Semiparametric Regression Analysis of Multiple Right- and Interval-Censored Events

Fei Gao, Donglin Zeng, David Couper, and D. Y. Lin

Department of Biostatistics, University of North Carolina, Chapel Hill, NC

ABSTRACT

Health sciences research often involves both right- and interval-censored events because the occurrence of a symptomatic disease can only be observed up to the end of follow-up, while the occurrence of an asymptomatic disease can only be detected through periodic examinations. We formulate the effects of potentially time-dependent covariates on the joint distribution of multiple right- and interval-censored events through semiparametric proportional hazards models with random effects that capture the dependence both within and between the two types of events. We consider nonparametric maximum likelihood estimation and develop a simple and stable EM algorithm for computation. We show that the resulting estimators are consistent and the parametric components are asymptotically normal and efficient with a covariance matrix that can be consistently estimated by profile likelihood or nonparametric bootstrap. In addition, we leverage the joint modelling to provide dynamic prediction of disease incidence based on the evolving event history. Furthermore, we assess the performance of the proposed methods through extensive simulation studies. Finally, we provide an application to a major epidemiological cohort study. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received August 2017

Revised May 2018

KEYWORDS

Dynamic prediction; Joint models; Nonparametric likelihood; Proportional hazards; Random effects; Semiparametric efficiency

1. Introduction

Many clinical and epidemiological studies are concerned with multiple types of diseases, which may be symptomatic or asymptomatic. Time to the development of a symptomatic disease is right-censored if the disease does not occur during the follow-up, whereas time to the development of an asymptomatic disease is typically interval-censored because the disease occurrence can only be monitored periodically using biomarkers. In the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators 1989), for instance, subjects were followed for up to 27 years for symptomatic cardiovascular diseases, such as myocardial infarction (MI) and stroke, through reviews of hospital records; they were also examined over five clinic visits, with the first four at approximately 3-year intervals, for occurrences of asymptomatic diseases, such as diabetes and hypertension.

There is a large body of literature on the joint analysis of correlated right-censored events (Kalbfleisch and Prentice 2002, chap. 10; Hougaard 2012), as well as a growing body of literature on correlated interval-censored events (Goggins and Finkelstein 2000; Kim and Xue 2002; Wen and Chen 2013; Chen et al. 2014; Zeng, Gao, and Lin 2017). In addition, there is a considerable amount of literature on competing risks and semi-competing risks (Fine and Gray 1999; Fine, Jiang, and Chappell 2001; Kalbfleisch and Prentice 2002, chap. 8). However, the existing literature has treated right-censored and interval-censored events separately. Joint modeling of the two kinds of data would allow investigators to evaluate the effects of covariates on both kinds of events and to predict the occurrence of a symptomatic disease given the history of asymptomatic diseases.

In this article, we relate potentially time-dependent covariates to the joint distribution of multiple types of right- and interval-censored event times through semiparametric proportional hazards models with random effects. Specifically, we assume a shared random effect for the interval-censored events, which affects the right-censored events with unknown coefficients. We assume an additional shared random effect for the right-censored events to capture their own dependence. The proposed models allow semi-competing risks and are reminiscent of selection models for joint modeling of survival and longitudinal data (Hogan and Laird 1997).

We estimate the model parameters through nonparametric maximum likelihood estimation, under which the baseline hazard functions are completely nonparametric. We develop a simple EM algorithm that converges stably for arbitrary sample sizes, even with time-dependent covariates. We show that the resulting estimators are consistent and the parametric components are asymptotically normal and asymptotically efficient. We also show that the covariance matrix of the parametric components can be estimated consistently with profile likelihood or nonparametric bootstrap. We pay special attention to the estimation of the conditional distribution function given the event history, which can be used to predict disease occurrence dynamically. Finally, we assess the performance of the proposed numerical and inferential procedures through extensive simulation studies and provide a substantive application to the ARIC data on diabetes, hypertension, stroke, MI, and death.

CONTACT Donglin Zeng dzeng@bios.unc.edu Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.

2. Methods

2.1. Data, Models, and Likelihood

Suppose that there are K_1 types of asymptomatic events occurring at times $T_1, \ldots, T_{K_1}$ and K_2 types of symptomatic events occurring at times $T_{K_1+1}, \ldots, T_K$, where $K = K_1 + K_2$. Let $\boldsymbol{X}_k(\cdot)$ be a p-vector of possibly time-dependent external covariates for the event time T_k. For $k = 1, \ldots, K_1$, the hazard function of T_k conditional on covariate $\boldsymbol{X}_k$ and random effect b_1 is given by

$$\lambda_k(t; \boldsymbol{X}_k, b_1) = e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_k(t)+b_1}\lambda_k(t), \tag{1}$$

where $\boldsymbol{\beta}$ is a set of unknown regression parameters, $\lambda_k(\cdot)$ is an arbitrary baseline hazard function, and b_1 is a latent normal random variable with mean zero and variance σ_1^2. For $k = K_1 + 1, \ldots, K$, the hazard function of T_k conditional on covariates $\boldsymbol{X}_k$ and random effects b_1 and b_2 is given by

$$\lambda_k(t; \boldsymbol{X}_k, b_1, b_2) = e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_k(t)+\gamma_k b_1+b_2}\lambda_k(t), \tag{2}$$

where $\lambda_k(\cdot)$ is an arbitrary baseline hazard function, $\boldsymbol{\gamma} \equiv (\gamma_{K_1+1}, \ldots, \gamma_K)^{\mathrm{T}}$ is a set of unknown coefficients, and b_2 is a latent normal random variable with mean zero and variance σ_2^2. Write $\boldsymbol{\Sigma} = (\sigma_1^2, \sigma_2^2)$. By letting $\boldsymbol{X}_k$ depend on k, models (1) and (2) allow the regression parameters to be different among the K events by appropriate definitions of dummy variables; see Lin (1994).

We implicitly assume that K_1 and K_2 are greater than one; otherwise, some of the parameters need to be fixed to ensure identifiability. For example, if $K_1 = K_2 = 1$, we require $\sigma_2^2 = 0$ and $\gamma_1 = 1$; if $K_1 > 1$ and $K_2 = 1$, we require $\sigma_2^2 = 0$; and if $K_1 = 1$ and $K_2 > 1$, we require one of the γ_k's to be 1. In the last scenario, we may set different γ_k to 1 and choose the model that yields the largest value of the likelihood function.

Remark 1. The random effects b_1 and b_2 characterize the underlying health conditions for the asymptomatic and symptomatic events, respectively. The random effect for the asymptomatic events affects the kth symptomatic event through the unknown coefficient γ_k. For example, in the ARIC study, b_1 represents the common pathways for diabetes and hypertension, such as obesity, inflammation, oxidative stress, and insulin resistance, which also serve as potential risk factors for MI, stroke, and death. The random effect b_2 represents the underlying propensity for major cardiovascular diseases and death.

Suppose that the asymptomatic event time T_k $(k = 1, \ldots, K_1)$ is monitored at a sequence of positive time points $U_{k1} < \cdots < U_{k,M_k}$ and is known to lie in the interval $(L_k, R_k]$, where $L_k = \max\{U_{kl} : U_{kl} < T_k, l = 0, \ldots, M_k\}$, and $R_k = \min\{U_{kl} : U_{kl} \geq T_k, l = 1, \ldots, M_k + 1\}$, with $U_{k0} = 0$ and $U_{k,M_k+1} = \infty$. Let C_k denote the censoring time on the symptomatic event time T_k $(k = K_1 + 1, \ldots, K)$ such that we observe $Y_k = \min(T_k, C_k)$ and $\Delta_k = I(T_k \leq C_k)$, where $I(\cdot)$ is the indicator function. For a random sample of n subjects, the data consist of $\{\mathcal{O}_i : i = 1, \ldots, n\}$, where

$$\mathcal{O}_i = \{L_{ik}, R_{ik}, \boldsymbol{X}_{ik}(\cdot) : k = 1, \ldots, K_1\} \cup \times \{Y_{ik}, \Delta_{ik}, \boldsymbol{X}_{ik}(\cdot) : k = K_1 + 1, \ldots, K\}.$$

We assume that $\{U_{ikl} : k = 1, \ldots, K_1; l = 1, \ldots, M_{ik}\}$ and $\{C_{ik} : k = K_1 + 1, \ldots, K\}$ are independent of $\{T_{ik} : k = 1, \ldots, K\}$ and $\boldsymbol{b}_i \equiv (b_{i1}, b_{i2})$ conditional on $\{\boldsymbol{X}_{ik}(\cdot) : k = 1, \ldots, K\}$. Then, the likelihood concerning the parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ and $\mathcal{A} \equiv (\Lambda_1, \ldots, \Lambda_K)$ is

$$\begin{aligned}
&\prod_{i=1}^{n} \int_{\boldsymbol{b}_i} \prod_{k=1}^{K_1} \left[\exp\left\{ -\int_0^{L_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ik}(s)+b_{i1}} d\Lambda_k(s) \right\} \right. \\
&\quad \left. - \exp\left\{ -\int_0^{R_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ik}(s)+b_{i1}} d\Lambda_k(s) \right\} \right] \\
&\quad \times \prod_{k=K_1+1}^{K} \left[\left\{ e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ik}(Y_{ik})+\gamma_k b_{i1}+b_{i2}} \lambda_k(Y_{ik}) \right\}^{\Delta_{ik}} \right. \\
&\quad \left. \times \exp\left\{ -\int_0^{Y_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ik}(s)+\gamma_k b_{i1}+b_{i2}} d\Lambda_k(s) \right\} \right] \psi(\boldsymbol{b}_i; \boldsymbol{\Sigma}) d\boldsymbol{b}_i,
\end{aligned}$$

where $\psi(\boldsymbol{b}_i; \boldsymbol{\Sigma}) = \prod_{j=1}^{2} \phi(b_{ij}; \sigma_j^2)$, $\phi(b_{ij}; \sigma_j^2) = (2\pi\sigma_j^2)^{-1/2} \exp\{-b_{ij}^2/(2\sigma_j^2)\}$, $\Lambda_k(t) = \int_0^t \lambda_k(s) ds$, and $\exp\{-\int_0^{\infty} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ik}(s)+b_{i1}} d\Lambda_k(s)\} = 0$.

In some studies, one of the symptomatic events is terminal (e.g., death), such that we have a semi-competing risks set-up (Fine, Jiang, and Chappell 2001), where the occurrence of the terminal event precludes the development of the other events but not vice versa. Without loss of generality, suppose that the Kth event is terminal. Then the monitoring times for T_k $(k \leq K_1)$ consist of the U_{kl}'s that are smaller than T_K, and the censoring time for T_k $(k = K_1 + 1, \ldots, K - 1)$ is $\min(C_k, T_K)$. Conditional on (b_1, b_2), the event times $T_1, \ldots, T_{K-1}$ are mutually independent and are independent of the monitoring times and censoring times. Thus, for any set $\mathcal{S}_k$ that may depend on the monitoring times and censoring times, the joint probability of $T_1 \in \mathcal{S}_1, \ldots, T_{K-1} \in \mathcal{S}_{K-1}$ conditional on (b_1, b_2) is equal to $\prod_{k=1}^{K-1} P(T_k \in \mathcal{S}_k | b_1, b_2)$ with $\mathcal{S}_k$ as a deterministic set. Therefore, the likelihood remains the same as before.

2.2. Estimation Procedure

We adopt the nonparametric maximum likelihood estimation approach. For $k = 1, \ldots, K_1$, let $0 = t_{k0} < t_{k1} < t_{k2} < \cdots < t_{k,m_k} < \infty$ be the ordered sequence of all L_{ik} and R_{ik} with $R_{ik} < \infty$. For $k = K_1 + 1, \ldots, K$, let $0 = t_{k0} < t_{k1} < t_{k2} < \cdots < t_{k,m_k} < \infty$ be the ordered sequence of all Y_{ik} with $\Delta_{ik} = 1$. The estimator for Λ_k $(k = 1, \ldots, K)$ is a step function that jumps only at $t_{k1}, \ldots, t_{k,m_k}$ with respective jump sizes $\boldsymbol{\lambda}_k \equiv (\lambda_{k1}, \ldots, \lambda_{k,m_k})$. We maximize the objective function

$$\begin{aligned}
L_n(\boldsymbol{\theta}, \mathcal{A}) = &\prod_{i=1}^{n} \int_{\boldsymbol{b}_i} \left\{ \prod_{k=1}^{K_1} g_{ik}^{(1)}(b_{i1}; \boldsymbol{\beta}, \boldsymbol{\lambda}_k) \right\} \\
&\times \left\{ \prod_{k=K_1+1}^{K} g_{ik}^{(2)}(\boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\lambda}_k) \right\} \psi(\boldsymbol{b}_i; \boldsymbol{\Sigma}) d\boldsymbol{b}_i,
\end{aligned}$$

over $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$, where

$$g_{ik}^{(1)}(b_{i1}; \boldsymbol{\beta}, \boldsymbol{\lambda}_k) = \exp\left(-\sum_{t_{kl} \le L_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}} \lambda_{kl}\right) - I(R_{ik} < \infty) \exp\left(-\sum_{t_{kl} \le R_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}} \lambda_{kl}\right),$$

$$g_{ik}^{(2)}(\boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\lambda}_k) = \left[\Lambda_k\{Y_{ik}\} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ik}(Y_{ik}) + \gamma_k b_{i1} + b_{i2}}\right]^{\Delta_{ik}} \times \exp\left(-\sum_{t_{kl} \le Y_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + \gamma_k b_{i1} + b_{i2}} \lambda_{kl}\right),$$

$\boldsymbol{X}_{ikl} = \boldsymbol{X}_{ik}(t_{kl})$ for $k = 1, \ldots, K$ and $l = 1, \ldots, m_k$, and $\Lambda_k\{Y_{ik}\}$ is the jump size of Λ_k at Y_{ik}.

Direct maximization of the objective function is difficult due to the lack of analytical expressions for $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$. We introduce latent Poisson random variables to form a likelihood equivalent to the objective function such that the maximum likelihood estimators can be easily obtained via a simple EM algorithm. For $k = 1, \ldots, K_1$, we denote $R_{ik}^* = I(R_{ik} = \infty) L_{ik} + I(R_{ik} < \infty) R_{ik}$ and introduce independent Poisson random variables W_{ikl} $(l = 1, \ldots, m_k, t_{kl} \le R_{ik}^*)$ with means $\lambda_{kl} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1})$. Conditional on b_{i1}, the likelihood function of $\{W_{ikl}; l = 1, \ldots, m_k, t_{kl} \le R_{ik}^*\}$ is

$$\prod_{l=1, t_{kl} \le R_{ik}^*}^{m_k} \left\{\frac{1}{W_{ikl}!} \left(\lambda_{kl} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}}\right)^{W_{ikl}} \exp\left(-\lambda_{kl} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}}\right)\right\}.$$

Let $A_{ik} = \sum_{t_{kl} \le L_{ik}} W_{ikl}$ and $B_{ik} = I(R_{ik} < \infty) \sum_{L_{ik} < t_{kl} \le R_{ik}} W_{ikl}$. The observed-data likelihood for $A_{ik} = 0$ and $B_{ik} > 0$ given b_{i1} is equal to

$$\exp\left(-\sum_{t_{kl} \le L_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}} \lambda_{kl}\right) - I(R_{ik} < \infty) \exp\left(-\sum_{t_{kl} \le R_{ik}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}} \lambda_{kl}\right),$$

which is the same as $g_{ik}^{(1)}(b_{i1}; \boldsymbol{\beta}, \boldsymbol{\lambda}_k)$. Therefore, the objective function $L_n(\boldsymbol{\theta}, \mathcal{A})$ can be viewed as the observed-data likelihood for $\{A_{ik} = 0, B_{ik} > 0 : i = 1, \ldots, n; k = 1, \ldots, K_1\} \cup \{Y_{ik}, \Delta_{ik} : i = 1, \ldots, n; k = K_1 + 1, \ldots, K\}$ with $(W_{ikl}, \boldsymbol{b}_i)$ $(i = 1, \ldots, n; k = 1, \ldots, K_1; l = 1, \ldots, m_k, t_{kl} \le R_{ik}^*)$ as latent variables. In view of the foregoing results, we propose an EM algorithm treating W_{ikl} and $\boldsymbol{b}_i$ as missing data.

In the M-step, we maximize the conditional expectation of the complete-data log-likelihood given the observed data so as to update the parameters. In particular, the conditional expectation of the complete-data log-likelihood is

$$\sum_{i=1}^{n} \widehat{E}\left(\sum_{k=1}^{K_1}\left[\sum_{l=1}^{m_k} I(t_{kl} \le R_{ik}^*)\left\{W_{ikl}\left(\log \lambda_{kl} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}\right) - \lambda_{kl} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1})\right\}\right] + \sum_{k=K_1+1}^{K}\left[\Delta_{ik}\left\{\log \Lambda_k\{Y_{ik}\} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ik}(Y_{ik}) + \gamma_k b_{i1} + b_{i2}\right\} - \sum_{t_{kl} \le Y_{ik}} \lambda_{kl} \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + \gamma_k b_{i1} + b_{i2}\right)\right]\right), \tag{3}$$

where $\widehat{E}(\cdot)$ denotes the conditional expectation given the observed data $\widetilde{\mathcal{O}}_i$ $(i = 1, \ldots, n)$, with $\widetilde{\mathcal{O}}_i = \{A_{ik} = 0, B_{ik} > 0, \boldsymbol{X}_{ik}(\cdot) : k = 1, \ldots, K_1\} \cup \{Y_{ik}, \Delta_{ik}, \boldsymbol{X}_{ik}(\cdot) : k = K_1 + 1, \ldots, K\}$. To update the parameters, we first differentiate (3) with respect to λ_{kl} $(k = 1, \ldots, K; l = 1, \ldots, m_k)$ to obtain the updating formulas for $\boldsymbol{\lambda}_k$:

$$\lambda_{kl} = \frac{\sum_{i=1}^{n} I(t_{kl} \le R_{ik}^*) \widehat{E}(W_{ikl})}{\sum_{i=1}^{n} I(t_{kl} \le R_{ik}^*) \widehat{E}\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}\right)\right\}} \tag{4}$$

for $k = 1, \ldots, K_1$ and $l = 1, \ldots, m_k$ and

$$\lambda_{kl} = \frac{\sum_{i=1}^{n} \Delta_{ik} I(Y_{ik} = t_{kl})}{\sum_{i=1}^{n} I(Y_{ik} \ge t_{kl}) \widehat{E}\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + \gamma_k b_{i1} + b_{i2}\right)\right\}} \tag{5}$$

for $k = K_1 + 1, \ldots, K$ and $l = 1, \ldots, m_k$. We then update $\boldsymbol{\beta}$ by solving the equation

$$\sum_{i=1}^{n}\left\{\sum_{k=1}^{K_1}\sum_{l=1}^{m_k} \widehat{E}(W_{ikl}) I(t_{kl} \le R_{ik}^*)\left[\boldsymbol{X}_{ikl} - \frac{\sum_{j=1}^{n} \boldsymbol{X}_{jkl} I(t_{kl} \le R_{jk}^*) \widehat{E}\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jkl} + b_{j1}\right)\right\}}{\sum_{j=1}^{n} I(t_{kl} \le R_{jk}^*) \widehat{E}\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jkl} + b_{j1}\right)\right\}}\right] + \sum_{k=K_1+1}^{K} \Delta_{ik}\left(\boldsymbol{X}_{ik}(Y_{ik}) - \frac{\sum_{j=1}^{n} I(Y_{jk} \ge Y_{ik}) \boldsymbol{X}_{jk}(Y_{ik}) \widehat{E}\left[\exp\left\{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jk}(Y_{ik}) + \gamma_k b_{j1} + b_{j2}\right)\right\}\right]}{\sum_{j=1}^{n} I(Y_{jk} \ge Y_{ik}) \widehat{E}\left[\exp\left\{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jk}(Y_{ik}) + \gamma_k b_{j1} + b_{j2}\right\}\right]}\right)\right\} = \mathbf{0}$$

and update γ_k by solving the equation

$$\sum_{i=1}^{n} \Delta_{ik}\left(\widehat{E}(b_{i1}) - \frac{\sum_{j=1}^{n} I(Y_{jk} \ge Y_{ik}) \widehat{E}\left[b_{j1} \exp\left\{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jk}(Y_{ik}) + \gamma_k b_{j1} + b_{j2}\right)\right\}\right]}{\sum_{j=1}^{n} I(Y_{jk} \ge Y_{ik}) \widehat{E}\left[\exp\left\{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{jk}(Y_{ik}) + \gamma_k b_{j1} + b_{j2}\right\}\right]}\right) = 0.$$

The two equations are obtained by differentiating (3) with respect to $\boldsymbol{\beta}_k$ or γ_k and replacing λ_{kl} by the right hand side of (4) or (5). Finally, we update σ_j^2 by $\sigma_j^2 = \sum_{i=1}^n \widehat{E}(b_{ij}^2)/n$ for $j = 1, 2$.

In the E-step, we evaluate the conditional expectation of W_{ikl} $(k = 1, \ldots, K_1; l = 1, \ldots, m_k, t_{kl} \le R_{ik}^*)$ and the other terms of $\boldsymbol{b}_i$ given the observed data $\widetilde{\mathcal{O}}_i$ for $i = 1, \ldots, n$. Specifically, the conditional expectation of W_{ikl} $(k = 1, \ldots, K_1; l = 1, \ldots, m_k, t_{kl} \le R_{ik}^*)$ given $\widetilde{\mathcal{O}}_i$ and $\boldsymbol{b}_i$ is

$$I(L_{ik} < t_{kl} \le R_{ik} < \infty) \frac{\lambda_{kl} \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl} + b_{i1}\right)}{1 - \exp\left(-\sum_{L_{ik} < t_{kl'} \le R_{ik}} \lambda_{kl'} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ikl'} + b_{i1}}\right)}.$$

Note that the density of $\boldsymbol{b}_i$ given $\widetilde{\mathcal{O}}_i$ is proportional to $\{\prod_{k=1}^{K_1} g_{ik}^{(1)}(b_{i1}; \boldsymbol{\beta}, \boldsymbol{\lambda}_k)\} \times \{\prod_{k=K_1+1}^{K} g_{ik}^{(2)}(\boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\lambda}_k)\} \psi(\boldsymbol{b}_i; \boldsymbol{\Sigma})$. We evaluate the conditional expectation of W_{ikl} and the other terms through numerical integration over $\boldsymbol{b}_i$ with Gauss–Hermite quadratures.

We iterate between the E-step and M-step until convergence. In the M-step, the high-dimensional nuisance parameters λ_{kl} $(k = 1, \ldots, K; l = 1, \ldots, m_k)$ are calculated explicitly, such that inversion of high-dimensional matrices is avoided. We denote the final estimators for $\boldsymbol{\theta}$ and $\mathcal{A}$ as $\widehat{\boldsymbol{\theta}} \equiv (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Sigma}})$ and $\widehat{\mathcal{A}} \equiv (\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_K)$.

2.3. Asymptotic Theory

We establish the asymptotic properties of $(\widehat{\boldsymbol{\theta}}, \widehat{\mathcal{A}})$ under the following regularity conditions.

Condition 1. The true value of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_0 \equiv (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_0)$, belongs to the interior of a known compact set $\Theta \equiv \mathcal{B} \times \mathcal{G} \times \mathcal{S}$, where $\mathcal{B} \subset \mathbb{R}^p$, $\mathcal{G} \subset \mathbb{R}^{K_2}$, and $\mathcal{S} \subset (0, \infty) \times (0, \infty)$.

Condition 2. For $k = 1, \ldots, K$, the true value $\Lambda_{k0}(\cdot)$ of $\Lambda_k(\cdot)$ is strictly increasing and continuously differentiable in $[0, \tau_k]$ with $\Lambda_{k0}(0) = 0$.

Condition 3. For $k = 1, \ldots, K_1$, the monitoring times have finite support $\mathcal{U}_k$ with the least upper bound τ_k. The number of potential monitoring times M_k is positive with $E(M_k) < \infty$. There exists a positive constant η such that $\Pr\{\min_{1 \le k \le K_1, 0 \le m < M_k} (U_{k,m+1} - U_{km}) \ge \eta | M_k, \boldsymbol{X}_k\} = 1$. In addition, there exists a probability measure μ_k in $\mathcal{U}_k$ such that the bivariate distribution function of $(U_{km}, U_{k,m+1})$ conditional on $(M_k, \boldsymbol{X}_k)$ is dominated by $\mu_k \times \mu_k$ and its Radon–Nikodym derivative, denoted by $\widetilde{f}_{km}(u, v; M_k, \boldsymbol{X}_k)$, can be expanded to a positive and twice-continuously differentiable function in the set $\{(u, v) : 0 \le u \le \tau_k, 0 \le v \le \tau_k, v - u \ge \eta\}$.

Condition 4. For $k = K_1 + 1, \ldots, K$, let τ_k denote the study duration time and $\mathcal{U}_k = [0, \tau_k]$. There exists a positive constant δ such that $\Pr(C_k \ge \tau_k | \boldsymbol{X}_k) = \Pr(C_k = \tau_k | \boldsymbol{X}_k) \ge \delta$ almost surely.

Condition 5. With probability 1, $\boldsymbol{X}_k(\cdot)$ has bounded total variation in $\mathcal{U}_k$. If there exists a constant vector $\boldsymbol{a}_1$ and a deterministic function $a_{2k}(t)$ such that $\boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{X}_k(t) + a_{2k}(t) = 0$ for any $t \in \mathcal{U}_k$ and any $k \in \{1, \ldots, K\}$ with probability 1, then $\boldsymbol{a}_1 = \boldsymbol{0}$ and $a_{2k}(t) = 0$ for any $t \in \mathcal{U}_k$ and any $k \in \{1, \ldots, K\}$.

Remark 2. Conditions 1, 2, and 5 are standard conditions for failure time regression with time-dependent covariates. Condition 3 pertains to the joint distribution of monitoring times of the asymptomatic events. It requires that two adjacent monitoring times are separated by at least η; otherwise, the data may contain exact observations, which require a different theoretical treatment. The dominating measure μ_k is chosen as the Lebesgue measure if the monitoring times are continuous random variables and as the counting measure if monitorings occur only at a finite number of time points. The number of potential monitoring times M_k can be fixed or random, is possibly different among study subjects and event types, and is allowed to depend on covariates. Condition 4 implies that there is a positive probability for the kth symptomatic event to be observed in the time interval $[0, \tau_k]$.

We state the strong consistency of $(\widehat{\boldsymbol{\theta}}, \widehat{\mathcal{A}})$ and the weak convergence of $\widehat{\boldsymbol{\theta}}$ in two theorems.

Theorem 1. Under Conditions 1–5, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \to_{a.s.} 0$, and $\|\widehat{\Lambda}_k - \Lambda_{k0}\|_{l^\infty(\mathcal{U}_k)} \to_{a.s.} 0$, where $\|\cdot\|_{l^\infty(\mathcal{U}_k)}$ denotes the supremum norm on $\mathcal{U}_k$ for $k = 1, \ldots, K$.

Theorem 2. Under Conditions 1–5, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges weakly to a $(p + K_2 + 2)$-dimensional zero-mean normal random vector with a covariance matrix that attains the semiparametric efficiency bound.

The proofs of all theorems are provided in the Section S.1 of the supplementary materials.

We propose two approaches to estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$. The first approach makes use of the profile likelihood (Murphy and Van der Vaart 2000). Specifically, we define the profile log-likelihood function

$$pl_n(\boldsymbol{\theta}) = \max_{\mathcal{A} \in \mathcal{C}_1 \times \cdots \times \mathcal{C}_K} \log L_n(\boldsymbol{\theta}, \mathcal{A}),$$

where $\mathcal{C}_k$ is the set of step functions with nonnegative jumps at t_{kl} $(k = 1, \ldots, K; l = 1, \ldots, m_k)$. We estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the inverse of

$$\sum_{i=1}^{n} \begin{pmatrix} \dfrac{pl_i(\widehat{\boldsymbol{\theta}} + h_n e_1) - pl_i(\widehat{\boldsymbol{\theta}})}{h_n} \\ \vdots \\ \dfrac{pl_i(\widehat{\boldsymbol{\theta}} + h_n e_{p+K_2+2}) - pl_i(\widehat{\boldsymbol{\theta}})}{h_n} \end{pmatrix}^{\otimes 2},$$

where pl_i is the ith subject's contribution to pl_n, e_j is the jth canonical vector in $\mathbb{R}^{p+K_2+2}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$, and h_n is a constant of order $n^{-1/2}$. To evaluate the profile likelihood, we use the EM algorithm of Section 2.2 but only update $\Lambda_1, \ldots, \Lambda_K$ in the M-step.

Alternatively, we approximate the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ by bootstrapping the observations. In particular, we draw a simple random sample of size n with replacement from the observed data $\{\mathcal{O}_i : i = 1, \ldots, n\}$. Let $\widehat{\boldsymbol{\theta}}^*$ be the estimator of $\boldsymbol{\theta}$ in the bootstrap sample. The empirical distribution of $\widehat{\boldsymbol{\theta}}^*$ can be used to approximate the distribution of $\widehat{\boldsymbol{\theta}}$. Confidence intervals for $\boldsymbol{\theta}_0$ can be constructed by the Wald method (with the variance of $\widehat{\boldsymbol{\theta}}^*$) or from the empirical percentiles of $\widehat{\boldsymbol{\theta}}^*$.

The following theorem states the asymptotic properties of $\widehat{\boldsymbol{\theta}}^*$, thereby validating the bootstrap procedure.

Theorem 3. Under Conditions 1−5, the conditional distribution of $n^{1/2}(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}})$ given the data converges weakly to the asymptotic distribution of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

2.4. Dynamic Prediction

Given the fitted joint model, we can predict future events by updating the event history. For a subject with covariates $\boldsymbol{X}$, let $\mathcal{O}(t)$ denote the event history at time $t > 0$, which includes the interval-censored observations of the asymptomatic events $\{L_k(t), R_k(t) : k = 1, \ldots, K_1\}$, and the right-censored observations of the symptomatic events $\{Y_k(t), \Delta_k(t) : k = K_1 + 1, \ldots, K\}$.

If no event history is available, the density of the random effect $\boldsymbol{b}$ can be estimated by $\psi(\boldsymbol{b}; \widehat{\boldsymbol{\Sigma}})$. We estimate the survival function of T_k, denoted by $P(T_k \geq t|\boldsymbol{X})$, by

$$\int_{\boldsymbol{b}} s_k(t; \boldsymbol{X}, \boldsymbol{b}) \psi(\boldsymbol{b}; \widehat{\boldsymbol{\Sigma}}) d\boldsymbol{b},$$

where

$$s_k(t; \boldsymbol{X}, \boldsymbol{b}) = \begin{cases} \exp\left\{-\int_0^t e^{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_k(u) + b_1} d\widehat{\Lambda}_k(u)\right\} & k = 1, \ldots, K_1 \\ \exp\left\{-\int_0^t e^{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_k(u) + \widehat{\gamma}_k b_1 + b_2} d\widehat{\Lambda}_k(u)\right\} & k = K_1 + 1, \ldots, K \end{cases},$$

and the integral is evaluated by numerical integration with Gauss–Hermite quadratures. Here, the function $s_k(t; \boldsymbol{X}, \boldsymbol{b})$ can be interpreted as the conditional survival probability of T_k at time t given $\boldsymbol{b}$ and $\boldsymbol{X}$.

In the semi-competing risks set-up, where one of the symptomatic events is terminal, it is more meaningful to use the cumulative incidence function to predict the event time of interest. Without loss of generality, we assume the Kth event is terminal. The cumulative incidence function of the kth event $(k = 1, \ldots, K-1)$ is given by

$$\begin{aligned} &P(T_k \leq t, T_k \leq T_K|\boldsymbol{X}) \\ &= \int_{\boldsymbol{b}} \{P(T_k \leq t \leq T_K|\boldsymbol{X}, \boldsymbol{b}) + P(T_k \leq T_K < t|\boldsymbol{X}, \boldsymbol{b})\} \psi(\boldsymbol{b}; \boldsymbol{\Sigma}) d\boldsymbol{b} \\ &= \int_{\boldsymbol{b}} \Big\{ P(T_k \leq t|T_K \geq t, \boldsymbol{X}, \boldsymbol{b}) P(T_K \geq t|\boldsymbol{X}, \boldsymbol{b}) \\ &\quad + \int_0^t P(T_k \leq u|T_K = u, \boldsymbol{X}, \boldsymbol{b}) dP(T_K \leq u|\boldsymbol{X}, \boldsymbol{b}) \Big\} \psi(\boldsymbol{b}; \boldsymbol{\Sigma}) d\boldsymbol{b}, \end{aligned}$$

which can be estimated by

$$\begin{aligned} &\int_{\boldsymbol{b}} \Big[\{1 - s_k(t; \boldsymbol{X}, \boldsymbol{b})\} s_K(t; \boldsymbol{X}, \boldsymbol{b}) \\ &\quad + \int_0^t \{1 - s_k(u; \boldsymbol{X}, \boldsymbol{b})\} s_K(u; \boldsymbol{X}, \boldsymbol{b}) e^{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_K(u) + \widehat{\gamma}_K b_1 + b_2} d\widehat{\Lambda}_K(u) \Big] \\ &\quad \times \psi(\boldsymbol{b}; \widehat{\boldsymbol{\Sigma}}) d\boldsymbol{b}. \end{aligned}$$

Here, the function $s_k(t; \boldsymbol{X}, \boldsymbol{b})$ can be interpreted as the conditional survival probability of T_k at time t given $T_K \geq t$, $\boldsymbol{b}$, and $\boldsymbol{X}$.

At time $t_0 > 0$, we update the posterior density of $\boldsymbol{b}$ given the event history $\mathcal{O}(t_0)$ so as to perform dynamic prediction. Note that the posterior density of $\boldsymbol{b}$ is proportional to

$$\begin{aligned} J(\boldsymbol{b}; t_0, \boldsymbol{X}) \equiv &\prod_{k=1}^{K_1} \{s_k(L_k(t_0); \boldsymbol{X}, \boldsymbol{b}) - s_k(R_k(t_0); \boldsymbol{X}, \boldsymbol{b})\} \\ &\times \prod_{k=K_1+1}^{K} \Big(s_k(Y_k(t_0); \boldsymbol{X}, \boldsymbol{b}) \\ &\times \left[\widehat{\Lambda}_k\{Y_k(t_0)\} e^{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_k\{Y_k(t_0)\} + \widehat{\gamma}_k b_1 + b_2} \right]^{\Delta_k(t_0)} \Big) \psi(\boldsymbol{b}; \widehat{\boldsymbol{\Sigma}}). \end{aligned}$$

If the subject has not developed the kth event or the terminal event by time t_0, that is, $Y_k(t_0) = Y_K(t_0) = t_0$ and $\Delta_k(t_0) = \Delta_K(t_0) = 0$, we estimate the conditional cumulative incidence function of the kth event, $P(T_k \leq t, T_k \leq T_K|\mathcal{O}(t_0), \boldsymbol{X})$, by

$$\begin{aligned} &\int_{\boldsymbol{b}} \frac{J(\boldsymbol{b}; t_0, \boldsymbol{X})}{s_k(t_0; \boldsymbol{X}, \boldsymbol{b}) s_K(t_0; \boldsymbol{X}, \boldsymbol{b}) \int_{\boldsymbol{b}'} J(\boldsymbol{b}'; t_0, \boldsymbol{X}) d\boldsymbol{b}'} \\ &\quad \times \Big[\{s_k(t_0; \boldsymbol{X}, \boldsymbol{b}) - s_k(t; \boldsymbol{X}, \boldsymbol{b})\} s_K(t; \boldsymbol{X}, \boldsymbol{b}) \\ &\quad + \int_{t_0}^t \{s_k(t_0; \boldsymbol{X}, \boldsymbol{b}) - s_k(u; \boldsymbol{X}, \boldsymbol{b})\} \\ &\quad \times s_K(u; \boldsymbol{X}, \boldsymbol{b}) e^{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_K(u) + \widehat{\gamma}_K b_1 + b_2} d\widehat{\Lambda}_K(u) \Big] d\boldsymbol{b}. \end{aligned}$$

In practice, it is desirable to identify subjects who are at increased risk as the event history is accumulating. In the same vein as the risk score under the standard proportional hazards model, we use the risk score $\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_k(t_0) + \widehat{\gamma}_k \widehat{b}_1(t_0) + \widehat{b}_2(t_0)$ to dynamically predict the kth event $(k = K_1 + 1, \ldots, K)$, where $\widehat{\boldsymbol{b}}(t_0) \equiv (\widehat{b}_1(t_0), \widehat{b}_2(t_0))$ is a suitable estimator of $\boldsymbol{b}$ given the event history $\mathcal{O}(t_0)$. The estimator $\widehat{\boldsymbol{b}}(t_0)$ can be the posterior mean or mode of $\boldsymbol{b}$ or an imputed value from the posterior distribution. For example, the risk score using the posterior mean is given by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_k(t_0) + \frac{\int_{\boldsymbol{b}} (\widehat{\gamma}_k b_1 + b_2) J(\boldsymbol{b}; t_0, \boldsymbol{X}) d\boldsymbol{b}}{\int_{\boldsymbol{b}} J(\boldsymbol{b}; t_0, \boldsymbol{X}) d\boldsymbol{b}}.$$

The risk score quantifies the subject-specific risk and can be very useful to both individual patients and clinicians when making decisions about lifestyle modifications and preventive medical treatments.

3. Simulation Studies

We conducted simulation studies to assess the performance of the proposed methods. We considered one time-independent covariate $X_1 \sim \mathrm{Unif}(0, 1)$ and one time-dependent covariate $X_2(t) = I(t \leq V)B_1 + I(t > V)B_2$, where B_1 and B_2 are independent Bernoulli(0.5), $V \sim \mathrm{Unif}(0, \tau)$, and $\tau = 4$. We considered two asymptomatic events and two symptomatic events. We set $\boldsymbol{X}_k = e_k \otimes (X_1, X_2)^{\mathrm{T}}$, where e_k is the kth canonical vector in $\mathbb{R}^4$, and $\otimes$ denotes the Kronecker product. We set $\boldsymbol{\beta} = (0.5, 0.4, 0.5, -0.2, -0.5, 0.5, -0.5, 0.5)^{\mathrm{T}}$, $\Lambda_1(t) = 0.5t$, $\Lambda_k(t) = \log\{1 + t/(k-1)\}$ for $k = 2, 3, 4$, $\gamma_3 = \gamma_4 = 0.25$, and $\sigma_1^2 = \sigma_2^2 = 1$. Both symptomatic events were censored by $C \sim \mathrm{Unif}(2\tau/3, \tau)$, such that the censoring rates are 33% and 39%, respectively. The series of monitoring times were generated sequentially, with $U_m = U_{m-1} + 0.1 + \mathrm{Unif}(0, 0.5)$

Table 1. Summary statistics for the simulation studies without a terminal event.

	$n = 100$						$n = 200$					
			Profile		Bootstrap				Profile		Bootstrap	
	Bias	SE	SEE	CP	SEE	CP	Bias	SE	SEE	CP	SEE	CP
β_{11}	0.006	0.585	0.597	0.961	0.627	0.967	0.027	0.405	0.399	0.947	0.412	0.953
β_{12}	0.029	0.327	0.321	0.941	0.348	0.960	0.019	0.222	0.216	0.949	0.228	0.953
β_{21}	0.015	0.623	0.609	0.946	0.648	0.963	0.014	0.410	0.409	0.951	0.424	0.958
β_{22}	−0.005	0.341	0.329	0.940	0.355	0.962	−0.002	0.225	0.222	0.951	0.233	0.961
β_{31}	−0.022	0.617	0.635	0.957	0.610	0.949	−0.004	0.416	0.428	0.960	0.420	0.948
β_{32}	−0.002	0.319	0.338	0.965	0.322	0.949	0.009	0.221	0.229	0.958	0.222	0.947
β_{41}	−0.012	0.623	0.651	0.969	0.629	0.955	0.006	0.449	0.440	0.947	0.431	0.942
β_{42}	0.004	0.330	0.348	0.967	0.332	0.950	−0.001	0.231	0.235	0.955	0.229	0.945
γ_3	−0.012	0.227	0.252	0.979	0.260	0.971	−0.012	0.159	0.171	0.962	0.170	0.960
γ_4	−0.013	0.237	0.260	0.976	0.266	0.972	−0.016	0.162	0.173	0.966	0.177	0.963
σ_1^2	0.062	0.445	0.751	0.978	0.493	0.956	0.031	0.317	0.482	0.982	0.318	0.946
σ_2^2	−0.102	0.413	0.510	0.993	0.482	0.971	−0.062	0.297	0.335	0.987	0.312	0.974

NOTE: SE and SEE denote, respectively, the empirical standard error and mean standard error estimator. CP stands for the empirical coverage probability of the 95% confidence interval based on the Wald method for the profile-likelihood approach and the 95% symmetric confidence interval for the bootstrap approach. For $\gamma_3, \gamma_4, \sigma_1^2$, and σ_2^2, bias and SEE are based on the median instead of the mean, and SE is based on the mean absolute deviation. For σ_1^2 and σ_2^2, the confidence intervals are based on the log transformation.

for $m \geq 1$ and $U_0 = 0$. The last monitoring time is the largest U_m that is smaller than C. We set $n = 100$ or 200 and simulated 2000 replicates. For each dataset, we applied the proposed EM algorithm by setting the initial value of $\boldsymbol{\beta}$ to $\mathbf{0}$, the initial values of γ_k and σ_k^2 to 1 and the initial value of λ_{kl} to $1/m_k$. We used 20 quadrature points for integration with respect to each random effect and set the convergence threshold to 10^{-3}. For variance estimation, we set $h_n = 5n^{-1/2}$ for profile likelihood and used 100 bootstrap samples.

Table 1 summarizes the simulation results. The biases for all parameter estimators are small, especially for $n = 200$. Both the profile-likelihood and bootstrap variance estimators for $\widehat{\boldsymbol{\beta}}$ are accurate, especially for $n = 200$. Both variance estimators for $\widehat{\boldsymbol{\gamma}}$ tend to overestimate the true variabilities, but the coverage probabilities of the confidence intervals get closer to the nominal level as sample size increases. The profile-likelihood variance estimators for $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ overestimate the true variabilities, while the bootstrap variance estimators for $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ accurately reflect the true variabilities. Figure S.1(a) of the Supplemental Materials shows the estimation of the baseline survival functions with sample size $n = 200$. The estimators are virtually unbiased.

We considered a second setup with an additional terminal event. We set $\boldsymbol{X}_k = e_k \otimes (X_1, X_2)^{\mathrm{T}}$, where e_k is the kth canonical vector in $\mathbb{R}^5$. In addition, we set $\boldsymbol{\beta} = (0.5, 0.4, 0.5, -0.2, -0.5, 0.5, -0.5, 0.5, 0.3, -0.2)^{\mathrm{T}}$, $\Lambda_5(t) = \log(1 + t/4)$, and $\gamma_5 = 0.25$. The terminal event was also censored by C. The censoring rates for the right-censored events are 51%, 58%, and 43%, respectively. The results are shown in Table S.1 and Figure S.1(b) of the supplemental materials. The conclusions are similar to the case of no terminal event.

We assessed the performance of dynamic prediction based on the conditional cumulative incidence function in the setting with a terminal event. Suppose that at the first monitoring time $t_0 = 1$, event 2 has occurred but events 1, 3, and 4 have not. Figure 1 shows the estimation of the baseline cumulative incidence functions (pertaining to $\boldsymbol{X} = \mathbf{0}$) for events 3 and 4 given the event history at time $t_0 = 1$. The estimators slightly underestimate the true values at the right tail, but the biases get smaller as n increases.

To investigate the performance of the proposed dynamic prediction methods under misspecified models, we conducted another set of simulation studies where the event times were

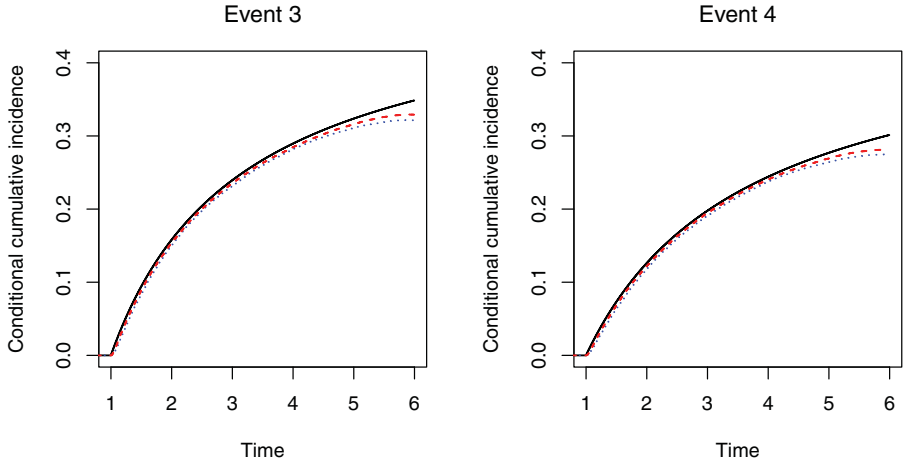

Figure 1. Estimation of the baseline cumulative incidence function conditional on the event history. The solid black curve, dotted blue curve, and dashed red curve pertain, respectively, to the true value and the mean estimates from the proposed method with $n = 100$ and $n = 200$.

Table 2. Estimation results for the regression parameters of the asymptomatic events in the ARIC study.

	Diabetes			Hypertension		
Covariate	Estimate	Std error	*p*-value	Estimate	Std error	*p*-value
Forsyth County, white	−0.5332	0.1817	0.0033	−0.5032	0.0615	<0.0001
Jackson, black	−0.1356	0.1806	0.4530	−0.1075	0.0673	0.1104
Minneapolis, white	−0.9415	0.1802	<0.0001	−0.5747	0.0579	<0.0001
Washington County, white	−0.3778	0.1778	0.0336	−0.3798	0.0592	<0.0001
Age	−0.0093	0.0057	0.1025	0.0166	0.0036	<0.0001
Male	−0.0655	0.0593	0.2694	−0.2329	0.0396	<0.0001
BMI	0.0911	0.0059	<0.0001	0.0254	0.0044	<0.0001
Glucose	0.1075	0.0033	<0.0001	0.0004	0.0023	0.8744
Systolic blood pressure	0.0096	0.0026	0.0003	0.0780	0.0022	<0.0001
Smoker	0.4576	0.0674	<0.0001	0.3134	0.0468	<0.0001

NOTE: The blacks in Forsyth County form the reference group for the cohort × race variables.

generated from the proportional odds, instead of the proportional hazards, models with random effects. As shown in Section S.2 of the supplemental materials, the dynamic prediction is still quite accurate.

4. ARIC Study

ARIC is a perspective epidemiological cohort study conducted in four U.S. communities: Forsyth County, NC; Jackson, MS; Minneapolis, MN; and Washington County, MD. A total of 15792 participants received a baseline examination between 1987 and 1989 and four subsequent examinations in 1990–1992, 1993–1995, 1996–1998, and 2011–2013. At each examination, medical data were collected, such that interval-censored observations for diabetes and hypertension were obtained. The participants were also followed for cardiovascular diseases through reviews of hospital records, such that potentially right-censored observations on MI, stroke, and death were collected.

We related the disease incidence to race, sex, and five baseline risk factors: age, body mass index (BMI), glucose level, systolic blood pressure, and smoking status. Since the Jackson cohort is composed of black subjects only, and neither Minneapolis nor Washington County cohorts contain black subjects, we included the cohort × race indicators as predictors. We excluded subjects with prevalent cases at baseline or missing covariate values to obtain a total of 8728 subjects. During the study, 17.3%, 46.8%, 8.3%, and 5.1% of the subjects developed diabetes, hypertension, MI, and stroke, respectively, while 28.7% died.

We jointly modeled the asymptomatic and symptomatic events in the ARIC study with equations (1) and (2). For variance estimation, we used the profile likelihood approach with $h_n = 5n^{-1/2}$. Tables 2 and 3 show the estimation results for the regression parameters. Several characteristics and baseline risk factors are found to be predictive of the events. Older subjects have higher risks of hypertension, MI, stroke, and death than younger subjects. Males have lower risk of hypertension but higher risks of MI, stroke, and death than females. Smokers have significantly higher risks for all events than non-smokers. In addition, higher baseline BMI increases the risks of diabetes, hypertension, and MI; higher baseline glucose level increases the risks of diabetes, stroke, and death; and higher baseline value of systolic blood pressure increases the risks of all considered events.

The estimation results for the remaining parametric components are shown in Table S.2 of the supplemental materials. The variance components σ_1^2 and σ_2^2 are significantly larger than zero, indicating strong correlation among the asymptomatic events and among the symptomatic events. The parameters γ_{MI}, γ_{Stroke}, and γ_{Death} are also significantly larger than zero, reflecting the strong positive dependence of the symptomatic events on the asymptomatic events. The Akaike information criterion (AIC) for the proposed model is 108852.8. For comparisons, we also fit a model with one random effect shared by all events. The corresponding AIC is 109000.6, and the *p*-value for the likelihood ratio test is less than 0.0001, indicating that the proposed model provides a much better fit to the data than the model with one shared random effect.

To evaluate the performance of the proposed prediction methods, we randomly divided the study cohort into training and testing sets with equal numbers of subjects. We analyzed the training set to obtain parameter estimates, based on which we calculated the risk scores for subjects in the testing set, where the

Table 3. Estimation results for the regression parameters of the symptomatic events in the ARIC study.

	MI			Stroke			Death		
Covariate	Estimate	Std error	*p*-value	Estimate	Std error	*p*-value	Estimate	Std error	*p*-value
Forsyth County, white	0.0467	0.2477	0.8504	0.1308	0.3688	0.7228	−0.2475	0.1049	0.0183
Jackson, black	−0.3121	0.2681	0.2444	0.6622	0.3755	0.0778	0.1871	0.1118	0.0941
Minneapolis, white	−0.1052	0.2476	0.6710	0.0507	0.3688	0.8907	−0.3262	0.1040	0.0017
Washington County, white	0.1953	0.2457	0.4266	0.5013	0.3653	0.1700	−0.1194	0.1032	0.2471
Age	0.0805	0.0078	<0.0001	0.1121	0.0099	<0.0001	0.1465	0.0054	<0.0001
Male	0.9279	0.0901	<0.0001	0.4050	0.1071	0.0002	0.6108	0.0545	<0.0001
BMI	0.0273	0.0101	0.0068	−0.0010	0.0123	0.9356	0.0080	0.0060	0.1847
Glucose	0.0059	0.0046	0.2007	0.0215	0.0057	0.0002	0.0104	0.0030	0.0006
Systolic blood pressure	0.0135	0.0036	0.0002	0.0192	0.0047	<0.0001	0.0089	0.0022	0.0001
Smoker	1.2378	0.0888	<0.0001	1.0023	0.1127	<0.0001	1.3045	0.0599	<0.0001

NOTE: See the Note to Table 2.

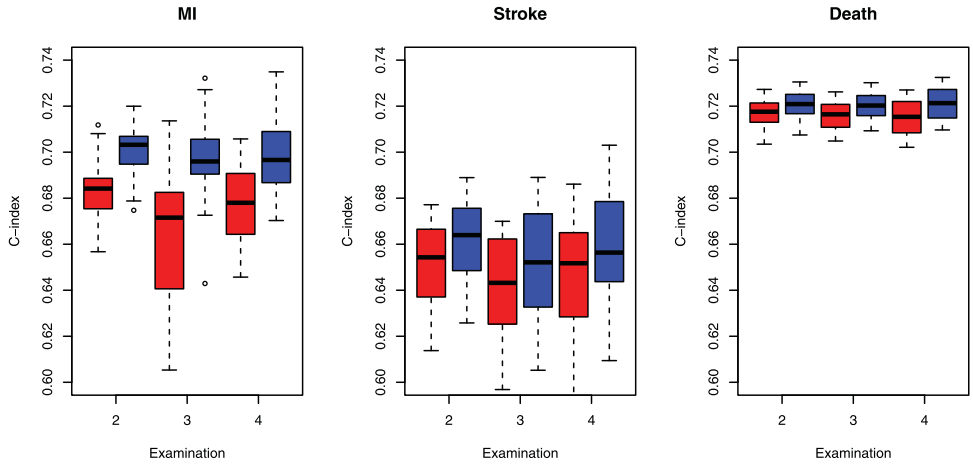

Figure 2. Boxplots of the estimates of the C-index at each examination in the ARIC study. The red boxes pertain to the univariate model of Fine and Gray (1999) for MI and stroke and the standard proportional hazards model for death. The blue boxes pertain to the proposed joint model.

posterior means of the random effects were used. Specifically, at examinations 2, 3, and 4, we calculated the risk scores of MI or stroke for subjects who have not developed the disease. We evaluated the performance of the prediction using C-index (Uno et al. 2011) and compared it with that of the risk scores based on the standard models. In particular, for MI and stroke, we considered the univariate model of Fine and Gray (1999) with death as a competing risk; for death, we considered the standard proportional hazards model. The values of the C-index based on twenty randomly divided training/test tests are shown in Figure 2. The proposed risk score performs better than the risk score of the standard model at all examinations for all symptomatic events.

Figure 3 shows the estimated conditional cumulative incidence functions of MI and stroke for two smokers and two non-smokers who have different event histories at year 3 but with the same values of other risk factors. The risks of MI and stroke are considerably higher for the smokers than the non-smokers with the same event history. The estimated conditional probabilities for the subjects who have developed both diabetes and hypertension are higher than those who have not developed diabetes or hypertension.

Figures S.2(a) and S.2(b) in the supplementary materials illustrate the estimation of the conditional cumulative incidence functions of stroke given different event histories. We estimated the cumulative incidence functions at time zero when only baseline covariates are available and then updated them at two examinations at year 3 and year 6 using the event histories. The development of diabetes, hypertension, and MI substantially increases the incidence of stroke, whereas the history of no diabetes, hypertension, or MI over the first six years entails lower incidence of stroke. For comparison, we show in Figures S.2(c) in the supplementary materials the estimated cumulative incidence function of stroke under the univariate model of Fine and Gray (1999), which does not condition on the event history and thus reflects the population average. This estimate lies between the two previous conditional estimates, as expected.

5. Discussion

In this article, we formulated the joint distribution of multiple right- and interval-censored events with proportional hazards models with random effects. We characterized the correlation structure of the asymptomatic and symptomatic events through two independent random effects and used unknown coefficients to capture the effects of the asymptomatic events on the symptomatic events. To our knowledge, no such modeling approach has been previously adopted.

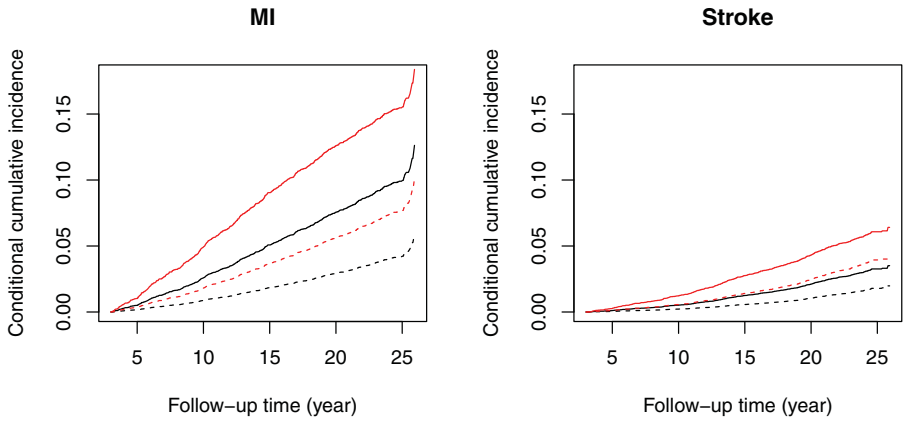

Figure 3. Estimation of the conditional cumulative incidence functions of MI and stroke for a 50-year-old white female residing in Forsyth County, NC, with BMI 40 kg/m^2, glucose 98 mg/dL, and systolic blood pressure 113 mmHg. The solid curves pertain to smokers, while the dashed curves pertain to non-smokers. The black curves pertain to subjects who have not developed diabetes or hypertension by year 3. The red curves pertain to subjects who have developed both diabetes and hypertension by year 3.

We studied efficient nonparametric maximum likelihood estimation of the proposed joint model and established the asymptotic properties of the estimators through innovative use of modern empirical process theory. We showed the Glivenko–Cantelli and Donsker properties for the classes of functions of interest by carefully evaluating their bracketing numbers. The estimators of the cumulative baseline hazard functions for the symptomatic and asymptomatic events converge at different ($n^{1/2}$ and $n^{1/3}$) rates, such that separate treatments were required in the proofs.

The proposed EM algorithm performed well in both the simulation studies and the real example. There was no occurrence of nonconvergence in any of the simulated or real dataset. It took 2.5 or 12 minutes to analyze a simulated dataset with $K = 5$ events and sample sizes $n = 100$ or 200, respectively. It took ten days to analyze the ARIC data, which involves 8728 subjects with 10 covariates and 2232, 2291, 701, 431, and 2130 distinct jump times for diabetes, hypertension, MI, stroke, and death, respectively. We can alleviate the computational burden for such large studies by grouping or subsampling the examination times so as to reduce the number of distinct time points. In particular, the computing time was shortened to two days when the distinct values were reduced to 154, 162, 276, 229, and 311 by rounding the examination times to the nearest months in the ARIC data.

We proposed nonparametric bootstrap for variance estimation as an alternative to the conventional profile-likelihood approach. We established the validity of the bootstrap procedure and showed through simulation studies that bootstrap yields more accurate estimators of the variabilities for the variance components. To our knowledge, bootstrap with interval-censored data has not been rigorously studied. In large studies, bootstrap may be overly time-consuming. It would be worthwhile to develop other versions of bootstrap, such as subsampling bootstrap, to reduce computational burden.

In models (1) and (2), we distinguish asymptomatic from symptomatic events when modeling the correlation structures because it is of particular interest to study the effects of asymptomatic diseases, which are typically interval-censored, on symptomatic diseases, which are typically right-censored, and to use the former to predict the latter. We show in Section S.3 of the Supplementary Materials that our framework can be modified to allow any of the K event times to be interval- or right-censored.

ARIC is one of many epidemiological cohort studies with multiple symptomatic and asymptomatic events. Such events are also available in electronic health records. Indeed, other types of outcomes, such as longitudinal repeated measures and recurrent events, may also be available. The proposed joint model can be extended to accommodate additional multivariate outcomes and improve dynamic prediction.

The authors thank the staff and participants of the ARIC study for their important contributions.

Supplementary Materials

The online supplementary materials contain the appendices for the article.

Funding

This work was supported by the National Institutes of Health awards R01GM047845, R01AI029168, R01CA082659, and P01CA142538. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C).

References

Chen, M. H., Chen, L. C., Lin, K. H., and Tong, X. (2014), "Analysis of Multivariate Interval Censoring by Diabetic Retinopathy Study," *Communications in Statistics: Simulation and Computation*, 43, 1825–1835. [1232]

Fine, J. P., and Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509. [1232,1239]

Fine, J. P., Jiang, H., and Chappell, R. (2001), "On Semi-Competing Risks Data," *Biometrika*, 88, 907–919. [1232,1233]

Goggins, W. B., and Finkelstein, D. M. (2000), "A Proportional Hazards Model for Multivariate Interval-Censored Failure Time Data," *Biometrics*, 56, 940–943. [1232]

Hogan, J. W., and Laird, N. M. (1997), "Model-Based Approaches to Analysing Incomplete Longitudinal and Failure Time Data," *Statistics in Medicine*, 16, 259–272. [1232]

Hougaard, P. (2012), *Analysis of Multivariate Survival Data*, New York: Springer. [1232]

Kalbfleisch, J. D., and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Hoboken, NJ: Wiley. [1232]

Kim, M. Y., and Xue, X. (2002), "The Analysis of Multivariate Interval-Censored Survival Data," *Statistics in Medicine*, 21, 3715–3726. [1232]

Lin, D. Y. (1994), "Cox Regression Analysis of Multivariate Failure Time Data: the Marginal Approach," *Statistics in Medicine*, 13, 2233–2247. [1233]

Murphy, S. A., and Van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465. [1235]

The ARIC Investigators., (1989), "The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives," *American Journal of Epidemiology*, 129, 687–702. [1232]

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011), "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures With Censored Survival Data," *Statistics in Medicine*, 30, 1105–1117. [1239]

Wen, C. C., and Chen, Y. H. (2013), "A Frailty Model Approach for Regression Analysis of Bivariate Interval-Censored Survival Data," *Statistica Sinica*, 23, 383–408. [1232]

Zeng, D., Gao, F., and Lin, D. (2017), "Maximum Likelihood Estimation for Semiparametric Regression Models With Multivariate Interval-Censored Data," *Biometrika*, 104, 505–525. [1232]