Published in final edited form as:

Stat Med. 2017 February 28; 36(5): 738–753. doi:10.1002/sim.7161.

Analyzing semi-competing risks data with missing cause of informative terminal event

Renke Zhou^a, Hong Zhu^{b,*†}, Melissa Bondy^a, and Jing Ning^c

^aDuncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA

^bDivision of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA

^cDepartment of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Abstract

Cancer studies frequently yield multiple event times that correspond to landmarks in disease progression, including non-terminal events (i.e., cancer recurrence) and an informative terminal event (i.e., cancer-related death). Hence, we often observe semi-competing risks data. Work on such data has focused on scenarios in which the cause of the terminal event is known. However, in some circumstances, the information on cause for patients who experience the terminal event is missing; consequently, we are not able to differentiate an informative terminal event from a non-informative terminal event. In this article, we propose a method to handle missing data regarding the cause of an informative terminal event when analyzing the semi-competing risks data. We first consider the nonparametric estimation of the survival function for the terminal event time given missing cause-of-failure data via the Expectation-Maximization (EM) algorithm. We then develop an estimation method for semi-competing risks data with missing cause of the terminal event, under a pre-specified semiparametric copula model. We conduct simulation studies to investigate the performance of the proposed method. We illustrate our methodology using data from a study of early-stage breast cancer.

Keywords

Copula model; EM algorithm; informative censoring; missing cause of failure; semi-competing risks

1. Introduction

Understanding the disease process in cancer patients is essential in clinical, epidemiological and translational research. Cancer studies frequently yield multiple event times that correspond to landmarks in disease progression. The events may be non-terminal events, such as disease recurrence, or a terminal event, such as death. Semi-competing risks data [1,

*Correspondence to: Hong Zhu, Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

†hong.zhu@utsouthwestern.edu

2] are encountered when the terminal event censors the non-terminal event, but the non-terminal event does not prevent the subsequent observation of the terminal event. An example can be found in a retrospective study of women diagnosed with stage I or II breast cancer, for which a database is maintained as the Early Stage Breast Cancer Repository (ESBCR) at The University of Texas MD Anderson Cancer Center [3]. For this study, breast cancer relapse is the non-terminal event and cancer-related death is the terminal event of interest. Besides estimating the survival function of cancer-related death, it is also crucial to evaluate the survival times of women who experience breast cancer relapse following the initial treatment regimen. This evaluation will inform both predictions of the risk of relapse for patients undergoing treatment and future treatment decisions. However, two challenges arise in analyzing such data. The first is the informative censoring issue in the semi-competing risks data. Specifically, the occurrence of cancer-related death is expected to be positively correlated with cancer relapse and cannot be considered as independent censoring. Consequently, the Kaplan-Meier estimator [4] will tend to overestimate the true survival probability associated with cancer relapse in the presence of cancer-related death. The second challenge comes from missing information on the cause of death. Among 2425 patients with early-stage breast cancer who experienced cancer relapse and for whom there is complete information on vital status information in the ESBCR database, there were 763 observed death events. Within this group, 188 (24.64%) deaths were defined as cancer-related, 70 (9.17%) as cancer-unrelated deaths, and 505 cases (66.19%) had missing cause of death. The cancer-unrelated deaths could be regarded as independent censoring. For the subgroup of patients who died without recorded causes, the question is how to differentiate cancer-related deaths from the deaths from other causes when analyzing the semi-competing risks data.

Analyses of survival data with missing cause of failure have long received considerable interest. As pointed out by Racine-Poon and Hoel [5], forced cause-of-death data can produce substantial bias. Therefore, methods have been developed to modify the Kaplan-Meier procedure [6, 7, 8, 9]. Under the missing-at-random (MAR) assumption, Subramanian [10] introduced an estimator of a survival function in the missing censoring indicator model of random censorship. In particular, competing risks data with missing information regarding the type of failure have been considered by many authors when modeling the cause-specific hazard of interest. Goetghebeur and Ryan [11] considered the regression problem using the partial likelihood for two types of failures under the assumption that the cause-specific hazards for the two failure types are proportional. Lu and Tsiatis [12] proposed a regression analysis method for competing risks data where the cause-specific hazard for the cause of interest is related to covariates through a proportional hazards relationship and when the cause of failure is missing for some individuals. Dewanji and Sengupta [13] developed nonparametric estimation of the different cause-specific hazards, based on data in which a set of possible failure types containing the true type is observed, and without the assumption of proportional hazards due to different types. Sen et al. [14] proposed a semiparametric Bayesian approach for analyzing competing risks survival data with missing cause of death. However, the statistical literature indicates that relatively little attention has been given to analyzing semi-competing risks data in the presence of missing cause of failure for the terminal event. For analyzing such data, one objective is to

nonparametrically estimate the marginal survival function of the terminal event of interest (i.e., cancer-related death). In the ESBCR database, some patients are known to have died but the information on the cause of death was not collected or was lost. Without such information, we cannot distinguish the informative terminal event (cancer-related death) from the non-informative terminal event (cancer-unrelated death), which can be treated as random censoring. As a result, we are not able to use standard methods such as the Kaplan-Meier estimator. Note that simply excluding the missing-cause observations from the analysis or treating them as censored may yield biased results. The other objective is to estimate the marginal survival function of the non-terminal event (i.e., cancer relapse) and the dependence between the non-terminal and terminal events.

The remainder of the article is organized as follows. In Section 2, under the MAR assumption, we propose a nonparametric estimator of the survival function for missing cause-of-failure data via the Expectation-Maximization (EM) algorithm. In Section 3, under a pre-specified semiparametric copula model, we simultaneously estimate the marginal survival function of the non-terminal event and the association parameter for semi-competing risks data with missing cause of the terminal event. Section 4 presents simulation studies to assess the finite-sample performance of the proposed method. We illustrate our method by an application to the ESBCR data in Section 5, and provide concluding remarks in Section 6. We provide details of the asymptotic properties in the Appendix.

2. Nonparametric Estimation of Survival Function for Missing Cause-of-failure Data

2.1. EM algorithm

Let T and C denote the failure time (e.g., time to cancer-related death) and censoring time (e.g., time to cancer-unrelated death or random loss to follow-up), respectively. Let $\delta = I(T \leq C)$ denote the censoring indicator. Assume that T is independent of C . Let $t_1 < \dots < t_J$ be the distinct observed failure times, and λ_j be the hazard rate at the time t_j for $j = 1, \dots, J$. For conventional right-censored survival data, the likelihood function is proportional to

$$L = \prod_{j=1}^J \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j},$$

where d_j is the number of failures at t_j and $n_j = \sum_{i=1}^n I(t_i \geq t_j)$ is the number of individuals at risk prior to t_j . The maximum likelihood estimator (MLE) of the hazard rate is $\hat{\lambda}_j = d_j/n_j$ and the Kaplan-Meier estimator for the survival function is

$$\hat{S}_t = \prod_{j: t_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j: t_j \leq t} (1 - \hat{\lambda}_j). \quad (1)$$

In the presence of missing cause-of-failure data, the censoring indicator δ is not observed for all subjects; hence, the Kaplan-Meier method cannot be applied directly. Let σ denote the observed event type: 0 for independent censoring, 1 for the failure event of interest, and 2 for the failure event with missing cause. Let p_j denote the probability of a failure of interest when the cause of failure is missing at t_j , and $d_{\sigma j}$ denote the number of individuals with the observed event type σ at t_j . Let ξ be the unknown parameter vector $\xi = (\lambda_1, \dots, \lambda_J, p_1, \dots, p_J)$. Let (Y_{obs}, Y_{miss}) denote the complete data, where Y_{obs} and Y_{miss} represent random vectors of observed variables and unobserved common factors, respectively. We assume that the data are missing at random, that is, the probability that the cause of failure is missing depends only on the observed information on all subjects. The likelihood function of the complete data can be written as

$$L = P(Y_{obs}, Y_{miss} | \xi) = \prod_{j=1}^J \lambda_j^{d_{1j} + p_j d_{2j}} (1 - \lambda_j)^{n_j - d_{1j} - p_j d_{2j}}.$$

The corresponding log-likelihood function is

$$l = \sum_{j=1}^J \{ (d_{1j} + p_j d_{2j}) \log \lambda_j + (n_j - d_{1j} - p_j d_{2j}) \log (1 - \lambda_j) \}.$$

Note that both λ_j 's and p_j 's are unknown. We can use the EM algorithm [15] to estimate λ_j by assuming that these p_j 's are observed in the complete data. The MLE of λ_j is

$$\hat{\lambda}_j = \frac{d_{1j} + p_j d_{2j}}{n_j}.$$

The E-step involves taking the conditional expectation of the log-likelihood function given the incomplete observed data. At the s th iteration of the E-step, we have

$$E[p_j^{(s)} | d_{0j}, d_{1j}, d_{2j}] = \frac{\lambda_j^{(s)}}{\lambda_j^{(s)} + \lambda_{cen,j}^{(s)}},$$

where $\lambda_{cen,j}$ is the hazard rate of the censoring time C at t_j . The M-step maximizes the conditional expectation of the log-likelihood function by

$$\lambda_j^{(s+1)} = \frac{d_{1j} + p_j^{(s)} d_{2j}}{n_j}, \quad \lambda_{cen,j}^{(s+1)} = \frac{d_{0j} + (1 - p_j^{(s)}) d_{2j}}{n_j}.$$

The initial $\hat{\lambda}_j^{(0)}$ and $\hat{\lambda}_{cen,j}^{(0)}$ are

$$\hat{\lambda}_j^{(0)} = \frac{d_{1j} + p_j^* d_{2j}}{n_j}, \quad \hat{\lambda}_{cen,j}^{(0)} = \frac{d_{0j} + (1 - p_j^*) d_{2j}}{n_j},$$

respectively, where p_j^* is the initial value for the probability of a failure of interest when observing a failure with missing cause at t_j and is typically set as 0.5. Then we have

$$\hat{d}_j^{(0)} = d_{1j} + \hat{p}_j^{(0)} d_{2j}, \quad \hat{d}_{cen,j}^{(0)} = d_{0j} + (1 - \hat{p}_j^{(0)}) d_{2j},$$

where

$$p_j^{(0)} = \frac{\hat{\lambda}_j^{(0)}}{\hat{\lambda}_j^{(0)} + \hat{\lambda}_{cen,j}^{(0)}}.$$

In the M-step,

$$\hat{\lambda}_j^{(1)} = \frac{\hat{d}_j^{(0)}}{n_j}, \quad \hat{\lambda}_{cen,j}^{(1)} = \frac{\hat{d}_{cen,j}^{(0)}}{n_j}.$$

The E-step and M-step are alternately repeated until convergence is achieved to obtain $\hat{\lambda}_j^{(P)}$.

Then the marginal survival function is calculated by (1) as $\hat{S}_t = \prod_{j:t_j \leq t} (1 - \hat{\lambda}_j^{(P)})$.

2.2. Convergence of the EM algorithm

The hazard rate λ_j and the probability of a failure of interest for the missing-cause failure p_j are specified as piecewise constant models. Inference based on the EM algorithm is valid only if data are missing at random. Under the MAR assumption, we have

$$P(Y_{obs}, Y_{miss} | \xi) = P(Y_{obs} | \xi) P(Y_{miss} | Y_{obs}, \xi),$$

where $\xi = (\lambda_1, \dots, \lambda_J, p_1, \dots, p_J)$. This implies that

$$-\frac{\partial^2}{\partial \xi \partial \xi^T} \log P(Y_{obs} | \xi) = -\frac{\partial^2}{\partial \xi \partial \xi^T} \log P(Y_{obs}, Y_{miss} | \xi) + \frac{\partial^2}{\partial \xi \partial \xi^T} \log P(Y_{miss} | Y_{obs}, \xi).$$

Integrating both sides with respect to $P(Y_{miss} | Y_{obs}, \xi)$, we can obtain

$$-\frac{\partial^2}{\partial \xi \partial \xi^T} \log P(Y_{obs}|\xi) = -\frac{\partial^2}{\partial \xi \partial \xi^T} Q(\xi|\xi') + \frac{\partial^2}{\partial \xi \partial \xi^T} H(\xi|\xi'),$$

where

$$\begin{aligned} Q(\xi|\xi') &= \int \log P(Y_{obs}, Y_{miss}|\xi') P(Y_{miss}|Y_{obs}, \xi') dY_{miss}, \\ H(\xi|\xi') &= \int \log P(Y_{miss}|Y_{obs}, \xi') P(Y_{miss}|Y_{obs}, \xi') dY_{miss}. \end{aligned}$$

Then, the observed log-likelihood can be expressed as

$$l_{obs}(\xi) = Q(\xi|\xi') - H(\xi|\xi').$$

We can use the results of Wu [16] to show that the EM algorithm converges to a stationary point for the case with piecewise constant hazards and probabilities of a failure of interest for the missing-cause failure. Let $\Omega \in \mathbb{R}^{2J}$ denote the parameter space.

Lemma—For any ξ_0 such that $l_{obs}(\xi_0) > -\infty$, the set $\Omega_{\xi_0} = \{\xi \in \Omega : l_{obs}(\xi) = l_{obs}(\xi_0)\}$ is compact in $\mathbb{R}^{2J}$.

Proof: For any time t_j , $\frac{d_{1j}}{n_j} \leq \lambda_j \leq \frac{d_{1j}+d_{2j}}{n_j}$ and $0 < p_j < 1$. Since $l_{obs}(\xi)$ is continuous, the set $\Omega_{\xi_0} = \{\xi \in \Omega : l_{obs}(\xi) = l_{obs}(\xi_0)\}$ is a closed subset of a compact set, and therefore is also compact.

The above Lemma implies that the conditions (5)–(7) in Wu [16] are satisfied. Further, $Q(\xi|\xi')$ is continuous in both ξ and ξ' . By Theorem 2 in Wu [15], it implies that the limit points of any instance $\xi^{(k)}$ of the EM iteration are stationary points of l_{obs} , and $l_{obs}(\xi^{(k)})$ converges monotonically to $l_{obs}(\xi^*)$ for some stationary point ξ^* . The key assumptions and theoretical results in Wu [16] are stated in the Appendix.

2.3. Asymptotic Variance Estimation

The EM algorithm defines a mapping $\xi^{(k+1)} = M(\xi^{(k)})$, for the iteration $k = 0, 1, \dots$. Meng and Rubin [17] showed that the desired variance matrix $V = \text{var}(\hat{\xi})$ is given by

$$V = I_{oc}^{-1} + I_{oc}^{-1} DM(I - DM)^{-1},$$

where I_{oc} is the complete information matrix, and

$$DM = \left\{ \frac{\partial M_j(\xi)}{\partial \xi_i} \right\} \Big|_{\xi = \hat{\xi}_{MLE}}$$

is the Jacobian matrix of the map M evaluated at $\xi = \hat{\xi}_{MLE}$, and I is the identity matrix. Although DM cannot be obtained analytically in most applications, we can determine its elements by running a series of restricted EM algorithms, following Meng and Rubin [17]. The procedure is computationally convenient because each of these constrained EM algorithms is similar to the one in the original problem but uses different initial values for $\xi^{(0)}$.

In our case, the complete data log-likelihood is proportional to

$$l = \sum_{j=1}^J \{ (d_{1j} + p_j d_{2j}) \log \lambda_j + (n_j - d_{1j} - p_j d_{2j}) \log (1 - \lambda_j) \}.$$

The complete data score function at t_j is,

$$S(\lambda_j | d_j) = \frac{\partial l(\lambda_j | d_j)}{\partial \lambda_j} = \frac{d_j}{\lambda_j} - \frac{n_j - d_j}{1 - \lambda_j}$$

We also have that

$$\frac{\partial^2 l(\lambda_j | d_j)}{\partial \lambda_j^2} = -\frac{d_j}{\lambda_j^2} - \frac{n_j - d_j}{(1 - \lambda_j)^2}.$$

Then

$$\begin{aligned} I_c(\lambda_j | d_{j,obs}) &= -E \left\{ \frac{\partial^2 l(\lambda_j | d_{j,obs}, d_{j,miss})}{\partial \lambda_j^2} | d_{j,obs}, \hat{\lambda}_j \right\} \\ &= \frac{d_{1j} + p_j^{(P)} d_{2j}}{\hat{\lambda}_j^2} + \frac{n_j - (d_{1j} + p_j^{(P)} d_{2j})}{(1 - \hat{\lambda}_j)^2}, \end{aligned}$$

and

$$\begin{aligned} I_m(\lambda_j | d_{j,obs}) &= \text{Var} \left\{ S(\lambda_j | d_{j,obs}, \lambda_j = \hat{\lambda}_j) \right\} \\ &= \{ \hat{\lambda}_j^{-2} - (1 - \hat{\lambda}_j)^{-2} \} d_{2j} p_j^{(P)} (1 - p_j^{(P)}), \end{aligned}$$

where $p_j^{(P)}$ is the p_j value at the last iteration. The observed information is

$$I(\lambda_j | d_{j,obs}) = I_c(\lambda_j | d_{j,obs}) - I_m(\lambda_j | d_{j,obs}).$$

Now we have that the asymptotic variance of $\hat{\lambda}_j$ is $I(\hat{\lambda}_j | Y_{j,obs})^{-1}$. To estimate the variance of $\hat{S}(t)$, we first estimate the variance of $\log \hat{S}(t) = \sum_{j:t_j \leq t} \log(1 - \hat{\lambda}_j)$. Based on the Delta method, an asymptotic variance estimator of $\log \hat{S}(t)$ is

$$\begin{aligned}\widehat{Var}[\log \hat{S}(t)] &= \sum_{j:t_j \leq t} (1 - \hat{\lambda}_j)^{-2} \widehat{Var}(1 - \hat{\lambda}_j) \\ &= \sum_{j:t_j \leq t} (1 - \hat{\lambda}_j)^{-2} \widehat{Var}(\hat{\lambda}_j).\end{aligned}$$

We can then estimate the asymptotic variance of $\hat{S}(t)$ by the Delta method,

$$\widehat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j:t_j \leq t} (1 - \hat{\lambda}_j)^{-2} \widehat{Var}(\hat{\lambda}_j).$$

3. Semiparametric Copula Model with Missing Cause of Informative Terminal Event

Let T_1 be the time from the initial event (i.e., initial treatment) to the non-terminal event (i.e., cancer relapse), and T_2 be the time from the initial event to the terminal event of interest (i.e., cancer-related death). The random variable T_2 may dependently censor T_1 , but not vice versa. Let C be the independent censoring time of both T_1 and T_2 , such as an administrative loss to follow-up. Define $Z = \min(T_1, T_2)$, $X = \min(Z, C)$, $Y = \min(T_2, C)$, $\delta_1 = I(T_1 \leq Y)$, $\delta_2 = I(T_2 \leq C)$, and $\delta_Z = I(Z \leq C)$. The observed data can be denoted as $\{(X_i, \delta_{1i}, Y_i, \delta_{2i}, Z_i, \delta_{Zi}), i = 1, \dots, n\}$. In the presence of missing cause of the terminal event, the censoring indicator δ_2 is not observed for all the subjects and the marginal survival function of T_2 cannot be estimated by the Kaplan-Meier estimator.

With semi-competing risks data, the dependence structure needs to be determined to ensure the validity for the estimation of the marginal survival function of T_1 and is usually of biological interest by itself. In the presence of missing information on the cause of the terminal event (death), we are not able to directly use the diagnosis method of Chen and Bandeen-Roche [18]; hence, we consider two extreme scenarios: (1) all deaths with missing causes are treated as cancer-related; and (2) all deaths with missing causes are treated as independent censoring and are excluded. As shown in Figure 1, under both scenarios, the estimated cross ratio generally increases with an increase in the estimated survival function. This suggests that the positive stable copula would capture the underlying dependence structure for the ESBCR data. Note that the distribution of (T_1, T_2) is identifiable only on the upper triangle where $T_1 \leq T_2$. Without loss of generality, we assume the positive stable copula model for the bivariate survival function of (T_1, T_2) on the upper triangle as

$$S(t_1, t_2) = \exp \left(- \left[\{-\log S_1(t_1)\}^{\frac{1}{\alpha}} + \{-\log S_2(t_2)\}^{\frac{1}{\alpha}} \right]^{\alpha} \right), 0 < \alpha < 1, 0 \leq t_1 \leq t_2.$$

Given the observed data, the log-likelihood function under the positive stable copula model is

$$\begin{aligned} l(\alpha, S_1, S_2) = & \sum_{i=1}^n -\delta_{1i} \log S_1(X_i) - \delta_{2i} \log S_2(Y_i) + (\alpha^{-1} - 1) [\delta_{1i} \log \{-\log S_1(X_i)\} \\ & + \delta_{2i} \log \{-\log S_2(Y_i)\}] - [\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}}]^{\alpha} \\ & + \{\alpha(\delta_{1i} + \delta_{2i} - \delta_{1i}\delta_{2i}) - \delta_{1i} - \delta_{2i}\} \log [\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}}] \\ & + \delta_{1i}\delta_{2i} \log \left([\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}}]^{\alpha} - 1 + \alpha^{-1} \right). \end{aligned}$$

The marginal survival function of T_1 can be expressed as

$$S_1(t) = g\{S_2(t), S_Z(t), \alpha\} = \exp \left(- \left[\{-\log S_Z(t)\}^{\frac{1}{\alpha}} - \{-\log S_2(t)\}^{\frac{1}{\alpha}} \right]^{\alpha} \right),$$

where $g(u, v, \alpha) = \exp \left[- \left\{ (-\log v)^{\frac{1}{\alpha}} - (-\log u)^{\frac{1}{\alpha}} \right\}^{\alpha} \right]$. Then the log-likelihood function of α can be re-written as

$$l^g(\alpha, S_2, S_Z) = l\{\alpha, g(S_2, S_Z, \alpha), S_2\}. \quad (2)$$

Since the censoring indicators δ_2 and δ_Z are not always observable, S_2 and S_Z cannot be directly estimated by Kaplan-Meier estimators. By the method proposed in Section 2 for right-censored survival data with missing cause of failure, the hazard functions for T_2 and Z can be estimated using the EM algorithm. On the basis of the estimated hazard functions, the consistent product-limit estimators for marginal survival functions S_2 and S_Z can be obtained and are denoted as $\hat{S}_2$ and $\hat{S}_Z$, respectively. Further, in the log-likelihood function, δ_2 may not be observed for some subjects, and we replace the unknown δ_2 by the estimated probability of a failure of interest when the cause of failure is missing, obtained from the EM algorithm. Then, the estimator $\hat{\alpha}$ is the solution to the estimating equation

$$U_{\alpha}(\alpha, \hat{S}_2, \hat{S}_Z) = \frac{\partial l^g(\alpha, \hat{S}_2, \hat{S}_Z)}{\partial \alpha} = \sum_i \frac{\partial l_0^g(\alpha, \hat{S}_2, \hat{S}_Z)}{\partial \alpha} = 0.$$

Under some regularity conditions, we prove that $\hat{\alpha}$ is consistent and asymptotically normally distributed. The asymptotic properties of $\hat{\alpha}$ are summarized in Theorem 1.

Theorem 1—The estimator $\hat{\alpha}$ is a consistent estimator of the true value α_0 , and as $n \rightarrow \infty$, $n^{1/2}(\hat{\alpha} - \alpha_0)$ converges weakly to a normal distribution with a mean of zero and variance $\rho^2 = (\rho_1^2 + \rho_2^2) / \rho_1^4$.

Details of the regularity conditions, definitions of ρ_1^2 and ρ_2^2 , and the discussion of Theorem 1 are provided in the Appendix. For the marginal survival function $S_1(t)$, a natural estimator is obtained as $\hat{S}_1(t) = g\{\hat{S}_2(t), \hat{S}_Z(t), \hat{a}\}$. We have that $\hat{S}_2(t)$ and $\hat{S}_Z(t)$ are strongly consistent with $S_2(t)$ and $S_Z(t)$, uniformly for $t \in [0, \tau]$ [19]. Since function g has bounded derivatives, a continuous mapping theorem gives the uniform convergence of $\hat{S}_1(t)$ to $S_1(t)$.

4. Simulation Studies

We carry out a set of simulations to evaluate the performance of the EM algorithm for estimating the survival function for right-censored survival data in the presence of missing cause-of-failure data. We generate T from a unit exponential distribution. The independent censoring time C follows a uniform distribution of $U(0, 2.25)$, giving 40% censoring. We simulate 500 datasets with sample sizes $n = 250$ and 500. For each simulated dataset, we (A) randomly select 70% censoring and 50% terminal failure and treat them as the missing cause of terminal events, and correspondingly, the censoring rate, failure rate and missing cause rate are 12%, 30% and 58%, respectively; (B) randomly select 40% censoring and 20% terminal failure and treat them as the missing cause of terminal events, and correspondingly, the censoring rate, failure rate and missing cause rate are 24%, 48% and 28%, respectively. Table 1 presents the estimation results of the survival function in various simulation scenarios. At each time point $T = 0.357, 0.693$ and 1.204 , which correspond to the true survival functions $S(t) = 0.7, 0.5$ and 0.3 , we calculate the proposed estimator $\hat{S}(t)^a$ by the EM algorithm, its model-based standard error and empirical standard error, the Kaplan-Meier (KM) estimator $\hat{S}(t)^b$ if all events with missing causes are treated as failures, the KM estimator $\hat{S}(t)^c$ if all events with missing causes are treated as independent censoring, and the KM estimator $\hat{S}(t)^d$ if we exclude all events with missing causes. Compared to the KM estimators $\hat{S}(t)^b$, $\hat{S}(t)^c$ and $\hat{S}(t)^d$, the proposed estimator $\hat{S}(t)^a$ is generally closer to the true value. The model-based standard errors are small and get closer to the empirical standard errors as the uncertainty rate decreases or the sample size increases, suggesting a good performance of the proposed inference procedure. Note that the coverage probabilities of the 95% confidence interval for $\hat{S}(t)^a$ depend on both the empirical biases and model-based standard errors of $\hat{S}(t)^a$. The model-based standard errors tend to be conservative and may overestimate the variance, thus the coverage probabilities are not exactly 95% but are in a reasonable range. Figure 2 shows the estimated survival functions in simulations with $n = 500$. Plotted in this figure are the true survival function $S(t)$, the proposed estimator by the EM algorithm $\hat{S}(t)^a$, the KM estimators if all events with missing causes are treated as failures $\hat{S}(t)^b$, or as independent censoring $\hat{S}(t)^c$, or if all events with missing causes are excluded $\hat{S}(t)^d$. Compared to the KM estimators, the proposed estimator obtained by the EM algorithm is closer to the true survival curve in most situations, indicating better performance of the proposed method.

We conduct a second set of simulations to evaluate the performance of the proposed estimators for semi-competing risks data with missing cause of the informative terminal event. We generate n pairs of (T_1, T_2) from the positive stable copula with different strengths of the solicitation by choosing different values of α . The marginal survival functions S_1 and S_2 are unit exponentials. We generate the independent censoring time C from $U(0, 2.25)$,

giving 40% independent censoring. For each combination of $\alpha = 0.2, 0.4, 0.6, 0.8$ and $n = 250, 500$, we simulate 500 datasets. Again, we randomly select 70% censoring and 50% terminal failure or 40% censoring and 20% terminal failure as the missing cause of terminal events. Table 2 presents the estimated association parameter $\hat{\alpha}$ in different simulation scenarios. Generally, the estimator $\hat{\alpha}$ performs well with small bias. The bias decreases as the sample size increases. The stronger the association between the non-terminal and terminal events (smaller value of α), the smaller is the bias. Figure 3 shows the estimated marginal survival functions of the non-terminal event time for the simulated semi-competing risks data with missing cause of terminal event ($n = 500$), when 40% censoring and 20% terminal failure are randomly selected as the missing cause of the terminal events. In the figure, we plot the true margin S_1 , the proposed estimators $\hat{S}_1$, and the naive KM estimator of S_1 . Figure 3 shows that the proposed estimator $\hat{S}_1$ agrees well with the true margin. The naive KM estimator of S_1 departs far from the true survival curve, suggesting that the ignorance of dependent censoring from the terminal event could result in substantial overestimation of the true marginal survival function, especially when the association between the non-terminal and terminal events is strong for a small value of α .

5. Application to ESBCR Data

The ESBCR was created for a retrospective cohort study of 2470 women diagnosed with stage I or II breast cancer (according to American Joint Committee on Cancer pathologic staging criteria) who were treated at The University of Texas MD Anderson Cancer Center between January 1, 1985 and December 31, 2000 [3]. The inclusion criteria for the ESBCR study were female patients with stage I or II breast cancer, who were residents of the state of Texas, underwent surgical treatment for breast cancer at MD Anderson, had tissue samples available, and for whom a complete assessment of the medical records was available.

Among the 2470 ESBCR patients, 2425 patients had experienced cancer relapse and had complete vital status information, which was 98.2% of the total patients in the study. The mean age at diagnosis was 54.4 years, with a standard deviation of 12.7 years. The majority of women were white (77.73%). African Americans were 9.57% of the study population, Hispanics were 9.44% and others were 3.26%. About 42.06% of the patients had stage I breast cancer at diagnosis and 57.94% had stage II. The non-terminal event was the first breast cancer relapse, which was defined as the occurrence of cancer in a local lymph node or a breast tumor recurrence; metastasis to the contralateral breast, chest wall, or other body sites; or a second primary breast cancer (including ductal carcinoma in situ). In the analysis, T_1 was defined as the number of years from the earliest treatment (date of definitive surgical procedure or start of neoadjuvant chemotherapy) to the date of first breast cancer relapse. There were 562 patients who reported disease relapse, with a mean time of 4.7 years and standard deviation of 3.6 years, and 763 patients who died by the end of the study. Among the 763 patients, 188 (24.64%) deaths were cancer-related, which was the terminal event of interest. The time to cancer-related death was denoted by T_2 , and the mean of the observed T_2 was 5.8 years, with a standard deviation of 4 years. There were 70 (9.17%) patients who died from cancer-unrelated causes, which could be considered as independent censoring, and 505 (66.19%) patients who had missing information on the cause of death, which are the

terminal events with missing cause. The censoring rate for breast cancer relapse is 76.8% and that for cancer-related death is 71.4%.

We analyze the data using the proposed semi-competing risks model with missing cause of the informative terminal event. The estimated association parameter $\hat{a}$ in the positive stable copula is 0.574, with the bootstrap standard error of 0.019 from 200 resamples, suggesting a moderate positive association between breast cancer relapse and cancer-related death in the ESBCR data. We also estimate the cross ratio [20], $\theta(t_1, t_2) = \lambda(t_1|T_2 = t_2)/\lambda(t_1|T_2 > t_2)$, as another association measure. The estimated cross ratio is not constant over time, but decreases with the time to cancer relapse and death. Table 3 shows the estimation results for the marginal survival functions, bivariate survival function and cross ratio of cancer relapse and cancer-related death at 5, 10, 15 and 20 years after the initial treatment in the ESBCR data. The estimated cross ratio $\hat{\theta}$ describes the strength of dependence between cancer relapse and cancer-related death. For example, $\hat{\theta}(t_1 = 5, t_2 = 5) = 4.978$ indicates that the risk of cancer-related death among patients who have cancer relapse at 5 years after the initial treatment is 4.978 times that of those who are free of cancer at that time. The estimated cross ratio $\hat{\theta}$ at 10, 15, and 20 years after the initial treatment is respectively 3.428, 2.783 and 2.450, showing a trend of decreasing association over time. The association between breast cancer relapse and cancer-related death in the ESBCR data is presented by a plot of the estimated cross ratio versus the estimated bivariate survival function based on the positive copula model, in Figure 4. Figure 5 presents the estimated survival function $\hat{S}_1$ for the time to cancer relapse with the 95% point-wise confidence intervals, the estimated survival function $\hat{S}_2$ for the time to cancer-related death, and the naive KM estimator of S_1 when considering the cancer-related death as independent censoring. The substantial difference between the proposed estimator $\hat{S}_1$ and the naive KM estimator suggests that ignoring dependent censoring may provide misleading results and overestimate the survival function of cancer relapse. For example, the estimated disease-free survival rate at 20 years obtained by the proposed method is 0.651, with the 95% confidence interval of (0.611, 0.687), while that obtained by the naive KM estimator is 0.682, with the 95% confidence interval of (0.653, 0.713).

6. Discussion

In this article, we proposed a flexible modeling framework for analyzing semi-competing risks data with missing cause of the informative terminal event. Missing information regarding the cause of death is commonly encountered when analyzing data from large registries and population-based studies. We used an EM algorithm to develop the nonparametric estimator of the survival function for right-censored survival data with missing cause of failure. On this basis, we proposed simultaneous inference on the marginal survival function of the non-terminal event time and the association between the non-terminal and terminal event times under a pre-specified semiparametric copula model for semi-competing risks data. Although we focused on the positive stable copula for the dependence structure of the bivariate event times, the inferential procedure we proposed is applicable to the general Archimedean copula models. The appropriate selection of the copula model largely depends on the observed data from specific applications.

In many cancer studies, it is of interest to determine the impact of pre-treatment characteristics or treatments on cancer progression and survival. A relevant issue is how to incorporate covariates and formulate their effects on the survival functions of the event times via regression models. In the context of missing information on the cause of the terminal event, the problem may not be simple but is worthy of further investigation in a future study.

Acknowledgments

This work was supported in part by grants from the National Institutes of Health and the Cancer Prevention Research Institute of Texas. Hong Zhu is partially supported by the Cancer Center Support Grant from the National Cancer Institute (2P30CA142543) and the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001105). Jing Ning is partially supported by grants from the National Cancer Institute (R01CA193878 and P30CA016672). Renke Zhou is partially supported by a predoctoral fellowship grant from the Cancer Prevention Research Institute of Texas (RP140103).

References

1. Fine J, Jiang H, Chappell R. On semi-competing risks data. *Biometrika*. 2001; 88:907–919.
2. Peng, L.; Jiang, H.; Chappell, R.; Fine, J. *Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*. John Wiley and Sons; New Jersey: 2007.
3. Brewster A, Do K, Thompson P, Hahn K, Sahin A, Cao Y, Stewart M, Murray J, Hortobagyi G, Bondy M. Relationship between epidemiologic risk factors and breast cancer recurrence. *Journal of Clinical Oncology*. 2007; 25:4438–4444. [PubMed: 17785707]
4. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
5. Racine-Poon A, Hoel D. Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics*. 1984; 40:1151–1158. [PubMed: 6534414]
6. Dinse G. Nonparametric estimation for partially-complete time and type of failure data. *Biometrics*. 1982; 38:417–431. [PubMed: 6896831]
7. Miyakawa M. Analysis of incomplete data in competing risks model. *IEEE Transactions in Reliability*. 1984; 33:293–296.
8. Lo S. Estimating a survival function with incomplete cause-of-death data. *Journal of Multivariate Analysis*. 1991; 39:217–235.
9. Mukerjee H, Wang J. Nonparametric maximum likelihood estimation of increasing hazard rate for uncertain cause-of-death data. *Scandinavian Journal of Statistics*. 1993; 20:17–33.
10. Subramanian S. Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Statistical Methodology*. 2006; 3:125–136. [PubMed: 18953423]
11. Goetghebuer E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika*. 1995; 82:821–833.
12. Lu K, Tsiatis A. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*. 2001; 57:1191–1197. [PubMed: 11764260]
13. Dewanji A, Sengupta D. Estimation of competing risks with general missing pattern in failure type. *Biometrics*. 2003; 59:1063–1070. [PubMed: 14969486]
14. Sen A, Banerjee M, Li Y, Noone A. A Bayesian approach to competing risks analysis with masked cause of death. *Statistics in Medicine*. 2010; 29:1681–1695. [PubMed: 20575048]
15. Dempster A, Laird N, Rubin D. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–22.
16. Wu C. On the convergence of the EM algorithm. *The Annals of Statistics*. 1983; 11:95–103.
17. Meng X, Rubin D. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*. 1991; 86:899–909.

18. Chen MC, Bandeen-Roche K. A diagnostic for association in bivariate survival models. *Lifetime Data Analysis*. 2005; 11:245–264. [PubMed: 15940823]
19. Fleming, T.; Harrington, D. *Counting Processes and Survival Analysis*. John Wiley and Sons; New Jersey: 2005.
20. Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*. 1989; 84:487–493.
21. Shih J, Louis T. Inferences on the association parameters in copula models for bivariate survival data. *Biometrics*. 1995; 51:1384–1399. [PubMed: 8589230]

Appendix

A.1 Key assumptions and theoretic results in Wu (1983)

In the following, we state the key assumptions and theoretical results in Wu (1983) that are used in Section 2.2 (Convergence of the EM algorithm).

Let X denote the unobserved complete data with the density function $f(x|\phi)$, where $\phi \in \Omega$. Let Y denote the observed incomplete data, and its density function be given by $g(y|\phi) = \int_{\mu(y)} f(x|\phi) dx$, where $\mu(y) = \{x : y(x) = y\}$.

Let $k(x|y, \phi) = f(x|\phi)/g(y|\phi)$ be the conditional density of X given Y and ϕ . Then the log-likelihood is

$$L(\phi') = \log g(y|\phi') = Q(\phi'|\phi) - H(\phi'/\phi),$$

where $Q(\phi'|\phi) = E\{\log f(x|\phi')|y, \phi\}$ and $H(\phi'|\phi) = E\{\log k(x|y, \phi')|y, \phi\}$ are assumed to exist for all pairs (ϕ, ϕ') . The EM iteration $\phi_p \rightarrow \phi_{p+1} \in M(\phi_p)$ is defined as follows: (i) E-step, determine $Q(\phi|\phi_p)$; (ii) M-step, choose ϕ_{p+1} to be any value of $\phi \in \Omega$ which maximizes $Q(\phi|\phi_p)$. Note that M is a point-to-set map, i.e., $M(\phi_p)$ is the set of ϕ values which maximizes $Q(\phi|\phi_p)$ over $\phi \in \Omega$.

The following assumptions were made in Wu (1983):

- (5) Ω is a subset in the r -dimensional Euclidean space R^r ,
- (6) $\Omega_{\phi_0} = \{\phi \in \Omega : L(\phi) \geq L(\phi_0)\}$ is compact for any $L(\phi_0) > -\infty$,
- (7) L is continuous in Ω and differentiable in the interior of Ω .

As a consequence of (5), (6), (7), we have

- (8) $\{L(\phi_p)\}_{p=0}^{\infty}$ is bounded above for any $\phi_0 \in \Omega$.

Theorem 2 (Wu, 1983)

Suppose $Q(\psi|\phi)$ is continuous in both ψ and ϕ . Then all the limit points of any instance $\{\phi_p\}$ of an EM algorithm are stationary points of L and $L(\phi_p)$ converges monotonically to $L^* = L(\phi^*)$ for some stationary point ϕ^* .

A.2 Details of Theorem 1

We introduce two regularity conditions for the asymptotic properties of $\hat{a}$ as follows:

- i. The standard regularity conditions for the maximum likelihood estimator hold.
- ii. The functions $W_{\alpha}\{\alpha, S_2(t_2), S_Z(z)\}$, $V_{\alpha}\{\alpha, S_2(t_2), S_Z(z)\}$, $V_{\alpha,1}\{\alpha, S_2(t_2), S_Z(z)\}$, and $V_{\alpha,2}\{\alpha, S_2(t_2), S_Z(z)\}$ are continuous and bounded for $(z, t_2) \in \mathcal{A} = [0, t_{02}] \times [0, z_0]$, where

$$W_{\alpha}(\alpha, u, v) = \frac{\partial l_0^g(\alpha, u, v)}{\partial \alpha}, \quad V_{\alpha}(\alpha, u, v) = \frac{\partial^2 l_0^g(\alpha, u, v)}{\partial \alpha^2},$$

$$V_{\alpha,1}(\alpha, u, v) = \frac{\partial^2 l_0^g(\alpha, u, v)}{\partial \alpha \partial u}, \quad V_{\alpha,2}(\alpha, u, v) = \frac{\partial^2 l_0^g(\alpha, u, v)}{\partial \alpha \partial v},$$

$$t_{02} = \sup\{t: \Pr(T_2 > t, C > t) > 0\}, \quad z_0 = \sup\{z: \Pr(Z > z, C > z) > 0\}, \text{ and we write } (u, v) \text{ for } (S_2, S_Z).$$

The specification of $\mathcal{A}$ is required to ensure that $\hat{S}_2(t_2)$ is consistent for $S_2(t_2)$ with $t_2 \in [0, t_{02}]$ and $\hat{S}_Z(z)$ for $S_Z(z)$ with $z \in [0, z_0]$. The proof of Theorem 1 generally follows the proof of Theorem 2 in Shih and Louis [21]. The formulas for ρ_1^2 and ρ_2^2 are specified as

$$\rho_1^2 = E\{-V_{\alpha}(\alpha_0, S_2, S_Z)\} = \int_{\mathcal{A}} -V_{\alpha}\{\alpha_0, S_2(t_2), S_Z(z)\} dJ_{\alpha_0}(t_2, z, \delta_2, \delta_z),$$

$$\rho_2^2 = E[\{I_1(Y_1, \delta_{21}, \alpha_0) + I_2(Z_1, \delta_{z1}, \alpha_0)\}^2]$$

$$= \int_{\mathcal{A}} \{I_1(t_2, \delta_2, \alpha_0) + I_2(z, \delta_z, \alpha_0)\}^2 dJ_{\alpha_0}(t_2, z, \delta_2, \delta_z),$$

where J_{α_0} is the joint distribution of (Y_i, δ_{2i}) and (Z_i, δ_{zi}) . For $i = 1, \dots, n$, I_1 and I_2 are defined by

$$I_1(T_{2i}, \delta_{2i}, \alpha_0) = \int_{\mathcal{A}} V_{\alpha,1}\{\alpha_0, S_2(t_2), S_Z(z)\} I_1^0(Y_i, \delta_{2i})(t_2) dJ_{\alpha_0}(t_2, z, \delta_2, \delta_z),$$

$$I_2(Z_i, \delta_{zi}, \alpha_0) = \int_{\mathcal{A}} V_{\alpha,2}\{\alpha_0, S_2(t_2), S_Z(z)\} I_2^0(Z_i, \delta_{zi})(z) dJ_{\alpha_0}(t_2, z, \delta_2, \delta_z),$$

where

$$I_1^0(Y_i, \delta_{2i})(t_2) = -S_2(t_2) \left\{ \int_0^{t_2} \frac{dN_{2i}(u)}{p(T_2 \geq u, C \geq u)} - \int_0^{t_2} \frac{I(Y_i \geq u) d\Lambda_2(u)}{p(T_2 \geq u, C \geq u)} \right\},$$

$$I_2^0(Z_i, \delta_{zi})(z) = -S_Z(z) \left\{ \int_0^z \frac{dN_{zi}(u)}{p(Z \geq u, C \geq u)} - \int_0^z \frac{I(Z_i \geq u) d\Lambda_z(u)}{p(Z \geq u, C \geq u)} \right\},$$

$$N_{2i}(u) = I(Y_i \leq u, \delta_{2i} = 1), \quad N_{zi}(u) = I(Z_i \leq u, \delta_{zi} = 1), \quad i = 1, 2, \dots, n,$$

and $\Lambda_2(u)$ and $\Lambda_z(u)$ are the cumulative hazard functions for T_2 and Z , respectively. A variance estimator of $\hat{a}$ may be obtained as $\hat{\rho}^2 = (\hat{\rho}_1^2 + \hat{\rho}_2^2) / \hat{\rho}_1^4$, by replacing J_{α_0} by its empirical distribution function J_n , and replacing S_2 , S_Z , and α_0 by the proposed estimators

$\hat{S}_2$, $\hat{S}_Z$, and $\hat{\alpha}$, respectively, in ρ_1^2 and ρ_2^2 . The consistency of J_n , $\hat{S}_2$, $\hat{S}_Z$ and $\hat{\alpha}$ implies that $\hat{\rho}^2$ is a consistent estimator of ρ^2 . Given the established weak convergence under the regularity conditions, a bootstrap resampling procedure can be used to estimate the asymptotic variance of $\hat{\alpha}$.

A.4 Additional simulation studies for evaluation of the proposed methods

First, we conduct additional simulations in which the bivariate survival data were generated from the Clayton copula model with exponential margins. The simulation results are presented in Table 4 and Figure 6, which indicate good finite sample performances of the proposed methods for the case when the bivariate data are generated from the Clayton copula model.

Second, we carry out additional simulations where S_1 and S_2 were generated from the Weibull distribution with a scale parameter of 1 and a shape parameter of 3 (increasing hazard). The censoring time was from $U(0, 1.5)$, giving 59% independent censoring. The simulation results are presented in Table 5 and Figure 7.

Lastly, we conduct additional simulation studies to evaluate the impact of the initial value p^* on the estimation results in the EM-algorithm. Table 6 summarizes the simulation results for the estimation of the survival function of right-censored survival data in the presence of missing information on the cause of failure with $p^* = 0.3$ and $p^* = 0.7$. Table 7 summarizes the simulation results for the estimation of the association parameter α in the positive stable copula model for semi-competing risks data with missing cause of the terminal event, with $p^* = 0.3$ and $p^* = 0.7$. It shows that the proposed methods work well and are robust to different initial values of p^* .

Figure 1. Relationship between the estimated cross ratio and the estimated bivariate survival function for the ESBCR data. Left: scenario (1); Right: scenario (2).

Figure 2.

Estimated survival functions for simulated survival data with missing cause of failure ($n = 500$). Plotted are the true survival function $S(t)$, the proposed estimator by the EM algorithm $\hat{S}(t)^a$, the KM estimators if all events with missing causes are treated as failures $\hat{S}(t)^b$, or as independent censoring $\hat{S}(t)^c$, or if all events with missing causes are excluded $\hat{S}(t)^d$. (A) 70% censoring and 50% failure as missing cause of failure; (B) 40% censoring and 20% failure as missing cause of failure.

Figure 3.

Estimated marginal survival functions of the non-terminal event time for simulated semi-competing risks data with missing cause of the terminal event ($n = 500$).

Figure 4.
Plot of the association between cancer relapse and cancer-related death in the ESBCR data.

Figure 5.

Estimated marginal survival functions of the time to breast cancer relapse and the time to cancer-related death in the ESBCR data with missing cause of death. The proposed estimator $\hat{S}_1$ with the 95% point-wise confidence intervals, the proposed estimator $\hat{S}_2$ and the naive KM estimator of S_1 are plotted.

Figure 6.

Estimated marginal survival functions of the non-terminal event time for simulated semi-competing risks data with the Clayton (gamma) copula ($\alpha = 2$) and Exponential margins ($n = 500$).

Figure 7. Estimated marginal survival functions of the non-terminal event time for simulated semi-competing risks data with Weibull margins ($n = 500$).

Table 1

Estimation results of survival function in simulated survival data with missing cause-of-failure data. $k(\cdot)$, the empirical bias of the estimator of $S(\rho)$; $\hat{S}(\rho)^a$, the proposed estimator by the EM algorithm; $\hat{S}(\rho)^b$, the KM estimator if all events with missing causes are treated as failures; $\hat{S}(\rho)^c$, the KM estimator if all events with missing causes are treated as independent censoring; $\hat{S}(\rho)^d$, the KM estimator if all events with missing causes are excluded; SEM, model-based standard error; SEE, empirical standard error; CP, coverage probability of the 95% confidence interval for $\hat{S}(\rho)^a$; Cens%, censoring rate; Fail%, failure rate; Uncert%, rate of the uncertain/missing cause.

(A) 70% censoring and 50% failure as missing cause of failure												
Cens%	Fail%	Uncert%	n	T	S(t)	$b(\hat{S}(\rho)^a)$	SEM	SEE	$b(\hat{S}(\rho)^b)$	$b(\hat{S}(\rho)^c)$	CP	
12%	30%	58%	250	0.357	0.7	0.031	0.037	0.025	-0.079	0.136	-0.048	93.2%
				0.693	0.5	0.040	0.044	0.031	-0.111	0.207	-0.068	92.6%
				1.204	0.3	0.031	0.048	0.032	-0.124	0.245	-0.078	98.4%
			500	0.357	0.7	0.031	0.026	0.019	-0.079	0.136	-0.048	91.0%
				0.693	0.5	0.038	0.031	0.022	-0.114	0.206	-0.070	90.0%
				1.204	0.3	0.030	0.034	0.023	-0.125	0.247	-0.077	94.6%
(B) 40% censoring and 20% failure as missing cause of failure												
Cens%	Fail%	Uncert%	n	T	S(t)	$k(\hat{S}(\rho)^a)$	SEM	SEE	$k(\hat{S}(\rho)^b)$	$k(\hat{S}(\rho)^c)$	CP	
24%	48%	28%	250	0.357	0.7	0.009	0.034	0.028	-0.047	0.050	-0.029	97.4%
				0.693	0.5	0.012	0.039	0.034	-0.068	0.073	-0.042	97.0%
				1.204	0.3	0.008	0.042	0.034	-0.080	0.079	-0.047	98.0%
			500	0.357	0.7	0.011	0.024	0.021	-0.045	0.052	-0.028	95.4%
				0.693	0.5	0.012	0.028	0.024	-0.068	0.074	-0.042	96.2%
				1.204	0.3	0.010	0.030	0.025	-0.079	0.082	-0.046	98.0%

Simulation results of the association parameter α in the positive stable copula model for semi-competing risks data with missing cause of the terminal event. SEE, empirical standard error.

Table 2

(A) 70% censoring and 50% terminal failure as missing cause							
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
250	68%	12%	30%	58%	0.2	0.266	0.027
	66%				0.4	0.460	0.045
	64%				0.6	0.649	0.054
	62%				0.8	0.825	0.059
500	68%	12%	30%	58%	0.2	0.266	0.018
	66%				0.4	0.460	0.033
	64%				0.6	0.654	0.039
	62%				0.8	0.831	0.039

(B) 40% censoring and 20% terminal failure as missing cause							
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
250	68%	24%	48%	28%	0.2	0.234	0.025
	66%				0.4	0.419	0.044
	64%				0.6	0.608	0.056
	62%				0.8	0.800	0.061
500	68%	24%	48%	28%	0.2	0.221	0.017
	66%				0.4	0.420	0.030
	64%				0.6	0.618	0.041
	62%				0.8	0.803	0.043

Table 3

Estimations for the marginal survival function, bivariate survival function, and cross ratio at 5, 10, 15, and 20 years after the initial treatment in the ESBCR data.

t	5 years	10 years	15 years	20 years
$\hat{S}_1(t)$	0.846	0.765	0.696	0.651
$\hat{S}_2(t)$	0.934	0.884	0.842	0.788
$\hat{S}(t_1 = t, t_2 = t)$	0.830	0.737	0.660	0.599
$\hat{\alpha}(t_1 = t, t_2 = t)$	4.978	3.428	2.783	2.450

Table 4
Simulation results of the association parameter α in the Clayton copula model with Exponential margins.

(A) 70% censoring and 50% terminal failure as missing cause							
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
250	66%	12%	30%	58%	2	2.504	0.341
	68%				5	4.587	0.522
500	66%	12%	30%	58%	2	2.497	0.222
	68%				5	4.593	0.375

(B) 40% censoring and 20% terminal failure as missing cause							
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
250	66%	24%	48%	28%	2	2.365	0.393
	68%				5	5.227	0.748
500	66%	24%	48%	28%	2	2.369	0.257
	68%				5	5.145	0.490

Table 5

Simulation results of the association parameter α in the positive stable copula model with Weibull margins. Cens%, censoring rate; Fail%, failure rate; Uncert%, rate of the uncertain cause; SEE, empirical standard error.

(A) 70% censoring and 50% terminal failure as missing cause						
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	$\hat{\alpha}$	SEE
250	78%	18%	20%	62%	0.2	0.220
	77%				0.4	0.388
	76%				0.6	0.576
	75%				0.8	0.762
500	78%	18%	20%	62%	0.2	0.217
	77%				0.4	0.391
	76%				0.6	0.581
	75%				0.8	0.768

(B) 40% censoring and 20% terminal failure as missing cause						
n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	$\hat{\alpha}$	SEE
250	78%	36%	33%	32%	0.2	0.213
	77%				0.4	0.403
	76%				0.6	0.590
	75%				0.8	0.776
500	78%	36%	33%	32%	0.2	0.206
	77%				0.4	0.395
	76%				0.6	0.582
	75%				0.8	0.774

Estimation results of survival function in simulated survival data with missing cause-of-failure data, for different initial value p^* . 40% censoring and 20% terminal failure were randomly selected as the missing cause of terminal events.

Table 6

p^*	Cens%	Fail%	Uncert%	n	T	$S(t)$	$b(\hat{S}(t)^{pr})$	SEM	SEE
0.3	24%	48%	28%	250	0.357	0.7	0.025	0.034	0.029
					0.693	0.5	0.035	0.041	0.033
				500	1.204	0.3	0.037	0.045	0.035
					0.357	0.7	0.026	0.024	0.021
				500	0.693	0.5	0.037	0.029	0.023
					1.204	0.3	0.037	0.032	0.025
p^*	Cens%	Fail%	Uncert%	n	T	$S(t)$	$b(\hat{S}(t)^{pr})$	SEM	SEE
0.7	24%	48%	28%	250	0.357	0.7	-0.006	0.033	0.029
					0.693	0.5	-0.009	0.038	0.034
				500	1.204	0.3	-0.015	0.039	0.033
					0.357	0.7	-0.008	0.023	0.021
				500	0.693	0.5	-0.013	0.027	0.024
					1.204	0.3	-0.018	0.027	0.024

Simulation results of the association parameter α in the positive stable copula model for semi-competing risks data with missing cause of the terminal event, for different initial value p^* : 40% censoring and 20% terminal failure were randomly selected as the missing cause of terminal events.

Table 7

p^*	n	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
0.3	250	68%	24%	48%	28%	0.2	0.228	0.024
		66%				0.4	0.423	0.045
		64%				0.6	0.615	0.054
	500	62%				0.8	0.795	0.066
		68%	24%	48%	28%	0.2	0.224	0.018
		66%				0.4	0.423	0.030
0.7	250	64%				0.6	0.621	0.038
		62%				0.8	0.802	0.043
	500	Cens(T_1)%	Cens(T_2)%	Fail%	Uncert%	α	$\hat{\alpha}$	SEE
		68%	24%	48%	28%	0.2	0.219	0.024
		66%				0.4	0.416	0.044
		64%				0.6	0.608	0.055
0.7	250	62%				0.8	0.795	0.062
	500	68%	24%	48%	28%	0.2	0.217	0.018
		66%				0.4	0.417	0.030
		64%				0.6	0.610	0.039
	500	62%				0.8	0.798	0.044