Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2019 April ; 81(2): 409–429. doi:10.1111/rssb.12308.

Semiparametric Model for Bivariate Survival Data Subject to Biased Sampling

Jin Piao,

The University of Southern California, Los Angeles, USA.

Jing Ning,

The University of Texas MD Anderson Cancer Center, Houston, USA.

Yu Shen

The University of Texas MD Anderson Cancer Center, Houston, USA.

Summary.

To better understand the relationship between patient characteristics and their residual survival after an intermediate event such as the local cancer recurrence, it is of interest to identify patients with the intermediate event and then analyze their residual survival data. One challenge in analyzing such data is that the observed residual survival times tend to be longer than those in the target population, since patients who die before experiencing the intermediate event are excluded from the identified cohort. We propose to jointly model the ordered bivariate survival data using a copula model and appropriately adjusting for the sampling bias. We develop an estimating procedure to simultaneously estimate the parameters for the marginal survival functions and the association parameter in the copula model, and use a two-stage expectation-maximization algorithm. Using empirical process theory, we prove that the estimators have strong consistency and asymptotic normality. We conduct simulations studies to evaluate the finite sample performance of the proposed method. We apply the proposed method to two cohort studies to evaluate the association between patient characteristics and residual survival.

Keywords

Bivariate survival outcomes; Copula model; EM algorithm; Prevalent cohort; Sampling bias; Two-stage estimation

1. Introduction

The management of breast cancer has made great progress due to early detection and improvements in locoregional and targeted therapy (Siegel et al., 2014). Some patient characteristics, including intrinsic cancer subtype, are important prognostic and predictive factors in breast cancer (Goldhirsch et al., 2011). Up to 30% of node-negative and up to 70% of node-positive breast cancer will relapse, and patients with early-stage breast cancer who

are disease-free after five years still have substantial risk of cancer recurrence (Brewster et al., 2008; Cardoso et al., 2012). The management of relapsed breast cancer is complex, and it is important to understand survival distributions and the association between patient characteristics and survival after cancer recurrence. As noted in the literature (Høst and Lund, 1986; Dent et al., 2014), patients' characteristics may have different effects on the survival time measured from the initial diagnosis and the survival time following disease recurrence. Using the Surveillance, Epidemiology and End Results (SEER)-Medicare database, we investigate the association between patient characteristics and survival time for patients who have an intermediate event such as the local and regional recurrence so that we may better predict the residual survival time and inform treatment decisions.

A population-based cancer registry, the SEER-Medicare database, provides data for prevalent cohorts, which consists of patients who have already experienced the initiating event (e.g., diagnosis of breast cancer) but have not experienced the failure event (e.g., death) at the time of ascertainment. From the SEER-Medicare claim data, we identify patients with local and regional recurrence, defined as having a subsequent mastectomy more than 10 month after their initial breast conserving surgery (Albert et al., 2012; Yao et al., 2013). It is well known that the patients in the prevalent cohort often have longer survival times (Asgharian et al., 2014; Cox, 1969; Gail and Benichou, 2000; Neyman, 1955; Qin and Shen, 2010; Rothman et al., 2008; Vardi, 1982, 1985, 1989; Wang et al., 1993; Wolfson et al., 2001; Zelen and Feinleib, 1969; Zelen, 2006) and their distribution of covariates may not be representative of the target population (i.e., incident cohort) (Begg and Gray, 1987; Bergeron et al., 2008) because patients with longer survival times are preferentially sampled.

Determining the best way to adjust for the sampling bias from the prevalent cohort has been a longstanding statistical problem. Statistical methods for analyzing prevalent sampling survival data have been widely studied by Wang et al. (1986); Asgharian et al. (2002, 2005); Shen et al. (2009); Tsai (2009); Qin et al. (2011); Kim et al. (2013); Ning et al. (2014); Liu et al. (2016) and others. Most of the methods used in the aforementioned publications dealt with a special type of prevalent data under the stationary assumption that the disease incidence is constant over time (stable disease model). However, over the past decades, the incidence of breast cancer has varied, the mortality rate has decreased, and consequently the observed biased data are unlikely to be length-biased data. Developing models to accommodate this general biased sampling is essential in cancer studies. As indicated in Figure 1, another major challenge is that the primary outcome (the time measured from cancer recurrence to death) is positively associated with the overall survival time measured from the time at which the cancer was diagnosed, and then is associated with the truncation probability, which is determined by the overall survival time and truncation time. In other words, the association between the residual survival time measured from the cancer recurrence and the overall survival time measured from the initial cancer diagnosis causes the biased sampling in our observed data. This unique data structure imposes a statistical challenge for joint modeling of the bivariate survival times when they are subject to biased sampling.

The copula model has been widely used to analyze correlated bivariate survival data, in which the marginal survival functions do not depend on the choice of the copula (i.e.,

dependency structure) (Nelsen, 2007). This feature implies that the association of the bivariate survival times and the marginal distributions can be modeled and estimated separately. However, different copula choices may induce quite different dependence structures (Shih and Louis, 1995); hence, a number of methods have addressed the selection within the copula family to avoid incorrect association modeling (Genest and Rivest, 1993; Chen and Bandeen-Roche, 2005; Chen and Fan, 2007). We extend the copula model within the context of biased sampling and propose estimating procedures.

The remainder of the article is organized as follows. In Section 2, we introduce the notations and the copula model in which the marginal distributions are modeled by the semiparametric proportional hazards models. In Section 3, we present the proposed estimating procedure, in which we use a two-stage expectation-maximization (EM) algorithm to iteratively estimate the marginal distributions and the association parameter. This algorithm converts the biased sampling problem to a missing data framework to facilitate the computation. In Section 4, we establish the asymptotic properties of the proposed estimators. In Section 5, we assess the empirical performance of the proposed estimators in simulation studies. We present two applications in Section 6 and provide concluding remarks in Section 7. Technical details can be found in the Supplementary Materials.

2. Notations and Model

For the target population, $\tilde{U}$ and $\tilde{V}$ denote the time from the disease diagnosis to disease recurrence and the time from the disease recurrence to death, respectively, and $\tilde{A}$ denotes the time from the disease diagnosis to the study sampling time (see Figure 1). Under biased sampling, patients have to survive until the sampling time to be enrolled, i.e., one can observe the bivariate times only if $\tilde{U} + \tilde{V} > \tilde{A}$. Figure 1 illustrates the sampling mechanism for the biased bivariate survival data. Let (U, V, A) be the observed biased counterparts of unbiased $(\tilde{U}, \tilde{V}, \tilde{A})$. Let $\mathbf{X}$ and $\mathbf{Z}$ respectively denote the covariate vectors associated with $\tilde{U}$ and $\tilde{V}$. Usually, the covariate vector $\mathbf{Z}$ is identical to the baseline covariate vector $\mathbf{X}$ or consists of $\mathbf{X}$ and additional information collected at the disease recurrence (e.g., salvage treatment information). Define C to be the residual censoring time measured from the disease recurrence. With potential right censoring, the observed data are $\{(a_i, u_i, y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \dots, n\}$, where $y_i = \min(v_i, c_i)$ is the observed time from the disease recurrence until death or right censoring, and $\delta_i = I\{v_i < c_i\}$ is the death indicator. Throughout this article, we assume that (i) $\tilde{A}$ and $(\tilde{U}, \tilde{V})$ are conditionally independent given covariates $\mathbf{X}$ and $\mathbf{Z}$, and (ii) C is conditionally independent of $(\tilde{A}, \tilde{U}, \tilde{V})$ given covariates $\mathbf{X}$ and $\mathbf{Z}$.

Under the copula model and the assumption that $\tilde{U}$ is independent of $\mathbf{Z}$ given $\mathbf{X}$ and $\tilde{V}$ is independent of $\mathbf{X}$ given $\mathbf{Z}$, the bivariate survival and density functions of $(\tilde{U}, \tilde{V})$ given the covariates $(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ can be written as

$$\begin{aligned} S(u, v | \mathbf{x}, \mathbf{z}) &= P(\tilde{U} > u, \tilde{V} > v | \mathbf{x}, \mathbf{z}) = B_{\alpha}\{S_1(u | \mathbf{x}), S_2(v | \mathbf{z})\}, \\ f(u, v | \mathbf{x}, \mathbf{z}) &= b_{\alpha}\{S_1(u | \mathbf{x}), S_2(v | \mathbf{z})\} f_1(u | \mathbf{x}) f_2(v | \mathbf{z}), u, v \geq 0, \end{aligned} \quad (1)$$

where B_α is the bivariate distribution function, with density b_α on $[0,1]^2$ for $\alpha \in \mathcal{R}$ (Oakes, 1982). Here, (S_1, S_2) and (f_1, f_2) are the marginal survival and density functions, respectively. The association parameter α is related to the cross-ratio (Oakes, 1989). This nice connection has been used for model selection in copulas (Chen and Bandeen-Roche, 2005; Wang and Wells, 2000).

For simplicity of notation, we use the same vector of covariate $\mathbf{X}$ in the two marginal models by assuming $\mathbf{Z} = \mathbf{X}$, but it is easy to accommodate different subsets of covariates in the two marginal models. To determine the association between the survival outcomes and covariates, Cox proportional hazards models are used for the marginal distributions:

$$\lambda_1(u|\mathbf{X}, \boldsymbol{\beta}) = \lambda_{01}(u)\exp(\mathbf{X}^T \boldsymbol{\beta}) \quad (2)$$

$$\lambda_2(v|\mathbf{X}, \boldsymbol{\gamma}) = \lambda_{02}(v)\exp(\mathbf{X}^T \boldsymbol{\gamma}), \quad (3)$$

where $\lambda_{01}(u)$ and $\lambda_{02}(v)$ are unspecified baseline hazard functions for the marginal distributions. For the purpose of model identifiability, for the truncation time $\tilde{A}$, we assume $H_{\boldsymbol{\theta}}(\cdot)$ is known up to finitely many unknown parameters. Let $h_{\boldsymbol{\theta}}(\cdot)$ be the density function of $H_{\boldsymbol{\theta}}(\cdot)$ (Wang, 1989).

3. Estimation Procedure

Given covariate $\mathbf{X} = \mathbf{x}$, the probability of a subject from the target population being selected to a study is

$$P(\tilde{U} + \tilde{V} > \tilde{A} | \mathbf{X} = \mathbf{x}) = \int_0^\infty \int_0^\infty H_{\boldsymbol{\theta}}(s+t)f(s, t|\mathbf{x})dsdt.$$

Under the biased sampling mechanism, a joint model for the bivariate survival data is required to correct the sampling bias. Given the conditional independence assumptions in Section 2, the full likelihood function of the observed data conditional on $\mathbf{X}_i = \mathbf{x}_i$ is proportional to

$$\begin{aligned} & L(\boldsymbol{\theta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_{01}, \lambda_{02}) \\ &= \prod_{i=1}^n \left\{ \frac{f(u_i, y_i | \mathbf{x}_i) \times h_{\boldsymbol{\theta}}(a_i)}{\int_0^\infty \int_0^\infty H_{\boldsymbol{\theta}}(s+t)f(s, t|\mathbf{x}_i)dsdt} \right\}^{\delta_i} \left\{ \frac{\partial S(u_i, y_i | \mathbf{x}_i) / \partial u_i \times h_{\boldsymbol{\theta}}(a_i)}{\int_0^\infty \int_0^\infty H_{\boldsymbol{\theta}}(s+t)f(s, t|\mathbf{x}_i)dsdt} \right\}^{1-\delta_i}, \end{aligned} \quad (4)$$

where $f(u_i, y_i | \mathbf{x}_i)$ and $S(u_i, y_i | \mathbf{x}_i) / u_i$ involve marginal density functions

$$f_1(u_i|\mathbf{x}_i) = \lambda_{01}(u_i)\exp(\boldsymbol{\beta}^T \mathbf{x}_i)\exp(-\Lambda_1(u_i)\exp(\boldsymbol{\beta}^T \mathbf{x}_i))$$

and

$$f_2(v_i|\mathbf{x}_i) = \lambda_{02}(v_i)\exp(\boldsymbol{\gamma}^T \mathbf{x}_i)\exp(-\Lambda_2(v_i)\exp(\boldsymbol{\gamma}^T \mathbf{x}_i)).$$

Note that the joint density function in the double integrals involves two infinite-dimensional parameters $\lambda_{01}(\cdot)$ and $\lambda_{02}(\cdot)$; hence, directly maximizing the observed likelihood is computationally prohibitive. On the other hand, the double integrals make it even more challenging to implement the profile likelihood method. For computational stability, we propose an estimating procedure that uses the pseudo-likelihood together with the two-step EM algorithm to estimate the parameters in the joint density function by plugging in a consistent estimator for the parameters in the left-truncated distribution. Inspired by the two-stage estimation of Shih and Louis (1995), we propose an estimating procedure to separately estimate the association parameter and the parameters in the marginal functions, in which we develop the EM algorithm to estimate the infinite-dimensional parameters and finite-dimensional parameters simultaneously. We describe this procedure in greater detail in the following sections.

3.1. Estimation of the truncated data distribution

We utilize the conditional likelihood to estimate the unknown parameters associated with the truncated data distribution H . Under the assumption that the truncation time is independent of the covariates and independent of the censoring given the covariates, the conditional density of $\tilde{A}$ given $(\tilde{U} = u, \tilde{V} = v, \mathbf{X} = \mathbf{x}, \delta = 1, \tilde{A} < u + v)$ is

$$f_{\tilde{A}}(a|\tilde{U} = u, \tilde{V} = v, \mathbf{X} = \mathbf{x}, \delta = 1, \tilde{A} < u + v) = \frac{h_{\boldsymbol{\theta}}(a)f(u, v|\mathbf{X} = \mathbf{x})P(C > v|\mathbf{X} = \mathbf{x})}{\int_0^{u+v} h_{\boldsymbol{\theta}}(t)dt f(u, v|\mathbf{X} = \mathbf{x})P(C > v|\mathbf{X} = \mathbf{x})} = \frac{h_{\boldsymbol{\theta}}(a)}{H_{\boldsymbol{\theta}}(u+v)}.$$

The conditional likelihood for estimating $\boldsymbol{\theta}$ is

$$\prod_{i=1}^n \left\{ \frac{h_{\boldsymbol{\theta}}(a_i)}{H_{\boldsymbol{\theta}}(u_i + y_i)} \right\}^{\delta_i}, \quad (5)$$

with the corresponding score function as

$$U(\boldsymbol{\theta}) = \sum_{i=1}^n \delta_i \left\{ \frac{\partial \log h_{\boldsymbol{\theta}}(a_i)}{\partial \boldsymbol{\theta}} - \frac{\partial \log H_{\boldsymbol{\theta}}(u_i + y_i)}{\partial \boldsymbol{\theta}} \right\}. \quad (6)$$

Let $\hat{\boldsymbol{\theta}}$ be the solution of $U(\boldsymbol{\theta}) = 0$ and $\boldsymbol{\theta}_0$ be the true value.

Theorem 1: Under the assumptions given in the Supplementary Materials, $\hat{\theta}_n \rightarrow \theta$ in probability as $n \rightarrow \infty$. Moreover, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to $N(0, \Sigma_{\theta_0})$ in distribution as $n \rightarrow \infty$, where Σ_{θ_0} is given in the Supplementary Materials.

A sketch of the proof and the estimating formula for Σ_{θ_0} is given in the Supplementary Materials.

3.2. Estimation of the bivariate distribution

Given the estimator, $\hat{\theta}$, we define the pseudo-likelihood to approximate the likelihood defined in Equation (4), i.e. $L^*(\alpha, \beta, \gamma, \lambda_{01}, \lambda_{02}) = L(\hat{\theta}, \alpha, \beta, \gamma, \lambda_{01}, \lambda_{02})$. Let $u_1 < u_2 < \dots < u_{k_1} < \infty$ be the distinct observed points for the time from disease diagnosis to recurrence (U), and $u_1 < u_2 < \dots < u_{k_2} < \infty$ be the distinct observed points for the time from the disease recurrence to death or censoring (Y). Let $\lambda_{1j} \equiv d\Lambda_1(u_j)$ be the positive masses of the discrete cumulative baseline function Λ_1 for the time-lag U with $j = 1, \dots, k_1$, and let $\lambda_{2j'} \equiv d\Lambda_2(v_{j'})$ be the positive masses of the discrete cumulative baseline function Λ_2 for the time-lag V with $j' = 1, \dots, k_2$. Denote $\lambda_1 = (\lambda_{11}, \dots, \lambda_{1k_1})$ and $\lambda_2 = (\lambda_{21}, \dots, \lambda_{2k_2})$.

With biased sampling, for subject i , it can be considered as sampling the unbiased time $(\tilde{U}, \tilde{V}, \tilde{A})$ for a random m_i times until the sampling constraint $(\tilde{U} + \tilde{V} > \tilde{A})$ is satisfied. This random integer m_i follows a Geometric distribution with parameter $\pi_i = P(\tilde{U} + \tilde{V} > \tilde{A} | \mathbf{X} = \mathbf{x}_i)$. Denote the observed data as $\mathcal{O}_i = (A_i, U_i, Y_i, \mathbf{X}_i, \delta_i)$, and denote the latent truncated (unobserved) data as $\mathcal{O}_i^* = \{(A_{il}^*, U_{il}^*, V_{il}^*, \mathbf{X}_i), U_{il}^* + V_{il}^* < A_i^*\}, l = 1, \dots, m_i, i = 1, \dots, n$. Following the principle of the EM algorithm, we treat $(\mathcal{O}, \mathcal{O}^*)$ as the ‘complete data’.

Denoting the parameters of interest as $\psi = \{\alpha, \beta, \gamma, \Lambda_1, \Lambda_2\}$, the log-pseudo-likelihood for the complete data can be written as

$$l_c^*(\psi) = \sum_{i=1}^n \left[\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} I(U_i = u_j, V_i = v_{j'}) \left\{ \delta_i \log f(u_j, v_{j'} | \mathbf{X}_i; \psi) + (1 - \delta_i) \log \frac{\partial S(u_j, v_{j'} | \mathbf{X}_i; \psi)}{\partial u_j} \right\} \right. \\ \left. + \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \sum_{l=1}^{m_i} I(U_{il}^* = u_j, V_{il}^* = v_{j'}) \log f(u_j, v_{j'} | \mathbf{X}_i; \psi) \right].$$

We then use the EM algorithm to obtain the parameter estimates. In the E-step, we denote the current parameter value in the p th step by $\psi^{(p)} = \{\alpha^{(p)}, \beta^{(p)}, \gamma^{(p)}, \Lambda_1^{(p)}, \Lambda_2^{(p)}\}$. Given $\hat{\theta}$, $\psi^{(p)}$ and the observed data, the expectation of m_i is

$$E(m_i | \mathcal{O}_i; \hat{\theta}, \psi^{(p)}) = \frac{1 - \hat{\pi}_i}{\hat{\pi}_i} = \frac{1 - \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} H_{\hat{\theta}}(u_j + v_{j'}) f(u_j, v_{j'} | \mathbf{X}_i; \psi^{(p)})}{\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} H_{\hat{\theta}}(u_j + v_{j'}) f(u_j, v_{j'} | \mathbf{X}_i; \psi^{(p)})}.$$

We denote $w_{ijj'}$ as the expected total number of observations of $(U_{il}^*, V_{il}^*) = (u_j, v_{j'})$ among the m_i observations, then we have

$$\begin{aligned} w_{ijj'} &= E \left[\sum_{l=1}^{m_i} I(U_{il}^* = u_j, V_{il}^* = v_{j'}) | \mathcal{O}_i; \hat{\theta}, \psi^{(p)} \right] \\ &= E \left[E \left[\sum_{l=1}^{m_i} I(U_{il}^* = u_j, V_{il}^* = v_{j'}) | m_i \right] | \mathcal{O}_i; \hat{\theta}, \psi^{(p)} \right] \\ &= E(m_i | \mathcal{O}_i; \hat{\theta}, \psi^{(p)}) E \left[I(U_{il}^* = u_j, V_{il}^* = v_{j'}) | \mathcal{O}_i; \hat{\theta}, \psi^{(p)} \right] \\ &= \frac{1 - \hat{\pi}_i}{\hat{\pi}_i} \frac{f(U_{il}^* = u_j, V_{il}^* = v_{j'}, A_i^* > u_j + v_{j'} | \mathcal{O}_i; \hat{\theta}, \psi^{(p)})}{1 - \hat{\pi}_i} \\ &= \left(1 - H_{\hat{\theta}}(u_j + v_{j'}) \right) f(u_j, v_{j'} | \mathbf{X}_i; \psi^{(p)}) / \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} H_{\hat{\theta}}(u_j + v_{j'}) f(u_j, v_{j'} | \mathbf{X}_i; \psi^{(p)}). \end{aligned}$$

We calculate the expected complete data log-pseudo-likelihood conditional on the observed data $\mathcal{O}$:

$$\begin{aligned} l_E(\psi) &= \sum_{i=1}^n \left[\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} I(U_i = u_j, V_i = v_{j'}) \left\{ \delta_i \log f(u_j, v_{j'} | \mathbf{X}_i; \psi) \right. \right. \\ &\quad \left. \left. + (1 - \delta_i) \log \frac{\partial S(u_j, v_{j'} | \mathbf{X}_i; \psi)}{\partial u_j} \right\} + \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} w_{ijj'} \log f(u_j, v_{j'} | \mathbf{X}_i; \psi) \right], \end{aligned} \quad (7)$$

In the M-step, it is computationally non-trivial to maximize the expected complete data log-pseudo-likelihood, which involves infinite-dimensional parameters λ_1 and λ_2 . Additionally, the association between the bivariate times adds another layer of computational complexity such that the profile likelihood principle cannot be used here. Motivated by these challenges, we decompose equation (7) as the following three parts and propose the following iterative steps to facilitate the computation. Specifically, we have

$$l_E(\psi) = l_E^U(\beta, \lambda_1) + l_E^V(\gamma, \lambda_2) + l_E^B(\psi),$$

where

$$l_E^U(\boldsymbol{\beta}, \lambda_1) = \sum_{i=1}^n \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \left\{ I(U_i = u_j) + w_{ijj'} \right\} \left\{ \log \lambda_{1j} + \boldsymbol{\beta}^T \mathbf{X}_i - \Lambda_1(u_j) \exp(\boldsymbol{\beta}^T \mathbf{X}_i) \right\},$$

$$l_E^V(\boldsymbol{\gamma}, \lambda_2) = \sum_{i=1}^n \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \left[I(V_i = v_{j'}) \left\{ \delta_i \left(\log \lambda_{2j'} + \boldsymbol{\gamma}^T \mathbf{X}_i \right) - \Lambda_2(v_{j'}) \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) \right\} \right. \\ \left. + w_{ijj'} \left\{ \log \lambda_{2j'} + \boldsymbol{\gamma}^T \mathbf{X}_i - \Lambda_2(v_{j'}) \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) \right\} \right],$$

$$l_E^B(\boldsymbol{\psi}) = \sum_{i=1}^n \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} \left[I(U_i = u_j, V_i = v_{j'}) \left\{ \delta_i \log b_{\alpha}(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi}) + (1 - \delta_i) \log \frac{\partial S(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi})}{\partial u_j} \right. \right. \\ \left. \left. - (1 - \delta_i) \log S_2(v_{j'} | \mathbf{X}_i; \boldsymbol{\gamma}) - (1 - \delta_i) \log f_1(u_j | \mathbf{X}_i; \boldsymbol{\beta}) \right\} + w_{ijj'} \log b_{\alpha}(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi}) \right].$$

Here, $l_E^U(\boldsymbol{\beta}, \lambda_1)$ and $l_E^V(\boldsymbol{\gamma}, \lambda_2)$ are the expected complete-data log-marginal-likelihood functions for U and V , respectively, and $l_E^B(\boldsymbol{\psi})$ is the expected log-likelihood corresponding to the association parameter.

M Step 1—Calculate the probability masses of $f(u_j, v_{j'} | \mathbf{X}_i)$ based on the current estimator $\boldsymbol{\psi}^{(p)}$ on the grids shown in Figure 2, denoted as $p_{ijj'}^{(p)}$. More details can be found in the Appendix. Then we can derive the weight function on the grids as

$$w_{ijj'} = \left\{ 1 - H_{\hat{\boldsymbol{\theta}}}(u_j + v_{j'}) \right\} p_{ijj'}^{(p)} \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} H_{\hat{\boldsymbol{\theta}}}(u_j + v_{j'}) p_{ijj'}^{(p)}.$$

M Step 2—Estimate the marginal function of $\tilde{U}$ by maximizing the following component

$$l_E^U(\boldsymbol{\beta}, \lambda_1) = \sum_{i=1}^n \left[\sum_{j=1}^{k_1} \left\{ I(U_i = u_j) + w_{ij, U} \right\} \left\{ \log \lambda_{1j} + \boldsymbol{\beta}^T \mathbf{X}_i - \Lambda_1(u_j) \exp(\boldsymbol{\beta}^T \mathbf{X}_i) \right\} \right],$$

where $\Lambda_1(u_j) = \sum_{h=1}^j \lambda_{1h}$ and $w_{ij, U} = \sum_{j'=1}^{k_2} w_{ijj'}$. The maximizer of λ_{1j} expressed by $P_{ijj'}$ now can be written as a function of $\boldsymbol{\beta}$,

$$\lambda_{1j}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^n \{I(U_i = u_j) + w_{ij,U}\}}{\sum_{i=1}^n \sum_{h=j}^{k_1} \{I(U_i = u_h) + w_{ih,U}\} \exp(\boldsymbol{\beta}^T \mathbf{X}_i)}. \quad (8)$$

The score equation with respect to $\boldsymbol{\beta}$ of $l_E^U(\boldsymbol{\beta}, \lambda_1)$ is

$$\sum_{i=1}^n \left[\sum_{j=1}^{k_1} \{I(U_i = u_j) + w_{ij,U} - \{I(U_i = u_j) + w_{ij,U}\} \Lambda_1(u_j) \exp(\boldsymbol{\beta}^T \mathbf{X}_i)\} \mathbf{X}_i \right] = 0. \quad (9)$$

By plugging in the closed form of $\lambda_{1j}(\boldsymbol{\beta})$, equation (9) can be rewritten as

$$\begin{aligned} U_1(\boldsymbol{\beta}) = \sum_{i=1}^n \left[\sum_{j=1}^{k_1} I(U_i = u_j) + w_{ij,U} - \{I(U_i = u_j) + w_{ij,U}\} \right. \\ \left. \sum_{l=1}^j \frac{\sum_{i'=1}^n \{I(U_{i'} = u_l) + w_{i'l,U}\}}{\sum_{i'=1}^n \sum_{h=1}^{k_1} \{I(U_{i'} = u_h) + w_{i'h,U}\} \exp(\boldsymbol{\beta}^T \mathbf{X}_{i'})} \exp(\boldsymbol{\beta}^T \mathbf{X}_i) \right] \mathbf{X}_i = 0. \end{aligned} \quad (10)$$

Hence, the updated estimate of $\boldsymbol{\beta}$ can be obtained by solving equation (10), and then the estimator of λ_{1k} can be updated by equation (8).

M Step 3—Estimate the marginal function of $\tilde{V}$ by maximizing the following component

$$\begin{aligned} l_E^V(\boldsymbol{\gamma}, \lambda_2) = \sum_{i=1}^n \sum_{j'=1}^{k_2} \left[I(V_i = v_{j'}) \{ \delta_i (\log \lambda_{2j'} + \boldsymbol{\gamma}^T \mathbf{X}_i) - \Lambda_2(v_{j'}) \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) \} \right. \\ \left. + w_{ij',V} \{ \log \lambda_{2j'} + \boldsymbol{\gamma}^T \mathbf{X}_i - \Lambda_2(v_{j'}) \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) \} \right], \end{aligned} \quad (11)$$

where $\Lambda_2(v_{j'}) = \sum_{h=1}^{j'} \lambda_{2h'}$ and $w_{ij',V} = \sum_{j=1}^{k_1} w_{ijj'}$. Following a similar strategy in M step 2, we can estimate λ_2 and $\boldsymbol{\gamma}$ using the follow equations:

$$\lambda_{2j'}(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \{I(V_i = v_{j'}) \delta_i + w_{ij',V}\}}{\sum_{i=1}^n \sum_{h=j'}^{k_2} \{I(V_i = v_h) + w_{ih,V}\} \exp(\boldsymbol{\gamma}^T \mathbf{X}_i)} \quad (12)$$

and

$$U_2(\boldsymbol{\gamma}) = \sum_{i=1}^n \left[\sum_{j'=1}^{k_2} I(V_i = v_{j'}) \delta_i + w_{ij',V} - \{I(V_i = v_{j'}) + w_{ij',V}\} \right. \\ \left. \sum_{l=1}^{j'} \frac{\sum_{i'=1}^n \{I(V_{i'} = v_l) \delta_{i'} + w_{i'l,V}\}}{\sum_{i'=1}^n \sum_{h=1}^{k_2} \{I(V_{i'} = v_h) + w_{i'h,V}\} \exp(\boldsymbol{\gamma}^T \mathbf{X}_{i'})} \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) \right] \mathbf{X}_i = 0. \quad (13)$$

M Step 4—Giving the estimators of $(\boldsymbol{\beta}, \lambda_1, \boldsymbol{\gamma}, \lambda_2)$ from M steps 2–3, the maximization with respect to the association parameter is achieved by solving the following score Equation (14). The association parameter can be updated by plugging in the estimator for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_1, \lambda_2)$ from M steps 2 and 3.

$$U_3(\alpha) = \partial l_E(\boldsymbol{\psi}) / \partial \alpha = \sum_{i=1}^n \left[\sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} w_{ijj'} \frac{\partial \log f(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi})}{\partial \alpha} \right. \\ \left. + \sum_{j=1}^{k_1} \sum_{j'=1}^{k_2} I(U_i = u_j, V_i = v_{j'}) \left\{ \delta_i \frac{\partial \log f(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi})}{\partial \alpha} + (1 - \delta_i) \frac{\partial \log \frac{\partial S(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi})}{\partial u_j}}{\partial \alpha} \right\} \right]. \quad (14)$$

3.3. Implementation of M steps

The M steps for the updating parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of the marginal distribution can be easily implemented by the existing program for the Cox model with right-censored data.

Specifically, we generate a data set that represents the unobserved subjects in which the failure times are created by repeating the observed unique failure time n times, i.e.,

$U_{nk_1} = (u_1, \dots, u_{k_1}, \dots, u_1, \dots, u_{k_1})$. The corresponding death indicator is a vector of 1, denoted

as $\Delta_{nk_1} = (1, \dots, 1)$. The covariates are matched with the failure times, with

$\mathbf{X}_{nk_1} = (\mathbf{X}_1, \dots, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots, \mathbf{X}_n)$. Then we combine the generated unobserved data

with the observed data to obtain the analyzed data set. By using the function `coxph()` with the “weight” option in R, we can estimate $\boldsymbol{\beta}$ by

$$\text{coxph}(\text{Surv}(U, \Delta_U) \sim \mathbf{X}_U, \text{weight1})$$

where $U^T = (u_1, \dots, u_n, U_{nk_1})$, $\Delta_U^T = (1, \dots, 1, \Delta_{nk_1})$, $\mathbf{X}_U^T = (\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}_{nk_1})$ and the weight equals $(1, \dots, 1, w_{11, U}, w_{12, U}, \dots, w_{1k_1, U}, \dots, w_{n1, U}, w_{n2, U}, \dots, w_{nk_1, U})^T$. Note that the first n elements have a weight of 1 since they are associated with the observed data, and have a recurrence indicator of 1 since there is no censoring for the time-lag U .

Similar to M step 2, we can estimate $(\boldsymbol{\gamma}, \lambda_2)$ using the existing R function,

$$\text{coxph}(\text{Surv}(V, \Delta_V) \sim \mathbf{X}_V, \text{weight2})$$

where $V^T = (y_1, \dots, y_n, V_{nk_2})$, $V_{nk_2} = (v_1, \dots, v_{k_2}, \dots, v_1, \dots, v_{k_2})$, $\Delta_V^T = (\delta_1, \dots, \delta_n, 1, \dots, 1)$, $\mathbf{X}_V^T = (\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}_{nk_2})$, $\mathbf{X}_{nk_2} = (\mathbf{X}_1, \dots, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots, \mathbf{X}_n)$

the weight equals $(1, \dots, 1, w_{11, V}, w_{12, V}, \dots, w_{1k_2, V}, \dots, w_{n1, V}, w_{n2, V}, \dots, w_{nk_2, V})^T$. Note that the time-lag V is subject to potential right censoring, thus the first n elements of Δ_V^T are set as the death indicator from the observed data.

In summary, the maximization in the M-step cycles through $\boldsymbol{\beta}, \lambda_1, \boldsymbol{\gamma}, \lambda_2$ and α until a pre-specified criterion is met. We repeat the E and M steps until convergence is achieved.

4. Theoretical Properties

4.1. Convergence of the EM algorithm

The M-step of the proposed EM algorithm is different from that of the traditional EM algorithm, in which the parameters are estimated simultaneously by maximizing the expected complete data likelihood. In the proposed algorithm, we maximize the marginal likelihoods to estimate the parameters in the marginal Cox models in M steps 2–3. In Theorem 2, we establish the ascent property, which implies the convergence of the proposed computation algorithm. More details can be found in the Supplementary Materials.

Theorem 2 (Ascent property): The likelihood function of the observed data increases with the iterations, i.e., $I_O(\boldsymbol{\psi}^{(p+1)}) \geq I_O(\boldsymbol{\psi}^{(p)})$, where $I_O(\boldsymbol{\psi})$ be the observed-data log-likelihood function.

4.2. Asymptotic Properties

In this section, we establish the asymptotic properties of the proposed estimators, the maximizer of the pseudo-likelihood $L^*(\boldsymbol{\psi})$, denoted by $\hat{\boldsymbol{\psi}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\alpha}_n, \hat{\Lambda}_{1n}, \hat{\Lambda}_{2n})$. Here the subscript n is used to emphasize that the estimators are based on a sample of size n . Denote the true value of the parameters by $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \alpha_0, \Lambda_{10}, \Lambda_{20})$. The log-pseudo-likelihood function is

$$l_n(\boldsymbol{\psi}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \left[\delta_i \log \left[\frac{f(u_i, y_i | \mathbf{X}_i; \boldsymbol{\psi}) \times h_{\hat{\boldsymbol{\theta}}}(a_i)}{\pi_i} \right] + (1 - \delta_i) \log \left[\frac{\partial S(u_i, y_i | \mathbf{X}_i; \boldsymbol{\psi}) / \partial u_i \times h_{\hat{\boldsymbol{\theta}}}(a_i)}{\pi_i} \right] \right] \\ \propto \sum_{i=1}^n \left[\delta_i \left[\log f(u_i, y_i | \mathbf{X}_i; \boldsymbol{\psi}) - \log \frac{\partial S(u_i, y_i | \mathbf{X}_i; \boldsymbol{\psi})}{\partial u_i} \right] + \log \frac{\partial S(u_i, y_i | \mathbf{X}_i; \boldsymbol{\psi})}{\partial u_i} - \log \pi_i \right],$$

where $\pi_i = \int_0^{\tau_2} \int_0^{\tau_1} H_{\hat{\boldsymbol{\theta}}}(s+t) f(s, t | \mathbf{X}_i; \boldsymbol{\psi}) ds dt$, and τ_1 and τ_2 are the upper bounds of the support for $\tilde{U}$ and $\tilde{V}$, respectively.

Let $N_{i1}(s) = I(U_i \leq s)$, $N_{i2}(t) = I(Y_i \leq t) \delta_i$, and the log-pseudo-likelihood function can be expressed in the counting process formulations as

$$l_n(\boldsymbol{\psi}; \hat{\boldsymbol{\theta}}) \propto \sum_{i=1}^n \left[\int_0^{\tau_2} \left\{ \log \frac{B_{\alpha}^{11}(u_i, t | \mathbf{X}_i; \boldsymbol{\psi})}{B_{\alpha}^{10}(u_i, t | \mathbf{X}_i; \boldsymbol{\psi})} + \boldsymbol{\gamma}^T \mathbf{X}_i + \log d \Lambda_2(t) - \int_0^t \exp(\boldsymbol{\gamma}^T \mathbf{X}_i) d \Lambda_2(v) \right\} dN_{i2}(t) \right. \\ \left. + \int_0^{\tau_1} \left\{ \log B_{\alpha}^{10}(s, y_i | \mathbf{X}_i; \boldsymbol{\psi}) + \boldsymbol{\beta}^T \mathbf{X}_i + \log d \Lambda_1(s) - \int_0^s \exp(\boldsymbol{\beta}^T \mathbf{X}_i) d \Lambda_1(u) \right\} dN_{i1}(s) \right. \\ \left. - \log \pi_i \right].$$

$$\text{where } B_{\alpha}^{mn}(u, y | \mathbf{x}) = \frac{\partial^{m+n} B_{\alpha}(u, y | \mathbf{x})}{\partial S_1(u | \mathbf{x})^m \partial S_2(y | \mathbf{x})^n}, \quad m, n = 0 \text{ or } 1.$$

We prove strong consistency by the classical Kullback-Leibler information approach. As $\hat{\boldsymbol{\psi}}_n$ maximizes the log-pseudo-likelihood function, the empirical Kullback-Leibler information $l_n(\hat{\boldsymbol{\psi}}_n; \hat{\boldsymbol{\theta}}) - l_n(\boldsymbol{\psi}_0; \hat{\boldsymbol{\theta}}) \geq 0$ must always be nonnegative. If $\hat{\boldsymbol{\psi}}_n$ converges, say, to $\boldsymbol{\psi}^*$, then following the uniform law of large numbers (Lehmann, 1999, definition 2.6.1), we can show $l_n(\hat{\boldsymbol{\psi}}_n; \hat{\boldsymbol{\theta}}) - l_n(\boldsymbol{\psi}_0; \hat{\boldsymbol{\theta}})$ converges to the negative Kullback-Leibler distance between $P_{\boldsymbol{\psi}^*}$ and $P_{\boldsymbol{\psi}_0}$ where $P_{\boldsymbol{\psi}}$ is the probability measure under the parameter $\boldsymbol{\psi}$. As the Kullback-Leibler information is always non-negative, it implies that $P_{\boldsymbol{\psi}^*} = P_{\boldsymbol{\psi}_0}$ almost surely. Under the regularity conditions listed in the Supplementary Materials, the model $P_{\boldsymbol{\psi}}$ is identifiable, implying that $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$.

Theorem 3: Under the regularity conditions listed in the Supplementary Materials, the estimators $\hat{\boldsymbol{\psi}}_n$ are consistent: $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{a}_n)$ converges almost surely to $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, a_0)$,

$\sup_{0 < u < \tau_1} |\hat{\Lambda}_{1n}(u) - \Lambda_{10}(u)|$ and $\sup_{0 < u < \tau_2} |\hat{\Lambda}_{2n}(v) - \Lambda_{20}(v)|$ converge almost surely to 0 as $n \rightarrow \infty$.

The score equations of $l_n(\boldsymbol{\psi}, \hat{\boldsymbol{\theta}})$ consist of infinite-dimensional estimation equations, therefore we apply the Z-theorem for infinite-dimensional estimating equations to prove the weak convergence of the estimators (Van der Vaart and Wellner, 1996, Theorem 3.3.1, page 310).

By the von Mises method for semiparametric maximum likelihood estimators (Gill et al., 1989), the score functions are derived by taking the derivative of $l_n(\boldsymbol{\psi}, \hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, a submodel $d\Lambda_{\eta_1}(\cdot) = \{1 + \eta_1\phi_1(\cdot)\}d\Lambda_1(\cdot)$ and a submodel $d\Lambda_{\eta_2}(\cdot) = \{1 + \eta_2\phi_2(\cdot)\}d\Lambda_2(\cdot)$. Here, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are bounded and integrable functions, and η_1 and η_2 are two positive constants.

We denote the infinite-dimensional score functions of the log-pseudo-likelihood function $l(\boldsymbol{\psi}, \hat{\boldsymbol{\theta}})$ by

$$\mathbf{U}_n^*(\boldsymbol{\psi}) = \{U_{1n}^*(\boldsymbol{\psi}), U_{2n}^*(\boldsymbol{\psi}), U_{3n}^*(\boldsymbol{\psi}), U_{4n}^*(u, \boldsymbol{\psi}), U_{5n}^*(v, \boldsymbol{\psi})\},$$

and its expectation under true values $\boldsymbol{\psi}_0$ by

$$\begin{aligned} \mathbf{U}_0^*(\boldsymbol{\psi}) &= \{U_{10}^*(\boldsymbol{\psi}), U_{20}^*(\boldsymbol{\psi}), U_{30}^*(\boldsymbol{\psi}), U_{40}^*(u, \boldsymbol{\psi}), U_{50}^*(v, \boldsymbol{\psi})\} \\ &= \{E_0\{U_{1n}^*(\boldsymbol{\psi})\}, E_0\{U_{2n}^*(\boldsymbol{\psi})\}, E_0\{U_{3n}^*(\boldsymbol{\psi})\}, E_0\{U_{4n}^*(u, \boldsymbol{\psi})\}, E_0\{U_{5n}^*(v, \boldsymbol{\psi})\}\}. \end{aligned}$$

Note that $\mathbf{U}_n$ is the score function of the log-likelihood function $\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\theta})$ and U_n^* is the score function of the log-pseudo-likelihood function $l(\boldsymbol{\psi}, \hat{\boldsymbol{\theta}})$ with estimated parameter $\hat{\boldsymbol{\theta}}$. The score function, U_n^* and its expectation U_0^* are defined on the parameter set $\mathcal{A} \times \mathcal{B}$, where set $\mathcal{A}$ is assumed to be compact in $\mathcal{R}^{2p+1}$, and the set $\mathcal{B}$ consists of nondecreasing functions in the space of functions with bounded variation. The true value $\boldsymbol{\psi}_0$ satisfies the population score function $\mathbf{U}_0(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0) = 0$. The estimating functions evaluated at the true value $\boldsymbol{\psi}_0$ can be written as an empirical process

$$\sqrt{n}\mathbf{U}_n^*(\boldsymbol{\psi}_0) = \sqrt{n}\{\mathbf{U}_n(\boldsymbol{\psi}_0, \hat{\boldsymbol{\theta}}) - \mathbf{U}_n(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0) + \mathbf{U}_n(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0) - \mathbf{U}_0(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0)\}$$

By a lemma for the central limit theorem for processes of bounded variation (Van der Vaart and Wellner, 1996, Example 2.11.16), it can be shown that $\sqrt{n}\{\mathbf{U}_n(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0) - \mathbf{U}_0(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0)\}$ converges weakly to $(\mathcal{W}_1, \mathcal{W}_2)$, where $\mathcal{W}_1$ is a Gaussian random vector, and $\mathcal{W}_2$ is a tight Gaussian process. Furthermore $\sqrt{n}\{\mathbf{U}_n(\boldsymbol{\psi}_0, \hat{\boldsymbol{\theta}}) - \mathbf{U}_n(\boldsymbol{\psi}_0, \boldsymbol{\theta}_0)\}$ converges to $\mathcal{W}_3$ by the asymptotic behavior of $\hat{\boldsymbol{\theta}}$, where $\mathcal{W}_3$ is a Gaussian random vector with mean 0 and variance matrix $\mathcal{I}(\boldsymbol{\theta}_0) = -E\{\dot{\mathcal{L}}(\boldsymbol{\theta})/\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\theta}}'\}$. Denote $\mathcal{W} = (\mathcal{W}_1, \mathcal{W}_2) + \mathcal{W}_3$, and the Fréchet derivative of $\mathbf{U}_0(\boldsymbol{\psi})$ evaluated at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ as $\dot{\mathbf{U}}_0$. In the Supplementary Materials, we outline the proof of the three main conditions for using the Z-theorem: Fréchet differentiability and invertibility, weak convergence of $\sqrt{n}\mathbf{U}_n^*(\boldsymbol{\psi}_0)$ and a stochastic approximation condition of the estimating equations.

Theorem 4: Under the regularity conditions listed in the Supplementary Materials, $\sqrt{n}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0)$ converges weakly to a tight, mean zero Gaussian process $-\dot{\mathbf{U}}_0^{-1}(\mathcal{W})$.

Although the asymptotic variance of the estimators is completely determined by the tightness of $-\dot{U}_0^{-1}(\mathcal{W})$ and its marginal covariance function, direct calculation of the variance-covariance matrix is extremely difficult as there is no explicit form. Alternatively, the bootstrap procedure can be used to estimate the asymptotic variance matrix, as the rate of convergence for $\hat{\psi}_n$ is shown to be $n^{-1/2}$. More details about regularity conditions, proof of Theorem 3 and Theorem 4 are shown in the Supplementary Materials.

5. Simulation studies

We conducted simulations to evaluate the performance of the proposed method and compare it with that of the method that ignores the sampling bias, which we refer to as the naive method. We generated the unbiased bivariate survival times $(\tilde{U}, \tilde{V})$ from the Clayton copula with association parameter α . We considered the Cox models for the two marginal survival functions with covariates (X_1, X_2) and (Z_1, Z_2) , respectively, where X_1 was a binary variable following a Bernoulli distribution with probability 0.5, X_2 was a continuous variables from a Uniform distribution on $U[-0.5, 0.5]$, Z_1 was a binary covariate following a Bernoulli distribution with probability 0.4, and Z_2 was a continuous covariate from a normal distribution on $N(0, 0.5)$. We set $\beta = \gamma = (0.5, 1)$, $\lambda_{01}(u) = u$ and $\lambda_{02}(v) = v$. The underlying truncation time $\tilde{A}$ was generated from an exponential distribution with the parameter θ with a truncation probability of 65%. Following the truncation sampling mechanism, we kept only subjects with $\tilde{U} + \tilde{V} < \tilde{A}$, and thus subjects with $\tilde{U} + \tilde{V} \geq \tilde{A}$ were excluded from the observed data set. The censoring time C was generated independently from a uniform distribution on the interval $U[0, \tau]$. We chose different values of τ to have different censoring rates in the simulation study. Each scenario comprised 500 repetitions.

Table 1 lists the average of the estimators (MEAN), empirical standard errors (ESD), and bootstrap standard errors (ASE) based on 100 resamples of the proposed method and the naive method under four scenarios (I) sample size is 100 and censoring rate is 10%; (II) sample size is 100 and censoring rate is 20%; (III) sample size is 200 and censoring rate is 10%; and (IV) sample size is 200 and censoring rate is 20%. All the proposed estimators achieved good accuracy: the empirical biases were less than 10% when $n=100$ and less than 4% when $n=200$. As expected, the empirical standard errors decreased with the increasing sample sizes. The bootstrap and empirical standard errors agreed well, indicating that the bootstrap procedure can accurately capture the variability of the proposed estimators. Figure 3 shows the estimation of the cumulative baseline hazard functions from the proposed method and the naive method. For all scenarios, the estimated curves obtained from the proposed method were very close to the true curves, with indistinguishable biases.

In contrast, the naive method yielded severely biased estimates for the association parameter and cumulative hazard functions, and would overestimate the effect size, as the biased sampling mechanism was ignored. The bias was relatively small for the effect size, but still greater than the bias from the proposed method. More importantly, the naive method consistently underestimated the cumulative hazard functions. For example, when $n = 100$ and the censoring percentage was 20%, the empirical biases were around 25% for the covariate effects and over 40% for the cumulative hazard functions from the naive method.

Such bias could result in a misleading prediction of the residual survival time after disease recurrence.

We also conducted simulation studies shown in Supplementary Materials (G) to demonstrate that the exploratory diagnostic method (Chen and Bandeen-Roche, 2005) could effectively help the selection of an appropriate copula model even under biased-sampling. Although the diagnostic method could not accurately estimate the cross ratio, it can adequately detect the association pattern, from which we could make a decision of which copula model to choose.

6. Data Applications

6.1. Study of Breast Cancer Recurrence

We first apply the proposed method to estimate the survival distribution and evaluate its risk factors after breast cancer recurrence. Breast cancer recurrence is defined as having a subsequent mastectomy occurred more than 10 months after their initial breast conserving surgery, from the SEER-Medicare database. This algorithm of defining breast cancer recurrence has been used in literature (Albert et al., 2012; Yao et al., 2013). The identified cohort included 386 patients aged 65 years and older who had cancer recurrence and were alive in 2004. We chose year 2004 as the enrollment time since around that time Trastuzumab was showed to increase the clinical benefit for patients with HER2-positive (Marty et al., 2005; Slamon et al., 2001). Patients in the identified cohort were followed until death or the last observation time as of December 31, 2011. We extracted each patient's baseline covariates, including race, age, tumor biology including hormone receptor (HR) status, and American Joint Committee on Cancer (AJCC) stages. We retrieved information on the date of the initial diagnosis of breast cancer for each patient in the cohort. Biased sampling occurred because the observed time intervals from the breast cancer diagnosis to death tended to be longer than those in the general population as illustrated in Figure 1.

Table 2 summarizes the patient characteristics. Two Cox proportional hazard models were assumed for the time from the breast cancer diagnosis to disease recurrence and time from the recurrence to death, respectively. As shown in Figure S3 of the Supplementary Materials, the estimated cross ratio stayed approximately constant, which suggested a Clayton copula was appropriate for the underlying dependent structure between time from diagnosis to recurrence and time from recurrence to death. We then assumed a Clayton copula for the dependence structure and a Weibull distribution for the truncation time. The estimated hazard ratio from the proposed method along with 95% confidence intervals (CIs), standard errors (SEs) and p-values are provided in the left panel of Table 3. The results from the naive method, which ignores the sampling bias, are shown in the right panel of Table 3. For the residual survival time, the proposed method indicates that age (> 75), HR (+) and advanced AJCC stage (II, III or IV) were significantly associated with higher risk of death, while the naive method suggests that AJCC stage did not have a significant effect on the residual survival time. The estimated association parameter α from the copula model was 0.10, with SE of 0.05. This suggests a mildly positive association between the survival time from diagnosis to cancer recurrence and the survival time from recurrence to death. Note that although the data suggest a mild association between time to cancer recurrence ($\tilde{U}$) and

residual survival time ($\tilde{V}$), the association between the overall survival time from cancer diagnosis ($\tilde{T}$) and the residual survival time ($\tilde{V}$) could be strong. By the naive method, the estimated association parameter α was -0.15 , with SE of 0.07 , which mistakenly suggests a negative correlation between $\tilde{T}$ and $\tilde{V}$.

In Figure 4, we provide the estimated survival functions for time to cancer recurrence and the residual survival functions for patients with different profiles, using both the proposed method and the naive method. The median survival times for subgroups I ~ IV as determined by age and HR status were 41.52, 74.95, 114.67 and infinity months, respectively, from the naive method and were 24.96, 36.93, 56.30 and 90.77 months, respectively, from the proposed method. The substantial differences between the estimated survival functions indicate that the use of the naive approach, which ignores biased sampling, will lead to severe overestimation of the survival distributions.

6.2. Nun Study of Aging and Alzheimer's Disease

We next illustrate the proposed method by analyzing data from a longitudinal cohort study, the Nun Study of Aging and Alzheimer's Disease. The design of this cohort study and its main results have been fully described (Snowdon et al., 1997; Snowdon, 2003). The data were collected from Catholic sisters of Notre Dame, who were 75 to 106 years of age. We aim to assess the residual survival time measured from a dementia diagnosis, and the relationship between risk factors and the residual survival time. By excluding subjects who did not show signs of dementia during the study follow-up, a total of 145 subjects were included in the analysis. The observed data were subject to biased sampling because the subjects who died before their diagnosis of dementia were excluded from the study.

We assumed a Clayton copula for the dependence structure, two Cox proportional hazard models for the time from birth to dementia diagnosis and time from the dementia diagnosis to death, and a Weibull distribution for the truncation time. Table 4 lists the estimated covariate coefficients for the Nun Study obtained from the proposed method (left panel) and the naive method (right panel), which ignores the sampling bias. Figure S1 in the Supplementary Materials provides the estimated survival curves according to different risk factor profiles. In terms of the effects of the risk factors on the time from dementia diagnosis until death, the conclusions from the proposed method and naive method were fairly consistent: both the presence of APOE4 and high education level did not have significant effects on the residual survival time from the onset of dementia. However, as seen in Figure S1, the naive method consistently overestimated the survival curves for both the time to dementia diagnosis and residual survival time to death for all subgroups. For example, for the four subgroups determined by the APOE4 status and education level, the median residual survival times were 3.33, 3.21, 2.60 and 2.46 years by the naive method, and 2.10, 1.92, 1.72 and 1.64 years by the proposed method after properly adjusting for the sampling bias. The estimated degree of association between the time to dementia diagnosis and the residual survival time from the onset of dementia were different: the proposed method suggested a mildly negative association ($\hat{\alpha} = -0.11$, SE = 0.04); whereas the naive method suggested a stronger negative association ($\hat{\alpha} = -0.18$, SE = 0.06).

7. Discussion

In this paper, we propose a bivariate survival model for exploiting the association between covariates and time-to-event outcomes from a prevalent cohort in which the data are subject to left truncation. To the best of our knowledge, there are no semiparametric models and inference available in the literature for analyzing bivariate survival data from a prevalent cohort. We developed a modified EM algorithm to improve computational efficiency, in which we converted the biased sampling problem to a missing data framework. We implemented a two-stage estimation procedure to further decrease the computational burden: in the first stage, we estimate marginal distributions, and in the second stage, we estimate the association parameter by fixing the marginal distributions from stage 1. To take into account of the sampling bias, we can impose a bivariate model either on $(\tilde{T}, \tilde{V})$ or on $(\tilde{U}, \tilde{V})$. If we choose the former bivariate model, we have an additional constraint $\tilde{T} > \tilde{V}$, which brings challenges in the model identifiability and computational complexity in the EM algorithm. It motivated us to use $\tilde{U}$ as a bridge to simplify the model and estimating procedure when handling the sampling bias.

To model bivariate survival data, a known copula function is assumed to facilitate the dependence structure. Determining the best way to choose an appropriate copula model from the observed biased data is essential. For traditional bivariate survival data without sampling bias, Chen and Bandeen-Roche (2005) proposed an exploratory method for accessing the appropriateness of a copula model. Even though the data from the prevalent cohort are subject to truncation, we can still implement their method to carry out an exploratory analysis and to choose a reasonable copula function. As expected, the diagnostic method could not accurately estimate the cross-ratio, but it detected the association pattern for the model selection.

In the proposed method, the truncation mechanism is assumed to be known up to a finite parametric form. For the univariate survival outcome, Wang (1989) showed that if the distributions of the survival time and truncation time are both nonparametric, the joint model is not identifiable. We expect similar identifiability issues for the bivariate survival outcomes. To have a more robust model for the truncation time, we can use a flexible class of parametric models, such as a truncated generalized Gamma distribution with three parameters (Stacy, 1962). We performed simulation studies to investigate the performance when using the truncated generalized Gamma distribution as the assumed truncation distribution under different scenarios, including a misspecification scenario. Details can be found in the online Supplementary Materials. In summary, the simulation confirmed that the use of the truncated generalized Gamma distribution as the assumed truncation distribution had stable and robust performance when estimating the parameters of interest (i.e., parameters under the bivariate survival model).

For unknown parameter θ associated with the truncated distribution, we used the conditional likelihood to estimate it before conducting the EM algorithm. Note that the maximum likelihood estimator of θ can also be simultaneously estimated in the EM algorithm by updating the values of $\hat{\theta}^{(p)}$ along with other parameters in each iteration of the EM algorithm until convergence is achieved. Although theoretically the estimators obtained by maximizing

the full likelihood should be the most statistically efficient, our simulation studies suggested that the performance of the pseudo-likelihood method was more stable than that of the iteration method for the full likelihood in this setting. One possible reason is the challenge in numerically identifying the global maximum of the full likelihood. This finding motivates us to separately estimate the parameter associated with the truncated distribution by using the conditional likelihood. Another model assumption is the proportionality assumption on the two marginal hazard functions. Standard diagnostic tools, such as the method by Lin et al. (1993), cannot be used directly to assess the proportionality due to the biased sampling issue. Developing rigorous model-checking tools to accommodate the sampling bias is beyond the scope of this paper and is worthy of future research.

Different from our proposed method, the method for semi-competing risk data provides an alternative tool for studying the disease progression process (Fine et al., 2001; Peng and Fine, 2007; Zhou et al., 2016). When applying the method for semi-competing risk data, the endpoints are usually measured from the disease diagnosis. While such a method is preferred when studying the disease progression process, including the period from disease diagnosis to disease recurrence/progression, our proposed method focuses on understanding the residual survival time after disease recurrence/progression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by grants from the National Cancer Institute (R01CA193878 and P30CA016672) and the Andrew Sabin Family Fellowship.

8.: Appendix: Details of M Step 1

In M step1, we calculate the probability masses of $f(u_j, v_j | \mathbf{X}_i)$ based on the current estimator $\psi^{(p)}$ on the grids shown in Figure 2, denoted as $p_{ijj'}^{(p)}$. Specifically, we let

$$\begin{aligned}
 p_{ik_1k_2}^{(p)} &= S(u_{k_1}, v_{k_2} | \mathbf{X}_i; \psi^{(p)}) \\
 p_{ik'_1k_2}^{(p)} &= S(u_{k'_1}, v_{k_2} | \mathbf{X}_i; \psi^{(p)}) - p_{ik_1k_2}^{(p)}, \quad k'_1 = k_1 - 1 \\
 p_{ik_1k'_2}^{(p)} &= S(u_{k_1}, v_{k'_2} | \mathbf{X}_i; \psi^{(p)}) - p_{ik_1k_2}^{(p)}, \quad k'_2 = k_2 - 1 \\
 p_{ik'_1k'_2}^{(p)} &= S(u_{k'_1}, v_{k'_2} | \mathbf{X}_i; \psi^{(p)}) - p_{ik_1k_2}^{(p)} - p_{ik'_1k_2}^{(p)} - p_{ik_1k'_2}^{(p)}, \quad k'_1 = k_1 - 1, k'_2 = k_2 - 1 \\
 &\dots \\
 p_{ijj'}^{(p)} &= S(u_j, v_{j'} | \mathbf{X}_i; \psi^{(p)}) - \sum_{l=j}^{k_1} \sum_{q=j'}^{k_2} \{1 - I(l=j, q=j')\} p_{ilq}^{(p)}, \quad j = k_1, \dots, 1; j' = k_2, \dots, 1,
 \end{aligned}$$

where

$$S(u_j, v_{j'} | \mathbf{X}_i; \boldsymbol{\psi}^{(p)}) = B_{\alpha^{(p)}} \{ S_1(u_j | \mathbf{X}_i; \boldsymbol{\psi}^{(p)}), S_2(v_{j'} | \mathbf{X}_i; \boldsymbol{\psi}^{(p)}) \}.$$

$$S_1(u_j | \mathbf{X}_i; \boldsymbol{\psi}^{(p)}) = \exp \left\{ - \sum_{l=1}^j \lambda_{1l}^{(p)} \exp(\mathbf{X}_i^T \boldsymbol{\beta}^{(p)}) \right\},$$

and

$$S_2(v_{j'} | \mathbf{X}_i; \boldsymbol{\psi}^{(p)}) = \exp \left\{ - \sum_{l=1}^{j'} \lambda_{2l}^{(p)} \exp(\mathbf{X}_i^T \boldsymbol{\gamma}^{(p)}) \right\}.$$

References

- Albert JM, Liu DD, Shen Y, Pan I-W, Shih Y-CT, Hoffman KE, Buchholz TA, Giordano SH, and Smith BD (2012). Nomogram to predict the benefit of radiation for older patients with breast cancer treated with conservative surgery. *Journal of Clinical Oncology* 30(23), 2837–2843. [PubMed: 22734034]
- Asgharian M, M'Lan CE, and Wolfson DB (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association* 97(457), 201–209.
- Asgharian M, Wolfson C, and Wolfson DB (2014). Analysis of biased survival data: the canadian study of health and aging and beyond. In: *Statistics in action: a Canadian outlook*, 193–208.
- Asgharian M, Wolfson DB, et al. (2005). Asymptotic behavior of the unconditional npml of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics* 33(5), 2109–2131.
- Begg CB and Gray RJ (1987). Methodology for case-control studies with prevalent cases. *Biometrika* 74 (1), 191–195.
- Bergeron P-J, Asgharian M, and Wolfson DB (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association* 103(482), 737–742.
- Brewster AM, Hortobagyi GN, Broglio KR, Kau S-W, Santa-Maria CA, Arun B, Buzdar AU, Booser DJ, Valero V, Bondy M, et al. (2008). Residual risk of breast cancer recurrence 5 years after adjuvant therapy. *Journal of the National Cancer Institute* 100(16), 1179–1183. [PubMed: 18695137]
- Cardoso F, Costa A, Norton L, Cameron D, Cufer T, Fallowfield L, Francis P, Gligorov J, Kyriakides S, Lin N, et al. (2012). 1st international consensus guidelines for advanced breast cancer (abc 1). *The Breast* 21 (3), 242–252. [PubMed: 22425534]
- Chen MC and Bandeen-Roche K (2005). A diagnostic for association in bivariate survival models. *Lifetime Data Analysis* 11 (2), 245–264. [PubMed: 15940823]
- Chen X and Fan Y (2007). A model selection test for bivariate failure-time data. *Econometric Theory* 23(03), 414–439.
- Cox D (1969). Some sampling problems in technology in *New Developments in Survey Sampling*, eds. Johnson and Smith, New York: Wiley.
- Dent R, Valentini A, Hanna W, Rawlinson E, Rakovitch E, Sun P, and Narod S(2014). Factors associated with breast cancer mortality after local recurrence. *Current Oncology* 21 (3), 418–425.
- Fine J, Jiang H, and Chappell R (2001). On semi-competing risks data. *Biometrika* 88 (4), 907.
- Gail MH and Benichou J (2000). *Encyclopedia of epidemiologic methods*. John Wiley & Sons.
- Genest C and Rivest L-P (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association* 88(423), 1034–1043.
- Gill RD, Wellner JA, and Præstgaard J (1989). Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics* 16(2), 97–128.

- Goldhirsch A, Wood W, Coates A, Gelber R, Thürlimann B, Senn H-J, et al. (2011). Strategies for subtypes dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of oncology* 22(8), 1736–1747. [PubMed: 21709140]
- Høst H and Lund E (1986). Age as a prognostic factor in breast cancer. *Cancer* 57(11), 2217–2221. [PubMed: 3697919]
- Kim JP, Lu W, Sit T, and Ying Z (2013). A unified approach to semiparametric transformation models under general biased sampling schemes. *Journal of the American Statistical Association* 108(501), 217–227.
- Lehmann EL (1999). *Elements of large-sample theory*. Springer Science & Business Media.
- Lin DY, Wei L-J, and Ying Z (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* 80(3), 557–572.
- Liu H, Ning J, Qin J, and Shen Y (2016). Semiparametric maximum likelihood inference for truncated or biased-sampling data. *Statistica Sinica* 26, 1087–1115.
- Marty M, Cognetti F, Maraninchi D, Snyder R, Mauriac L, Tubiana-Hulin M, Chan S, Grimes D, Antón A, Lluch A, et al. (2005). Randomized phase ii trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the m77001 study group. *Journal of Clinical Oncology* 23(19), 4265–4274. [PubMed: 15911866]
- Nelsen RB (2007). *An introduction to copulas*, Volume 139 Springer Science & Business Media.
- Neyman J (1955). Statisticsservant of all sciences. *Science* 122(3166), 401–416. [PubMed: 13246647]
- Ning J, Qin J, and Shen Y (2014). Score estimating equations from embedded likelihood functions under accelerated failure time model. *Journal of the American Statistical Association* 109 (508), 1625–1635. [PubMed: 25663727]
- Oakes D (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (3), 414–422.
- Oakes D (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84(406), 487–493.
- Peng L and Fine JP (2007). Regression modeling of semicompeting risks data. *Biometrics* 63(1), 96–108.
- Qin J, Ning J, Liu H, and Shen Y (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association* 106 (496), 1434–1449. [PubMed: 22323840]
- Qin J and Shen Y (2010). Statistical methods for analyzing right-censored length-biased data under cox model. *Biometrics* 66(2), 382–392. [PubMed: 19522872]
- Rothman KJ, Greenland S, and Lash TL (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.
- Shen Y, Ning J, and Qin J (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association* 104 (487), 1192–1202. [PubMed: 21057599]
- Shih JH and Louis TA (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 51 (4), 1384–1399. [PubMed: 8589230]
- Siegel R, Ma J, Zou Z, and Jemal A (2014). *Cancer statistics, 2014*. CA: A Cancer Journal for Clinicians 64 (1), 9–29. [PubMed: 24399786]
- Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, et al. (2001). Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *New England Journal of Medicine* 344 (11), 783–792. [PubMed: 11248153]
- Snowdon DA (2003). Healthy aging and dementia: findings from the nun study. *Annals of Internal Medicine* 139(5_Part_2), 450–454. [PubMed: 12965975]
- Snowdon DA, Greiner LH, Mortimer JA, Riley KP, Greiner PA, and Markesbery WR (1997). Brain infarction and the clinical expression of alzheimer disease: the nun study. *JAMA* 277(10), 813–817. [PubMed: 9052711]

- Stacy EW. (1962). A generalization of the gamma distribution. *The Annals of mathematical statistics* 33(3), 1187–1192.
- Tsai WY (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* 96(3), 601–615. [PubMed: 22422175]
- Van der Vaart AW and Wellner JA (1996). *Weak convergence and empirical processes*. Springer, New York.
- Vardi Y (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* 10(2), 616–620.
- Vardi Y (1985). Empirical distributions in selection bias models. *The Annals of Statistics* 13 (1), 178–203.
- Vardi Y (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* 76(4), 751–761.
- Wang M-C (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* 84 (407), 742–748.
- Wang M-C, Brookmeyer R, and Jewell NP (1993). Statistical models for prevalent cohort data. *Biometrics* 49 (1), 1–11. [PubMed: 8513095]
- Wang M-C, Jewell NP, and Tsai W-Y (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics* 14 (4), 1597–1605.
- Wang W and Wells MT (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* 95 (449), 62–72.
- Wolfson C, Wolfson DB, Asgharian M, M 'Lan CE, Østbye T, Rockwood K, and Hogan D. f. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine* 344 (15), 1111–1116. [PubMed: 11297701]
- Yao N, Mackley HB, Anderson RT, and Recht A (2013). Survival after partial breast brachytherapy in elderly patients with nonmetastatic breast cancer. *Brachytherapy* 12(4), 293–302.
- Zelen M (2006). Forward and backward recurrence times and length biased sampling: age specific models In *Probability, Statistics and Modelling in Public Health*, pp. 1–11. Springer.
- Zelen M and Feinleib M (1969). On the theory of screening for chronic diseases. *Biometrika* 56(3), 601–614.
- Zhou R, Zhu H, Bondy M, and Ning J (2016). Semiparametric model for semi-competing risks data with application to breast cancer study. *Lifetime Data Analysis* 22 (3), 456–471. [PubMed: 26340889]

Fig. 1.

Bivariate survival data from biased sampling. Patients with ID (1), (2) and (3) are sampled, whereas patient (4) is excluded.

Fig. 2.
Flow chart of calculating the probability mass on the grids.

Fig. 3.

Cumulative baseline hazard function estimation of two marginal models ($\tilde{U}$ and $\tilde{V}$) for sample-biased bivariate survival data. Red line: true curve; blue solid line: estimate from the proposed method; blue dashed line: 95% CI from the proposed method; black solid line: estimate from the naive method; black dotted line: 95% CI from naive method.

I: age ≥ 75 , HR(+); II: age ≥ 75 , HR(-); III: age < 75 , HR(+); IV: age < 75 , HR(-)

Fig. 4.

Survival plot along with 95% confidence interval (shaded) for patients of AJCC stage 0 or I with different profiles. BC: breast cancer; HR: hormone receptor

Table 1.

Summary of simulation studies from 500 runs with sample-biased bivariate survival data.

Censor	n	PAR	TRUE	Proposed Method			Naive Method		
				MEAN	ESD	ASE	MEAN	ESD	ASE
10%	100	θ	0.25	0.260	0.124	0.112	-	-	-
		α	0.5	0.494	0.168	0.188	0.460	0.185	0.176
		β_1	0.5	0.531	0.202	0.197	0.621	0.227	0.224
		β_2	1.0	1.094	0.361	0.351	1.277	0.399	0.396
		γ_1	0.5	0.539	0.211	0.212	0.643	0.243	0.244
		γ_2	1.0	1.050	0.228	0.224	1.250	0.251	0.253
		$\Lambda_1(l)$	1.0	0.939	0.174	0.152	0.583	0.113	0.110
		$\Lambda_2(l)$	1.0	0.942	0.165	0.149	0.558	0.102	0.104
	200	θ	0.25	0.255	0.084	0.080	-	-	-
		α	0.5	0.487	0.100	0.095	0.466	0.131	0.124
		β_1	0.5	0.513	0.135	0.131	0.611	0.159	0.152
		β_2	1.0	1.040	0.235	0.232	1.238	0.272	0.268
		γ_1	0.5	0.512	0.132	0.140	0.622	0.158	0.164
		γ_2	1.0	1.025	0.150	0.149	1.241	0.170	0.171
		$\Lambda_1(l)$	1.0	0.963	0.120	0.106	0.579	0.074	0.075
		$\Lambda_2(l)$	1.0	0.973	0.120	0.105	0.560	0.072	0.072
20%	100	θ	0.25	0.261	0.142	0.127	-	-	-
		α	0.5	0.512	0.208	0.211	0.458	0.191	0.185
		β_1	0.5	0.533	0.204	0.198	0.621	0.227	0.223
		β_2	1.0	1.097	0.354	0.352	1.277	0.399	0.395
		γ_1	0.5	0.539	0.219	0.221	0.647	0.255	0.257
		γ_2	1.0	1.045	0.243	0.238	1.257	0.264	0.270
		$\Lambda_1(l)$	1.0	0.935	0.178	0.154	0.583	0.113	0.110
		$\Lambda_2(l)$	1.0	0.954	0.178	0.156	0.558	0.105	0.107
	200	θ	0.25	0.255	0.100	0.094	-	-	-
		α	0.5	0.491	0.109	0.098	0.465	0.138	0.130
		β_1	0.5	0.512	0.135	0.132	0.611	0.159	0.152
		β_2	1.0	1.039	0.237	0.233	1.238	0.272	0.268
		γ_1	0.5	0.508	0.139	0.145	0.625	0.168	0.172
		γ_2	1.0	1.012	0.159	0.158	1.246	0.181	0.182
		$\Lambda_1(l)$	1.0	0.963	0.121	0.107	0.579	0.074	0.075
		$\Lambda_2(l)$	1.0	0.988	0.124	0.109	0.560	0.074	0.074

Table 2.

Patient characteristics in the cohort of SEER-Medicare data.

Characteristics		N	%
Age	75	247	64.0
	>75	139	36.0
AJCC	0 or I	257	66.6
	II,III or IV	99	25.6
	NA	30	7.8
Race	Non-Hispanic White	348	90.2
	Other	38	9.8
HR	–	171	44.3
	+	215	55.7

Table 3.

Analysis results the cohort from SEER-Medicare data

Characteristics	Proposed Method			Naive Method		
	HR	SE (95% CI)	p-value	HR	SE (95% CI)	p-value
Survival model for time from disease diagnosis to recurrence						
Age (75 v.s. <75)	1.52	0.16 (1.24, 1.86)	<0.01	1.46	0.17 (1.17, 1.82)	<0.01
AJCC (II,III or IV v.s. 0 or I)	1.48	0.16 (1.20, 1.83)	<0.01	1.49	0.18 (1.17, 1.89)	<0.01
HR(+ v.s. -)	0.92	0.08 (0.77, 1.10)	0.34	0.88	0.10 (0.71, 1.10)	0.26
Survival model for time from disease recurrence to death						
Age (75 v.s. <75)	1.99	0.22 (1.59, 2.47)	<0.01	2.47	0.39 (1.81, 3.37)	<0.01
AJCC (II,III or IV v.s. 0 or I)	1.36	0.15 (1.09, 1.68)	0.01	1.33	0.23 (0.96, 1.86)	0.09
HR(+ v.s. -)	1.41	0.18 (1.09, 1.81)	0.01	1.76	0.30 (1.27, 2.45)	<0.01

Table 4.

Analysis results of the Nun Study.

Characteristics	Proposed Method			Naive Method		
	HR	SE (95% CI)	p-value	HR	SE (95% CI)	p-value
Survival model for time from birth to dementia						
APOE4(presence)	1.39	0.24 (0.99, 1.95)	0.06	1.58	0.31 (1.08, 2.33)	0.02
Education(graduate level)	1.21	0.16 (0.94, 1.57)	0.14	1.24	0.22 (0.88, 1.75)	0.21
Survival model from dementia to death						
APOE4(presence)	0.88	0.14 (0.65, 1.19)	0.40	0.77	0.16 (0.51, 1.17)	0.23
Education(graduate level)	0.96	0.15 (0.70, 1.31)	0.78	0.92	0.17 (0.63, 1.33)	0.66