Statistica Sinica **18**(2008), 219-234

ANALYSIS OF COMPETING RISKS DATA WITH MISSING CAUSE OF FAILURE UNDER ADDITIVE HAZARDS MODEL

Wenbin Lu and Yu Liang

North Carolina State University and SAS Institute Inc.

Abstract: Competing risks data arise when study subjects may experience several different types of failure. It is common that the cause of failure is missing due to various reasons. Analysis of competing risks data with missing cause of failure has received considerable attention recently (Goetghebeur and Ryan (1995), Lu and Tsiatis (2001), Gao and Tsiatis (2005), among others). In this article, we study the semiparametric additive hazards model for analysis of competing risk data with missing cause of failure. Different estimating equation approaches using the inverse probability weighted and double robust techniques are proposed for estimating the regression parameters of interest. The resulting estimators have closed forms and their theoretical properties are established for inference. Simultaneous confidence bands of survival curves are constructed using a resampling technique. Simulations and an example show that the proposed approach is appropriate for practical use.

Key words and phrases: Additive hazards model, Competing risks data, Double robust, Estimating equation, Inverse probability weight, Missing cause of failure.

1. Introduction

Competing risks data arise in medical and public health studies when subjects may experience several different types of failure. The data typically include, for each subject, the failure time, possibly censored, the cause of failure when an failure is observed, and some covariates. In the literature, cause-specific hazards have been widely used to assess the covariate effects on the failure times of main interest (Prentice and Kalbfleisch (1978) and Cox and Oakes (1984)). In most applications, the causes of observed failures are assumed to be known.

In practice, however, the cause of failure for some subjects may be missing or uncertain. For example, documentation containing the information needed for attributing the cause of failure may be lost or not collected, or the cause of disease for some patients may be difficult to determine (Andersen, Goetghebeur and Ryan (1996)). Analysis using only the failures with known causes, i.e., the complete-case analysis, may lead to substantial bias. A number of statistical methods have been proposed for analysis of competing risks data with missing cause of failure,

e.g., Dinse (1982, 1986), Racine-Poon and Hoel (1984) and Goetghebeur and Ryan (1990). More recently, semiparametric survival models have been used to study the effects of covariates for such data. In particular, Goetghebeur and Ryan (1995) and Lu and Tsiatis (2001) studied the proportional hazards model, and Gao and Tsiatis (2005) considered linear transformation models. In addition, Craiu and Duchesne (2004) and Craiu and Reiser (2006) investigated the competing risks model with masked causes of failure.

For survival data, the additive hazards model is another useful framework for describing the association between risk factors and failure time. Compared with the proportional hazards model (Cox (1972)), the covariate effects in the additive hazards model are assumed to be additive instead of multiplicative to the baseline hazard function. In this article, we study competing risks data with missing cause of failure under the additive hazards model. Estimating equation approaches based on the inverse probability weighted (IPW) and double robust (DR) techniques are proposed for estimating the regression parameters. The resulting estimators have closed forms and are easy to compute. Their theoretical properties are derived for inference. The weak convergence properties of the estimated baseline cumulative hazard function and the corresponding survival function are established. A resampling technique is proposed for constructing simultaneous confidence bands for the survival curve of a given subject.

The remainder of the article is organized as follows. The next section briefly reviews the method proposed by Lin and Ying (1994) for conventional right-censored survival data under the additive hazards model. The IPW and DR estimating equations are developed and the asymptotic properties of the corresponding estimators are established in Section 3. Construction of simultaneous confidence bands of survival curve is also discussed here. Section 4 is devoted to numerical studies. Some conclusions and discussions are given in Section 5. Major technical derivations are contained in the Appendix.

2. Notation and Model Specification

Consider a study involving n independent subjects. Without loss of generality, we assume that each study subject might experience two types of failure: types 1 and 2. For subject i, let $Z_i(\cdot)$ be a p-dimensional vector of possibly time-varying covariates, and let T_{i1} and T_{i2} denote the potential failure times from types 1 and 2, respectively, $i = 1, \ldots, n$. Here, instead of observing T_{i1} and T_{i2}, we observe the minimum of T_{i1}, T_{i2} and the censoring time C_i, i.e. $T_i = \min(T_{i1}, T_{i2}, C_i)$. Define $\Delta_i = 1$ if $T_i = T_{i1}$, $\Delta_i = 2$ if $T_i = T_{i2}$, and 0 otherwise. Throughout the paper, we assume that $Z_i(\cdot)$ is an external covariate process (Kalbfleish and Prentice (2002)) and that given $Z_i(\cdot)$, C_i is independent of T_{i1} and T_{i2}. If there were no missing causes of failure, the observed data

would consist of $\{T_i, \Delta_i, Z_i(t) : 0 < t \leq T_i\}$, $i = 1, \ldots, n$. Suppose that our interest focuses on assessing the association between the first type of failure and the covariates. The cause-specific hazard function $\lambda_1^*(t|Z)$ of type 1 failure for a subject with covariate Z is given by

$$\lambda_1^*(t|Z) = \lim_{h \to 0} h^{-1} P(t \leq T < t + h, \Delta = 1 | T \geq t, Z). \tag{1}$$

The additive hazards model assumes that

$$\lambda_1^*(t|Z) = \lambda_1(t) + \beta' Z(t), \tag{2}$$

where $\lambda_1(t)$ is the completely unspecified baseline hazard function and β a p-dimensional regression parameter vector. Lin and Ying (1994) proposed the following estimating equation for β:

$$\sum_{i=1}^{n} \int_0^{\infty} \{Z_i(t) - \bar{Z}(t)\}\{dN_i(t) - Y_i(t)\beta' Z_i(t)dt\} = 0, \tag{3}$$

where $N_i(t) = I(\Delta_i = 1)I(T_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$ are the usual counting and at-risk processes. The resulting estimator for β can be written as

$$\hat{\beta} = \Big[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dt\Big]^{-1} \Big[\sum_{i=1}^{n} \int_0^{\infty} \{Z_i(t) - \bar{Z}(t)\} dN_i(t)\Big],$$

where $\bar{Z}(t) = \sum_{j=1}^{n} Z_j(t)Y_j(t) / \sum_{j=1}^{n} Y_j(t)$. It was also shown that $\hat{\beta}$ is consistent and asymptotically normal.

As discussed before, the cause of failure may not be observed for every subject, i.e., Δ_i is not always observed. If this is the case, equations (3) cannot be used directly for parameter estimation. Let R_i denote the missing indicator for the cause of failure, i.e., $R_i = 1$ if the cause of failure for subject i is observed, and 0 otherwise. Here we define $R_i = 1$ when $\Delta_i = 0$, i.e., the failure time of subject i is censored. Then the observed data consist of $\{R_i,\ T_i,\ I(\Delta_i = 0),\ R_i I(\Delta_i = 1),\ R_i I(\Delta_i = 2),\ G_i, Z_i(t) : 0 < t \leq T_i\}$, $i = 1, \ldots, n$, where the G_i's are some auxiliary covariates collected for every subject.

3. Estimating Equations and Theoretical Results

In this article, we assume that the cause of failure is missing at random (MAR) (Rubin (1976)). That is, given $\Delta_i > 0$ and $X_i = (T_i, Z_i, G_i)$, the probability that the cause of failure for subject i is missing depends only on the observed quantities X_i, but not on the unobserved Δ_i. In other words,

$$P(R_i = 1 | \Delta_i, \Delta_i > 0, X_i) = P(R_i = 1 | \Delta_i > 0, X_i) \equiv \pi(X_i). \tag{4}$$

The MAR assumption has been widely used for handling missing data problems in the statistical literature. The collection of auxiliary covariates G_i's is to ensure the validity of such assumption. When there are only two possible failure causes, the MAR assumption is equivalent to the symmetry assumption discussed in the literature (Craiu and Duchesne (2004)).

3.1. Inverse probability weighted estimating equations

Following the inverse selection probability idea of Horvitz and Thompson (1952), we consider two estimating equations for β and Λ_1:

$$\sum_{i=1}^{n} \frac{R_i}{\pi^*(X_i, \Delta_i, \hat{\gamma})}\{dN_i(t) - Y_i(t)\beta' Z_i(t)dt - Y_i(t)d\Lambda_1(t)\} = 0, \;\; t > 0, \tag{5}$$

$$\sum_{i=1}^{n} \int_0^{\infty} \frac{R_i}{\pi^*(X_i, \Delta_i, \hat{\gamma})} Z_i(t)\{dN_i(t) - Y_i(t)\beta' Z_i(t)dt - Y_i(t)d\Lambda_1(t)\} = 0, \tag{6}$$

where $\pi^*(X_i, \Delta_i, \gamma) = I(\Delta_i > 0)\pi(X_i, \gamma) + I(\Delta_i = 0)$ with $\pi(X_i, \gamma)$ being a parametric model posited for the missing cause probability $\pi(X_i)$. For example, since R_i is binary, logistic regression can be used for $\pi(X_i, \gamma)$. An estimator $\hat{\gamma}$ may be obtained by maximizing the likelihood based on uncensored data, i.e., $\hat{\gamma}$ maximizes

$$\prod_{i=1}^{n} \{\pi(X_i, \gamma)\}^{R_i I(\Delta_i > 0)} \{1 - \pi(X_i, \gamma)\}^{(1-R_i) I(\Delta_i > 0)}.$$

If the parametric model $\pi(X_i, \gamma)$ is correctly specified, $\hat{\gamma}$ consistently estimates γ_0, the true value of γ. As before, the resulting IPW estimator for β has the closed form

$$\hat{\beta}_{IPW} = \Big[\sum_{i=1}^{n} \int_0^{\infty} \frac{R_i}{\pi^*(X_i, \Delta_i, \hat{\gamma})} Y_i(t)\{Z_i(t) - \bar{Z}^*(t, \hat{\gamma})\}^{\otimes 2} dt\Big]^{-1}$$
$$\times \Big[\sum_{i=1}^{n} \int_0^{\infty} \frac{R_i}{\pi^*(X_i, \Delta_i, \hat{\gamma})} \{Z_i(t) - \bar{Z}^*(t, \hat{\gamma})\} dN_i(t)\Big],$$

where $\bar{Z}^*(t, \gamma) = \sum_{j=1}^{n} [R_j/(\pi^*(X_j, \Delta_j, \gamma))] Z_j(t) Y_j(t) / \sum_{j=1}^{n} [R_j/(\pi^*(X_j, \Delta_j, \gamma))] Y_j(t)$.

3.2. Double robust estimating equations

The validity of the inverse probability weighted estimator $\hat{\beta}_{IPW}$ depends on the correct specification of the parametric model $\pi(X_i, \gamma)$. If it is misspecified, $\hat{\beta}_{IPW}$ may be biased. In addition, since the calculation of $\hat{\beta}_{IPW}$ only uses the complete-case data, it may lose efficiency. To improve the robustness as well

as the efficiency of $\hat{\beta}_{IPW}$, we construct estimating equations based on the double robust technique developed by Robins, Rotnitzky and Zhao (1994). Define $h(Q_i) = P(\Delta_i = 1|\Delta_i > 0, R_i = 0, X_i)$, where $Q_i = (T_i, Z_i)$. Under the assumption of missing at random, we have

$$\begin{aligned} h(Q_i) &= P(\Delta_i = 1|\Delta_i > 0, R_i = 0, X_i) = P(\Delta_i = 1|\Delta_i > 0, R_i = 1, X_i) \\ &= P(\Delta_i = 1|\Delta_i > 0, X_i) = P(\Delta_i = 1|\Delta_i > 0, Q_i). \end{aligned}$$

In addition, $h(Q)$ can be determined by the ratio of the cause-specific hazard functions of T_1 and T_2, namely,

$$\frac{h(Q)}{1 - h(Q)} = \frac{\lambda_1^*(T|Z)}{\lambda_2^*(T|Z)}, \tag{7}$$

where $\lambda_2^*(T|Z)$ is the cause-specific hazard function for type 2 failure (see Dewanji (1992) and Lu and Tsiatis (2001)). Here, instead of directly estimating h based on (7) which involves the estimation of the two nonparametric cause-specific hazard functions $\lambda_1^*(t|Z)$ and $\lambda_2^*(t|Z)$, we posit a parametric model $h(Q, \theta)$ for $h(Q)$. For example, logistic regression can be used for $h(Q, \theta)$ for convenience, though other parametric models can also be accommodated easily. Due to the MAR assumption, the parameters θ can be estimated by maximizing the complete-case data likelihood

$$\prod_{i=1}^{n} \{h(Q_i, \theta)\}^{R_i I(\Delta_i = 1)} \{1 - h(Q_i, \theta)\}^{R_i I(\Delta_i = 2)}.$$

Let $\hat{\theta}$ denote the corresponding maximizer. It is known that when $h(Q, \theta)$ is correctly specified, $\hat{\theta}$ consistently estimates θ_0, the true value of θ. The double robust estimating equations can then be constructed as

$$\begin{aligned} \sum_{i=1}^{n} \Big\{ & \frac{R_i}{\pi(X_i, \hat{\gamma})} dN_i(t) - \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} h(Q_i, \hat{\theta}) dN_i^*(t) \\ & - Y_i(t) \beta' Z_i(t) dt - Y_i(t) d\Lambda_1(t) \Big\} = 0, \ \ t > 0, \end{aligned} \tag{8}$$

$$\begin{aligned} \sum_{i=1}^{n} \int_0^\infty Z_i(t) \Big\{ & \frac{R_i}{\pi(X_i, \hat{\gamma})} dN_i(t) - \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} h(Q_i, \hat{\theta}) dN_i^*(t) \\ & - Y_i(t) \beta' Z_i(t) dt - Y_i(t) d\Lambda_1(t) \Big\} = 0, \end{aligned} \tag{9}$$

where $N_i^*(t) = I(\Delta_i > 0) I(T_i \le t)$. From (8) and (9), we also have the closed-form DR estimator of β,

$$\hat{\beta}_{DR} = \Big[\sum_{i=1}^{n} \int_0^\infty Y_i(t) \{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dt \Big]^{-1}$$

$$\times\Big[\sum_{i=1}^{n}\int_0^{\infty}\{Z_i(t)-\bar{Z}(t)\}\Big\{\frac{R_i}{\pi(X_i,\hat{\gamma})}dN_i(t)-\frac{R_i-\pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})}h(Q_i,\hat{\theta})dN_i^*(t)\Big\}\Big].$$

Remark 1. Let γ^* and θ^* denote the limits of $\hat{\gamma}$ and $\hat{\theta}$, respectively. When either the parametric model $\pi(X_i,\gamma)$ or $h(Q_i,\theta)$ is correctly specified, we have, for any $t>0$,

$$E\Big[\frac{R_i}{\pi(X_i,\gamma^*)}N_i(t)-\frac{R_i-\pi(X_i,\gamma^*)}{\pi(X_i,\gamma^*)}h(Q_i,\theta^*)N_i^*(t)-\int_0^t Y_i(s)\{\beta_0'Z_i(s)ds+d\Lambda_{01}(s)\}\Big]=0, \qquad (10)$$

where β_0 and Λ_{01} denote the true values of β and Λ_1, respectively.

The proof of (10) can be derive from the representation

$$\begin{aligned}&\frac{R_i}{\pi(X_i,\gamma^*)}N_i(t)-\frac{R_i-\pi(X_i,\gamma^*)}{\pi(X_i,\gamma^*)}h(Q_i,\theta^*)N_i^*(t)-\int_0^t Y_i(s)\{\beta_0'Z_i(s)ds+d\Lambda_{01}(s)\}\\&\quad=M_i(t)-\frac{R_i-\pi(X_i,\gamma^*)}{\pi(X_i,\gamma^*)}\{I(\Delta_i=1)-h(Q_i,\theta^*)\}N_i^*(t),\end{aligned}$$

where $M_i(t)=N_i(t)-\int_0^t Y_i(s)\{\beta_0'Z_i(s)ds+d\Lambda_{01}(s)\}$ is a martingale process (Fleming and Harrington (1991)) with $E\{M_i(t)\}=0$. For the second term, if the parametric model $\pi(X_i,\gamma)$ is correctly specified ($\gamma^*=\gamma_0$), conditioning on $(X_i,\Delta_i,\Delta_i>0)$, we have

$$\frac{E(R_i|X_i,\Delta_i>0)-\pi(X_i,\gamma^*)}{\pi(X_i,\gamma^*)}\{I(\Delta_i=1)-h(Q_i,\theta^*)\}N_i^*(t)=0.$$

On the other hand, if the parametric model $h(Q_i,\theta)$ is correctly specified ($\theta^*=\theta_0$), conditioning on $(X_i,R_i,\Delta_i>0)$, we have

$$\{P(\Delta_i=1|Q_i,\Delta_i>0)-h(Q_i,\theta^*)\}\frac{R_i-\pi(X_i,\gamma^*)}{\pi(X_i,\gamma^*)}N_i^*(t)=0.$$

Therefore (10) holds.

3.3. Theoretical results

For simplicity, we only develop the asymptotic properties for $\hat{\beta}_{DR}$. The large sample results for $\hat{\beta}_{IPW}$ can be similarly derived and are omitted here. Define

$$\begin{aligned}M_i^*(t,\beta,\gamma,\theta,\Lambda_1)=&\frac{R_i}{\pi(X_i,\gamma)}N_i(t)-\frac{R_i-\pi(X_i,\gamma)}{\pi(X_i,\gamma)}h(Q_i,\theta)N_i^*(t)\\&-\int_0^t Y_i(s)\{\beta'Z_i(s)ds+d\Lambda_1(s)\},\end{aligned}$$

$$\phi_i = \int_0^\infty \{Z_i(t) - \bar{z}(t)\} dM_i^*(t, \beta_0, \gamma^*, \theta^*, \Lambda_{01}) - B_\gamma I_\gamma^{-1} S_{\gamma i} - B_\theta I_\theta^{-1} S_{\theta i},$$

and $A = E[\int_0^\infty Y_1(t)\{Z_1(t) - \bar{z}(t)\}^{\otimes 2} dt]$, where $\bar{z}(t) = s^{(1)}(t)/s^{(0)}(t)$ with $s^{(k)}(t) = E[\{Z_1(t)\}^k Y_1(t)]$, $k = 0, 1$, I_γ and $S_{\gamma i}$ are, respectively, the Fisher information matrix and score function derived from the parametric model $\pi(X_i, \gamma)$, and I_θ and $S_{\theta i}$ are similarly defined for the parametric model $h(Q_i, \theta)$. The formulations of B_γ, I_γ, $S_{\gamma i}$, B_θ, I_θ and $S_{\theta i}$ are given in the Appendix. The following theorem establish the theoretical properties of $\hat{\beta}_{DR}$.

Theorem 1. *Under the regularity conditions given in the Appendix,*

$$n^{\frac{1}{2}}(\hat{\beta}_{DR} - \beta_0) = A^{-1}\Big(n^{-\frac{1}{2}} \sum_{i=1}^n \phi_i\Big) + o_p(1).$$

Based on Theorem 1, when either the parametric model $\pi(X, \gamma)$ or $h(Q, \theta)$ is correctly specified, $n^{1/2}(\hat{\beta}_{DR} - \beta_0)$ is asymptotically a sum of independent random vectors with zero mean. Thus, by the Central Limit Theorem, it converges in distribution to a normal random vector with zero mean and variance-covariance matrix $A^{-1}E(\phi_1\phi_1')(A^{-1})'$. In addition, the variance-covariance matrix can be consistently estimated by $\hat{A}^{-1}(n^{-1}\sum_{i=1}^n \hat{\phi}_i\hat{\phi}_i')(\hat{A}^{-1})'$ where $\hat{A} = (1/n)\sum_{i=1}^n \int_0^\infty Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dt$ and

$$\hat{\phi}_i = \int_0^\infty \{Z_i(t) - \bar{Z}(t)\} dM_i^*(t, \hat{\beta}_{DR}, \hat{\gamma}, \hat{\theta}, \hat{\Lambda}_1) - \hat{B}_\gamma \hat{I}_\gamma^{-1} \hat{S}_{\gamma i} - \hat{B}_\theta \hat{I}_\theta^{-1} \hat{S}_{\theta i}.$$

Here $\hat{B}_\gamma$, $\hat{I}_\gamma$, $\hat{S}_{\gamma i}$, $\hat{B}_\theta$, $\hat{I}_\theta$ and $\hat{S}_{\theta i}$ are obtained by substituting $(\hat{\beta}_{DR}, \hat{\gamma}, \hat{\theta})$ for $(\beta_0, \gamma^*, \theta^*)$ and replacing the expectation E by its empirical counterpart, and $\hat{\Lambda}_1(t) = \hat{\Lambda}_1(t, \hat{\beta}_{DR})$ is the Nelson-Aalen estimator for $\Lambda_{01}(t)$, where

$$\hat{\Lambda}_1(t, \beta) = \int_0^t \frac{\sum_{i=1}^n \{\frac{R_i}{\pi(X_i, \hat{\gamma})} dN_i(s) - \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} h(Q_i, \hat{\theta}) dN_i^*(s) - \beta' Z_i(s) Y_i(s) ds\}}{\sum_{i=1}^n Y_i(s)}.$$

Furthermore, define

$$\psi_i(t) = \int_0^t \frac{dM_i^*(u, \beta_0, \gamma^*, \theta^*, \Lambda_{01})}{s^{(0)}(u)} - \xi'(t) A^{-1} \phi_i - C_\gamma'(t) I_\gamma^{-1} S_{\gamma i} - C_\theta'(t) I_\theta^{-1} S_{\theta i},$$

where $\xi(t) = \int_0^t \bar{z}(u) du$, $\dot{\pi}_\gamma(\cdot, \gamma) = \partial \pi(\cdot, \gamma)/\partial \gamma$, $\dot{h}_\theta(\cdot, \theta) = \partial h(\cdot, \theta)/\partial \theta$ and

$$C_\gamma(t) = E\left[\int_0^t \frac{R_1 \dot{\pi}_\gamma(X_1, \gamma^*)}{\pi^2(X_1, \gamma^*) s^{(0)}(u)} \{dN_1(u) - h(Q_1, \theta^*) dN_1^*(u)\}\right],$$

$$C_\theta(t) = E\left\{\int_0^t \frac{R_1 - \pi(X_1, \gamma^*)}{\pi(X_1, \gamma^*) s^{(0)}(u)} \dot{h}_\theta(Q_1, \theta^*) dN_1^*(u)\right\}.$$

The following theorem establishes the asymptotic properties for the baseline cumulative hazard estimator $\hat{\Lambda}_1(t)$.

Theorem 2. *Under the conditions of Theorem* 1*,*

$$V(t) \equiv n^{\frac{1}{2}}\{\hat{\Lambda}_1(t) - \Lambda_{01}(t)\} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_i(t) + o_p(1).$$

By Theorem 2, when either the parametric model $\pi(X, \gamma)$ or $h(Q, \theta)$ is correctly specified, $V(t)$ converges weakly to a zero-mean Gaussian process. The asymptotic covariance function of $V(\cdot)$ at (s, t) is then $E\{\psi_1(s)\psi_1(t)\}$, which can be consistently estimated by $n^{-1}\sum_{i=1}^{n} \hat{\psi}_i(s)\hat{\psi}_i(t)$, where

$$\hat{\psi}_i(t) = \int_0^t \frac{dM_i^*(u, \hat{\beta}_{DR}, \hat{\gamma}, \hat{\theta}, \hat{\Lambda}_1)}{\sum_{i=1}^{n} Y_i(u)/n} - \hat{\xi}'(t)\hat{A}^{-1}\hat{\phi}_i - \hat{C}'_\gamma(t)\hat{I}_\gamma^{-1}\hat{S}_{\gamma i} - \hat{C}'_\theta(t)\hat{I}_\theta^{-1}\hat{S}_{\theta i}.$$

Here $\hat{\xi}(t)$, $\hat{C}_\gamma(t)$, $\hat{C}_\theta(t)$ are obtained by substituting $(\hat{\gamma}, \hat{\theta})$ for (γ^*, θ^*) and replacing the expectation E by its empirical counterpart.

For a subject with the covariate vector $z_0(t)$, the cumulative hazard function can be estimated by $\hat{\Lambda}_1(t; z_0) = \hat{\Lambda}_1(t, \hat{\beta}_{DR}) + \int_0^t \hat{\beta}'_{DR} z_0(u)du$, and the survival function by $\hat{S}_1(t; z_0) = \exp\{-\hat{\Lambda}_1(t; z_0)\}$. Define $\Lambda_{01}(t, z_0) = \Lambda_{01}(t) + \int_0^t \beta_0 z_0(u)du$ and $S_{01}(t, z_0) = \exp\{-\Lambda_{01}(t, z_0)\}$.

Theorem 3. *Under the conditions of Theorem* 1*,*

$$V_{z_0}(t) \equiv n^{1/2}\{\hat{\Lambda}_1(t; z_0) - \Lambda_{01}(t; z_0)\} = n^{-1/2} \sum_{i=1}^{n} \psi_i^{z_0}(t) + o_p(1),$$

where $\psi_i^{z_0}(t) = \psi_i(t) + \{\int_0^t z_0(u)du\}A^{-1}\phi_i$.

The proof of Theorem 3 is omitted since it is similar to that of Theorem 2. When either the parametric model $\pi(X, \gamma)$ or $h(Q, \theta)$ is correctly specified, $V_{z_0}(t)$ converges weakly to a zero-mean Gaussian process and the asymptotic covariance function of $V_{z_0}(\cdot)$ at (s, t) can be consistently estimated by $n^{-1}\sum_{i=1}^{n} \hat{\psi}_i^{z_0}(s)\hat{\psi}_i^{z_0}(t)$, where $\hat{\psi}_i^{z_0}(t) = \hat{\psi}_i(t) + \{\int_0^t z_0(u)du\}\hat{A}^{-1}\hat{\phi}_i$. By the functional delta method, the normalized survival process $n^{1/2}\{\hat{S}_1(t; z_0) - S_{01}(t; z_0)\}$ converges weakly to a zero-mean Gaussian process, and the covariance function at (s, t) can be consistently estimated by

$$\hat{S}_1(s; z_0)\hat{S}_1(t; z_0)\Big\{\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_i^{z_0}(s)\hat{\psi}_i^{z_0}(t)\Big\}.$$

A point-wise confidence interval for the survival function $S_{01}(t; z_0)$ can be easily constructed using the above result. But construction of simultaneous

confidence bands for all t involves functionals of the limiting distribution of $V_{z_0}(t)$, which does not have an independent increment structure. We use a resampling technique of Lin, Fleming and Wei (1994) to approximate the distribution of $V_{z_0}(t)$. A similar method was also used by Yin and Cai (2004) to construct simultaneous confidence bands of survival curves for multivariate survival data under the additive hazards model. To be specific, define $\hat{V}_{z_0}(t) = (1/\sqrt{n}) \sum_{i=1}^{n} \hat{\psi}_i^{z_0}(t) W_i$, where $W_i, i = 1, \ldots, n$, are n i.i.d. random variables from the standard normal distribution and are independent of the observed data.

Theorem 4. *Conditional on the observed data, $\hat{V}_{z_0}(t)$ converges weakly to the same zero-mean Gaussian process as that of $V_{z_0}(t)$ when either the parametric model $\pi(X, \gamma)$ or $h(Q, \theta)$ is correctly specified.*

The proof of Theorem 4 easily follows the conditional multiplier central limit theorem in van der Vaart and Wellner (1996, Thm. 2.9.6) and the steps given in Appendix 3 of Yin and Cai (2004). The simultaneous confidence band for $S_{01}(t; z_0)$ can then be easily constructed based on the perturbed process $\hat{V}_{z_0}(t)$ along the lines of Yin and Cai (2004).

4. Numerical Studies

4.1. Simulation studies

The properties of our proposed estimators are assessed in a series of simulation studies under practical settings. The additive hazards model (2) is used to generate failure time T_1 due to the cause of interest. Two independent covariates Z_1 and Z_2 are considered, with Z_1 following a uniform distribution on $(0, 1)$ and Z_2 a Bernoulli distribution with success probability 0.5. The regression parameter $\beta = (\beta_1, \beta_2)' = (1, -1)'$ and the baseline hazard function $\lambda_1(t) \equiv 1.3$. Failure time T_2 from the other cause is generated from a Gompertz distribution with hazard function $\lambda_2(t|Z) = \exp(a + bt)$, where $a = -1$ and $b = 1$. Censoring time C is generated from a uniform distribution on $(0, c)$, where c is chosen to yield a desired censoring level, $P(\Delta = 0)$ is either 15% or 40%. In the above setting, when the censoring level is 15%, we have, on average, approximately 55% type 1 failures, while when the censoring level is 40%, approximately 42% failures are of type 1.

The missing cause indicator R for a failure is generated from a logistic model: $\pi(X) = P(R = 1|\Delta > 0, X) = \exp(\gamma' X)/\{1 + \exp(\gamma' X)\}$, where $X = (1, T, Z_1, Z_2)'$, $\gamma = (-2.5, 2, 2, 2)'$ and $T = \min(T_1, T_2, C)$. Then, under the 15% censoring level, nearly 44% of the failures have missing causes, i.e., $R_i = 0$, while 50% of the failures have missing causes under the 40% censoring level.

Here three different methods are used for estimating the regression parameter β. The first method directly applies Lin and Ying (1994)'s equation (3) to complete-case data. Denote the resulting estimator by $\hat{\beta}_{CC}$. In the second and third methods we consider the inverse probability weighted estimator $\hat{\beta}_{IPW}$ and the double robust estimator $\hat{\beta}_{DR}$ discussed in Sections 3.1 and 3.2, respectively. To derive the inverse probability weighted and the double robust estimators, we posit two different parametric models for $\pi(X)$: one is the true logistic model $\pi(X,\gamma) = \exp(\gamma' X)/\{1+\exp(\gamma' X)\}$; the other is the misspecified constant model π_0, where π_0 is a constant between 0 and 1. Furthermore, in the settings we consider above, the true model for $h(X)$ is given by $h(X)/(1-h(X)) = (1.3 + Z_1 - Z_2)/\exp(-1+T)$. Since the cause-specific hazard functions for the two types of failure are usually unknown in practice, instead of estimating $h(X)$ directly from the data, we posit two different parametric models for $h(X)$: one is the logistic model $h(X,\theta) = \exp(\theta' X)/\{1+\exp(\theta' X)\}$, and the other is the constant model $h_0 \in (0,1)$. Both of them are thus misspecified. 500 runs with sample size $n = 200$ are used under each scenario. All simulations are done with R codes. Simulation results of the above three estimators are summarized in Table 1 and Table 2, where Table 1 presents the results for the 15% censoring level and Table 2 does so for the 40% censoring level.

Results from Tables 1 and 2 indicate that the complete-case estimators show large bias in all the settings, and the inverse probability weighted estimators are essentially unbiased only when the parametric model for $\pi(X)$ is correctly specified. But when $\pi(X)$ is misspecified, they also show large bias. The double

Table 1. Simulation results for the 15% censoring level.

Parameters	$\beta_1 = 1$				$\beta_2 = -1$			
Estimates	Bias	SD	SE	CP	Bias	SD	SE	CP
CC	0.131	0.416	0.391	94.8	0.575	0.285	0.274	41.6
IPW_1	-0.005	0.657	0.610	94.2	-0.011	0.396	0.399	95.6
IPW_2	0.179	0.447	0.484	96.4	0.525	0.306	0.286	51.4
DR_1	0.002	0.539	0.513	94.4	-0.016	0.331	0.321	94.8
DR_2	0.007	0.523	0.485	93.4	-0.014	0.328	0.309	93.0
DR_3	0.133	0.447	0.425	93.8	0.147	0.268	0.256	89.4
DR_4	0.069	0.466	0.438	93.6	0.018	0.314	0.293	93.2

† SD, sample standard deviation; SE, mean of estimated standard errors; CP, empirical coverage probability of 95% Wald-type confidence interval; CC, the complete-case estimator; IPW_1 and IPW_2, the inverse probability weighted estimators when using the logistic model and the constant model for $\pi(X)$, respectively; DR_1, DR_2, DR_3 and DR_4, the double robust estimators when using the logistic model and the constant model, the logistic model and the logistic model, the constant model and the constant model, the constant model and the logistic model for $\pi(X)$ and $h(X)$, respectively.

Table 2. Simulation results for the 40% censoring level.

Parameters	$\beta_1 = 1$				$\beta_2 = -1$			
Estimates	Bias	SD	SE	CP	Bias	SD	SE	CP
CC	0.162	0.415	0.417	95.6	0.763	0.276	0.264	20.4
IPW_1	0.100	0.689	0.715	96.4	0.009	0.428	0.471	95.6
IPW_2	0.444	0.534	0.573	91.6	0.659	0.363	0.332	46.8
DR_1	0.040	0.585	0.592	96.2	-0.027	0.389	0.365	95.2
DR_2	0.051	0.586	0.580	94.8	-0.021	0.394	0.357	93.8
DR_3	0.121	0.513	0.511	94.2	0.064	0.333	0.308	93.0
DR_4	0.099	0.523	0.526	95.4	‘0.005	0.377	0.342	93.8

† The notations are those in Table 1.

robust estimators are essentially unbiased when $\pi(X)$ is correctly specified and they are more efficient than the corresponding inverse probability weighted estimators. Still, when $\pi(X)$ is misspecified, since in our simulations the two parametric models proposed for $h(X)$ are both misspecified, the double robust estimators are biased. However, the biases are relatively small compared to those of the complete-case and the inverse probability weighted estimators, especially when using the logistic models for $h(X)$. Furthermore, the estimated standard errors (SE) are quite close to the sample standard errors (SD) under each case, and the 95% confidence intervals have reasonable empirical coverage probabilities when the estimators are essentially unbiased.

4.2. Example of a breast cancer study

We studied a dataset from clinical trial E1178 (Cummings et al. (1993)), which compared two years of tamoxifen therapy to placebo in 167 breast cancer patients with age greater than or equal to 65, and positive axillary nodes. Endpoints of interest include recurrence of breast cancer and death without recurrence, which are two competing risks events. In this dataset, the causes of failures are all known. To illustrate our method, we artificially deleted some failure causes according to three missing mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). For the MCAR, the causes of failures were randomly selected for missing with probability 0.4; for the MAR, the non-missing probability was chosen as $\pi = \exp(1.0 - 0.2*T + 0.5*\text{trt})/(1+\exp(1.0-0.2*T+0.5*\text{trt}))$, where T denotes the observed failure or censoring time and trt denotes the treatment indicator (1 = tamoxifen therapy; 0 = placebo); for the NMAR, $\pi = \exp(1.0-0.2*T+0.5*\text{trt}+0.25*I(\Delta = 1))/(1+\exp(1.0-0.2*T+0.5*\text{trt}+0.25*I(\Delta=1)))$, where the failure cause $\Delta = 1$ corresponds to the recurrence of breast cancer and $\Delta = 2$ corresponds to the death without recurrence. In the MAR, the missing probability is about 45%, while in the NMAR, it is about 40%.

We apply the proposed estimating equation approach for the additive hazards model to the above datasets with two covariates: treatment assignment (trt) and number of positive nodes after log transformation (lnode). Our interest focused on the type 1 failure. For comparison, three methods were studied: (i) Lin and Ying (1994)'s method for the original dataset without missing; (ii) the complete-case method for the derived datasets with missing; (iii) the proposed double robust method. For the DR method, two logistic regressions were fitted, respectively, for the missing probability π and the conditional expectation h with T, trt and lnode included as covariates. We also present the corresponding results fitted using the proportional hazards model. For the first two cases, the maximum partial likelihood estimators were used; while for the third one, the double robust method of Gao and Tsiatis (2005) was applied. The results are summarized as follows:

		AH		PH	
Missing	Method	trt	lnode	trt	lnode
NONE	LY/Cox	-0.066 (0.021)	0.039 (0.011)	-0.67 (0.20)	0.38 (0.11)
MCAR	CC	-0.071 (0.022)	0.044 (0.013)	-0.87 (0.25)	0.54 (0.15)
MCAR	DR	-0.066 (0.026)	0.040 (0.014)	-0.66 (0.21)	0.39 (0.11)
MAR	CC	-0.134 (0.036)	0.054 (0.015)	-1.16 (0.27)	0.56 (0.15)
MAR	DR	-0.085 (0.031)	0.051 (0.016)	-0.86 (0.23)	0.49 (0.11)
NMAR	CC	-0.136 (0.037)	0.053 (0.015)	-1.20 (0.26)	0.39 (0.14)
NMAR	DR	-0.076 (0.031)	0.052 (0.017)	-0.75 (0.22)	0.39 (0.10)

Here AH stands for the additive hazards model, PH stands for the proportional hazards model and LY/Cox stands for Lin and Ying (1994)'s method for the AH model and Cox's partial likelihood method for the PH model. The numbers given in parentheses are the estimated standard errors. From the results, we can see that treatment and number of positive nodes are significant under both the AH and PH models. Both CC and DR methods work fine in the MCAR case compared to the corresponding estimators without missing in the cause of failure. But for both the MAR and NMAR cases, the CC method has a large bias, especially for the estimates of treatment effect, while the DR method corrected such bias sufficiently. In addition, the parameter estimates under the additive hazards model are much smaller than those fitted using the proportional hazards model. However, as discussed by Lin and Ying (1994), this is not surprising since the additive hazards model pertain to the hazard difference whereas the proportional hazards model pertain to the hazard ratio.

5. Concluding Remarks

We have proposed inverse probability weighted and double robust estimators for competing risks data with missing cause of failure under the additive hazards

model. The proposed estimators for the regression parameters have closed forms and are very easy to compute. The theoretical properties of them are also established for inference. The resulting asymptotic variance-covariance matrix can be consistently estimated by the usual plug-in method. The simultaneous confidence bands of survival curves are also constructed via a resampling technique. The proposed estimating equation methods for competing risks data with missing cause of failure can be extended to incorporate missing covariates along the lines of Robins, Rotnitzky and Zhao (1994) for regression problems with missing covariates.

In the context of competing risks data, it is common that the cause of failure may be group masked, i.e., the cause of failure is only known to belong to a certain subset of all possible failures (Craiu and Duchesne (2004) and Craiu and Reiser (2006)). For masked failure causes, a second-stage analysis is usually conducted, in which the true cause can be uniquely determined for a random sample of the masked failures. The generalization of the proposed double robust methods to handle competing risks data with masked failure causes needs further investigation.

Acknowledgement

The authors are grateful to an associate editor and an anonymous referee for their constructive suggestions that have led to considerable improvement of an earlier version. This research was partly supported by the National Science Foundation Grant DMS-0504269.

Appendix

In the following, we first give the formulations of B_γ, I_γ, $S_{\gamma i}$, B_θ, I_θ and $S_{\theta i}$.

$$B_\gamma = E\left[\int_0^\infty \{Z_1(t) - \bar{z}(t)\}\frac{R_1\dot{\pi}'_\gamma(X_1,\gamma^*)}{\pi^2(X_1,\gamma^*)}\{dN_1(t) - h(Q_1,\theta^*)dN_1^*(t)\}\right],$$

$$B_\theta = E\left[\int_0^\infty \{Z_1(t) - \bar{z}(t)\}\frac{R_1 - \pi(X_1,\gamma^*)}{\pi(X_1,\gamma^*)}\dot{h}'_\theta(Q_1,\theta^*)dN_1^*(t)\right],$$

$$S_{\gamma i} = \frac{I(\Delta_i > 0)\{R_i - \pi(X_i,\gamma^*)\}\dot{\pi}_\gamma(X_i,\gamma^*)}{\pi(X_i,\gamma^*)\{1-\pi(X_i,\gamma^*)\}},$$

$$S_{\theta i} = \frac{R_i I(\Delta_i > 0)\{I(\Delta_i = 1) - h(Q_i,\theta^*)\}\dot{h}_\theta(Q_i,\theta^*)}{h(Q_i,\theta^*)\{1-h(Q_i,\theta^*)\}},$$

$$I_\gamma = E\left\{S_{\gamma 1}S'_{\gamma 1} - I(\Delta_1 > 0)\frac{R_1 - \pi(X_1,\gamma^*)}{\pi(X_1,\gamma^*)\{1-\pi(X_1,\gamma^*)\}}\ddot{\pi}_{\gamma\gamma}(X_1,\gamma^*)\right\},$$

$$I_\theta = E\left\{S_{\theta 1}S'_{\theta 1} - R_1 I(\Delta_1 > 0)\frac{I(\Delta_1 = 1) - h(Q_1,\theta^*)}{h(Q_1,\theta^*)\{1-h(Q_1,\theta^*)\}}\ddot{h}_{\theta\theta}(Q_1,\theta^*)\right\},$$

where $\ddot{\pi}_{\gamma\gamma}(\cdot,\gamma) = \partial\dot{\pi}_\gamma(\cdot,\gamma)/\partial\gamma'$ and $\ddot{h}_{\theta\theta}(\cdot,\theta) = \partial\dot{h}_\theta(\cdot,\theta)/\partial\theta'$.

For some $\tau > 0$, we assume the following set of regularity conditions throughout the paper: $P\{Y_i(t) = 1,\, t \in [0,\tau]\} > 0$; $\Lambda_{01}(\tau) < \infty$; the covariate vector $Z_i(t)$ is bounded for $t \in [0,\tau]$; and A is positive definite.

1. Proof of Theorem 1. We know that for any given β, (8) has the following solution for Λ_1,

$$\hat{\Lambda}_1(t,\beta) = \int_0^t \frac{\sum_{i=1}^n \{\frac{R_i}{\pi(X_i,\hat{\gamma})} dN_i(s) - \frac{R_i - \pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})} h(Q_i,\hat{\theta}) dN_i^*(s) - \beta' Z_i(s) Y_i(s) ds\}}{\sum_{i=1}^n Y_i(s)}.$$

Plugging $\hat{\Lambda}_1(t,\beta)$ into (9), we have

$$\sum_{i=1}^n \int_0^\infty \{Z_i(t) - \bar{Z}(t)\}\Big\{\frac{R_i}{\pi(X_i,\hat{\gamma})} dN_i(t) - \frac{R_i - \pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})} h(Q_i,\hat{\theta}) dN_i^*(t) - Y_i(t)\beta' Z_i(t) dt\Big\} = 0. \quad (11)$$

Let $U(\beta,\hat{\gamma},\hat{\theta})$ denote the left-hand side of (11). By the Taylor expansion of $U(\beta_0,\hat{\gamma},\hat{\theta})$ around γ^* and θ^* and some empirical process approximation techniques (Yin and Cai (2004, Thm. 1), we can show that

$$U(\beta_0,\hat{\gamma},\hat{\theta}) = \sum_{i=1}^n \Big[\int_0^\infty \{Z_i(t) - \bar{z}(t)\} dM_i^*(t,\beta_0,\gamma^*,\theta^*,\Lambda_{01}) - B_\gamma I_\gamma^{-1} S_{\gamma i} - B_\theta I_\theta^{-1} S_{\theta i}\Big] + o_p(\sqrt{n}).$$

In addition,

$$\frac{1}{n}\frac{\partial U(\beta,\hat{\gamma},\hat{\theta})}{\partial\beta} = -\frac{1}{n}\sum_{i=1}^n \int_0^\infty Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dt = -A + o_p(1).$$

Thus, Theorem 1 holds.

2. Proof of Theorem 2. We have

$$n^{\frac{1}{2}}\{\hat{\Lambda}_1(t) - \Lambda_{01}(t)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{[\frac{R_i}{\pi(X_i,\hat{\gamma})} dN_i(s) - \frac{R_i - \pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})} h(Q_i,\hat{\theta}) dN_i^*(s) - Y_i(s)\{\hat{\beta}'_{DR} Z_i(s) ds + d\Lambda_{01}(s)\}]}{\frac{1}{n}\sum_{j=1}^n Y_j(s)}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{dM_i^*(u,\beta_0,\gamma^*,\theta^*,\Lambda_{01})}{s^{(0)}(u)} - \xi'(t)\sqrt{n}(\hat{\beta}_{DR} - \beta_0)$$

$$
\begin{aligned}
&-C'_\gamma(t)\sqrt{n}(\hat{\gamma}-\gamma^*) - C'_\theta(t)\sqrt{n}(\hat{\theta}-\theta^*) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\int_0^t \frac{dM_i^*(u,\beta_0,\gamma^*,\theta^*,\Lambda_{01})}{s^{(0)}(u)} - \xi'(t)A^{-1}\phi_i - C'_\gamma(t)I_\gamma^{-1}S_{\gamma i} - C'_\theta(t)I_\theta^{-1}S_{\theta i}\right\} \\
&\quad + o_p(1).
\end{aligned}
$$

Theorem 2 then follows.

References

Andersen, J., Goetghebeur, E. and Ryan, L. (1996). Missing cause of death information in the analysis of survival data. *Statist. Medicine* **15**, 2191-2201.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, New York.

Craiu, R. V. and Duchesne (2004). Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika* **91**, 543-558.

Craiu, R. V. and Reiser B. (2006) Inference for the dependent competing risks model with masked causes of failure. *Lifetime Data Analysis* **12**, 21-33.

Cummings, F. J., Gray, R., Tormey, D. C., Davis, T. E., Volk, H., Harris, J., Falkson, G. and Bennett, J. M. (1993). Adjuvant tamoxifen versus placebo in elderly women with node-positive breast cancer: long-term follow-up and causes of death. *J. Clinical Oncology* **11**, 29-35.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika* **79**, 855-857.

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and of failure data. *Biometrics* **38**, 417-431.

Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J. Amer. Statist. Assoc.* **81**, 328-336.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92**, 875-891.

Goetghebeur, E. and Ryan, L. (1990). A modified logrank test for competing risks with missing failure type. *Biometrika* **77**, 207-211.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821-834.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Kalbfleish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd edition. Wiley, New York.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61-71.

Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73-81.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57**, 1191-1197.

Prentice, R. L. and Kalbfleisch, J. D. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541-554.

Racine-Poon, A. H. and Hoel, D. G. (1984). Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics* **40**, 1151-1158.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91**, 801-818.

Department of Statistics, North Carolina State University, 2501 FoundersDrive, Raleigh, NC 27695, U.S.A.

E-mail: lu@stat.ncsu.edu

SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, U.S.A.

E-mail: yu.liang@sas.com

(Received May 2006; accepted August 2006)