

Semiparametric model for semi-competing risks data with application to breast cancer study

Renke Zhou^{1,2} · Hong Zhu³ · Melissa Bondy¹ ·
Jing Ning⁴

Received: 12 September 2014 / Accepted: 25 August 2015 / Published online: 5 September 2015
© Springer Science+Business Media New York 2015

Abstract For many forms of cancer, patients will receive the initial regimen of treatments, then experience cancer progression and eventually die of the disease. Understanding the disease process in patients with cancer is essential in clinical, epidemiological and translational research. One challenge in analyzing such data is that death dependently censors cancer progression (e.g., recurrence), whereas progression does not censor death. We deal with the informative censoring by first selecting a suitable copula model through an exploratory diagnostic approach and then developing an inference procedure to simultaneously estimate the marginal survival function of cancer relapse and an association parameter in the copula model. We show that the proposed estimators possess consistency and weak convergence. We use simulation

Renke Zhou and Hong Zhu have contributed equally to this work.

✉ Hong Zhu
hong.zhu@utsouthwestern.edu

Renke Zhou
rzhou@bcm.edu

Melissa Bondy
mbondy@bcm.edu

Jing Ning
jning@mdanderson.org

- ¹ Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA
- ² Division of Biostatistics, The University of Texas School of Public Health, Houston, TX 77030, USA
- ³ Division of Biostatistics, Department of Clinical Sciences, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
- ⁴ Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

studies to evaluate the finite sample performance of the proposed method, and illustrate it through an application to data from a study of early stage breast cancer.

Keywords Copula model · Informative censoring · Model diagnostic · Semi-competing risks · Simultaneous inference

1 Introduction

Medical research frequently yields multiple event times that correspond to landmarks in disease progression. The events may be non-terminal events, such as disease recurrence, defined as progressive disease, or a terminal event, such as death. Semi-competing risks data (Fine et al. 2001; Peng et al. 2007) are encountered when the terminal event censors the non-terminal event, but the non-terminal event does not prevent subsequent observation of the terminal event. An example of this can be found in a retrospective study of data collected from women diagnosed with stage I or II breast cancer, and maintained as the early stage breast cancer repository (ESBCR) at the University of Texas MD Anderson Cancer Center. For this study, breast cancer relapse is the non-terminal event and death is the terminal event. It is essential to evaluate the survival distribution for women who experience breast cancer relapse following the initial treatment regimen in order to predict the risk of relapse for ongoing patients and make further treatment decisions. However, this endeavor is a challenge because of the informative censoring issue in the semi-competing risks data. Specifically, the occurrence of death is expected to be positively correlated with cancer relapse and cannot be considered as independent censoring. Consequently, the Kaplan–Meier estimator (Kaplan and Meier 1958) will tend to overestimate the true survival probabilities associated with cancer relapse in the presence of death.

One approach to semi-competing risks data is to use crude quantities, such as the cause-specific hazard and cumulative incidence function due to the identifiability issue. The naive method ignores the information available after the non-terminal event by only considering the time and type of the first event. Xu et al. (2010) considers the crude quantity (cause-specific hazard) and developed an extension of the illness-death model with a shared frailty, which can take into account the information after the non-terminal event. Another approach focuses on net quantities, such as the marginal distribution of the non-terminal event time, which is not identifiable without further assumption on the dependence structure between the non-terminal and terminal event times (similar to how competing risks are handled). This approach can fully utilize the observed data and provide useful information on the non-terminal event, but is subject to theoretical and computational challenges due to informative censoring (Fine et al. 2001; Peng and Fine 2007; Hsieh et al. 2008; Ding et al. 2009; Hsieh and Huang 2012; Chen 2012). This paper investigates the semi-competing risks data based on the net quantities.

To handle informative censoring, it is common to assume that the bivariate distribution of non-terminal and terminal event times is a known copula. For example, a Gamma frailty copula (Clayton 1978), which assumes a constant association, has been used for modeling dependence structure (Fine et al. 2001; Jiang et al. 2005).

The estimating procedure proposed by Fine et al. (2001) requires a separate estimation for the association parameter using a concordance estimating function at the first stage. These inferences were extended to some other parametric copulas by Wang (2003). Peng and Fine (2007) and Hsieh and Huang (2012) developed copula-based models to account for the time-varying effects of a treatment on the marginal distribution of a non-terminal event. Chen (2012) proposed a nonparametric maximum likelihood estimation approach for the semiparametric transformational models for marginal regressions, where dependence is modeled under an assumed copula. However, in reality, a specific copula model specification does not fit most applications, so its use will probably lead to incorrect statistical inference. Thus, an immediate question is how to choose an appropriate copula model for the dependence structure based on observed bivariate survival data. Chen and Bandeen-Roche (2005) provided a diagnostic method to infer the choice of frailty distribution (copula) family for bivariate survival data, which can be implemented by exploiting the relationship between a local version of Kendall's tau and the frailty distribution. Specifically, a constant relationship may suggest the Gamma frailty copula, and an increasing one may prefer a positive stable copula (Gumbel 1960).

Instead of restricting to the Gamma frailty copula for modeling dependence between breast cancer relapse and death in the ESBCR data and estimating its association parameter separately at an earlier stage, we first use the exploratory diagnostic method (Chen and Bandeen-Roche 2005) to check the association pattern between cancer relapse and death. Then, we develop a semiparametric method for simultaneous inference on the marginal distribution of breast cancer relapse and the association parameter, based on a chosen copula model.

The remainder of the article is organized as follows. In Sect. 2, we introduce the notation and copula model for semi-competing risks data. In Sect. 3, a suitable copula model is selected by exploring the association pattern in the ESBCR data. Section 4 presents the simultaneous inference on the association parameter and the marginal distribution of the non-terminal event time. In Sect. 5, simulation studies are conducted to assess the performance of the proposed method. An application to the ESBCR data is presented as an illustration in Sect. 6. Some concluding remarks are given in Sect. 7. The proof of the asymptotic properties is provided in the Appendix.

2 Notation and model

Let T_1 be the time from the initial event (e.g., initial treatment) to a non-terminal event (e.g., disease relapse), and T_2 be the time from the initial event to the terminal event (e.g., 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 $\delta_0 = I(Z \leq C) = \delta_1 + \delta_2 - \delta_1\delta_2$. The observed data can be denoted as $\{(X_i, \delta_{1i}, Y_i, \delta_{2i}, Z_i, \delta_{0i}), i = 1, \dots, n\}$.

With semi-competing risks data, the dependence structure is usually of biological interest and is required to ensure the validity for the estimation of the marginal survival function of T_1 . As a commonly used model, a copula C_α is a bivariate distribution

function with density c_α on $[0, 1]^2$ for $\alpha \in \mathfrak{R}^1$ (Oakes 1982). Under a copula, the bivariate survival function and density function of (T_1, T_2) can be written as

$$S(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2) = C_\alpha \{S_1(t_1), S_2(t_2)\}, \quad 0 \leq t_1 \leq t_2,$$

$$f(t_1, t_2) = c_\alpha \{S_1(t_1), S_2(t_2)\} f_1(t_1) f_2(t_2), \quad 0 \leq t_1 \leq t_2,$$

where (S_1, S_2) and (f_1, f_2) are the corresponding marginal survival and density functions, respectively. One reason for the popularity of copulas is that the association parameter α is related to a local dependence measure, the cross ratio (Oakes 1989), which is defined as

$$\theta(t_1, t_2) = \frac{\lambda(t_1|T_2 = t_2)}{\lambda(t_1|T_2 > t_2)}, \tag{1}$$

where $\lambda(t_1|A)$ is a conditional hazard function,

$$\lambda(t_1|A) = \lim_{\epsilon \rightarrow 0} \frac{\Pr(T_1 < t_1 + \epsilon | T_1 \geq t_1, T_2 \in A)}{\epsilon}.$$

The choice of the copula C_α determines the expression of $\theta(t_1, t_2)$. Particularly, under a popular class of copulas, namely the Archimedean copula, the bivariate survival function can be expressed as

$$S(t_1, t_2) = \phi_\alpha [\phi_\alpha^{-1}\{S_1(t_1)\} + \phi_\alpha^{-1}\{S_2(t_2)\}],$$

where ϕ_α is a non-increasing convex function defined on $(0, 1]$ with $\phi_\alpha(0) = 1$. Examples of Archimedean copulas include the Clayton (Gamma frailty) copula,

$$S(t_1, t_2) = \left\{ S_1(t_1)^{1-\alpha} + S_2(t_2)^{1-\alpha} - 1 \right\}^{\frac{1}{1-\alpha}}, \quad \alpha > 1,$$

the positive stable copula,

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

and the Frank copula (Frank 1979) as

$$S(t_1, t_2) = \log_\alpha \left\{ 1 + \frac{(\alpha^{S_1(t_1)} - 1)(\alpha^{S_2(t_2)} - 1)}{\alpha - 1} \right\}.$$

The cross ratio under the Clayton copula is a constant $\theta(t_1, t_2) = \alpha$, while under the positive stable copula, it is a function of $S(t_1, t_2)$ and decreases with time $\theta(t_1, t_2) = 1 - (1 - \alpha)\{\alpha \log S(t_1, t_2)\}^{-1}$. Such a connection can be used to choose an appropriate copula model in applications.

2.1 A motivating study

The ESBCR was created for a retrospective cohort study of ($n=2470$) women diagnosed with American Joint Committee on Cancer pathologic stage I or II breast cancer who were treated at The University of Texas MD Anderson Cancer Center between January 1, 1985 and December 31, 2000 (Brewster et al. 2007). The inclusion criteria for the ESBCR 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, 785 had died before the end of the study, 1684 were alive, and one record was missing the vital status. There were 2425 patients with complete information of disease relapse status and dates, vital status, and death/last contact date. Among that group, 562 had data regarding disease relapse, which included local lymph node recurrence, local breast cancer recurrence, metastasis to the contralateral breast or chest wall, metastasis to other body sites, and a second primary breast cancer.

3 Model selection by association pattern in ESBCR

Many existing estimation methods for semi-competing risks data require imposition of the Gamma frailty model assumption on T_1 and T_2 (Fine et al. 2001; Jiang et al. 2005), which implies a constant association and may not always be biologically reasonable in practical applications. To avoid possible misspecification on the copula and the induction of misleading conclusions, we first aim to select an appropriate copula model based on the observed bivariate survival data. Specifically, we apply an easily implemented approach (Chen and Bandeen-Roche 2005) for diagnosing the association pattern and choose the proper copula model for breast cancer relapse and death in the ESBCR data. This diagnostic approach constructs an exploratory display that plots the cross ratio and visually conveys the fit of a given Archimedean copula model.

We base our approach on the fact that the cross ratio defined in Eq. (1) is a ratio of the probabilities that one pair of bivariate failure times is concordant or discordant, conditional on the component-wise minimum failure times (Oakes 1989). For any pair of bivariate times, $T_i = (T_{1i}, T_{2i})$ and $T_j = (T_{1j}, T_{2j})$, let (T_{1ij}, T_{2ij}) be the corresponding component-wise minimum of (T_i, T_j) , defined by $T_{1ij} = \min(T_{1i}, T_{1j})$ and $T_{2ij} = \min(T_{2i}, T_{2j})$. Define a pair of bivariate times (T_i, T_j) as concordant if $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0$, and discordant if $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0$. We have

$$\theta(t_{1ij}, t_{2ij}) = \frac{\Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0 \mid (T_{1ij}, T_{2ij}) = (t_{1ij}, t_{2ij})\}}{\Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0 \mid (T_{1ij}, T_{2ij}) = (t_{1ij}, t_{2ij})\}},$$

which provides an estimator of the cross ratio θ by using the counts of concordances and discordances. Chen and Bandeen-Roche (2005) suggested using a plot of $\hat{\theta}(t_{1ij}, t_{2ij})$ against $\hat{S}(t_{1ij}, t_{2ij})$ to approximate the functional relationship between the cross ratio and the bivariate survival function. Specifically, a horizontal line may support a Gamma

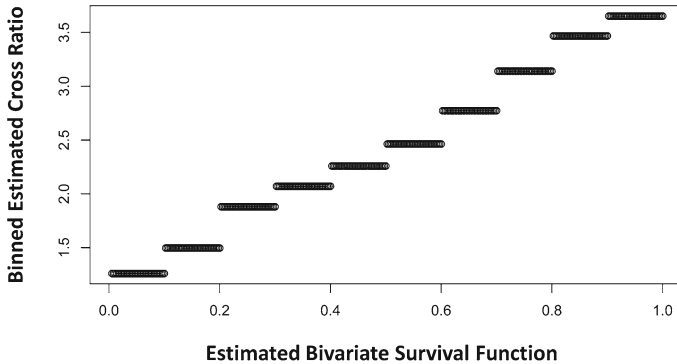


Fig. 1 Relationship between the estimated cross ratio and the estimated bivariate survival function using 10 bins for the ESBCR data

frailty model selection. If $\hat{\theta}(t_{1ij}, t_{2ij})$ increases as $\hat{S}(t_{1ij}, t_{2ij})$ increases, a positive stable model may be appropriate for the dependence structure.

In the ESBCR data, the time from the initial treatment for breast cancer to disease relapse is T_1 , and the time from the initial treatment to death is T_2 . To avoid unnecessary computational complexity, we used a subgroup of 350 patients for whom the observed times are available for both cancer relapse and subsequent death in order to count the concordances and discordances of all possible pairs. Based on this subgroup, we estimated the cross ratio within 10 bins and the corresponding bivariate survival function as well. As shown in Fig. 1, the estimated cross ratio increases with an increase in the estimated bivariate survival function, which suggests that a positive stable copula is appropriate for the underlying dependent structure between breast cancer relapse and death.

Note that the exploratory diagnosis is based on a subset of the sample with both non-terminal and terminal events observed, thus its effectiveness needs to be further evaluated. We realize that it is not feasible to estimate the cross ratio without the information on the joint survival function. Nevertheless, the exploratory diagnosis provides a tool for approximating the pattern of the dependent structure. Even though there are potential biases from using the subsample, it can still help us to distinguish the positive stable copula from the Clayton copula, where the latter assumes a constant correlation and has been used in many semi-competing risks data studies. Additional simulation studies have been conducted to show that the diagnostic method can sufficiently detect the association pattern, and are included in Sect. 5.

4 Inference on association parameter and marginal distribution

Without loss of generality, we develop simultaneous inference of the association parameter and marginal distribution of the non-terminal event under the positive stable copula, which is chosen by the visual diagnostic display of the ESBCR data. The inference procedure can be easily applied to other copulas. Under the positive stable copula, the bivariate survival function is

$$S(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2) = \exp \left(- \left[\{-\log S_1(t_1)\}^{\frac{1}{\alpha}} + \{-\log S_2(t_2)\}^{\frac{1}{\alpha}} \right]^\alpha \right). \tag{2}$$

Given the observed data, the likelihood function is

$$L(\alpha, S_1, S_2) = \prod_{i=1}^n f(X_i, Y_i)^{\delta_{1i}\delta_{2i}} \times \left\{ \frac{\partial S(X_i, Y_i)}{\partial X_i} \right\}^{\delta_{1i}(1-\delta_{2i})} \times \left\{ \frac{\partial S(X_i, Y_i)}{\partial Y_i} \right\}^{(1-\delta_{1i})\delta_{2i}} \times S(X_i, Y_i)^{(1-\delta_{1i})(1-\delta_{2i})}.$$

The corresponding log-likelihood function can be expressed as

$$l(\alpha, S_1, S_2) = \sum_{i=1}^n -\delta_{1i} \log S_1(X_i) - \delta_{2i} \log S_2(Y_i) + (\alpha^{-1} - 1) [\delta_{1i} \log \{-\log S_1(X_i)\} + \delta_{2i} \log \{-\log S_2(Y_i)\}] - \left[\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}} \right]^\alpha + \{\alpha(\delta_{1i} + \delta_{2i} - \delta_{1i}\delta_{2i}) - \delta_{1i} - \delta_{2i}\} \log \left[\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}} \right] + \delta_{1i}\delta_{2i} \log \left(\left[\{-\log S_1(X_i)\}^{\frac{1}{\alpha}} + \{-\log S_2(Y_i)\}^{\frac{1}{\alpha}} \right]^\alpha - 1 + \alpha^{-1} \right). \tag{3}$$

Maximizing a pseudo-likelihood by replacing S_1 and S_2 in (3) with their consistent estimators provides a consistent estimator of α (Shih and Louis 1995). To this end, one may employ the Kaplan–Meier estimators of S_1 and S_2 . However, with semi-competing risks data, the Kaplan–Meier estimator of S_1 does not generally converge to S_1 as $n \rightarrow \infty$, although S_2 can be consistently estimated by the Kaplan–Meier estimator. In fact, a consistent estimator of S_1 may not exist without an estimator of the association parameter α . Hereafter, we formulate an alternative pseudo-likelihood function of α using the unique feature of semi-competing risks data.

Under the copula model, the marginal survival function of $Z = \min(T_1, T_2)$ is

$$S_Z(t) = \Pr(T_1 > t, T_2 > t) = \exp \left(- \left[\{-\log S_1(t)\}^{\frac{1}{\alpha}} + \{-\log S_2(t)\}^{\frac{1}{\alpha}} \right]^\alpha \right). \tag{4}$$

Using Eq. (4), the marginal survival function of T_1 can be expressed as

$$S_1(t) = g\{S_2(t), S_Z(t), \alpha\} = \exp \left(- \left[\{-\log S_Z(t)\}^{\frac{1}{\alpha}} - \{-\log S_2(t)\}^{\frac{1}{\alpha}} \right]^\alpha \right), \tag{5}$$

where $g(u, v, \alpha) = \exp \left[- \left\{ (-\log v)^{\frac{1}{\alpha}} - (-\log u)^{\frac{1}{\alpha}} \right\}^\alpha \right]$. With this, the log-likelihood function of α becomes

$$l^g(\alpha, S_2, S_Z) = l\{\alpha, g(S_2, S_Z, \alpha), S_2\}. \tag{6}$$

Note that S_2 and S_Z can be consistently estimated by the Kaplan–Meier estimators \hat{S}_2 and \hat{S}_Z , respectively, because T_2 and Z are subject to independent censoring only by

C. Let $U_\alpha(\alpha, S_2, S_Z)$ be the score function of α , which is the derivative of (6) with respect to α . Then, estimator $\hat{\alpha}$ is the solution to the estimating equation

$$U_\alpha(\alpha, \hat{S}_2, \hat{S}_Z) = \frac{\partial l^g(\alpha, \hat{S}_2, \hat{S}_Z)}{\partial \alpha} = \sum_i \frac{\partial l_0^g\{\alpha, \hat{S}_2(Y_i), \hat{S}_Z(Z_i)\}}{\partial \alpha} = 0. \tag{7}$$

We prove that $\hat{\alpha}$ is consistent and asymptotically normally distributed under the following regularity conditions, C1 and C2.

- C1 The standard regularity conditions for the maximum likelihood estimator hold.
- C2 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

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

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

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

Theorem 1 summarizes these asymptotic properties.

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 $\rho^2 = (\rho_1^2 + \rho_2^2)/\rho_1^4$.*

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 is similar to that of Theorem 2 from Shih and Louis (1995). 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 $\hat{\alpha}$ 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 \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 simple bootstrap resampling procedure is valid for estimating the asymptotic variance of $\hat{\alpha}$.

For the marginal survival function S_1 , a natural estimator is obtained as $\hat{S}_1 = g\{\hat{S}_2, \hat{S}_Z, \hat{\alpha}\}$, where \hat{S}_2 and \hat{S}_Z are the Kaplan–Meier estimators for S_2 and S_Z , respectively, and $\hat{\alpha}$ is the solution to Eq. (7). 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 $t \in [0, \tau]$ (Flemming and Harrington 2005). Since function g has bounded derivatives, a continuous mapping theorem gives the uniform convergence of \hat{S}_1 to S_1 . The asymptotic properties of \hat{S}_1 are summarized in Theorem 2, and the proof is provided in the Appendix.

Theorem 2 *The estimator $\hat{S}_1(t)$ is consistent with the true survival function $S_1(t)$, and as $n \rightarrow \infty$, the process $n^{1/2}\{\hat{S}_1(t) - S_1(t)\}$ converges weakly to a zero-mean Gaussian process with covariance function $\left[\frac{\partial g\{S_2(t), S_Z(t), \alpha\}}{\partial \alpha} \right]^2 \rho^2 + \omega^2(t)$ for $t \in [0, \tau]$.*

Although the asymptotic variance of $\hat{S}_1(t)$ has a rather complicated form, the ordinary bootstrap method provides a direct and robust way to estimate the standard error. Thus, we adopt a nonparametric bootstrap resampling method to obtain the standard error of the estimated survival function in practice.

5 Simulation studies

We carried out simulation studies to evaluate the finite sample performances of the proposed estimators. We generate n pairs of (T_1, T_2) from the positive stable copula with a different strength of solicitation by choosing different values of parameter α . The marginal functions S_1 and S_2 are exponentials with parameter $(\lambda_1, \lambda_2) = (1, 1)$ or $(2, 1)$. We generate the independent censoring time C from $U(0, 4)$, giving 25% censoring of T_2 . For each set of (λ_1, λ_2) , each combination of $\alpha = 0.2, 0.4, 0.6, 0.8$ and $n = 250, 500$, we simulate 500 datasets. To improve the stability of the estimation procedure, we propose the following iteration algorithm.

Step 1: We set the initial estimator of the marginal survival function, $\hat{S}_1^{(0)}$, by the Kaplan–Meier estimator.

Table 1 Simulation results of the positive stable copula association parameter α

| | (λ_1, λ_2) | n | Cens(T_1) | Cens(T_2) | α | $\hat{\alpha}$ | SEE | SEB |
|--|--------------------------|-----|---------------|---------------|----------|----------------|-------|-------|
| Cens(T_1), censoring proportion for T_1 ; Cens(T_2), censoring proportion for T_2 ; SEE, empirical standard error; SEB, average bootstrap standard error | (1, 1) | 250 | 60% | 25% | 0.2 | 0.229 | 0.048 | 0.051 |
| | | | | | 0.4 | 0.422 | 0.041 | 0.048 |
| | | | | | 0.6 | 0.609 | 0.053 | 0.064 |
| | | | | | 0.8 | 0.809 | 0.064 | 0.065 |
| | 500 | 60% | 25% | 0.2 | 0.215 | 0.019 | 0.031 | |
| | | | | 0.4 | 0.411 | 0.028 | 0.033 | |
| | | | | 0.6 | 0.605 | 0.037 | 0.042 | |
| | | | | 0.8 | 0.804 | 0.041 | 0.045 | |
| | (2, 1) | 250 | 15% | 25% | 0.2 | 0.209 | 0.017 | 0.019 |
| | | | | | 0.4 | 0.406 | 0.033 | 0.033 |
| | | | | | 0.6 | 0.607 | 0.042 | 0.045 |
| | | | | | 0.8 | 0.805 | 0.052 | 0.051 |
| 500 | | 15% | 25% | 0.2 | 0.204 | 0.012 | 0.012 | |
| | | | | 0.4 | 0.403 | 0.022 | 0.022 | |
| | | | | 0.6 | 0.601 | 0.030 | 0.031 | |
| | | | | 0.8 | 0.802 | 0.035 | 0.036 | |

- Step 2: We obtain the Kaplan–Meier estimator \hat{S}_2 of the marginal survival function for the terminal event, and the Kaplan–Meier estimator \hat{S}_Z of S_Z .
- Step 3: We estimate the association parameter α based on the log-likelihood function (3) by replacing $\hat{S}_1^{(0)}$ and \hat{S}_2 , denoted as $\hat{\alpha}^{(0)}$.
- Step 4: Based on Eq. (5), we obtain an updated estimator $\hat{S}_1^{(1)}$ by plugging in \hat{S}_Z , \hat{S}_2 and $\hat{\alpha}^{(0)}$.
- Step 5: We can obtain an updated estimator $\hat{\alpha}^{(1)}$ using $\hat{S}_1^{(1)}$, \hat{S}_2 and the log-likelihood function (3).

The iteration between Steps 4 and 5 is continued until a specified criterion for convergence is achieved.

A simple nonparametric bootstrap procedure is used to obtain the standard errors for the estimators, based on 200 independent bootstrap samples. Table 1 presents simulation results of the association parameter α in different simulation scenarios. As expected, the empirical biases are very small and decrease as the sample size increases, and increase with increasing censoring rates. The bootstrap standard errors agree well with the empirical standard errors, implying that the inference procedure performs reasonably well. Figure 2 shows the estimated marginal survival function of the non-terminal event obtained in the simulations when $(\lambda_1, \lambda_2) = (1, 1)$. The dashed curve is the initial estimator $\hat{S}_1^{(0)}$, obtained by the Kaplan–Meier estimator. Seen from the plot, the initial estimator of S_1 departs far from the true curve, suggesting that ignorance of the informative censoring due to the terminal event may seriously overestimate the true curve. After the iterations, the proposed estimator \hat{S}_1 (dotted curve) is close to the true curve (solid curve) with indistinguishable empirical biases.

In addition, we conducted simulation studies to demonstrate that the exploratory diagnostic method using the subset of the sample with both non-terminal and terminal

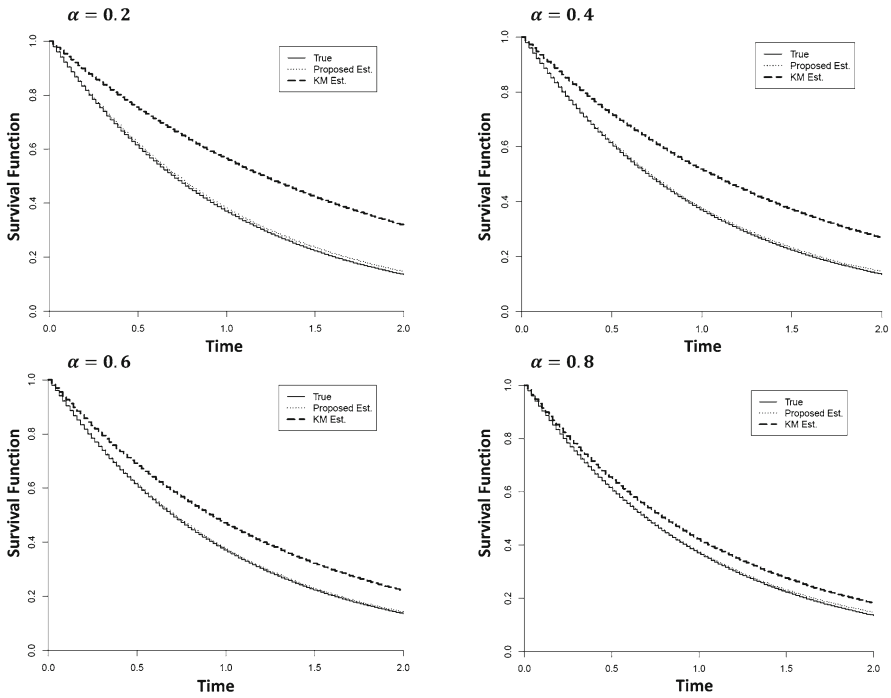


Fig. 2 Estimation of marginal survival function of the non-terminal event time in simulations. True margin S_1 , the proposed estimator \hat{S}_1 and the naive Kaplan–Meier estimator are plotted

events could effectively discriminate the association patterns between the positive stable copula and the Clayton copula. We generated $n = 250$ pairs of (T_1, T_2) from the positive stable copula/Clayton copula with different strength of association ($\alpha = 0.2, 0.4/\alpha = 3, 5$, respectively). The marginal functions S_1 and S_2 were exponential with parameter $(\lambda_1, \lambda_2) = (1, 1)$. We generated the independent censoring time C from $U(0,4)$, and simulated 250 datasets. Figure 3 shows the relationship between the estimated cross ratio curves and the estimated bivariate survival function in scenarios for the positive stable copula (A) and for the Clayton copula (B). The solid lines are for the true cross ratio curve. As expected, the diagnostic method could not accurately estimate the cross ratio, but it can sufficiently detect the association pattern. The estimated cross ratio increases with the increase in the estimated bivariate survival function in the positive stable copula but remains approximately constant in the Clayton copula. This suggests that our diagnostic approach performs well even when using the subsample.

6 Application to ESBCR data

As the most common malignant cancer among women in the United States, breast cancer has received tremendous research support, and as a result is associated with an excellent overall five-year survival rate. The majority of cases are diagnosed at the early

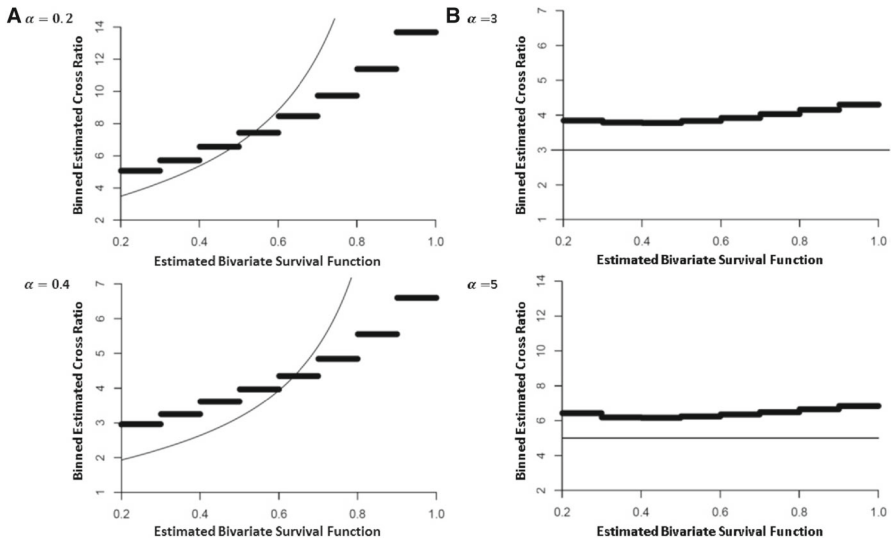


Fig. 3 Relationship between the estimated cross ratio and the estimated bivariate survival function using the upper wedge of the simulated semi-competing risks data with the positive stable copula **a** and the Clayton copula **b**

stage of disease, which is associated with 5-year relative survival rates ranging from 84 to 99 % (Siegel et al. 2014). Many of these female patients have the greatest risk of recurrence during the initial years after diagnosis (Brewster et al. 2007). Studying the survival probability of breast cancer relapse will help clinicians to make better treatment decisions and to appropriately manage the patient’s quality of life.

We applied the proposed estimation and inference procedure to the ESBCR data to evaluate the distribution of breast cancer relapse and its association with death. In ESBCR, a cohort of 2425 patients were identified after excluding patients for whom follow-up information was missing (Brewster et al. 2007). The follow-up time ranged from 0.02 to 22.5 years, with a median time of 9.35 years. Among the 2425 patients in the study, 350 experienced both breast cancer relapse and death during follow-up; 212 experienced only cancer relapse and remained alive throughout follow-up; 413 died without experiencing cancer relapse; and 1450 patients did not have either event observed. The censoring rate for breast cancer relapse was 76.8 % and that for death was 68.5 %.

The estimated association parameter $\hat{\alpha}$ in the positive stable copula model was 0.578, with a bootstrap standard error of 0.018 from 200 resamples. The value of $\hat{\alpha}$ suggests that there was a moderate positive association between breast cancer relapse and death in the ESBCR data. A further positive association measured by the cross ratio was not constant over time, but decreased with the time to breast cancer relapse and death. Figure 4 presents the estimated survival function \hat{S}_1 (solid curve) of the time to breast cancer relapse with 95 % point-wise confidence intervals (dotted curve), and the naive Kaplan–Meier estimator (dashed curve) considering death as independent censoring. The substantial difference between the solid and dashed curves suggests that methods which ignore the informative censoring will provide misleading results

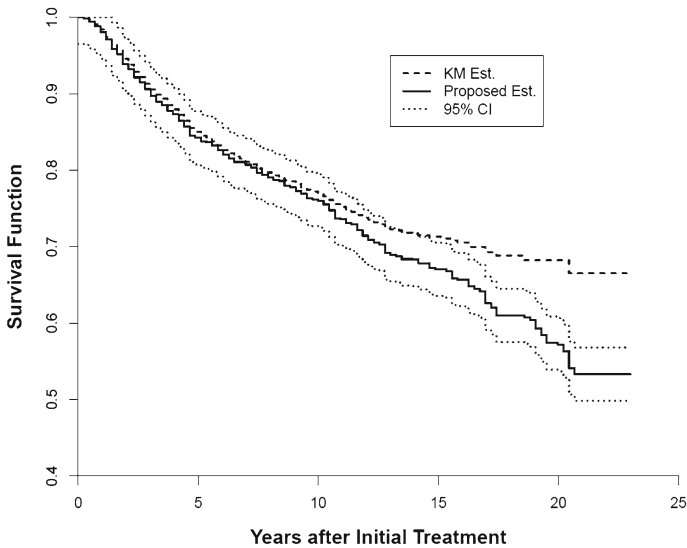


Fig. 4 Estimated marginal survival functions of the time to breast cancer relapse in the ESBCR data: Kaplan–Meier estimator and the proposed estimator \hat{S}_1 with 95 % point-wise confidence intervals

and overestimate the survival function of breast cancer relapse. For example, the estimated disease-free survival rate at 15 years obtained by the proposed joint model is 0.636, with 95 % confidence interval (0.670, 0.705), while that obtained by the naive Kaplan–Meier estimator is 0.711, with 95 % confidence interval (0.688, 0.734).

7 Concluding remarks

In this paper, we consider semi-competing risks data for which a non-terminal event may be dependently censored by a terminal event, but not vice versa. We propose simultaneous inference on the marginal survival function of the non-terminal event time and the association between the bivariate event times. To handle the informative censoring, a key component in the modeling strategies is to select a reasonable copula model for the dependence structure, which is implemented by an effective exploratory diagnostic approach. Although we focus on a specific class of positive stable copulas suggested by the model diagnostic, the proposed estimation methods are highly applicable to general Archimedean copula models.

Different to our proposed method, the illness-death model based on crude quantities provides an alternative tool for handling semi-competing risks data. For example, [Xu et al. \(2010\)](#) consider the illness-death model with a shared Gamma frailty. However, the strategy of selecting an appropriate shared frailty within the illness-death model framework for the semi-competing risks data is still under investigation. Following a reviewer’s suggestion, we also consider a Cox proportional hazards model of death by including the relapse status as a time-dependent covariate to analyze the ESBCR data. The estimated hazard ratio is 3.38 (95 % CI 2.45–4.70; P value < 0.001),

which suggests that breast cancer relapse significantly increases the risk of death and confirms our conclusion that there is a positive association between cancer relapse and death. However, the simple Cox model analysis could not describe the trend of the association over time. In contrast, our method provides a convenient estimation of the association parameter between cancer relapse and death as well as an estimation of the survival function of the relapse, once the joint survival function is established.

In addition to marginal survival functions and the association between event times, it is of interest in many studies to determine the impact of pretreatment characteristics or treatments on disease progression and survival. This can be addressed by incorporating covariates and formulating their effects on the survival functions of the event times via regression models. An interesting topic for future investigation is to extend the current copula model framework and develop estimation methods for time-varying effect regression models for semi-competing risks data.

Acknowledgments The authors thank the editor, the associate editor and two reviewers for their constructive comments that have greatly improved the initial version of this paper. This work was supported in part by Cancer Center Support Grants from the National Institutes of Health (CA142543 to Hong Zhu at UT Southwestern Medical Center and CA016672 to Jing Ning at UT MD Anderson Cancer Center) and by a predoctoral fellowship grant from the Cancer Prevention Research Institute of Texas (RP140103 to Renke Zhou).

8 Appendix

Proof of Theorem 2 Asymptotic results of $\hat{S}_1(t)$ are proved under the following regularity condition. Function $g(u, v, \alpha)$ is continuous and differentiable at u, v , and α , respectively, and the parameter α lies in a compact set.

First, we show the consistency of $\hat{S}_1(t)$. We have that $\hat{S}_2(t)$ converges in probability to $S_2(t)$ uniformly for $t \in [0, \tau]$, and $\hat{S}_Z(t)$ converges in probability to $S_Z(t)$ uniformly for $t \in [0, \tau]$. By Theorem 1, $\hat{\alpha}$ converges in probability to α_0 . Since the function $g(u, v, \alpha)$ is a continuous function of u, v and α , $g\{\hat{S}_2(t), \hat{S}_Z(t), \hat{\alpha}\}$ converges in probability to $g\{S_2(t), S_Z(t), \alpha_0\}$ uniformly for $t \in [0, \tau]$. Therefore, $\hat{S}_1(t)$ is a consistent estimator of $S_1(t)$.

Next, we illustrate the asymptotic distribution of $\hat{S}_1(t)$. Applying the functional delta method to $g\{\hat{S}_2(t), \hat{S}_Z(t), \hat{\alpha}\}$ around $S_2(t), S_Z(t)$ and α_0 , we have

$$\begin{aligned} & n^{1/2} [g\{\hat{S}_2(t), \hat{S}_Z(t), \hat{\alpha}\} - g\{S_2(t), S_Z(t), \alpha_0\}] \\ & \cong n^{1/2} \frac{\partial g\{S_2(t), S_Z(t), \alpha_0\}}{\partial u} (\hat{S}_2 - S_2)(t) \\ & \quad + n^{1/2} \frac{\partial g\{S_2(t), S_Z(t), \alpha_0\}}{\partial v} (\hat{S}_Z - S_Z)(t) \\ & \quad + n^{1/2} \frac{\partial g\{S_2(t), S_Z(t), \alpha_0\}}{\partial \alpha} (\hat{\alpha} - \alpha_0). \end{aligned} \tag{8}$$

Using martingale representations for \hat{S}_2 and \hat{S}_Z (Gill 1980), the sum of the first and second terms in (8) is asymptotically equivalent to

$$n^{-\frac{1}{2}} \left[\sum_{i=1}^n \frac{\partial g\{S_2(t), S_Z(t), \alpha_0\}}{\partial u} I_1^0(Y_i, \delta_{2i}) + \frac{\partial g\{S_2(t), S_Z(t), \alpha_0\}}{\partial v} I_2^0(Z_i, \delta_{zi}) \right], \tag{9}$$

which is a sum of n independent and identically distributed random variables, with $I_1^0(Y_i, \delta_{2i})$ and $I_2^0(Z_i, \delta_{zi})$ defined as in the previous section. Also the expectation of each term in (9) is zero. By the central limit theorem, (9) converges weakly to a normal distribution with mean zero and variance $\omega^2(t)$. By Theorem 1, $n^{1/2}(\hat{\alpha} - \alpha_0)$ converges weakly to a normal distribution with mean zero and variance ρ^2 . Therefore, the third term in (8) is asymptotically equivalent to

$$n^{-1/2} \frac{\partial g\{S_2(t), S_Z(t), \alpha\}}{\partial \alpha} \rho \sum_{i=1}^n \frac{\partial l_0^g\{\alpha, \hat{S}_2(Y_i), \hat{S}_Z(Z_i)\}}{\partial \alpha}, \tag{10}$$

which is a sum of n independent and identically distributed random variables.

Moreover, we have

$$\begin{aligned} & E \left(\left[\frac{\partial l_0^g\{\alpha, \hat{S}_2, \hat{S}_Z\}}{\partial \alpha} \right] \{I_1^0(Y_i, \delta_{2i}) + I_2^0(Z_i, \delta_{zi})\} \right) \\ &= E \left(\left\{ I_1^0(Y_i, \delta_{2i}) + I_2^0(Z_i, \delta_{zi}) \right\} E \left[\frac{\partial l_0^g\{\alpha, \hat{S}_2, \hat{S}_Z\}}{\partial \alpha} \mid Y_i, Z_i, \delta_{2i}, \delta_{zi} \right] \right) = 0, \end{aligned} \tag{11}$$

which means that (9) and (10) are asymptotically orthogonal. Therefore, (9), (10) and (11) imply that as $n \rightarrow \infty$, the process $n^{1/2}\{\hat{S}_1(t) - S_1(t)\}$ converges weakly to a zero-mean Gaussian process for $t \in [0, \tau]$ with covariance function $\left[\frac{\partial g\{S_2(t), S_Z(t), \alpha\}}{\partial \alpha} \right]^2 \rho^2 + \omega^2(t)$. □

References

Brewster AM, Do KA, Thompson PA, Hahn KM, Sahin AA, Cao Y, Stewart MM, Murray JL, Hortobagyi GN, Bondy ML (2007) Relationship between epidemiologic risk factors and breast cancer recurrence. *J Clin Oncol* 25:4438–4444

Chen MC, Bandeen-Roche K (2005) A diagnostic for association in bivariate survival models. *Lifetime Data Anal* 11:245–264

Chen YH (2012) Maximum likelihood analysis of semicompeting risks data with semiparametric regression models. *Lifetime Data Anal* 18:36–57

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65:141–151

Ding AA, Shi G, Wang W, Hsieh JJ (2009) Marginal regression analysis for semi-competing risks data under dependent censoring. *Scand J Stat* 36:481–500

Fine JP, Jiang H, Chappell RJ (2001) On semi-competing risks data. *Biometrika* 88:907–919

Fleming TR, Harrington DP (2005) Counting processes and survival analysis. Wiley, Chichester

Frank M (1979) On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$. *Aequ Math* 19:194–226

Gill RD (1980) Censoring and stochastic integrals. *Mathematisch Centrum*, Amsterdam

Gumbel EJ (1960) Bivariate exponential distributions. *J Am Stat Assoc* 55:698–707

Hsieh JJ, Wang W, Ding AA (2008) Regression analysis based on semicompeting risks data. *J R Stat Soc Series B Stat Methodol* 70:3–20

- Hsieh JJ, Huang YT (2012) Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime Data Anal* 103:302–320
- Jiang H, Fine JP, Kosorok MR, Chappell RJ (2005) Pseudo self-consistent estimation of a copula model with informative censoring. *Scand J Stat* 32:1–20
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53:457–481
- Oakes D (1982) A model for association in bivariate survival data. *J R Stat Soc B* 44:414–422
- Oakes D (1989) Bivariate survival models induced by frailties. *J Am Stat Assoc* 84:487–493
- Peng L, Fine JP (2007) Regression modeling of semicompeting risks data. *Biometrics* 63:96–108
- Peng L, Jiang H, Chappell RJ, Fine JP (2007) *Statistical advances in the biomedical sciences: clinical trials, epidemiology, survival analysis, and bioinformatics*. Wiley, Hoboken
- Shih JH, Louis TA (1995) Inferences on the association parameters in copula models for bivariate survival data. *Biometrics* 51:1384–1399
- Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. *CA Cancer J Clin* 64:9–29
- Wang W (2003) Estimating the association parameter for copula models under dependent censoring. *J R Stat Soc Series B Stat Methodol* 65:257–273
- Xu J, Kalbfleisch JD, Tai B (2010) Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* 66:716–725