i^{D*}(t) = I\{D_i \leqslant t\}$ and $N_i^D(t) = I\{T_i \leqslant t, \Delta_i = 1\}$, respectively. The observed recurrent event process is defined as $N_i^R(t) = \sum_{j=1}^{r_i} I\{R_{ij} \leqslant t, T_i \geqslant t\}$. The observed information of the ith subject at time t is $\mathbf{O}_i(t) = \{Y_i(u), N_i^R(u), N_i^D(u), Z_i, \boldsymbol{X_i}, 0 \leqslant u \leqslant t\}$, where Z_i is the covariate with a nonlinear effect and $\boldsymbol{X_i}$ is the covariate vector with linear coefficients in hazard functions (1). Define filtration $\mathcal{F}(t) = \sigma\{\mathbf{O_i}(u), 0 \leqslant u \leqslant t\}$.

We now formulate the intensity functions for recurrent events and terminal event. For recurrent events,

$$P(\mathrm{d}N_i^R(t) = 1|\mathcal{F}_t-) = Y_i(t)\mathrm{d}R_i(t),$$

and

$$\mathrm{d}R_i(t) = \exp\{f(Z_i) + \boldsymbol{X_i^T\beta} + \nu_i\}r_0(t)\mathrm{d}t, \tag{1}$$

where $r_0(t)$ is the baseline intensities of recurrent events and $R_0(t)$ the corresponding cumulative baseline intensity. We define by $f(\cdot)$ an unspecified (nonparametric) smooth function whose functional form is of interest.

For the terminal event,

$$P(\mathrm{d}N_i^D(t) = 1|\mathcal{F}_t-) = Y_i(t)\mathrm{d}\Lambda_i(t)$$

and

$$\mathrm{d}\Lambda_i(t) = \exp\{Z_i\alpha_1 + \boldsymbol{X_i^T\alpha_2} + \nu_i\}\lambda_0(t)\mathrm{d}t, \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard of terminal event and $\Lambda_0(t)$ the corresponding cumulative baseline hazard. The unobserved random effect (or frailty) ν_i is introduced to model the correlation among the recurrent events on the same subject and that between the recurrent and terminal events. In this paper, we assume ν_is to be independently Gaussian distributed with mean 0 and variance σ^2, denoted as $\phi_\sigma(\nu_i)$.

However, other distributions, for example, Gamma distribution, have been used in practice [15] [16]. And a more general model is to use two separate frailty terms ν_{i1} and ν_{i2} for each of the terminal and recurrent events as Cook and Lawless [1, Section 6.6.2] suggested.

For simplicity of presentation, we only present the case with a single nonparametric covariate effect. In principle, it is straightforward to extend the recurrent events model to allow for an additive nonparametric function with more than one covariate in the recurrent events component:

$$\mathrm{d}R_i(t) = \exp\left\{f_1(Z_{1i}) + f_2(Z_{2i}) + \cdots + f_q(Z_{qi}) + \boldsymbol{X_i^T\beta} + \nu_i\right\} r_0(t)\mathrm{d}t.$$

We also note here that both models can be extended to have time-dependent external covariates [17, p. 196]. A nonparametric covariate function can be incorporated into terminal event component too.

The complete likelihood of $\{(\mathbf{O}_i, \nu_i), i = 1, \cdots, n\}$ is

$$\begin{aligned} l &= \log \prod_{i=1}^{n} L(\mathbf{O}_i, \nu_i | Z_i, \mathbf{X}_i) \\ &= \sum_{i=1}^{n} l_i + \sum_{i=1} \log \phi_\sigma(\nu_i) \end{aligned}$$

where

$$\begin{aligned} l_i = &\left\{\sum_{j=1}^{r_i}[\nu_i + f(Z_i) + \boldsymbol{X_i^T\beta} + \log r_0(R_{ij})]\right. \\ &\left. - \int_0^\infty Y_i(t)\exp\{f(Z_i) + \boldsymbol{X_i^T\beta} + \nu_i\} r_0(t)\mathrm{d}t\right\} \\ &+ \left\{\Delta_i[\nu_i + Z_i\alpha_1 + \boldsymbol{X_i^T\alpha_2} + \log\lambda_0(T_i)]\right. \\ &\left. - \int_0^\infty Y_i(t)\exp\{Z_i\alpha_1 + \boldsymbol{X_i^T\alpha_2} + \nu_i\}\lambda_0(t)\mathrm{d}t\right\} \end{aligned}$$

The complete likelihood involves the unobserved frailty effect ν_i and a nonparametric covariate function with infinite number of parameters. We provide the estimation procedure in the next section.

3. Estimation method

We will employ the penalized spline method to describe the nonparametric function. For our model, the estimation is complicated because of the presence of random effects. In the absence of the nonparametric covariate function, there are two options to obtain the maximum likelihood estimation: Monte Carlo expectation–maximization algorithm [15] and Gaussian quadrature method [5]. However, the Monte Carlo expectation–maximization algorithm is difficult to implement, and the convergence is very slow. We can conveniently implement the Gaussian quadrature method in SAS PROC NLMIXED (SAS Institute Inc., Cary, NC, USA). However, it remains an issue to incorporate the penalty term in the penalized spline function into SAS PROC NLMIXED. To avoid the computational complexity and reduce the computational burden, we will use Laplace approximation to integrate out the random effect in the likelihood. Related works using Laplace approximation include [18, 19]. Next, we will give a brief introduction to the penalized spline and Laplace approximation.

3.1. *Penalized spline*

The complete likelihood involves a nonparametric covariate function $f(z)$ with an infinite number of parameters, which can be described by the penalized spline method. Specifically, we use a set of cubic spline basis functions $\{B_1(z), B_2(z), \cdots, B_M(z)\}$ to model $f(z) = \sum_{m=1}^{M} \eta_m B_m(z)$, where M is the number of spline basis functions. We can generate the cubic spline basis in the way detailed on page 23

of Green and Silverman [20]. We also include the R-code for spline basis generation using the R package fda in Appendix C. Similar to the Cox model where the intercept of regression function is not estimable, the $f(\cdot)$ is only estimable up to a constant in the joint model. For identifiability purpose, we fix the estimate of $f(z) = 0$ at the left end of z value by setting the intercept component on spline basis functions to be 0. We determine the number and the shape of the basis functions by the selected knots number and location. In the penalized spline setting, equally spaced knot locations are often used.

With penalized spline, one can choose a larger number of knots without introducing much variation. But there should be little advantage to use more than 10–20 knots as Gray [8] mentioned. We will use eight knots in our simulation study for demonstration. The penalty term is on the second derivative of the cubic spline. In practice, we can easily generate the cubic spline basis function and the penalty matrix by using the fda package in R or MATLAB (MathWorks, Natick, MA, USA). The penalized complete log-likelihood for $\{(\mathbf{O}_i, \nu_i), i = 1, \cdots, n\}$ is

$$\begin{aligned} l &= \log \prod_{i=1}^{n} L(\mathbf{O}_i, \nu_i | Z_i, X_i) \\ &= \sum_{i=1}^{n} l_i + \sum_{i=1} \log \phi_\sigma(\nu_i) - \frac{1}{2\rho} \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{P} \boldsymbol{\eta}, \end{aligned}$$

where $\boldsymbol{\eta} = \{\eta_1, \cdots, \eta_M\}^T$, $\mathbf{B}^{(2)}(s) = \{B_1^{(2)}(s), \cdots, B_M^{(2)}(s)\}^T$ is the vector of the second derivatives of B-spline basis, ρ is the smoothing parameter, and $\mathbf{P} = \int [\mathbf{B^{(2)}(s)B^{(2)}(s)^T}]^2 \mathbf{ds}$. The form of l_i is given as follows:

$$\begin{aligned} l_i = &\sum_{j=1}^{r_i} \left[\nu_i + \sum_{m=1}^{M} \eta_m B_m(Z_i) + \boldsymbol{X_i^T \beta} + \log r_0(R_{ij}) \right] \\ &- \int_0^\infty Y_i(t) \exp \left\{ \sum_{m=1}^{M} \eta_m B_m(Z_i) + \boldsymbol{X_i^T \beta} + \nu_i \right\} r_0(t) \mathrm{d}t \Bigg\} \\ &+ \Bigg\{ \Delta_i [\nu_i + Z_i \alpha_1 + \boldsymbol{X_i^T \alpha_2} + \log \lambda_0(T_i)] \\ &\qquad - \int_0^\infty Y_i(t) \exp\{Z_i \alpha_1 + \boldsymbol{X_i^T \alpha_2} + \nu_i\} \lambda_0(t) \mathrm{d}t \Bigg\} \end{aligned}$$

Note that the roughness of likelihood function is penalized by subtracting the integral of the second derivative function.

3.2. *Laplace approximation*

In this subsection, we propose to using Laplace approximation to evaluate the marginal penalized likelihood. To estimate the coefficients $\boldsymbol{\alpha}$s and $\boldsymbol{\eta} = \{\eta_1, \eta_2, \cdots, \eta_m\}^T$ in the complete likelihood, we have to integrate the frailty out. We obtain the logarithm of the marginal penalized hazard

$$ml_p = -\frac{1}{2\rho} \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{P} \boldsymbol{\eta} + \log \left[\int_{-\infty}^{\infty} \prod_{i=1}^{n} \exp\{l_i + \log \phi_\sigma(\nu_i)\} \mathrm{d}\boldsymbol{\nu} \right] \tag{3}$$

There is no closed form for the logarithm integral part in Equation (3). To evaluate Equation (3), we used the Laplace approximation for integral calculation by following Breslow and Clayton's derivation for generalized linear mixed models [18]. Ripatti and Palgrem [19] adopted a similar approach for frailty model. The approximation is

$$\log\{c|D|^{-1/2} \int e^{-K(\boldsymbol{\nu})} \mathrm{d}\boldsymbol{\nu}\} \approx -\frac{1}{2} \log|D| - \frac{1}{2} \log|K''(\tilde{\boldsymbol{\nu}})| - K(\tilde{\boldsymbol{\nu}}) \tag{4}$$

after ignoring a constant c, where $\tilde{\boldsymbol{\nu}}$ is the solution to the first derivative $K'(\boldsymbol{\nu}) = 0$ and $K''(\cdot)$ is the second derivative of $K(\cdot)$ with respect to $\boldsymbol{\nu}$. By applying the approximation on the logarithm part of

ml_p, we have

$$K(\boldsymbol{v}) = -\sum_{i=1}^{n} l_i + \frac{1}{2\sigma^2}\sum_{i=1}^{n} v_i^2$$

by noting that the $\phi_\sigma(v_i)$ is normally distributed with mean 0 and variance σ^2.

The logarithm part of Equation (3) can be further approximated by $-K(\tilde{\boldsymbol{v}})$ if the first two terms in Equation (4) vary little when $(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta})$ change as Breslow and Clayton indicated. Hence, the approximation of ml_p is

$$ml_p \approx \sum_{i=1}^{n} l_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} v_i^2 - \frac{1}{2\rho}\boldsymbol{\eta}^{\boldsymbol{T}}\mathbf{P}\boldsymbol{\eta}$$

3.3. *Estimating equations*

The approximation of log marginal likelihood can be further written as Equation (6) in Appendix A. We then simplified it by profiling out the the cumulative baseline intensities $R_0(t)$ and $\Lambda_0(t)$. By taking the derivative of the marginal likelihood 6 with respect to $r_0(R_{ij})$ and $\lambda_0(T_i)$, respectively, we obtain

$$r_0(R_{ij}) = \frac{1}{\sum_l I\{T_l \geqslant R_{ij}\} e^{\sum_{m=1}^{M}\eta_m B_m(Z_l) + \boldsymbol{X_l^T}\boldsymbol{\beta} + v_l}} \tag{5}$$

and

$$\lambda_0(T_i) = \frac{\Delta_i}{\sum_l I\{l : T_l \geqslant T_i\}\Delta_i e^{Z_l\alpha_1 + \boldsymbol{X_l^T}\boldsymbol{\alpha_2} + v_l}}. \tag{6}$$

Substituting these solutions back into the approximated marginal likelihood, we can profile out the baseline hazard functions and obtain the penalized partial likelihood

$$\begin{aligned} ml_p \approx & \sum_{i=1}^{n}\sum_{j=1}^{r_i}\left\{v_i + \sum_{m=1}^{M}\eta_m B_m(Z_i) + \boldsymbol{X_i^T}\boldsymbol{\beta} - \log\sum_l I\left\{T_l \geqslant R_{ij}\right\} e^{\sum_{m=1}^{M}\eta_m B_m(Z_l) + \boldsymbol{X_l^T}\boldsymbol{\beta} + v_l}\right\} \\ & + \sum_{i=1}^{n}\Delta_i\left\{v_i + Z_i\alpha_1 + \boldsymbol{X_i^T}\boldsymbol{\alpha_2} - \log\sum_l I\{T_l \geqslant T_i\} e^{Z_l\alpha_1 + \boldsymbol{X_l^T}\boldsymbol{\alpha_2} + v_l}\right\} \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{n} v_i^2 - \frac{1}{2\rho}\boldsymbol{\eta}^{\boldsymbol{T}}\mathbf{P}\boldsymbol{\eta} \end{aligned}$$

We then derive the estimating equations of $(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta})$ by taking the derivative of the approximated marginal likelihood with respect to $(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta})$ as follows:

$$\begin{aligned} S_{\boldsymbol{\eta},\boldsymbol{\beta}} = & \sum_{i=1}^{n}\sum_{j=1}^{r_i}\left\{(\boldsymbol{B(Z_i)^T}, \boldsymbol{X_i^T})^T - \frac{\sum_l I\{T_l \geqslant R_{ij}\} e^{\sum_{m=1}^{M}\eta_m B_m(Z_l) + \boldsymbol{X_l^T}\boldsymbol{\beta} + v_l}(\boldsymbol{B(Z_l)^T}, \boldsymbol{X_l^T})^T}{\sum_l I\{T_l \geqslant R_{ij}\} e^{\sum_{m=1}^{M}\eta_m B_m(Z_l) + \boldsymbol{X_l^T}\boldsymbol{\beta} + v_l}}\right\} \\ & - (\frac{1}{\rho}\boldsymbol{P\eta}, \mathbf{0})^T, \end{aligned} \tag{7}$$

where $\mathbf{B(z)} = \{B_1(z), \cdots, B_m(z)\}^T$;

$$S_\alpha = \sum_{i=1}^{n}\Delta_i\left\{(Z_i^T, \boldsymbol{X_i^T})^T - \frac{\sum_l I\{T_l \geqslant T_i\} e^{Z_l\alpha_1 + \boldsymbol{X_l^T}\boldsymbol{\alpha_2} + v_l}(Z_l^T, \boldsymbol{X_l^T})^T}{\sum_l I\{T_l \geqslant T_i\} e^{Z_l\alpha_1 + \boldsymbol{X_l^T}\boldsymbol{\alpha_2} + v_l}}\right\}; \tag{8}$$

and

$$S_v = (r_1, \cdots, r_n)^T - \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{\{I_{ij1}, \cdots, I_{ijn}\}^T}{\sum_l I_{ijl}} + (\Delta_1, \cdots, \Delta_n)^T \tag{9}$$
$$- \sum_{i=1}^{n} \Delta_i \frac{\left\{ I\{T_1 \geqslant T_i\} e^{Z_1\alpha_1 + \boldsymbol{X_1^T}\boldsymbol{\alpha_2} + v_1}, \cdots, I\{T_n \geqslant T_i\} e^{Z_n\alpha_1 + \boldsymbol{X_n^T}\boldsymbol{\alpha_2} + v_n} \right\}^T}{\sum_l I\{T_l \geqslant T_i\} e^{Z_l\alpha_1 + \boldsymbol{X_l^T}\boldsymbol{\alpha_2} + v_l}}$$
$$- \frac{\boldsymbol{v}}{\sigma^2},$$

where $I_{ijr} = I\{T_r \geqslant R_{ij}\} e^{\sum_{m=1}^{M} \eta_m B_m(Z_r) + \boldsymbol{X_r^T}\boldsymbol{\beta} + v_r}$.

We can obtain the estimates of regression coefficients by solving the estimating Equations (7)–(9) for fixed σ and ρ, using the Newton–Raphson algorithm. We derive the second derivatives of the likelihood in Appendix B. After obtaining the estimates of $\boldsymbol{\eta}$, the estimated nonparametric covariate function is then $\hat{f}(z) = \sum_{m=1}^{M} \hat{\eta}_m B_m(z)$. We propose to use $-I(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, the inverse of the negative second derivative of ml_p with respect to $(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, as their covariance estimate. We can then estimate the variance of $\hat{f}(\cdot)$ by $\widehat{V}ar\{\hat{f}(z)\} = \boldsymbol{B(z)^T V_{\eta\eta} B(z)}$, where $\boldsymbol{V_{\eta\eta}}$ is the submatrix of $-I(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ corresponding to η.

We can estimate the baseline intensity functions $r_0(t)$ and $\lambda_0(t)$ by plugging $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{v}}_{\boldsymbol{i}})$ back into Equations (5) and (6).

3.4. *Inferences on variance components and smoothing parameters*

To estimate the variance components σ^2, we employ a similar method to Ripatti and Palgrem [19]. For a shared frailty model where covariance of $\boldsymbol{v}$ is a diagonal matrix with σ^2 on the diagonal, we can estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{v}}'\hat{\boldsymbol{v}} + tr\left\{[K''(\hat{\boldsymbol{v}})^{-1}]\right\}}{n_c}, \tag{10}$$

where n_c is the number of clusters introduced by the frailty.

Note that $K''(\boldsymbol{v})$ involves the baseline intensity functions. In Equation (10), one can also use the second derivative from the penalized partial likelihood instead of $K''(\boldsymbol{v})$ to avoid calculating the baseline intensity functions. The performance is similar from our experiences. We will use penalized partial likelihood for demonstration in the simulation study.

One can usually carry out selection of smoothing parameter using cross validation. In the joint model of the survival data, the mathematical derivation is rather complicated (see the case for survival data in [21]). The cross validation method is also computationally challenging. Therefore, we propose an estimate by treating smoothing parameters as a variance component. For variance component σ^2, Equation (10) includes the number of parameter (random effects) in the denominator, the two parts are the sum of the square of estimates and the trace of the covariance matrix. Along the same lines, we can estimate the smoothing parameter by

$$\hat{\rho} = \frac{\sum_{i=1}^{n} \widehat{f}(Z_i)^2 + tr\left\{\mathrm{Cov}(\widehat{f}(Z_i))\right\}}{M},$$

where $\mathrm{Cov}(\widehat{f}(Z_i))$ is a covariance matrix of the nonparametric function and M is the number of basis functions for penalized spline. We will evaluate the performance of the nonparametric function estimate based on this smoothing parameter estimate using simulation.

We can estimate the variance of $\hat{\boldsymbol{v}}$ by using the corresponding part of the inverse of information matrix on page 1018 of Ripatti *et al.* [19]. But it should be noted that the standard error (SE) estimates for $\hat{\sigma}^2$ and $\hat{\rho}$ are not reliable using this method [19, 22]. Therneau and Grambsch suggested using bootstrapping method to estimate such terms and constructing confidence intervals [22]. We did pursue such approach further because of the intensive computation.

4. Simulation

In this section, we evaluate the performance of the proposed estimators through a simulation study. The R-code for fitting the proposed estimation procedure will be available upon request. We generate the recurrent events following a Weibull distribution:

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\left[-(\lambda t)^{\gamma}\right] \tag{11}$$

where the shape parameter $\gamma = 2$ and the scale parameter λ is $r_i = r_0\exp[f_1(Z_1)+\beta_1 X_1+\beta_2 X_2+\nu_i]$. We can write the conditional intensity function as follows:

$$r_i(t; f_1, \beta_1, \beta_2, \nu) = 2r_0 t\exp[f_1(Z_1) + \beta_1 X_1 + \beta_2 X_2 + \nu_i],$$

where $r_0 = 10/3$, $\nu_i \sim N(0,\theta)$ with $\theta = 0.5$. The hazard model for terminal event follow a Weibull distribution with a shape parameter equal to 1.5 and the scale parameter is

$$\lambda_i(t) = \lambda_0\left[\exp\left\{\alpha_0 Z_1 + \alpha_1 X_1 + \alpha_2 X_2 + \nu_i\right\}\right]^{1/1.5},$$

where $\lambda_0 = 10/3$

We run the simulation with two sets of nonparametric functions. In the first setting, we used a piecewise smooth function

$$f(z) = (0.5 - 0.5\times(z+1)^2)\times I(z<-1)) + 0.5\times I(z>=-1),$$

where z is uniformly distributed on $[-2,0]$, $\boldsymbol{\alpha} = (0.5,1,2)^T$, and $\boldsymbol{\beta} = (1,0.5)^T$. We generated X_1 as a binary variable taking values 0 and 1 with equal probability; we generated X_2 as a uniformly distributed random variable in $[-1,1]$. For each data set, we generated 400 subjects, and the follow-up time is 0.6. The percentage of censoring of the terminal events was about 20%. The average number of recurrent events per subject was about five. We generated 200 data sets. We then summarized the simulation results over the 200 replicates.

We fit the data using both the proposed procedure and naive methods, which models the recurrent and terminal events separately. The upper panel of Table I lists the estimate of parametric coefficients in the first setting for both recurrent events and terminal event components. The left part is the solution using the proposed joint model. In this setting, the biases of the parametric estimates for β_1 and β_2 are very small. The mean estimated SEs over 200 replicates are close to the empirical SE. The empirical coverage probabilities are close to the nominal 95%, using the proposed estimate SEs. The variance component is slightly underestimated. The bias of this scale is common in other studies using approximation for marginal likelihood estimation such as Ripatti and Palgrem [19]. For the terminal event component αs, the means of the estimates and estimated SE are close to the true values and the empirical SEs, respectively. The empirical coverage probabilities using the proposed SE approach the nominal 95%. The reduced model fitting has a much larger bias in both recurrent and terminal events, which results in a smaller than nominal level coverage probabilities.

In Figure 1, we present the estimate of the nonparametric covariate function $f(Z_1)$ for setting I using the joint model and reduced model. The solid line in the left panel is the true function, and the dotted line is the average of the estimated nonparametric function using joint model over 200 replicates. The estimated curve is quite close to the true function. The right panel shows the coverage probabilities of the 95% pointwise confidence interval (CI) for the nonparametric function. At each point, the coverage probability of joint model estimate (dotted line) is around 95%. The average coverage probability is 94.9%. The dash-dotted line is the nonparametric estimate when reduced model is used. The bias of the estimate is larger than that from the joint model. The coverage probability is 94.7%. This result shows that the proposed joint model estimate works well when the nonlinear function is piecewise smooth function, which is difficult to model when using polynomial function. We also generated the data using variance component $\theta = 1$, and the patterns of simulation results are similar to $\theta = 0.5$, hence not presented here.

In setting II, the nonlinear smooth function $f(z)$ as

$$f(z) = (2\times\beta(z/10,8,8) + \beta(z/10,5,5))/9,$$

where β is the density function of a beta distribution. The parametric coefficients $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and covariates X are the same as setting I. Covariate z is uniformly distributed on $[-1,1]$. The percentage of censoring of the terminal events was about 25%. The average number of recurrent events per subject was about 2. We

Table I. Simulation: parametric coefficient estimates.

	Parameter	Mean	SE*	SEM†	CP (%)◇	Mean	SE	SEM	CP (%)
Setting I			Joint	Model			Reduced	Model	
β_1	1	0.983	0.101	0.092	92.0	0.922	0.100	0.092	83.5
β_2	0.5	0.496	0.106	0.096	91.5	0.336	0.099	0.096	60.0
α_1	0.5	0.502	0.110	0.115	94.0	0.447	0.105	0.099	90.5
α_2	1	0.979	0.138	0.132	93.0	0.881	0.132	0.117	79.0
α_3	2	2.003	0.142	0.136	95.0	1.778	0.119	0.121	53.5
σ^2	0.5	0.469	0.046	0.037	N.A.	0.441	0.047	0.035	N.A.
Setting II			Joint	Model			Reduced	Model	
β_1	1	0.975	0.115	0.108	91.0	0.897	0.114	0.107	82.5
β_2	0.5	0.495	0.126	0.119	94.0	0.306	0.120	0.118	64.5
α_1	0.5	0.491	0.109	0.108	95.0	0.433	0.095	0.095	89.5
α_2	1	0.985	0.125	0.129	95.5	0.881	0.112	0.113	81.5
α_3	2	2.008	0.138	0.134	95.0	1.768	0.119	0.118	43.5
σ^2	0.5	0.477	0.057	0.036	N.A.	0.434	0.060	0.034	N.A.

*Empirical standard error.
†Mean of standard error.
◇Coverage probability.

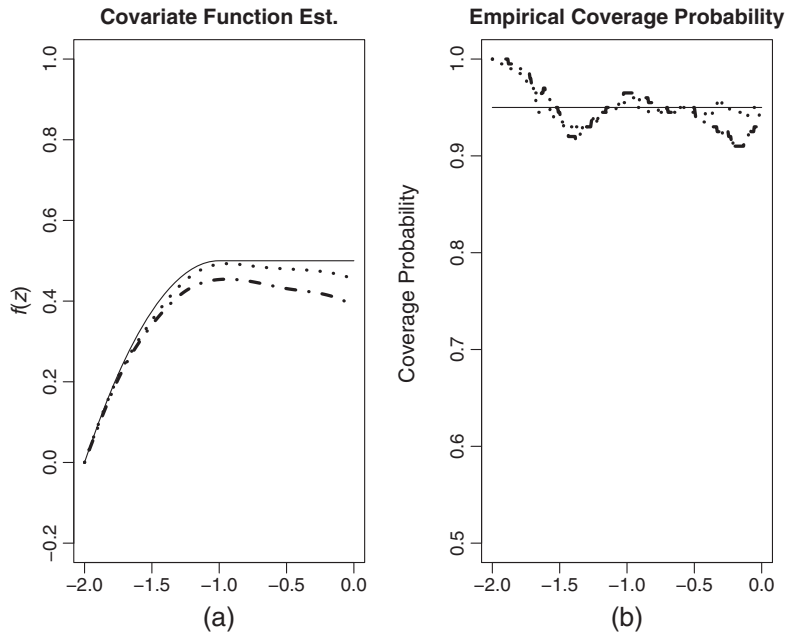

Figure 1. Setting I: simulation results for the nonparametric covariate function $f(z)$. (a) Penalized spline estimate of $\hat{f}(z)$: joint model estimate, dotted; reduced model estimate, dash-dotted; true $f(z)$, solid; (b) Empirical coverage probability of $f(z)$: joint model, dotted (mean coverage probability is 94.9%); reduced model estimate, dash-dotted (mean coverage probability is 94.7%).

generated 200 data sets with 400 subjects in each. For the joint model, the biases of the regression coefficient estimates are generally small. Results are present in the bottom of Table I. The averages of estimate SEs over 200 replicates are slightly smaller than the empirical SE, resulting in a minor undercoverage for some regression coefficients. The variance component is underestimated similarly to setting I. In the reduced model, the coefficient estimates, as shown in the lower right panel of Table I, are much more biased. The coverage probabilities are much lower than those in the joint model solution.

Figure 2 shows the performance of nonparametric function using the joint model and reduced model in setting II. In general, the estimated curve is close to the true curve when joint model is used. The average coverage probability is 97.7%. The dash-dotted line shows that the bias of the nonparametric

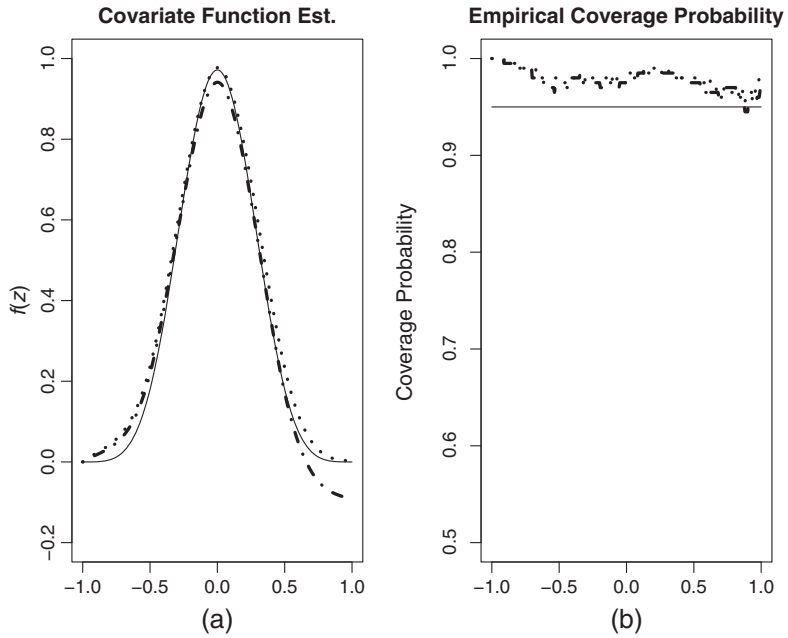

Figure 2. Setting II: simulation results for the nonparametric covariate function $f(z)$. (a) Penalized spline estimate of $\hat{f}(z)$: joint model estimate, dotted; reduced model estimate, dash-dotted; true $f(z)$, solid; (b) Empirical coverage probability of $f(z)$: joint model, dotted (mean coverage probability is 97.4%); reduced model estimate, dash-dotted (mean coverage probability is 97.4%).

estimate using the reduced model is substantial at boundary (around $z = 0$ and $z = 1.0$) and much larger than the joint model estimate. The mean coverage probability is 97.4%.

We have also run the simulation using the exponential distribution instead of Weibull distribution. The estimation of the joint model is also better than the reduced model. In summary, our estimation method yields small biases for the parametric coefficients and nonparametric covariate function in the joint model. The empirical CIs are close to the nominal level. The variance components are slightly underestimated, similarly to previous work in Gaussian frailty models. However, failure to account for the correlation between recurrent and terminal events and naively treating death event as independent censoring for recurrent events could lead to biases in both parametric and nonparametric components, as the reduced model results show.

Per request by one reviewer, we test the robustness of the proposed method to the misspecification of single frailty assumption (e.g., separate frailty terms for each of the recurrent events and terminal event as suggested in Section 6.6.2 of Cook and Lawless [1]). The result (not shown) indicates a minor bias for some parameters. This warrants future research on models accommodating separate frailty terms for each component.

5. Application

In this section, we apply the method to heart failure data. Congestive heart failure, or heart failure, is a chronic disease in which the heart cannot pump blood as it should. Over five million patients in the USA live with the disease, with 550,000 new cases diagnosed each year, making it the only cardiac disease growing in prevalence. Heart failure is the most common reason of hospitalizations for people 65 years and older, whereas hospitalization costs account for approximately 70% of total costs (Medicare Provider and Analysis Review File 1999).

Our data come from the CDR. It is a clinical information database to facilitate easy and flexible access to healthcare data at the University of Virginia Health System. In this study, we include patients 67 years and older at baseline who were first diagnosed and treated in 2004 with heart failure (International Classification of Diseases-Ninth Revision diagnosis code beginning with 428). We choose this age group to avoid the potential self-selection of hospitalization. Because Medicare covers the hospitalization cost after age 65 years, we expect there is a spike of hospitalizations right after age 65 years. Because we are

interested in the hospitalizations due to the natural disease course, not the availability of coverage, we exclude patients who were younger than 67 years at baseline. Because of data confidentiality, we remove all patients 90 years and older at baseline.

The study includes a total of 1172 patients. They were followed-up to July 31, 2006 or until death, whichever came first. Their records of hospitalization after the heart disease diagnosis were identified. The death records were extracted from death certificates filed with the Virginia Department of Vital Statistics. We deal with the ties using Breslow type approach (see page 105 of Kalbfleisch and Prentice [17]). Two hundred and ninety five (25.2%) patients died before the end of the follow-up period; others were censored. During the follow-up, 588 patients (50.2%) had at least one hospitalization during the follow-up (range of 1–17). The numbers of patients having one, two, three, four, or five or more hospitalizations were 300, 122, 61, 43, and 62, respectively. We are interested in describing the risk of hospitalization and death based on demographic information in this analysis. Specifically, we examine the effect of age at first diagnosis, sex, and race (white or others). There were 53% male and 72% white patients in the data set. The mean age at baseline is 76.1 years old with a standard deviation of 5.9.

To investigate the age effect on recurrent and the death events, we ran preliminary analysis for these two events separately using R package *coxph*, with *pspline* and *frailty* options. The analysis of recurrent events shows that the rate of hospitalization changes over age nonlinearly. The p-value of testing nonlinear age effect is 0.011. Analysis of terminal event shows that linear age effect on the hazard of death is significant. But the nonlinear effect of age on death hazard is not significant (p-value $= 0.86$). The hospitalization rate is dependent on the patients health status, so is the death event. Thus, it is not appropriate to treat the death events as independent censoring for the hospitalization intensity. We thus apply our joint frailty model, with nonlinear age effect in the recurrent events component and a linear age effect in the death event component. We estimate the nonlinear function using penalized spline with four knots. We also include covariates sex and race in the model.

Figure 3 shows the nonlinear function of age effect on the hospitalization rate. In general, the hospitalization rate increases with baseline age. However, such increase is not linear. We observe an initial elevation of hospitalization intensity from age 67 to 70 years. After that, the recurrent hospitalization intensity fluctuates then dips slightly up to age 83 years, followed by a steady increasing trend from 85 years.

Thanks to the advice of a reviewer, we also fit the data using a simple alternative approach: we assume a piecewise linear function for age effect, with slope changes at age $= 70$ and age $= 83$. These cutoff points were chosen based on the nonparametric estimate of age effect. The piecewise linear fit closely matches the nonlinear fit, validating our model fitting. Table II lists the estimation for the parametric coefficients using the proposed joint model, after adjusting for the age and gender effect, white patients had a significantly lower risk (p-value $= 0.030$) of hospitalization related to heart failure. By one year increment of age, the risk of death will increase by 3.9% (p-value $= 0.061$). White patients have a lower risk (p-value $= 0.290$) of death.

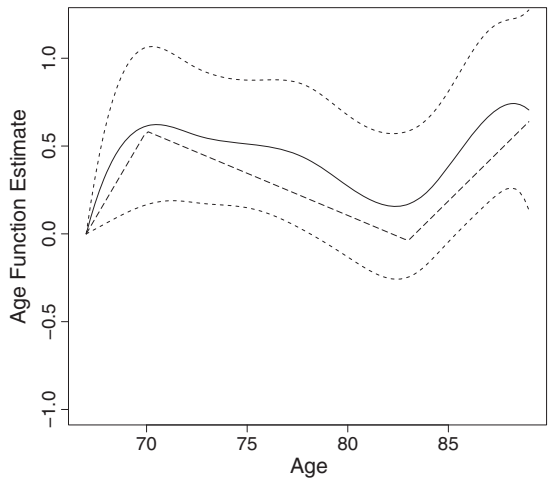

Figure 3. Application: penalized spline estimator $\hat{f}$(age), dotted; 95% confidence band, dash; piecewise linear fit, solid. Note that at age $= 67$ years, the estimate is fixed to be 0 for estimability similarly to other nonparametric regression estimates in Cox proportional hazards model.

Table II. Parametric coefficient estimates in hospitalization study analysis using the joint model.

Risk factor	Estimate	SE	Hazards ratio	p-value
Hospitalization				
Gender (Male)	−0.099	0.171	0.906	0.563
Race (White)	−0.417	0.192	0.659	0.030
Death				
Age	0.039	0.0207	1.039	0.061
Gender (Male)	0.052	0.256	1.053	0.840
Race (White)	−0.301	0.285	0.740	0.290
σ^2	1.107			

SE, standard error.

6. Discussion

Previous studies have demonstrated the values of joint modeling of recurrent events and a terminal event when terminal event is correlated with recurrent events [14, 15]. This work has further extended the flexibility of current joint models by incorporating into these models semiparametric components for the potential nonlinear predictors. For regression coefficients estimation, we use penalized cubic spline to model the nonlinear function in the recurrent event components and employ the Laplace approximation to evaluate the marginal likelihood without a closed form. The simulation study shows that the proposed estimates are generally unbiased for parametric coefficients and nonlinear function. The inference based on the proposed covariate is close to nominal level too. In summary, the estimation and inference procedures were shown to be valid. This warrants further investigation on large sample theory of spline estimate. For discussion of the asymptotic properties of such spline estimate in Cox proportional hazards model, see Gray [8].

Our model can be extended in several directions. First, nonparametric function can be included in the death hazard model as well. Second, our model can be generalized along the line of [23], where they proposed a joint model of recurrent events, repeated markers at recurrent events, and survival. Similar extension can also be applied to the joint model with gap time for recurrent events [14]. Further related work also includes extending our model to using separate frailty terms for the correlation among recurrent events and between recurrent events and the terminal event.

Appendix A. Approximated marginal likelihood

$$
\begin{aligned}
ml_p(\alpha,\eta,\beta) \approx \sum_{i=1}^{n}\Bigg\{\sum_{j=1}^{r_i}\Bigg[v_i + \sum_{m=1}^{M}\eta_m B_m(Z_i) + \boldsymbol{X_i}^T\beta + log\, r_0(R_{ij})\Bigg] \\
- \sum_{ls} I\{R_{ls} \leq T_i\} e^{\sum_{m=1}^{M}\eta_m B_m(Z_i) + \boldsymbol{X_i^T\beta} + v_i} r_0(R_{ls})\Bigg\} \\
+ \sum_{i=1}^{n}\Bigg\{\Delta_i\Big[v_i + Z_i\alpha_1 + \boldsymbol{X_i^T\alpha_2} + log\lambda_0(T_i)\Big] \\
- \sum_{l} I\{T_l \leq T_i\}\,\Delta_i e^{Z_l\alpha_1 + \boldsymbol{X_l^T\alpha_2} + v_i}\lambda_0(T_l)\Bigg\} \\
- \frac{1}{2\sigma^2}\sum_{i=1}^{n} v_i^2 - \frac{1}{2\rho}\boldsymbol{\eta}^T\mathbf{P}\boldsymbol{\eta}
\end{aligned}
$$

Appendix B. Second derivative of penalized partial likelihood

$$S^{(2)}_{\eta,\beta} = -\sum_{i=1}^{n}\sum_{j=1}^{r_i}\left[\frac{\sum_l I\{T_l \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_l)+\boldsymbol{X_l^T\beta}+\nu_l}(\boldsymbol{B(Z_l)^T, X_l^T})^{\otimes 2}}{S_0(R_{ij})} - \frac{\left\{\sum_l I\{T_l \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_l)+\boldsymbol{X_l(R_{ij})^T\beta}+\nu_l}(\boldsymbol{B(Z_l)^T, X_l^T})\right\}^{\otimes 2}}{S_0(R_{ij})^2}\right] - \frac{1}{\rho}P$$

where, $S_0(R_{ij}) = \sum_l I\{T_l \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_l)+\boldsymbol{X_l^T\beta}+\nu_l}$.

$$S^{(2)}_{\alpha} = -\sum_{i=1}^{n}\Delta_i\left[\frac{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}\{Z_l, X_l^T\}^{\otimes 2}}{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}} - \frac{\left\{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}\{Z_l, \boldsymbol{X_l^T}\}\right\}^{\otimes 2}}{\left\{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}\right\}^2}\right]$$

and

$$\begin{aligned}S^{(2)}_{\nu} = &-\sum_{i=1}^{n}\sum_{j=1}^{r_i}\left\{\frac{\mathrm{diag}\{I_{ij1},\cdots,I_{ijn}\}^T}{\sum_l I_{ijl}} - \frac{\left\{I\{T_1 \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_1)+\boldsymbol{X_1^T\beta}+\nu_1},\cdots,I\{T_n \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_n)+\boldsymbol{X_n^T\beta}+\nu_n}\right\}^{\otimes 2}}{\left\{\sum_l I\{T_l \geqslant R_{ij}\}e^{Z_l\sum_{m=1}^{M}\eta_m B_m(R_{ij})+\boldsymbol{X_l^T\alpha_2}+\nu_l}\right\}^2}\right\} \\ &-\sum_{i=1}^{n}\Delta_i\left\{\frac{\mathrm{diag}\left\{I\{T_1 \geqslant T_i\}e^{Z_1\alpha_1+\boldsymbol{X_1^T\alpha_2}+\nu_1},\cdots,I\{T_n \geqslant T_i\}e^{Z_n\alpha_1+\boldsymbol{X_n^T\alpha_2}+\nu_n}\right\}^T}{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}} - \frac{\left\{I\{T_1 \geqslant T_i\}e^{Z_1\alpha_1+\boldsymbol{X_1^T\alpha_2}+\nu_1},\cdots,I\{T_n \geqslant T_i\}e^{Z_n\alpha_1+\boldsymbol{X_n^T\alpha_2}+\nu_n}\right\}^{\otimes 2}}{\left\{\sum_l I\{T_l \geqslant T_i\}e^{Z_l\alpha_1+\boldsymbol{X_l^T\alpha_2}+\nu_l}\right\}}\right\} \\ &-\frac{1}{\sigma^2}\end{aligned}$$

where, $I_{ijr} = I\{T_r \geqslant R_{ij}\}e^{\sum_{m=1}^{M}\eta_m B_m(Z_r)+\boldsymbol{X_r^T\beta+\nu_r}}$.

Appendix C. R code for generating spline basis and penalty matrix

```
library(fda) # call the fda library
# obtain the min and max, and range of nonparametric covariate ;
minz=min(z)
maxz=max(z)
range.0=maxz-minz
knots=minz+c(0:nknots)*range.0/nknots # define the knots;
# generate the basis function based on knots and covariate z.
timebasis=bsplineS(z, knots, norder=4, nderiv=0)
# create bspline basis
basisobj = create.bspline.basis(rangeval=c(knots[1], knots[NROW(knots)]),breaks=knots)
# create the penalty matrix
penmat = bsplinepen(basisobj)
```

Acknowledgements

We are grateful to Dr Jason Lyman, Mr Mac Dent, and Mr Ken Scully at the CDR of the University of Virginia for preparing the medical cost data. A Department of Veteran Affairs grant (STR 03 - 168), an AHRQ grant (R03 HS016543), and an NIAAA grant (RC1 AA019274) partly supported this research.

References

1. Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events*. Springer: New York, 2007.
2. Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV patients. *Journal American Statistical Association* 1998; **93**:46–53.
3. 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.
4. Sinha D, Maiti T, Ibrahim JG, Ouyang B. Current methods for recurrent events data with dependent termination: a Bayesian perspective. *Journal American Statistical Association* 2008; **103**:866–878.
5. Liu L, Huang X. The use of Gaussian quadrature in frailty proportional hazards models. *Statisitics in Medicine* 2008; **27**:2665–2683.
6. Cook RJ, Lawless JF, Lakhal-Chaieb L, Lee KA. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *Journal American Statistical Association* 2009; **104**:60–75.
7. Zeng D, Lin D. Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* 2009; **65**:746–752.
8. Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal American Statistical Association* 1992; **87**:942–951.
9. Gray RJ. Spline-based tests in survival analysis. *Biometrics* 1994; **50**:640–652.
10. Hastie T, Tibshirani R. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* 1990; **46**:1005–1016.
11. Fan J, Gijbels I, King M. Local likelihood and local partial likelihood in hazard regression. *Annals of Statistics* 1997; **25**:1661–1690.
12. Duchateau L, Janssen P. Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. *Biometrics* 2004; **60**:608–614.
13. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.
14. Huang X, Liu L. A joint frailty model for survival and gap times between recurrents. *Biometrics* 2007; **63**:389–397.
15. Liu L, Wolfe R, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; **60**:747–756.
16. Liu L, Yu Z. Likelihood reformulation method in non-normal random effects models. *Statistics in Medicine* 2008; **27**:3105–3124.
17. Kalbfleisch JD, Prentice R. *The Statistical Analysis of Failure Time Data, Second Edition*. Wiley: New York, 2002.
18. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal American Statistical Association* 1993; **88**:9–25.
19. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 2000; **56**:1016–1022.
20. Green PJ, Silverman BW. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall/CRC, 1994.
21. O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* 1988; **9**:531–542.
22. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer: New York, 2000.
23. Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* 2008; **64**:950–958.