HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Am Stat Assoc. 2023 ; 118(542): 1282–1294. doi:10.1080/01621459.2021.1990766.

Evaluating Association Between Two Event Times with Observations Subject to Informative Censoring

Dongdong Li[a], **X. Joan Hu**[b], **Rui Wang**[a,c]

[a]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA;

[b]Department of Statistics and Actuarial Science, Simon Fraser University, British Columbia, Canada;

[c]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

Abstract

This article is concerned with evaluating the association between two event times without specifying the joint distribution parametrically. This is particularly challenging when the observations on the event times are subject to informative censoring due to a terminating event such as death. There are few methods suitable for assessing covariate effects on association in this context. We link the joint distribution of the two event times and the informative censoring time using a nested copula function. We use flexible functional forms to specify the covariate effects on both the marginal and joint distributions. In a semiparametric model for the bivariate event time, we estimate simultaneously the association parameters, the marginal survival functions, and the covariate effects. A byproduct of the approach is a consistent estimator for the induced marginal survival function of each event time conditional on the covariates. We develop an easy-to-implement pseudolikelihood-based inference procedure, derive the asymptotic properties of the estimators, and conduct simulation studies to examine the finite-sample performance of the proposed approach. For illustration, we apply our method to analyze data from the breast cancer survivorship study that motivated this research. Supplementary materials for this article are available online.

Keywords

Association parameter; Copula model; Marginal distribution; Pseudolikelihood estimation; Semiparametric modeling

✉**CONTACT** Rui Wang rwang@hsph.harvard.edu Harvard Pilgrim Health Care Institute, Population Medicine, 401 Park Dr, Boston, MA 02215.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Supplementary Materials

The supplementary materials contain an appendix file which details the conditions and proofs for Propositions 1–4 in Section 2.3 and supporting information including additional technical details for Section 2, additional simulation results for Section 3, and additional analysis results for the cancer survivorship study in Section 4.

1. Introduction

It is often of interest to study the association between two event times. For example, a breast cancer survivorship project (Davis et al. 2014) in British Columbia (BC), referred to as the cancer survivorship study, aims to examine the association between the time to relapse/second cancer (RSC) and the time to cardiovascular disease (CVD) among breast cancer survivors and to evaluate the covariate effects on the association. The study observations on the times to RSC and CVD are censored by either the end of the study follow-up or death. Because the death time is likely related to both the RSC and CVD times, conventional methods, such as the Kaplan–Meier estimator for the survival function of an event time, are not directly applicable. We must account for the potential dependence between the two event times of interest and the informative censoring. In addition, it is often impractical to confidently specify a parametric model for the joint distribution of the two event times. Motivated by the cancer survivorship study, this article focuses on the estimation of the conditional joint distribution of bivariate event time using a copula model. We simultaneously estimate the covariate effects on both the association between the two event times and their marginal distributions in the presence of informative censoring.

The potential informative censoring of event times due to death may be framed as a semi-competing risk (Fine, Jiang, and Chappell 2001): observations on the event times of interest can be censored by death, but not vice versa. Various methods have been proposed to address this type of informative censoring in situations with a *univariate* event time. Zheng and Klein (1995) proposed a nonparametric estimator for the marginal distributions for a given copula function assuming that its association parameter is known. Fine, Jiang, and Chappell (2001) analyzed the semi-competing risk data using a Clayton copula. Their approach was later extended to the Archimedean copula family by Wang (2003). Chen (2012) further extended the copula models to handle semiparametric regression analysis using the transformation Cox model. In the setting of clustered semi-competing risk data, Emura et al. (2017) and Peng, Xiang, and Wang (2018) proposed joint copula–frailty modeling approaches where the joint distribution of the nonterminal and terminal events was modeled using a copula model, and the dependence within clusters was modeled by random effects. Emura and Chen (2018) reviewed the use of copula-based approaches to deal with the informative censoring for *univariate* event times. However, methods handling the informative censoring with multivariate event times are lacking. In the cancer survivorship study, both the time to RSC and the time to CVD are likely censored by death informatively. We must account for the informative censoring to conduct valid inference on the joint distribution of the two event times.

Three approaches are commonly used to model multivariate event times, namely the marginal, frailty, and copula approaches. The marginal approach (Wei, Lin, and Weissfeld 1989) considers the marginal hazards and accounts for the dependence through the use of a robust variance estimator. Because it does not explicitly model the association between event times, it is not applicable to our setting, where the goal is to assess the covariate effects on the association parameters in the joint distribution of event times. The frailty approach models dependence of the event times through the incorporation of an unknown frailty variable in the conditional hazard functions (see, e.g., Wienke 2010). The copula

model approach directly links the marginal distribution of each event time through a copula dependence parameter; see the introductions to copulas in Joe (1997) and Nelsen (2006). The frailty and copula models (Archimedean copula in particular) have some connections, for example, the functional form of the inverse of the generator function for an Archimedean copula is identical to the Laplace transform of the frailty density function for the corresponding frailty model. However, they differ in the parameterization of the marginal distributions and their dependence (Wienke 2010; Goethals, Janssen, and Duchateau 2008). In a copula model, the marginal distributions do not have overlapping parameters with the dependence component, whereas such overlapping parameters are present in a frailty model (Hougaard 1986; Goethals, Janssen, and Duchateau 2008; Prenen, Braekers, and Duchateau 2016). We adapt the copula approach to simultaneously model the covariate effects on the association between the two event times and on the marginal distribution of each event time.

This article develops a semiparametric approach to analyzing the bivariate event time in the presence of informative censoring due to a terminal event. The ultimate goal is to estimate the association between the two event times to characterize their joint distribution, and to evaluate the covariate effects on both the marginal and joint distributions. We formulate the joint distribution of the bivariate event time and the informative censoring time by a nested copula function, which embeds a copula model for the joint distribution of the two event times in another copula function incorporating the dependence of the bivariate event time with the censoring time. This allows the association between the two event times of interest and their dependence on the informative censoring time to be different. Our approach is more flexible than that of Lo and Wilke (2010) and Li et al. (2019), which assumes that the joint distribution of the bivariate event time and the censoring time follows a multivariate copula model. Furthermore, our approach permits the assessment of covariate effects on the dependence parameters as elaborated below.

Most of the existing approaches using copula models for time-to-event studies treat the copula dependence parameter as a single constant. Lo and Wilke (2010) used a multivariate Archimedean copula to model multiple competing event times and proposed a nonparametric estimator for the marginal survival functions for a given copula function with a known constant dependence parameter. Sun and Ding (2019) proposed the use of a two-parameter copula family to model an interval-censored bivariate event time, where both copula parameters were considered as scalars. In the nonsurvival context, Nikoloulopoulos and Karlis (2008) introduced a regression component for the copula parameter by specifying the parameter conditional on the covariates. In the cancer survivorship study, the association between times to RSC and times to CVD was likely to be influenced by individual characteristics. This motivated us to extend the constant dependence parameter to functions of covariates. Doing so permits the direct assessment of the covariate effects on the association parameters.

A two-stage estimation procedure has been widely employed for inference on the parameters in copula models (Shih and Louis 1995; Genest, Ghoudi, and Rivest 1995). In the first stage, the marginal distributions are estimated, and the estimated distributions are used to estimate the dependence parameters in the copula model in the second stage. Glidden and Self (1999) and Glidden (2000) extended the approach of Shih and Louis (1995) to allow the failure

times to marginally follow the Cox proportional hazards (PH) model and a stratified Cox PH model, respectively. In our setting, because of the informative censoring, available and well-established approaches with the Cox PH and the accelerated failure time (AFT) models, for example, are not directly applicable. We propose an easy-to-implement pseudolikelihood estimation procedure, which adapts the two-stage procedure to accommodate informative censoring. As a byproduct of our approach, we obtain a semiparametric consistent estimator for the conditional distribution of an event time with observations subject to informative censoring.

The main contributions of this article are twofold. First, it proposes a strategy to address the challenges in the analysis of bivariate event times with informative censoring due to a terminating event. This approach can be adapted to other types of informative censoring. Second, it allows conditional modeling on both the marginal and association parameters in the joint distribution of the bivariate event times. To the best of our knowledge, conditional copula dependence modeling has not been studied formally in the context of multivariate event times.

The rest of this article is organized as follows. Section 2 presents the proposed approach, including the notation, the model formulation, and the inference procedure. Section 3 reports simulation studies conducted to examine the finite-sample performance of the proposed approach in terms of consistency, efficiency, robustness, flexibility, and the evaluation of covariate effects. Section 4 presents an analysis of the motivating breast cancer data, and Section 5 provides some final remarks. Additional technical details, simulation results, and data analysis results are provided in the supplementary materials. The programming code to implement the proposed approach is provided in the GitHub repository (https://github.com/dli-stats/bvic).

2. Semiparametric Estimation Based on Bivariate Observations Subject to Informative Censoring

2.1. Notation

Let T_1 and T_2 be the two event times of interest and $\boldsymbol{Z}$ the covariate vector. We use $S_{12}(t_1, t_2|\boldsymbol{Z}) = \Pr(T_1 \geq t_1, T_2 \geq t_2|\boldsymbol{Z})$ to denote the joint survival function of T_1 and T_2 conditional on $\boldsymbol{Z}$. Suppose that the observations on T_1 and T_2 are subject to right-censoring where the censoring time C is the minimum of the time to a terminating event D and the administrative end of follow-up time C_A, that is, $C = D \wedge C_A$. Let $S_D(d|\boldsymbol{Z})$ be the conditional survival function of D. Note that both the T_1 and T_2 observations can be are censored by the occurrence of D but not vice versa. Adopting the conventional notation, let Δ_D be the indicator $I\{D \leq C_A\}$, and $U_j = T_j \wedge C$ with $\Delta_j = I\{T_j \leq C\}$ for $j = 1, 2$. Suppose that the study data are n independent realizations of, $[(U_1, \Delta_1), (U_2, \Delta_2), (C, \Delta_D), \boldsymbol{Z}]$ denoted by

$$\text{Observed-Data} = \{[(u_{1i}, \delta_{1i}), (u_{2i}, \delta_{2i}), (c_i, \delta_{D_i}), \boldsymbol{z}_i] : i = 1, \ldots, n\}. \quad (1)$$

We consider in this article the distributions of the event times T_j over the intervals $[0, v_j]$ with v_j chosen to be slightly smaller than $\max_i\{u_{ji}\}$ for $j = 1, 2$.

2.2. Modeling and Inference

2.2.1. Joint Survival Function and Likelihood—We assume that the administrative end of follow-up time C_A is independent of the event times T_1 and T_2, and the time to the terminating event D. To specify the correlation between (T_1, T_2) and D given $\boldsymbol{Z}$, we embed the conditional bivariate survival function of (T_1, T_2) in a bivariate Archimedean copula model (e.g., Joe 1997; Nelsen 2006). A bivariate Archimedean copula is defined as $\mathscr{A}_{[2]}(u_1, u_2; \theta) = \psi^{-1}(\psi(u_1; \theta) + \psi(u_2; \theta); \theta)$, where for a fixed θ, $\psi(\cdot; \theta)$: $[0, 1] \to [0, \infty]$ is a continuous, strictly decreasing convex function with $\psi(1; \theta) = 0$, and $\psi^{-1}(\cdot; \theta)$ is its inverse function. Here θ is a parameter within parameter space Θ, and ψ is the generator function of the copula $\mathscr{A}_{[2]}(\cdot)$. The popularity of Archimedean copulas in statistical modeling is in part due to their ability to accommodate various dependence structures (Nelsen 2006).

Assume that the joint survival function of (T_1, T_2) and D conditional on $\boldsymbol{Z}$ is

$$\Pr(T_1 \geq t_1, T_2 \geq t_2, D \geq d \mid \boldsymbol{Z}) = \mathscr{A}_{[2]}(S_{12}(t_1, t_2 \mid \boldsymbol{Z}), S_D(d \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})). \quad (2)$$

Equation (2) employs the bivariate Archimedean copula $\mathscr{A}_{[2]}(u, v; \theta(\boldsymbol{Z}))$ for a fixed $\boldsymbol{Z}$ with $u = S_{12}(t_1, t_2|\boldsymbol{Z})$ and $v = S_D(d|\boldsymbol{Z})$. Since both $S_{12}(t_1, t_2|\boldsymbol{Z})$ and $S_D(d|\boldsymbol{Z})$ are survival functions, we have $0 \leq u \leq 1$ and $0 \leq v \leq 1$ for $t_1, t_2, d \geq 0$. Thus, Equation (2) presents a well-defined function of t_1, t_2, d, denoted by $S(t_1, t_2, d|\boldsymbol{Z})$. It can be verified that $S(t_1, t_2, d|\boldsymbol{Z})$ is a trivariate survival function (see supplementary material, Section S1.1). The copula function $A_{[2]}(\cdot)$ in Equation (2) is assumed to be invariant to the covariates $\boldsymbol{Z}$ in this article. However, it can be extended to be covariate-dependent as discussed in Section 5.

The association parameter $\theta(\boldsymbol{Z})$ characterizes the correlation between (T_1, T_2) and D conditional on $\boldsymbol{Z}$. The domain of $\theta(\boldsymbol{Z})$ depends on the corresponding copula family. For example, two widely used Archimedean copula families and their parameter spaces Θ are:

- Clayton copula: $\mathscr{A}_{[2]}(u, v, \theta) = \max\left(\left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}, 0\right)$, $\Theta = [-1, \infty)\backslash\{0\}$;
- Frank copula: $\mathscr{A}_{[2]}(u, v, \theta) = \frac{-1}{\theta}\log\left(1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right)$, $\Theta = \mathbb{R}\backslash\{0\}$.

The Kendall rank correlation coefficient (Kendall's tau) is a widely used scale-invariant measure of the correlation between variables and the metric is used in the presentation of the numerical results in Sections 3 and 4. Kendall's tau lies in [−1, 1], where the value 1 corresponds to perfect concordance and −1 corresponds to complete discordance. The correspondence between the association parameter θ in a bivariate Archimedean copula and Kendall's tau is $\tau = 4\int_0^1\int_0^1 \mathscr{A}_{[2]}(w_1, w_2; \theta)\mathscr{A}_{[2]}(dw_1, dw_2; \theta) - 1$.

Note that $S_{12}(t, 0|\boldsymbol{Z}) = \Pr(T_1 \geq t|\boldsymbol{Z})$ and $S_{12}(0, t|\boldsymbol{Z}) = \Pr(T_2 \geq t|\boldsymbol{Z})$ are the marginal survival functions of T_1 and T_2 conditional on $\boldsymbol{Z}$, respectively. Denote $S_j(t|\boldsymbol{Z}) = \Pr(T_j \geq t|\boldsymbol{Z})$ for $j = 1, 2$. The model in Equation (2) induces the joint conditional model of T_j and D:

$$\Pr(T_j \geq t_j, D \geq d \mid \boldsymbol{Z}) = \mathscr{A}_{[2]}(S_j(t_j \mid \boldsymbol{Z}), S_D(d \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})). \quad (3)$$

We further assume that the conditional joint distribution of (T_1, T_2) is

$$S_{12}(t_1, t_2 \mid \boldsymbol{Z}) = \mathscr{C}_{[2]}(S_1(t_1 \mid \boldsymbol{Z}), S_2(t_2 \mid \boldsymbol{Z}); \theta_{12}(\boldsymbol{Z})), \quad t_1, t_2 > 0, \tag{4}$$

where $\mathscr{C}_{[2]}(\cdot\,; \theta_{12}(\boldsymbol{Z}))$ is a bivariate function with a dependence parameter $\theta_{12}(\boldsymbol{Z})$. We note that $\mathscr{C}_{[2]}(\cdot)$ can be an Archimedean copula or a non-Archimedean copula such as a Gaussian copula. Because $S_1(\cdot|\boldsymbol{Z})$ and $S_2(\cdot|\boldsymbol{Z})$ are survival functions for a fixed $\boldsymbol{Z}$ and take values between 0 and 1, Equation (4) is well-defined. The domain of $\theta_{12}(\cdot)$ is specific to the choice of copula family $\mathscr{C}_{[2]}(\cdot)$. Let $\dot{\lambda}(r) = d\lambda(r)/dr$ for function $\lambda(r)$, and $\lambda^{(a_1, a_2)}(r_1, r_2; \phi) = \partial \lambda^{(a_1 + a_2)} / \partial r_1^{a_1} \partial r_2^{a_2}(r_1, r_2; \phi)$ for function $\lambda(r_1, r_2; \phi)$ with well-defined partial derivatives. The likelihood function based on the observed data in Equation (1) is

$$\begin{aligned} & L(S_1(\cdot \mid \cdot), S_2(\cdot \mid \cdot), S_D(\cdot \mid \cdot), \theta(\cdot), \theta_{12}(\cdot) \mid \text{Observed-Data}) \\ &= \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \dot{S}_D(c_i \mid \boldsymbol{z}_i)^{\delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathscr{A}_{[2]}^{(0, \delta_{Di})}}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} (S_{12}(u_{1i}, u_{2i} \mid \boldsymbol{z}_i), S_D(c_i \mid \boldsymbol{z}_i); \theta(\boldsymbol{z}_i)) \right\} \\ &= \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \dot{S}_D(c_i \mid \boldsymbol{z}_i)^{\delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathscr{A}_{[2]}^{(0, \delta_{Di})}}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \right. \\ & \left. \times (\mathscr{C}_{[2]}(S_1(u_{1i} \mid \boldsymbol{z}_i), S_2(u_{2i} \mid \boldsymbol{z}_i); \theta_{12}(\boldsymbol{z}_i)), S_D(c_i \mid \boldsymbol{z}_i); \theta(\boldsymbol{z}_i)) \right\}, \end{aligned} \tag{5}$$

where $\delta_{\cdot i} = \delta_{1i} + \delta_{2i}$. Directly maximizing (5) requires intensive computing because of the five unknown functions, namely $S_1(\cdot|\cdot)$, $S_2(\cdot|\cdot)$, $S_D(\cdot|\cdot)$, $\theta(\cdot)$, and $\theta_{12}(\cdot)$. Following the idea of the pseudolikelihood estimation procedure under a copula model (e.g., Lawless and Yilmaz 2011a; Li et al. 2019), we perform the estimation in two stages. In the first stage, we obtain consistent estimators of $S_j(t|\boldsymbol{Z})$, $j = 1, 2$, and $S_D(t|\boldsymbol{Z})$. In the second stage, we substitute them into the likelihood function to obtain the pseudo-MLE for the other parameters. We describe the estimation procedure in detail in the next subsections.

2.2.2. Estimation of Marginal Survival Functions $S_D(t|\boldsymbol{Z})$ and $S_j(t|\boldsymbol{Z})$ for $j = 1, 2$—

Since observations on the terminating event D are subject to noninformative censoring by C_A, well-established estimation procedures can be used to estimate $S_D(\cdot|\boldsymbol{Z})$ with the Cox PH model or the AFT model. In our simulation studies and the data application, we consider the Cox PH model,

$$S_D(t \mid \boldsymbol{Z}) = \exp\left\{ -H_{0D}(t) e^{\boldsymbol{\beta}_D' \boldsymbol{Z}} \right\}, \tag{6}$$

and estimate $\boldsymbol{\beta}_D$ by the partial-likelihood-based procedure and $H_{0D}(t)$ by the Breslow estimator (Cox 1972; Breslow 1972). Denote the estimated conditional survival function by $\tilde{S}_D(t \mid \boldsymbol{Z})$. In the absence of covariates $\boldsymbol{Z}$, we consider consistent estimators such as the Kaplan–Meier estimator to estimate the marginal survival function $S_D(t)$.

To estimate $S_j(t|\boldsymbol{Z})$ with the available data, we must account for the informative censoring due to the terminating event D. When the copula function $\mathscr{A}_{[2]}(\cdot\,; \theta(\boldsymbol{Z}))$ in Equation (2) is an Archimedean copula with a generator function $\psi(\cdot; \theta(\boldsymbol{Z}))$, the induced model (3) for the joint survival function of T_j and D is $\Pr(T_j \geq t, D \geq t|\boldsymbol{Z}) = \psi^{-1}(\psi(S_j(t|\boldsymbol{Z}); \theta(\boldsymbol{Z})) + \psi(S_D(t|\boldsymbol{Z}); \theta(\boldsymbol{Z})); \theta(\boldsymbol{Z}))$ by the definition of the Archimedean

copula. Let $T_j^* = T_j \wedge D$, the minimum of T_j and D, with conditional survival function $S_j^*(t \mid \boldsymbol{Z}) = \Pr(T_j^* \geq t \mid \boldsymbol{Z}) = \Pr(T_j \geq t, D \geq t \mid \boldsymbol{Z})$. Applying the ψ function to both sides of the equation above yields $\psi\{S_i^*(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})\} = \psi(S_j(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})) + \psi(S_D(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z}))$. Thus,

$$S_j(t \mid \boldsymbol{Z}) = \psi^{-1}\{\psi\{S_j^*(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})\} - \psi(S_D(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})); \theta(\boldsymbol{Z})\}, \tag{7}$$

denoted by $g\{S_j^*(t \mid \boldsymbol{Z}), S_D(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z})\}$ with $g(u_1, u_2; \theta) = \psi^{-1}(\psi(u_1;\theta) - \psi(u_2;\theta);\theta)$. Similarly to D, the observations on T_j^* are censored only by C_A, the noninformative administrative censoring time. Assuming

$$S_j^*(t \mid \boldsymbol{Z}) = \exp\left\{-H_{0j}^*(t)e^{\boldsymbol{\beta}_j^{*\prime}\boldsymbol{Z}}\right\}, \tag{8}$$

$\boldsymbol{\beta}_j^*$ and $H_{0j}^*(t)$ can be estimated through the conventional approach for noninformatively censored data. Denote the estimated conditional survival function by $\widetilde{S}_j^*(t \mid \boldsymbol{Z}), j = 1, 2$.

2.2.3. Estimation of Association Functions $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$—

Our modeling allows flexibility in specifying the association with parameters $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$ in either parametric or semiparametric forms.

Parametric specification.: Both $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$ can be specified up to a finite number of parameters (e.g., Nikoloulopoulos and Karlis 2008). We give examples below:

1. In conjunction with the Clayton copula, consider

$$\begin{aligned} \log(\theta(\boldsymbol{Z};\boldsymbol{\gamma}) + 1) &= \boldsymbol{\gamma}'(1, \boldsymbol{Z}')', \\ \log(\theta_{12}(\boldsymbol{Z};\boldsymbol{\beta}) + 1) &= \boldsymbol{\beta}'(1, \boldsymbol{Z}')'. \end{aligned} \tag{9}$$

   The $log(\cdot)$ link function is adopted because the parameter space Θ for θ and θ_{12} in the Clayton copula is $[-1, \infty)\backslash\{0\}$.

2. In conjunction with the Frank copula, consider

$$\begin{aligned} \theta(\boldsymbol{Z};\boldsymbol{\gamma}) &= \boldsymbol{\gamma}'(1, \boldsymbol{Z}')', \\ \theta_{12}(\boldsymbol{Z};\boldsymbol{\beta}) &= \boldsymbol{\beta}'(1, \boldsymbol{Z}')'. \end{aligned} \tag{10}$$

Here $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in Equations (9) and (10) are the finite-dimensional parameters to be estimated.

Semiparametric specification.: Suppose the covariate vector $\boldsymbol{Z} = (\boldsymbol{Z}^{*\prime}, W)'$, where $\boldsymbol{Z}^*$ is a d_1-dimensional vector of covariates, and W is a continuous covariate for which the effect is specified as an unknown function. We provide examples below.

1. In conjunction with the Clayton copula, consider

$$\begin{aligned} \log(\theta(\boldsymbol{Z}) + 1) &= \boldsymbol{\gamma}'(1, \boldsymbol{Z}^{*\prime})' + f(W), \\ \log(\theta_{12}(\boldsymbol{Z}) + 1) &= \boldsymbol{\beta}'(1, \boldsymbol{Z}^{*\prime})' + h(W). \end{aligned} \tag{11}$$

2. In conjunction with the Frank copula, consider

$$\theta(\boldsymbol{Z}) = \boldsymbol{\gamma}'(1, \boldsymbol{Z}^{*\prime})' + f(W),$$
$$\theta_{12}(\boldsymbol{Z}) = \boldsymbol{\beta}'(1, \boldsymbol{Z}^{*\prime})' + h(W). \quad (12)$$

The proposed model specifications for $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$ in Equations (11) and (12) assume additive effects of the covariate vector $\boldsymbol{Z}^*$ and a continuous covariate W. Alternatively, if $\boldsymbol{Z}^*$ is a vector of categorical covariates, we can consider a stratified model and approximate the effect of W separately for different levels of the categorical covariates. This gives more flexibility by allowing different spline specifications for different levels of the categorical covariates.

Here, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{d_1})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{d_1})'$ are $(d_1 + 1)$-dimensional regression parameters belonging to parameter spaces A_1 and A_2, respectively, where A_1 and A_2 are compact sets in $\mathbb{R}^{d_1+1}$. Both $f(\cdot)$ and $h(\cdot)$ are unspecified smooth functions belonging to space $\mathscr{M}$, which is a collection of real-valued functions defined on $[a, b]$ with bounded and continuous first- and second-order partial derivatives, and a and b are, respectively, the lower and upper bounds of the observation on the covariate W. For the ease of notation, we use f and h to denote $f(\cdot)$ and $h(\cdot)$, respectively. We consider approximating f and h using B-splines. Both f and h can be specified up to a set of finite-dimensional parameters (i.e., we fix the number and location of the knots of the splines), or we could use a sieve approach (e.g., Zhang, Hua, and Huang 2010; He, Xue, and Shi 2010; Lu, Zhang, and Huang 2007; Wellner and Zhang 2007; Xue, Lam, and Li 2004) that allows the number of knots and basis functions to increase with sample size.

For the sieve approach, we define the sieve space as follows. Following Zhang, Hua, and Huang (2010), we let $a = e_0 < e_1 < \cdots < e_{K_n+1} = b$ partition $[a, b]$ into $K \equiv (K_n + 1)$ intervals $I_{K_t} = [e_t, e_{t+1}], t = 0, \ldots, K_n$, where K_n is an integer that grows at a rate of n^{ν}, for $0 < \nu < 1$, and $\max_{1 \le k \le K_n+1} |e_k - e_{k-1}| = \mathcal{O}(n^{-\upsilon})$. Denote the set of partition points as $E_n = \{e_1, \ldots, e_{K_n}\}$. Let $\mathcal{S}_n(E_n, K_n, m)$ be the space of polynomial splines of order $m \ge 1$ consisting of functions satisfying the conditions given in (Zhang, Hua, and Huang 2010, p. 341). Let $\mathcal{B}_n = \{\boldsymbol{b}_j, 1 \le j \le q_n\}$ denote the B-spline for $\mathcal{S}_n(E_n, K_n, m)$, where $q_n = K_n + m$. We approximate $f(w)$ and $h(w)$ in the form $\sum_{j=1}^{q_n} \upsilon_j \boldsymbol{b}_j(w)$, and define

$$\mathscr{M}_n(E_n, K_n, m) = \left\{ m_n : m_n(w) = \sum_{j=1}^{q_n} \upsilon_j \boldsymbol{b}_j(w), \boldsymbol{\upsilon} \in B_n, w \in [a, b] \right\}$$

as the sieve space for $\mathscr{M}$, where $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_{q_n})'$ is the vector of spline coefficients, and $B_n \subseteq \mathbb{R}q_n$ is a feasible domain of the spline coefficients. We abbreviate $\mathscr{M}_n(E_n, K_n, m)$ as $\mathscr{M}_n$, let $f_n(\cdot) \in \mathscr{M}_n$ and $h_n(\cdot) \in \mathscr{M}_n$ be the spline approximations to f and h, and let $\boldsymbol{\kappa} \in B_n$ and $\boldsymbol{\pi} \in B_n$ be the spline coefficients for $f_n(\cdot)$ and $h_n(\cdot)$, respectively. Assuming that we are given the true marginals of $S_j^*(\cdot)$ and $S_D(\cdot)$, the remaining estimands in the likelihood are $(\theta(\cdot), \theta_{12}(\cdot))$, that is, we need to estimate the parameter $\boldsymbol{\zeta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, f, h)$ over the parameter space $\mathscr{H} = A_1 \times A_2 \times \mathscr{M} \times \mathscr{M}$. We look for $\hat{\boldsymbol{\zeta}}_n = \left(\hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\beta}}_n, \hat{f}_n, \hat{h}_n\right) \in \mathscr{H}_n = A_1 \times A_2 \times \mathscr{M}_n \times \mathscr{M}_n$ that

maximizes the likelihood, where $\mathscr{H}_n$ is a sieve space of $\mathscr{H}$. This is equivalent to finding ($\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, $\boldsymbol{\pi}$) that maximizes the likelihood over the parameter space $\mathscr{H}_n^- = A_1 \times A_2 \times B_n \times B_n$.

Without loss of generality, we use $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{12}$ to denote the parameters in the assumed models for $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$, respectively. Under the parametric specification of Equation (9) or (10), $\boldsymbol{\alpha} = \boldsymbol{\gamma}$ and $\boldsymbol{\alpha}_{12} = \boldsymbol{\beta}$; under the semiparametric specification of Equation (11) or (12), $\boldsymbol{\alpha} = (\boldsymbol{\gamma}, \boldsymbol{\kappa})$ and $\boldsymbol{\alpha}_{12} = (\boldsymbol{\beta}, \boldsymbol{\pi})$. In the absence of covariates $\boldsymbol{Z}$, the association functions $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$ reduce to two constant parameters θ and θ_{12}, and the proposed inference procedure is still applicable.

2.2.4. Pseudolikelihood Function and Two-Stage Estimation—As given in Equation (7), the marginal survival function $S_j(t|\boldsymbol{Z})$ under model (3) is a known function of the conditional survival functions $S_j^*(t \mid \boldsymbol{Z})$ and $S_D(t|\boldsymbol{Z})$ together with $\theta(\boldsymbol{Z})$. When $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$ are specified up to finite-dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{12}$ as described in Section 2.2.3, Equation (5) is equivalent to

$$\begin{aligned} &L\left(S_1^*(\cdot \mid \cdot), S_2^*(\cdot \mid \cdot), S_D(\cdot \mid \cdot), \boldsymbol{\alpha}, \boldsymbol{\alpha}_{12} \mid \text{Observed-Data}\right) \\ &= \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \dot{S}_D(c_i \mid \boldsymbol{z}_i)^{\delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathscr{A}_{[2]}^{(0,\delta_{Di})}}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \right. \\ &\left. \left(\mathscr{C}_{[2]}(S_1(u_{1i} \mid \boldsymbol{z}_i), S_2(u_{2i} \mid \boldsymbol{z}_i); \theta_{12}(\boldsymbol{z}_i; \boldsymbol{\alpha}_{12})), S_D(c_i \mid \boldsymbol{z}_i); \theta(\boldsymbol{z}_i; \boldsymbol{\alpha})\right) \right\} \end{aligned} \tag{13}$$

We propose a pseudo-MLE of ($\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_{12}$), the maximizer of Equation (13) with $\tilde{S}_D(t \mid \boldsymbol{Z})$ and $\tilde{S}_j^*(t \mid \boldsymbol{Z})$ from Section 2.2.2 plugged in. Specifically, with $\tilde{S}_j(t \mid \boldsymbol{Z}) = g\left(\tilde{S}_j^*(t \mid \boldsymbol{Z}), \tilde{S}_D(t \mid \boldsymbol{Z}); \theta(\boldsymbol{Z}; \boldsymbol{\alpha})\right)$ for $j = 1, 2$, the pseudolikelihood function is proportional to

$$L\left(\tilde{S}_1^*(\cdot \mid \cdot), \tilde{S}_2^*(\cdot \mid \cdot), \tilde{S}_D(\cdot \mid \cdot), \boldsymbol{\alpha}, \boldsymbol{\alpha}_{12} \mid \text{Observed-Data}\right) = \prod_{i=1}^{n} \Psi_i(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12}), \tag{14}$$

where $\Psi_i(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12})$ is given by

$$(-1)^{\delta_{\cdot i} + \delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathscr{A}_{[2]}^{(0,\delta_{Di})}}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \left(\mathscr{C}_{[2]}\left(\tilde{S}_1(u_{1i} \mid \boldsymbol{z}_i), \tilde{S}_2(u_{2i} \mid \boldsymbol{z}_i); \theta_{12}(\boldsymbol{z}_i; \boldsymbol{\alpha}_{12})\right), \tilde{S}_D(c_i \mid \boldsymbol{z}_i); \theta(\boldsymbol{z}_i; \boldsymbol{\alpha})\right),$$

for $i = 1, \ldots, n$. The corresponding partial derivatives in $\Psi_i(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12})$ are provided in Section S1.3 of the supplementary materials.

It is much easier to implement this pseudolikelihood estimation procedure than to directly maximize the likelihood (5) to obtain the MLE. There may be some expected loss of efficiency compared to the setting where $S_j(\cdot|\boldsymbol{Z})$ and $S_D(\cdot|\boldsymbol{Z})$ are known. Our simulation study in Section 3 indicates that the efficiency loss is acceptable in practice.

Given consistent estimators for $S_j^*(\cdot \mid \boldsymbol{Z})$ and $S_D(\cdot|\boldsymbol{Z})$, we maximize the pseudolikelihood in (14), or, equivalently, its log-transformation with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{12}$, and thus derive a pseudo-MLE:

$$(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\alpha}}_{12n}) = \text{argmax}_{(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12})} L\Big(\widetilde{S}_1(\cdot \mid \cdot), \widetilde{S}_2(\cdot \mid \cdot), \widetilde{S}_D(\cdot \mid \cdot), \boldsymbol{\alpha}, \boldsymbol{\alpha}_{12} \mid \text{Observed-Data}\Big). \tag{15}$$

An iterative algorithm to calculate $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\alpha}}_{12n})$ is detailed in Section S1.4 of the supplementary materials.

This yields $\hat{\theta}_n(\boldsymbol{Z}) = \theta(\boldsymbol{Z}; \widehat{\boldsymbol{\alpha}}_n)$ and $\hat{\theta}_{12n}(\boldsymbol{Z}) = \theta_{12}(\boldsymbol{Z}; \widehat{\boldsymbol{\alpha}}_{12n})$ in either the parametric or semiparametric formulation of the association parameters $\theta(\boldsymbol{Z})$ and $\theta_{12}(\boldsymbol{Z})$. Substituting $\hat{\theta}_n(\boldsymbol{Z})$, $\widetilde{S}_j^*(t \mid \boldsymbol{Z})$, and $\widetilde{S}_D(t \mid \boldsymbol{Z})$ into Equation (7) gives a natural estimator for the marginal survival function $S_j(\cdot|\boldsymbol{Z})$:

$$\widehat{S}_{jn}(t \mid \boldsymbol{Z}) = g\Big(\widetilde{S}_j^*(t \mid \boldsymbol{Z}), \widetilde{S}_D(t \mid \boldsymbol{Z}); \hat{\theta}_n(\boldsymbol{Z})\Big). \tag{16}$$

Further, an estimator for the joint survival function $S_{12}(t_1, t_2|\boldsymbol{Z})$ of (T_1, T_2) based on Equation (4) is

$$\widehat{S}_{12n}(t_1, t_2 \mid \boldsymbol{Z}) = \mathscr{C}_{[2]}\Big(\widehat{S}_{1n}(t_1 \mid \boldsymbol{Z}), \widehat{S}_{2n}(t_2 \mid \boldsymbol{Z}); \hat{\theta}_{12n}(\boldsymbol{Z})\Big). \tag{17}$$

2.3. Asymptotic Properties

The following propositions establish the consistency and asymptotic normality of the proposed estimator. The conditions and proofs of these propositions are detailed in the supplementary materials (Parts A–C of the Appendix). The proofs of Propositions 1–3 follow the arguments in Li et al. (2019) for deriving the asymptotic properties of their estimators in situations without the covariates $\boldsymbol{Z}$. Proposition 4 presents asymptotic properties for the resulting sieve estimators when $\theta(\cdot)$ and $\theta_{12}(\cdot)$ are specified semiparametrically, such as by Equations (11) and (12), so that the effect of a continuous covariate is approximated by B-splines and the number of bases increases with sample size. We provide its proof adapting the derivations in Zhang, Hua, and Huang (2010), Xue, Lam, and Li (2004), and Chen, Fan, Tsyrennikov (2006).

Proposition 1.

Under the regularity conditions (RC1)–(RC4) in the Appendix, and provided that $\widetilde{S}_j^*(t \mid \boldsymbol{z})$ and $\widetilde{S}_D(t \mid \boldsymbol{z})$ satisfy the condition (AC) in the appendix, $\Big(\hat{\theta}_n(\boldsymbol{z}), \hat{\theta}_{12n}(\boldsymbol{z})\Big) \xrightarrow{\text{a.s.}} (\theta(\boldsymbol{z}), \theta_{12}(\boldsymbol{z}))$ and $\sqrt{n}\Big\{\Big(\hat{\theta}_n(\boldsymbol{z}), \hat{\theta}_{12n}(\boldsymbol{z})\Big)' - (\theta(\boldsymbol{z}), \theta_{12}(\boldsymbol{z}))'\Big\} \xrightarrow{d} N(0, AV_{\theta, \theta_{12}}(\boldsymbol{z}))$ as $n \to \infty$, for a fixed $\boldsymbol{z}$,.

A natural variance estimator of $AV_{\theta, \theta_{12}}(z)$ is presented in the Appendix. It uses Huber's robust sandwich estimator (Huber 1967) for the finite-dimensional parameter estimators $\widehat{\boldsymbol{\alpha}}_n$ and $\widehat{\boldsymbol{\alpha}}_{12n}$. Similarly to Lawless and Yilmaz (2011a), we can also estimate the variance of $\Big(\hat{\theta}_n(\boldsymbol{z}), \hat{\theta}_{12n}(\boldsymbol{z})\Big)'$

via a bootstrap approach. The variance estimator can be used to construct a Wald-type $(1 - \alpha) \times 100\%$ confidence interval in the standard fashion or assess the independence through testing $H_0 : \theta(\boldsymbol{z}) = 0$ or $\theta_{12}(\boldsymbol{z}) = 0$ by a Wald test.

Proposition 2.

Under the regularity conditions (RC1)–(RC4) and provided that $\tilde{S}_j^*(\cdot \mid \boldsymbol{z})$ and $\tilde{S}_D(\cdot \mid \boldsymbol{z})$ satisfy the condition (AC), $\hat{S}_{jn}(t \mid \boldsymbol{z}) \xrightarrow{\text{a.s.}} S_j(t \mid \boldsymbol{z})$ uniformly, and $\sqrt{n}\left(\hat{S}_{jn}(t \mid \boldsymbol{z}) - S_j(t \mid \boldsymbol{z})\right) \xrightarrow{w} \mathcal{G}_j(t \mid \boldsymbol{z})$ with $t \in [0, \nu_j]$ for any $\boldsymbol{z}$ in the region of the covariates $\boldsymbol{Z}$ as $n \to \infty$. Here $\mathcal{G}_j(t \mid \boldsymbol{z})$ is a Gaussian process with mean zero and variance function $\sigma_j^2(t \mid \boldsymbol{z})$ as defined in (Andersen et al. 1993, p. 506).

Proposition 3.

Under the regularity conditions (RC1)–(RC4) and provided that $\tilde{S}_j^*(t \mid \boldsymbol{z})$ and $\tilde{S}_D(t \mid \boldsymbol{z})$ satisfy the condition (AC), $\hat{S}_n(t_1, t_2 \mid \boldsymbol{z}) \xrightarrow{\text{a.s.}} S(t_1, t_2 \mid \boldsymbol{z})$ uniformly, and $\sqrt{n}\left(\hat{S}_n(t_1, t_2 \mid \boldsymbol{z}) - S(t_1, t_2 \mid \boldsymbol{z})\right) \xrightarrow{w} \mathcal{G}(t_1, t_2 \mid \boldsymbol{z})$ with $t_1, t_2 \in [0, \nu_1] \times [0, \nu_2]$ for any $\boldsymbol{z}$ in the region of the covariates $\boldsymbol{Z}$ as $n \to \infty$. Here $\mathcal{G}(t_1, t_2 \mid \boldsymbol{z})$ is a Gaussian field with mean zero and variance function $\sigma^2(t_1, t_2 | \boldsymbol{z})$.

When $\theta(\cdot)$ and $\theta_{12}(\cdot)$ are specified semiparametrically and the effect of covariate W is approximated by B-splines (De Boor 1978, p. 145) using a sieve approach as described in Section 2.2.3, we establish the asymptotic properties as follows. Following the semiparametric model specification as given for Equation (11) or (12), first we define an L_2 metric on the parameter space $\mathcal{H}$ defined in Section 2.2.3. Let $\|\mathbf{a}\|$ be the Euclidean norm of a vector $\mathbf{a}$, and for a random vector $\boldsymbol{X} \sim \mathcal{P}$ where $\mathcal{P}$ is a probability measure, and $\left\|\lambda(\boldsymbol{X})\right\|_2 = \left(\int \lambda^2 d\mathcal{P}\right)^{\frac{1}{2}}$ be the $L_2(\mathcal{P})$ norm of a function λ. For any $\zeta_1, \zeta_2 \in \mathcal{H}$, define a distance $d(\zeta_1, \zeta_2)$ as

$$\begin{aligned} d(\zeta_1, \zeta_2) &= \|\zeta_1 - \zeta_2\|_{\mathcal{H}} \\ &= \left\{\|\gamma_1 - \gamma_2\|^2 + \|\beta_1 - \beta_2\|^2 + \|f_1 - f_2\|_2^2 + \|h_1 - h_2\|_2^2\right\}^{1/2}. \end{aligned}$$

Denote the sieve pseudo-MLE of $\zeta = (\gamma, \beta, f, h)$ by $\hat{\zeta}_n^{\text{ps}} = \left(\hat{\gamma}_n, \hat{\beta}_n, \hat{f}_n, \hat{h}_n\right)$. Let $K_n = \mathcal{O}_n(n^\nu)$, where ν satisfies $1/[2(1 + p)] < \nu < 1/(2p)$. Further, let $\mathcal{G}_j(t \mid \boldsymbol{z})$ and $\mathcal{G}(t_1, t_2 \mid \boldsymbol{z})$ be Gaussian processes with mean zero; the closed forms of the variance functions are generally not available (Chen, Fan, and Tsyrennikov 2006) but can be estimated via a bootstrap approach.

Proposition 4.

Under the regularity conditions (RC1)–(RC4) and (SC1)–(SC4) given in the Appendix, and provided that $\tilde{S}_j^*(t \mid \boldsymbol{z})$ and $\tilde{S}_D(t \mid \boldsymbol{z})$ satisfy the condition (AC), the following results hold as $n \to \infty$.

i. Let $\boldsymbol{\zeta}_0$ denote the true value of $\boldsymbol{\zeta}$. $d\left(\widehat{\boldsymbol{\zeta}}_n^{\text{ps}}, \boldsymbol{\zeta}_0\right) = \mathcal{O}_p\left(n^{-\min(pv,(1-v)/2)}\right)$, which implies that if $v = 1/(1+2p)$, $d\left(\widehat{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_0\right) = \mathcal{O}_p\left(n^{-p/(1+2p)}\right)$; this is the optimal convergence rate in the nonparametric regression setting.

ii. For any fixed $\boldsymbol{z}$, let $\hat{\theta}_n(\boldsymbol{z})$ and $\hat{\theta}_{12n}(\boldsymbol{z})$ be the plug-in estimators of the association parameters following their semiparametric model specification (e.g., Equation (11), (12)). Then we have $\left(\hat{\theta}_n(\boldsymbol{z}), \hat{\theta}_{12n}(\boldsymbol{z})\right) \xrightarrow{a.s} (\theta(\boldsymbol{z}), \theta_{12}(\boldsymbol{z}))$ and $\sqrt{n}\left\{\left(\hat{\theta}_n(\boldsymbol{z}), \hat{\theta}_{12n}(\boldsymbol{z})\right)' - (\theta(\boldsymbol{z}), \theta_{12}(\boldsymbol{z}))'\right\} \xrightarrow{d} N(0, AV_{\theta,\theta_{12}}(\boldsymbol{z}))$, where $AV_{\theta,\theta_{12}}(\boldsymbol{z})$ can be estimated via a bootstrap method.

iii. For any fixed $\boldsymbol{z}$, $\hat{S}_{jn}(t \mid \boldsymbol{z}) \xrightarrow{\text{a.s.}} S_j(t \mid \boldsymbol{z})$ uniformly, and $\sqrt{n}\left(\hat{S}_{jn}(t \mid \boldsymbol{z}) - S_j(t \mid \boldsymbol{z})\right) \xrightarrow{w} \mathcal{G}_j(t \mid \boldsymbol{z})$ for $t \in [0, v_j]$, $j = 1, 2$.

iv. For any fixed $\boldsymbol{z}$, $\hat{S}_n(t_1, t_2 \mid \boldsymbol{z}) \xrightarrow{\text{a.s.}} S(t_1, t_2 \mid \boldsymbol{z})$ uniformly, and $\sqrt{n}\left(\hat{S}_n(t_1, t_2 \mid \boldsymbol{z}) - S(t_1, t_2 \mid \boldsymbol{z})\right) \xrightarrow{w} \mathcal{G}(t_1, t_2 \mid \boldsymbol{z})$ for $t_1, t_2 \in [0, v_1] \times [0, v_2]$.

3. Simulation Studies

We conducted simulation studies to verify the consistency, efficiency, and robustness of the proposed pseudo-MLE. For comparison, we evaluated the MLE of $(\theta(\boldsymbol{z}), \theta_{12}(\boldsymbol{z}))$ derived from the likelihood function (5) using the true survival functions $S_j(\cdot|\boldsymbol{Z})$, $j = 1, 2$ and $S_D(\cdot|\boldsymbol{Z})$. We also calculated the naïve estimates obtained by maximizing (5) after replacing the marginal survival functions with their Kaplan–Meier estimates (in the absence of covariates) or estimates under the Cox PH model (in the presence of covariates). Note that, in the settings that we focus on, the MLE is not applicable since the true survival functions $S_j(\cdot|\boldsymbol{Z})$ and $S_D(\cdot|\boldsymbol{Z})$ are unknown in practice, and the naïve estimator can be biased because of the informative censoring.

3.1. Simulation Settings

We considered n independent units, and the primary outcome of the bivariate event time, denoted (T_1, T_2). We set the sample size n to 500, 1000, or 2000. The generated observations on (T_1, T_2) may be censored by either the terminating event time D or the administrative time C_A, whichever occurs first. That is, the study censoring time $C = D \wedge C_A$. We started with situations without covariates and then examined the performance in the presence of covariates. The following four settings were simulated. We report the results of Settings 1 and 4 in Section 3.2. Details for Settings 2 and 3 are presented in Sections S2.2 and S2.3 of the supplementary materials.

Setting 1 *(Study of consistency and efficiency):* The data were generated from nested Archimedean copula functions (Joe 1997) that allowed different association parameters in the bivariate Archimedean copula $\mathcal{A}_{[2]}(\cdot)$ in Equation (2) and in the bivariate copula $\mathcal{C}_{[2]}(\cdot)$ in Equation (4).

Setting 2 *(Study of robustness):* The data were generated from trivariate Gaussian copula functions to check that the proposed approach was robust to model misspecification when the true copula was non-Archimedean.

Setting 3 *(Study of flexibility):* We explored the flexibility of our approach in situations where the distribution of the bivariate event times followed a Gamma frailty model.

Setting 4 *(Evaluating covariate effects):* The data were generated in a regression setting to examine the performance of the estimators with parameters that represented the covariate effects on the association parameters and conditional survival functions.

3.2. Data Generation and Simulation Results

3.2.1. Setting 1: Study of Consistency and Efficiency—

The observed data $\{[(u_{1i}, \delta_{1i}), (u_{2i}, \delta_{2i}), (c_i, \delta_{Di})] : i = 1, \ldots, n\}$ were generated as follows:

Step (a). We generated independently the trivariate random variables (v_{1i}, v_{2i}, v_{3i}) for $i = 1, \ldots, n$ from one of the three Archimedean copula functions, namely the Clayton, Gumbel, or Frank copula, using the R package copula (Hofert and Mächler 2011). The values of the parameters (θ, θ_{12}) in $\mathscr{A}_{[2]}(\cdot)$ and $\mathscr{C}_{[2]}(\cdot)$ were chosen so that the corresponding Kendall's tau (τ, τ_{12}) was (0.4, 0.5) or (0.3, 0.8), to simulate different levels of dependence.

Step (b). Let $S_j(t_{ji}) = v_{ji} = g\left(S_j^*(t_{ji}), S_D(t_{ji}), \theta\right)$ with $g(\cdot)$ as defined in Equation (7) for $j = 1, 2$. Using the survival functions of the Weibull distributions $S_j^*(\cdot)$ and $S_D(\cdot)$, we solved for t_{ji}. The scale and shape parameters, together with the regression coefficients, were predetermined. Solving $S_D(d_i) = v_{3i}$, we obtained the terminating event time d_i for $i = 1, \ldots, n$.

Step (c). We generated the administrative censoring times c_{Ai} independently from (v_{1i}, v_{2i}, v_{3i}), from an exponential distribution with the parameter chosen to give an overall censoring rate of approximately 45%, 50%, and 20% for T_1, T_2, and D, respectively. We let $c_i = d_i \wedge c_{Ai}$ with the indicator $\delta_{Di} = I(d_i \leq c_{Ai})$ and $u_{ji} = t_{ji} \wedge c_i$ with the indicator $\delta_{ji} = I(t_{ji} \leq c_i)$.

Table 1 summarizes the resulting estimates using pseudo-MLE, the unachievable MLE, and the naïve estimator, based on 500 repetitions under the nested Clayton and the nested Frank models. The sample means of the pseudo-MLE and MLE estimates were close to the true parameter values, especially for large n. This verified the consistency of the pseudo-MLE and MLE approaches. On the other hand, we observed moderate but visible differences between the naïve estimates and the true value in general; these differences persisted as the sample size increased. When $n = 2000$, the absolute differences between the sample means of $\hat{\tau}$ and its true value were more than twice the corresponding sample standard errors. The naïve estimates of τ_{12} did not deviate substantially from the true values and were closer for the nested Frank model than the nested Clayton model. The moderate deviation associated with the naïve estimates may be in part due to the conversion from the estimated copula parameters θ and θ_{12} to τ and τ_{12}, respectively. The ranges of θ and θ_{12} were wider in

general than the [−1, 1] range of Kendall's tau, so the latter did not reflect the differences as much as the copula association parameters did. We report Kendall's tau because it allows us to compare associations across different copula families. The results for θ and θ_{12} are given in Table S1 in Section S2.1 of the supplementary materials. The sample means associated with the naïve estimator for θ and θ_{12} showed larger differences from the true values. The sample standard errors of the pseudo-MLE were slightly elevated compared to their MLE counterparts, suggesting that the efficiency loss of the pseudo-MLE was modest in the settings we examined. The average estimated standard errors for the pseudo-MLE estimators were in general close to their empirical counterparts. We also evaluated the Type I error and power of a Wald-type test for the null hypothesis of $H_0 : \theta = 0$ or $H_0 : \theta_{12} = 0$, and the results are given in Table S2 of the supplementary materials (Section S2.1).

Figure 1 presents estimates of the marginal survival function $S_1(\cdot)$ for varying sample sizes. The data were generated under the nested Clayton copula model with Kendall's τ = 0.4 for the outer copula $\mathscr{A}_{[2]}(\cdot)$ and τ_{12} = 0.5 for the inner copula $\mathscr{C}_{[2]}(\cdot)$. The estimated survival functions for the proposed pseudo-MLE and the naïve approach together with their approximate 95% confidence bands (CBs) are presented along with the true curves. The curve of the true $S_1(\cdot)$ was covered by the CB from the pseudo-MLE approach, but it deviated from those using the naïve approach in every case. Similar patterns appeared for the estimates of $S_2(\cdot)$. For τ = 0.3 and τ_{12} = 0.8, and when the data were generated from the nested Frank models, similar patterns were observed; see supplementary material, Section S2.1.

3.2.2. Setting 4: Evaluating Covariate Effects—Settings 1 to 3 examined the performance of our approach with constant association parameters θ and θ_{12}. Setting 4 assessed its performance in settings where covariates $\mathbf{Z}$ were incorporated into either the association parameters $\theta(\cdot|\mathbf{Z})$ and $\theta_{12}(\cdot|\mathbf{Z})$ or the marginal survival functions $S_J(\cdot|\mathbf{Z})$ and $S_D(\cdot|\mathbf{Z})$. We generated n independent samples by following the steps in Setting 1 with the following two modifications:

Modification 1.: We generated realizations of $\mathbf{Z}$ and obtained $\theta(\mathbf{z})$ and $\theta_{12}(\mathbf{z})$ as follows:

Case 1. We used $\theta_{12}(\mathbf{z}_i) = \exp(a_1 z_{1i} + a_2 z_{2i}) - 1$, $\theta(\mathbf{z}_i) = \exp(a_3 z_{1i} + a_4 z_{2i}) - 1$. We generated $\{(z_{1i}, z_{2i}), i = 1, \ldots, n\}$ from the population $\mathbf{Z} = (Z_1, Z_2)'$ with Z_1 following the Uniform [1, 2] distribution and Z_2 following the Bernoulli (1/3) distribution. We set $\boldsymbol{\alpha} = (a_1, a_2)' = (0.3, 1)'$ and $\boldsymbol{\alpha}_{12} = (a_3, a_4)' = (0.7, 2)'$.

Case 2. We used $\theta_{12}(z_i) = \exp\left(\sin\left(\frac{3}{2}\pi z_i\right)\right) + 4, \theta(z_i) = \exp(\sin(2\pi z_i)) + 1$, and independently generated z_i, $i = 1, \ldots, n$ from the Uniform [0, 1] distribution.

In Case 1, $\theta(\mathbf{z})$ and $\theta_{12}(\mathbf{z})$ were specified in the true functional form and estimated parametrically. In Case 2, $\theta(\mathbf{z})$ and $\theta_{12}(\mathbf{z})$ were estimated semiparametrically by B-spline approximation.

Modification 2.: In step (b) of Setting 1, we replaced marginal functions $S_j(t)$, $j = 1, 2$, by the conditional survival functions

$$S_j(t_{ji} \mid \boldsymbol{z}_i) = v_{ji} = g\left(S_{0j}^*(t_{ji})^{\exp\left(\boldsymbol{\beta}_j^{*\prime}\boldsymbol{z}_i\right)}, S_{0D}(t_{ji})^{\exp\left(\boldsymbol{\beta}_D'\boldsymbol{z}_i\right)}; \theta(\boldsymbol{z}_i)\right),$$

where $\beta_1^* = 2$, $\beta_2^* = 2$, $\boldsymbol{\beta}_D = 2$, $g(\cdot)$ was the function defined in (7) and $S_D(d_i \mid \boldsymbol{z}_i) = v_{3i} = S_{0D}(d_i)^{\exp\left(\boldsymbol{\beta}_D'\boldsymbol{z}_i\right)}$.

We evaluated the proposed estimators for $\theta(\cdot)$ and $\theta_{12}(\cdot)$ under four different modeling assumptions, where $\theta(\cdot)$ and/or $\theta_{12}(\cdot)$ are assumed to be scalar(s) or function(s) of the covariates $\boldsymbol{Z}$. That yielded four scenarios, labeled by Scenarios (I) to (IV). Assuming both $\theta(\cdot)$ and $\theta_{12}(\cdot)$ to be functions of $\boldsymbol{Z}$ (Scenario (IV)) corresponds to the real data-generation process. Assuming θ and/or θ_{12} to be scalars (Scenarios (I)–(III)) may be interpreted as estimating the average dependence strength across all levels of covariates, which is a common approach when a constant copula association parameter is assumed. For the parametric estimation, we used the true forms for $\theta(\cdot)$ and $\theta_{12}(\cdot)$; for the semiparametric estimation, we used a B-spline approximation.

For better visualization and ease of comparison, we converted $\theta(\cdot)$ and $\theta_{12}(\cdot)$ to Kendall's tau $\tau(\cdot)$ and $\tau_{12}(\cdot)$. In Figure 2 we present the pseudo-MLE obtained in the four scenarios above for Case 2. The estimated $\tau(\cdot)$ and $\tau_{12}(\cdot)$ in Scenario (IV) were close to the true functions. The true curves were fully covered by the CB associated with the pseudo-MLE. When the association parameters were treated as scalars in Scenarios (I) to (III), they represented an average of the association across the covariates. Similar patterns were observed for Case 1 (see Section S2.4.1 of the supplementary materials). The estimates of $\tau(\cdot)$ and $\tau_{12}(\cdot)$ using the naïve and MLE approaches together with the pseudo-MLE estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha_{12}}$ are reported in Section S2.4 of the supplementary materials.

Figure 3 presents the estimated conditional survival function $\hat{S}_{1n}(t \mid \boldsymbol{Z})$ given in (16), corresponding to semiparametric estimation for Case 2. The naïve estimates and the true survival functions are also presented. The naïve estimates were biased in all scenarios, while the pseudo-MLE-based CBs covered the true curve in Scenario (IV) when the models for $\theta(\cdot)$ and $\theta_{12}(\cdot)$ were correctly specified. The pseudo-MLEs of $\theta(\cdot)$ and $S_j(t|\boldsymbol{Z})$ were close to the true values in Scenario (III) when the model for $\theta(\cdot)$, the parameter that quantifies the association in model $\hat{S}_{jn}(t \mid \boldsymbol{Z})$, $j = 1, 2$, was correctly specified. However, the marginal estimates $\hat{S}_{jn}(t \mid \boldsymbol{Z})$ for Scenarios (I) and (II) were biased since $\theta(\cdot)$ in these scenarios is treated as a scalar for all values of $\boldsymbol{z}$. This confirms that incorrectly treating the copula dependence parameter as a constant may result in biased conditional marginal estimates. Additional results for Case 1 and Case 2 are given in Section S2.4.1 and S2.4.2 of the supplementary materials, respectively.

4. Analysis of Cancer Survivorship Study Data

4.1. Preliminary Analysis

The cancer survivorship study included adult females diagnosed with stage I, II, or III breast cancer between January 1, 1989 and December 31, 2011 in BC, Canada. The subjects were 18 years or older and residents of BC, and identified from the provincial cancer registry. The relevant demographic information, death, and RSC-related data to December 31, 2014 were extracted from the registry and clinical databases. The records of CVD-related hospitalizations from January 1, 1986 to December 31, 2013 were extracted from the hospital separations database of BC (Canadian Institute for Health Information 2011).

Taking each subject's date of breast cancer diagnosis as her time origin, we considered the event time T_1 to be the time from diagnosis to RSC, the event time T_2 to be the time from diagnosis to her first subsequent CVD-related hospitalization, and the death time D to be the time from diagnosis to death. Here T_1, T_2, and D had a common time origin, which was the time of diagnosis, and they were the lengths of the intervals from diagnosis to RSC, CVD, and death, respectively. The availability of information on T_1 and T_2 was subject to censoring by death or the end of follow-up. We formulated each subject's censoring time as $C = D \wedge C_A$, where D was the time to death and C_A was the time at the end of the administrative window.

Table 2 gives the descriptive summary information, overall and by age at cancer diagnosis (age), the cancer stage at diagnosis (*stage*: early (stage I or II) vs. late (stage III)), the type of treatment received in addition to surgical procedures (*treatment*: chemotherapy and radiation therapy, chemotherapy only, radiation therapy only, no chemotherapy or radiation therapy), and the year of birth (*era*: era 1 (1900–1927), era 2 (1928–1945), era 3 (1946–1986)).

In the preliminary analysis, we estimated $S_1^*(t \mid \boldsymbol{Z})$, $S_2^*(t \mid \boldsymbol{Z})$, and $S_D(t|\boldsymbol{Z})$ with the Cox PH model, that is, the conditional distribution of time to $T_1 \wedge D$, $T_2 \wedge D$, and D. Here, $S_j^*(t \mid \boldsymbol{Z})$ is a common metric in the biomedical literature for the composite endpoint $T_j \wedge D$, $j = 1$, 2. It provides some insight into survival time without the disease, RSC, or CVD. Table S11 in Section S3.1 of the supplementary materials lists the estimated regression coefficients. Diagnosis at a late stage, as expected, was associated with a decrease in overall survival (hazard ratio for death = $e^{0.879} = 2.40$), survival without RSC (hazard ratio = $e^{0.744} = 2.10$), and survival without CVD (hazard ratio = $e^{0.704} = 2.02$). The same pattern was observed for those treated with chemotherapy only. Those who were treated with radiation therapy only had better survival (both overall and without disease). Compared to receiving no chemotherapy or radiation therapy, receiving both was associated with a moderate decrease in overall survival and survival without CVD, but not with survival without RSC. This is likely a reflection of the disease severity since only low-risk patients or those too sick to tolerate chemotherapy are given radiation therapy alone.

4.2. Evaluation of Association Between Event Times

The main goal of the study was to examine the association between the times to CVD and RSC in the presence of the potential informative censoring caused by death. Using the

approach of Section 2, we performed inference on the association parameter function $\theta_{12}(\cdot)$, a measure of the association between RSC and CVD, and $\theta(\cdot)$, a measure of the association of the two event times with death. We carried out both parametric and semiparametric modeling of $\theta(\cdot)$ and $\theta_{12}(\cdot)$. To illustrate, we considered covariates $\boldsymbol{Z} = (Z_1, Z_2)$, where Z_1 is a continuous covariate and Z_2 is a categorical covariate. The parametric approach followed model (9) for the Clayton copula, with $\boldsymbol{Z}$ = (age, Z_2) where Z_2 was the categorical variable of interest and Z_1 was age at diagnosis. For example, when Z_2 was *stage* at diagnosis, the estimated association functions with the parametric specifications were

$$\log(\theta(\text{ age }) + 1) = \begin{cases} 4.177 - 0.039\text{age} & \text{for early stage} \\ 3.460 - 0.024\text{age} & \text{for late stage} \end{cases},$$

and

$$\log(\theta_{12}(\text{age}) + 1) = \begin{cases} 1.255 - 0.007\text{age} & \text{for early stage} \\ 1.199 + 0.002\text{age} & \text{for late stage.} \end{cases}$$

Our exploratory analysis indicated that the association was not increasing strictly linearly with age. We next considered a more flexible approach, where θ(age) and θ_{12}(age) given z_2 were approximated by cubic B-splines. Figure 4 shows the estimated curves for log(θ(age) + 1) and log(θ_{12}(age) + 1) for both the parametric and semiparametric specifications with a given level of Z_2 (*era*). Figure S16 in Section S3.2 of the supplementary materials displays the corresponding estimated functions for Kendall's tau $\tau(\cdot)$ and $\tau_{12}(\cdot)$ with CBs. The association τ between T_j (RSC and CVD times) and D appeared to be roughly the same for both early- and late-stage groups, exhibiting a decreasing trend as age increased. This indicated an informative censoring due to the terminating event *death* and highlighted the importance of considering informative censoring. Late stage at diagnosis appeared to be a significant risk factor for an increased association between the RSC time T_1 and the CVD time T_2. A stronger association of (T_1, T_2) with D was shown in the younger generation (*era* = 2 and 3) compared to those born earlier (*era* = 1). This is expected since younger women are less likely to die from age-related causes. The estimates of the association with Z_2 being *treatment* are provided in Section S3.2 of the supplementary materials.

4.3. Estimation of $S_j(t|\boldsymbol{Z})$ for j = 1, 2

We estimated the marginal survival functions for T_1 and T_2 via Equation (16). Figure 5 presents the estimates of $S_2(t|\boldsymbol{Z})$, the marginal survival function of the time to CVD, obtained using the proposed approach (pseudo-MLE) and the naïve approach, i.e., applying the Cox PH model directly to the semi-competing risk data $\{[(u_{ji}, \delta_{ji}), (c_i, \delta_{Di}), \boldsymbol{z}_i)] : i = 1, \ldots, n\}$ for j = 1, 2. Diagnosis at a late stage was associated with a shorter time to CVD. Estimates based on the proposed pseudo-MLE approach suggested that those who were treated by radiation therapy alone had the longest time to CVD; in contrast, the naïve approach indicated no differences in time to CVD across treatment groups. The estimates of $S_1(t|\boldsymbol{Z})$ are provided in Section S3.3 of the supplementary materials.

5. Concluding Remarks and Further Discussion

This article provides a flexible modeling of the association between two event times, and an easy-to-implement procedure for estimating the association and the marginal distributions when the observations on the event times are subject to potentially informative censoring caused by a terminal event. The proposed approach models the dependence between the bivariate event time and the informative censoring time through a bivariate Archimedean copula (as the outer copula function). The joint survival function of the two event times is specified through a copula, which can be different from the outer copula. An alternative formulation such as a frailty model can also be used. The proposed approach allows the association between the two event times and their dependence on the informative censoring time to be different. This provides desirable flexibility in the modeling, and reflects a more realistic analysis strategy for practical studies such as the cancer survivorship study, where the association between the bivariate event times is weaker than their dependence on the informative censoring time.

The function $\mathscr{C}_{[2]}(\cdot;\theta_{12}(\cdot))$ in Equation (4) can be modeled by the widely used bivariate parametric copula functions (e.g., Diao and Cook 2014). When $\mathscr{C}_{[2]}(\cdot)$ in model (4) is assumed to be the bivariate Archimedean copula $\mathscr{A}_{[2]}(\cdot)$ in model (2), the joint survival function of the trivariate event time (T_1, T_2, D) in Equation (2) becomes $\mathscr{A}_{[3]}(S_1(t_1 \mid \boldsymbol{Z}), S_2(t_2 \mid \boldsymbol{Z}), S_D(d \mid \boldsymbol{Z}); \theta(\boldsymbol{Z}))$. In such settings, the joint survival function of each pair of T_1, T_2, and D is the same, defined by the bivariate Archimedean copula $\mathscr{A}_{[2]}(\cdot)$ with the two indices equal to the corresponding univariate marginal survival functions. Statistical inference is easier to carry out with this approach, but the resulting model can be unrealistic; see Li et al. (2019) for more discussion.

The correct specification of the copula models is critical because different models lead to different tail dependence. We performed simulation studies to evaluate the robustness of the proposed methods under model misspecification. We found that under model misspecification the estimates obtained with the Frank copula were closer to the true values than those from other copulas. In future investigations, it would be useful to develop methods for the copula model selection. One potential approach is to use the likelihood ratio test following Lawless and Yilmaz (2011a), and to obtain the p-values via a bootstrap procedure.

It is possible to extend the proposed method to allow the Archimedean copula $\mathscr{A}_{[2]}(\cdot)$ in Equation (2) to be covariate-dependent. For a discrete covariate, we can choose covariate-specific copula models according to its categories, that is, postulating $\mathscr{A}_{e,[2]}$ for $e = 1, \ldots, E$. For a continuous covariate, we may discretize the variable and then proceed with category-specific copula models. Adopting such stratified models may lead to efficiency loss due to an increase in the number of parameters to estimate. Model selection procedures such as a likelihood ratio test may be adapted to choose among different model specifications.

In the cancer survivorship study, the administrative censoring time C_A is the length of the interval between the year of diagnosis (W_x) and the end of data collection (W_a). While W_a is independent of T_j for $j = 1, 2$, W_x can be correlated with the event times of interest. It

thus renders potential informative censoring due to C_A in the observations on the T_j's. The analysis in Section 4 accounted for this censoring by conditioning on the cancer stage at diagnosis.

We model the joint survival function of (T_j, D), $j = 1, 2$, accounting for the semi-competing-risk nature of D. It would be natural to posit a Clayton copula when the times are positively correlated. Since $T_j \leq D$, the joint survival function in Equation (3) is identifiable for $0 \leq t_j \leq d \leq \infty$, the upper wedge where the data are observable (Fine, Jiang, and Chappell 2001). If information about the cause of death is available, one may be interested in modeling sequentially observed survival times, say, (T_1, D_1) with D_1 the death time caused by RSC. In this case, extra care is necessary to account for the induced dependent censoring by D_1 and potential identifiability issues relating to the fact that D_1 is observable only if T_1 is uncensored (Lin 2000; Lin, Sun and Ying 1999; Schaubel and Cai 2004; Cook and Lawless 2007, sec. 4.4.1). To overcome these challenges, Lawless and Yilmaz (2011b) proposed a semiparametric approach based on a bivariate copula model; it included a cure rate feature for the marginal distribution of survival time T_1 to model the subpopulation who never have the first event T_1 (see, e.g., Tsodikov, Ibrahim, and Yakovlev 2003).

The type of informative censoring that we consider is a semicompeting risk scenario, motivated by our application where death is the informative censoring time. Our method is directly applicable to situations with general informative censoring. Specifically, denote the censoring time $C = C_1 \wedge C_2$, where C_1 is independent of the event time(s) of interest, and C_2 is potentially correlated with the event time(s). Here, C_1 and C_2 are, respectively, analogous to C_A, determined by the administrative extraction window, and D, the lifetime, potentially correlated to either of the times to RSC and CVD. The joint survival function of T_1, T_2, and C_2 can be modeled by $\Pr(T_1 \geq t_1, T_2 \geq t_2, C_2 \geq c_2 \mid Z) = \mathscr{A}_{[2]}(S_{12}(t_1, t_2 \mid Z), S_{C_2}(c_2 \mid Z); \theta(Z))$, and the joint distribution of T_1 and T_2 can be modeled by $S_{12}(t_1, t_2 \mid Z) = \mathscr{C}_{[2]}(S(t_1 \mid Z), S(t_2 \mid Z); \theta_{12}(Z))$. Denote $T_j^* = T_j \wedge C_2$ for $j = 1, 2$; then it follows that, analogous to Equation (7), $S_j(t \mid Z) = g\left(S_j^*(t \mid Z), S_{C_2}(t \mid Z); \theta(Z)\right)$. Since the observations on T_j^* and C_2 are censored only by C_1, their distributions can be estimated with conventional approaches.

Although motivated by the cancer survivorship study, our modeling framework is applicable to other settings that involve evaluating the covariate effects on the association parameters. It can also be extended to other multivariate settings with semi-competing risks. For example, in cluster randomized trials, the treatment may affect the dependence parameters (see, e.g., Chen, Tchetgen, and Wang 2019), and it would be of interest to produce treatment-specific association parameter estimates. We have studied the overall dependence of event times through their marginal survival functions, motivated by considerations for the cross-sectional nature of the times. Our approach is therefore not immediately applicable to situations with time-varying dependence. One may consider modeling the dependence of the two corresponding survival processes over time to accommodate such dependence. This is another topic for future investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank to the editor, the associate editor, and two referees for their insightful comments and suggestions, which led to an improved article. We also thank to Prof. John J. Spinelli for his thoughtful comments and suggestions about the article. Disclaimer: All inferences, opinions, and conclusions drawn in this manuscript are those of the authors, and do not reflect the opinions or policies of the Data Steward(s).

Funding

The authors gratefully acknowledge DG R611382 and DAS R611689 from the Natural Sciences and Engineering Research Council of Canada (NSERC), CRT 2017–2020 from the Canadian Statistical Sciences Institute (CANSSI), and R01 AI136947 and R37 AI51164 from the National Institute of Allergy and Infectious Disease.

References

Andersen P, Borgan O, Gill R and Keiding N (1993), Statistical Models Based on Counting Processes, Springer Series in Statistics, New York: Springer-Verlag.

Breslow N (1972), “Discussion on Professor Cox’s Paper,” Journal of the Royal Statistical Society, Series B, 34, 202–220.

Canadian Institute for Health Information (2011), Discharge Abstract Database (Hospital Separations). V2. Population Data BC. Data Extract. MOH (2011). Available at: http://www.popdata.bc.ca/data.

Chen T, Tchetgen EJT, and Wang R (2019), “A Stochastic Second-Order Generalized Estimating Equations Approach for Estimating Association Parameters,” Journal of Computational and Graphical Statistics, 29, 1–46. [PubMed: 33013150]

Chen X, Fan Y, and Tsyrennikov V (2006), “Efficient Estimation of Semiparametric Multivariate Copula Models,” Journal of the American Statistical Association, 101, 1228–1240.

Chen Y (2012), “Maximum Likelihood Analysis of Semicompeting Risks Data With Semiparametric Regression Models,” Lifetime Data Analysis, 18, 36–57. [PubMed: 21850528]

Cook RJ, and Lawless JF (2007), The Statistical Analysis of Recurrent Events, Statistics for Biology and Health, New York: Springer.

Cox DR (1972), “Regression Models and Life-Tables,” Journal of the Royal Statistical Society, Series B, 34, 187–202.

Davis M, Li D, Wai E, Tyldesley S, Simmons C, Baliski C and McBride M (2014), “Hospital-Related Cardiac Morbidity Among Survivors of Breast Cancer: Long-Term Risks and Predictors,” Canadian Journal of Cardiology, 30, S122–S123.

De Boor C (1978), A Practical Guide to Splines, Applied Mathematical Science, New York: Springer-Verlag.

Diao L, and Cook RJ (2014), “Composite Likelihood for Joint Analysis of Multiple Multistate Processes Via Copulas,” Biostatistics, 15, 690–705. [PubMed: 24719283]

Emura T, and Chen Y (2018), Analysis of Survival Data With Dependent Censoring: Copula-Based Approaches, Springer Briefs in Statistics, Singapore: Springer.

Emura T, Nakatochi M, Murotani K, and Rondeau V (2017), “A Joint Frailty-Copula Model Between Tumour Progression and Death for Meta-Analysis,” Statistical Methods in Medical Research, 26, 2649–2666. [PubMed: 26384516]

Fine JP, Jiang H, and Chappell R (2001), “On Semi-Competing Risks Data,” Biometrika, 88, 907–919.

Genest C, Ghoudi K, and Rivest L-P (1995), “A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions,” Biometrika, 82, 543–552.

Glidden D (2000), ‘A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model,” Lifetime Data Analysis, 6, 141–156. [PubMed: 10851839]

Glidden D, and Self S (1999), "Semiparametric Likelihood Estimation in the Clayton–Oakes Failure Time Model," Scandinavian Journal of Statistics, 26, 363–372.

Goethals K, Janssen P, and Duchateau L (2008), "Frailty Models and Copulas: Similarities and Differences," Journal of Applied Statistics, 35, 1071–1079.

He X, Xue H, and Shi N-Z (2010), "Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Inflated Poisson Models," Journal of Multivariate Analysis, 101, 2026–2038. [PubMed: 20671990]

Hofert M, and Mächler M (2011), "Nested Archimedean Copulas Meet R: The Nacopula Package," Journal of Statistical Software, 39, 1–20.

Hougaard P (1986), "A Class of Multivariate Failure Time Distributions," Biometrika, 73, 671–678.

Huber PJ (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, I, 221–233.

Joe H (1997), Multivariate Models and Multivariate Dependence Concepts, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, New York: Taylor & Francis.

Lawless JF, and Yilmaz YE (2011a), "Comparison of Semiparametric Maximum Likelihood Estimation and Two-Stage Semiparametric Estimation in Copula Models," Computational Statistics and Data Analysis, 55, 2446–2455.

—— (2011b), "Semiparametric Estimation in Copula Models for Bivariate Sequential Survival Times," Biometrical Journal, 53, 779–796. [PubMed: 21887793]

Li D, Hu XJ, McBride ML and Spinelli JJ (2019), "Multiple Event Times in the Presence of Informative Censoring: Modeling and Analysis by Copulas," Lifetime Data Analysis, 26, 573–602. [PubMed: 31732833]

Lin D (2000), "Linear Regression Analysis of Censored Medical Costs," Biostatistics, 1, 35–47. [PubMed: 12933524]

Lin D, Sun W, and Ying Z (1999), "Nonparametric Estimation of the Gap Time Distribution for Serial Events With Censored Data," Biometrika, 86, 59–70.

Lo SMS, and Wilke RA (2010), "A Copula Model for Dependent Competing Risks," Journal of the Royal Statistical Society, Series C, 59, 359–376.

Lu M, Zhang Y and Huang J (2007), "Estimation of the Mean Function With Panel Count Data Using Monotone Polynomial Splines," Biometrika, 94, 705–718.

Nelsen R (2006), An Introduction to Copulas, New York: Springer-Verlag.

Nikoloulopoulos A, and Karlis D (2008), "Multivariate Logit Copula Model With an Application to Dental Data," Statistics in Medicine, 27, 6393–6406. [PubMed: 18816583]

Peng M, Xiang L, and Wang S (2018), "Semiparametric Regression Analysis of Clustered Survival Data With Semi-Competing Risks," Computation Statistics and Data Analysis, 124, 53–70.

Prenen L, Braekers R, and Duchateau L (2016), "Extending the Archimedean Copula Methodology to Model Multivariate Survival Data Grouped in Clusters of Variable Size," Journal of the Royal Statistical Society, Series B, 79, 483–505.

Schaubel DE, and Cai J (2004), "Regression Methods for Gap Time Hazard Functions of Sequentially Ordered Multivariate Failure Time Data," Biometrika, 91, 291–303.

Shih J, and Louis T (1995), "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," Biometrics, 51, 1384–1399. [PubMed: 8589230]

Sun T, and Ding Y (2019), "Copula-Based Semiparametric Regression Method for Bivariate Data Under General Interval Censoring," Biostatistics, 124, 53–70.

Tsodikov AD, Ibrahim JG, and Yakovlev AY (2003), "Estimating Cure Rates From Survival Data," Journal of the American Statistical Association, 98, 1063–1078. [PubMed: 21151838]

Wang W (2003), "Estimating the Association Parameter for Copula Models Under Dependent Censoring," Journal of the Royal Statistical Society, Series B, 65, 257–273.

Wei LJ, Lin DY, and Weissfeld L (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," Journal of the American Statistical Association, 84, 1065–1073.

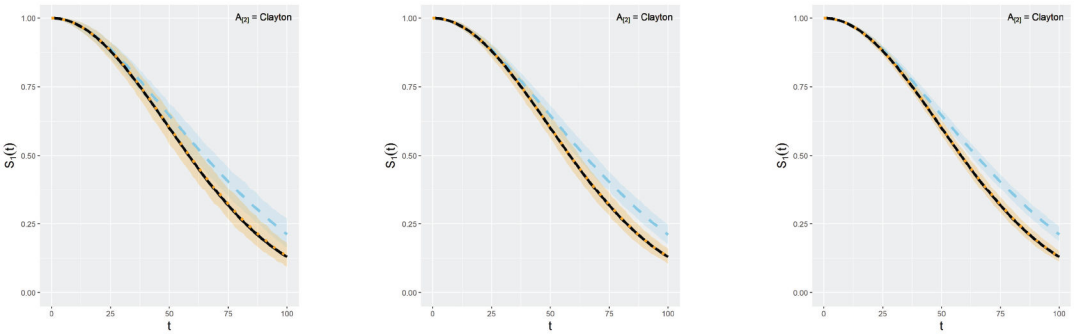
Wellner JA, and Zhang Y (2007), "Two Likelihood-Based Semiparametric Estimation Methods for Panel Count Data With Covariates," The Annals of Statistics, 35, 2106–2142.

Wienke A (2010), Frailty Models in Survival Analysis, Chapman & Hall/CRC Biostatistics Series, Boca Raton, FL: CRC Press.

Xue H, Lam KF, and Li G (2004), "Sieve Maximum Likelihood Estimator for Semiparametric Regression Models With Current Status Data," Journal of the American Statistical Association, 99, 346–356.

Zhang Y, Hua L and Huang J (2010), 'A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model With Interval-Censored Data," Scandinavian Journal of Statistics, 37, 338–354.

Zheng M, and Klein JP (1995), "Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula," Biometrika, 82, 127–138.



(a) $n = 500$ (b) $n = 1000$ (c) $n = 2000$

Figure 1.

Marginal survival functions: the true curve and estimates with data generated from the nested Clayton with $\tau = 0.4$, $\tau_{12} = 0.5$, and sample size n. The lines represent the true curve (solid black), naïve estimates (dashed blue), and the estimates from the proposed approach (dotted orange). Shaded areas correspond to confidence bands.

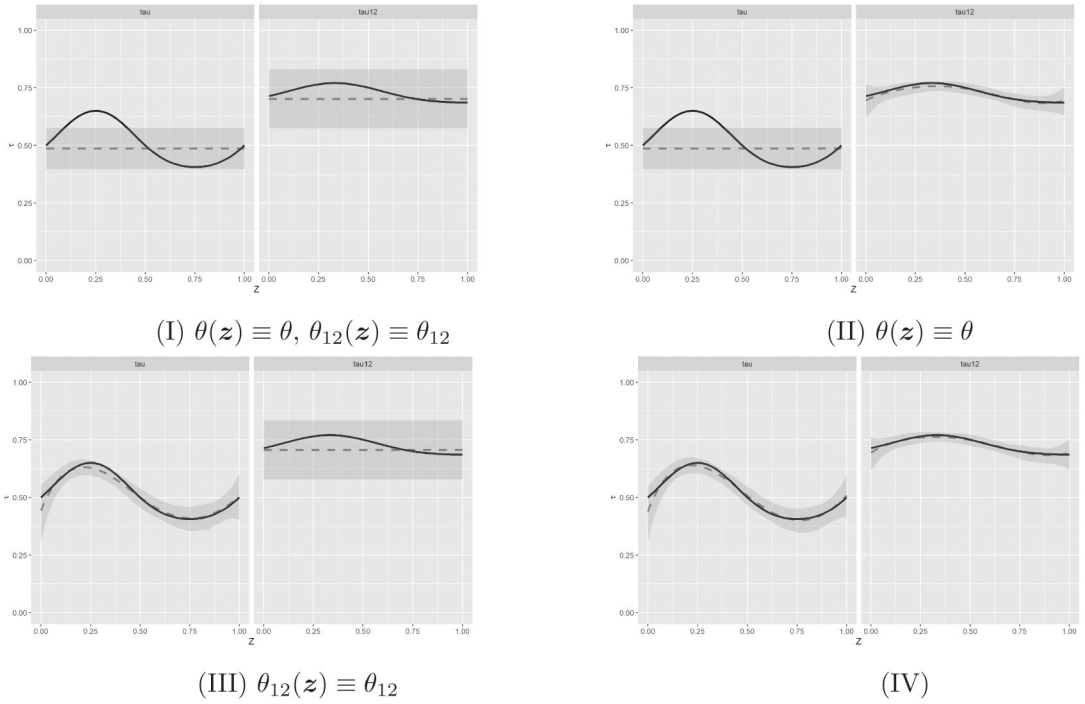


Figure 2.

Pseudo-MLEs of $\boldsymbol{\tau}(\boldsymbol{z})$ and $\boldsymbol{\tau}_{12}(\boldsymbol{z})$ and their confidence bands with data generated from Case 2, using B-spline approximation to $\boldsymbol{\theta}(\boldsymbol{z})$ and $\boldsymbol{\theta}_{12}(\boldsymbol{z})$. Scenarios (I)–(III) correspond to the cases where $\boldsymbol{\theta}(\cdot)$ and/or $\boldsymbol{\theta}_{12}(\cdot)$ were specified as scalar(s). Scenario (IV) corresponds to the case where both $\boldsymbol{\theta}(\cdot)$ and $\boldsymbol{\theta}12(\cdot)$ are approximated by B-spline functions. Solid: truth; dashed: estimated.

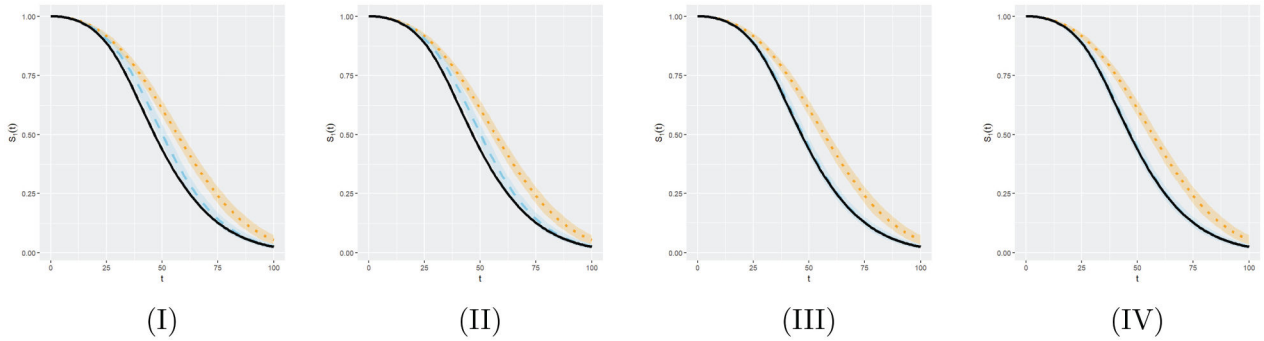
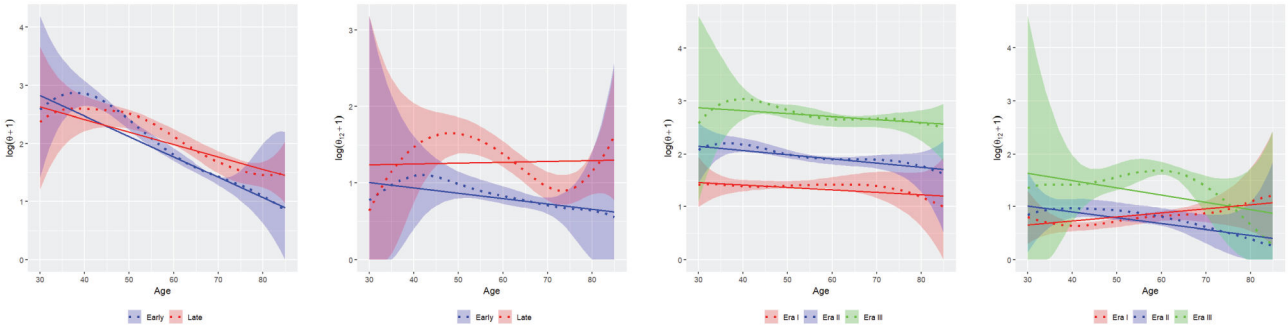


Figure 3.

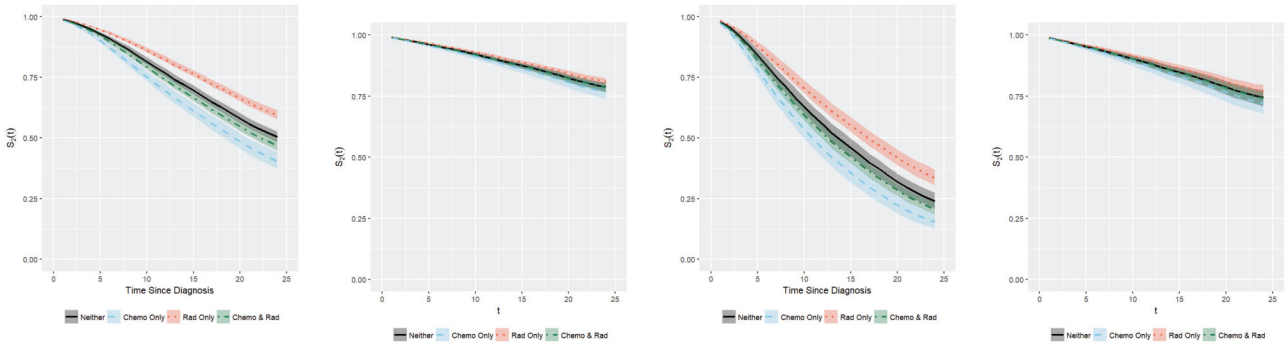
Estimates of $S_1(t|\boldsymbol{z})$ under Scenarios (I)–(IV) with data generated from Case 2, using B-spline approximation to $\theta(\boldsymbol{z})$ and $\theta_{12}(\boldsymbol{z})$. In all scenarios, $z = 0.3$. Scenario (I): $\theta(\boldsymbol{z}) \equiv \theta$, $\theta_{12}(\boldsymbol{z}) \equiv \theta_{12}$; (II): $\theta(\boldsymbol{z}) \equiv \theta$; (III): $\theta_{12}(\boldsymbol{z}) \equiv \theta_{12}$. Scenario (IV) corresponds to the case where $\theta(\cdot)$ and $\theta_{12}(\cdot)$ are specified by parametric functions. The curves are: truth (black solid), pseudo-MLE (blue dashed), and naïve (yellow dotted).



(a) $\log(\theta(age)+1)$ (b) $\log(\theta_{12}(age)+1)$ (c) $\log(\theta(age)+1)$ (d) $\log(\theta_{12}(age)+1)$

Figure 4.

Parametric and semiparametric estimation of association functions with real data. Solid lines represent parametric estimates, and dotted curves represent approximation by B-splines with corresponding confidence bands in shaded areas. (a),(b): Z_2 is *stage*. (c),(d): Z_2 is *era*.



(a) early stage, pseudo-MLE (b) early stage, naïve (c) late stage, pseudo-MLE (d) late stage, naïve

Figure 5.

Marginal survival function estimates of T_2 (time to CVD) for early and late stage at diagnosis with real data using the proposed approach and naïve Cox PH modeling. Each plot shows curves for four treatment groups among those diagnosed at age 49 and born in era 2. Shaded areas correspond to confidence bands.

J Am Stat Assoc. Author manuscript; available in PMC 2024 January 01.

Table 1.

Estimation of association parameters τ and τ_{12} with simulated data from nested Archimedean copulas with $\tau = 0.4$, $\tau_{12} = 0.5$ based on 500 repetitions.

		Nested Clayton			Nested Frank		
Parameter	**Estimates**	***n* = 500**	***n* = 1000**	***n* = 2000**	***n* = 500**	***n* = 1000**	***n* = 2000**
				Pseudo-MLE			
τ_{12}	smean†	0.49	0.49	0.50	0.50	0.50	0.50
	sse‡	0.027	0.019	0.013	0.024	0.016	0.012
	ese*	0.025	0.017	0.013	0.024	0.016	0.011
τ	smean	0.39	0.39	0.40	0.40	0.40	0.40
	sse	0.029	0.019	0.013	0.025	0.017	0.013
	ese	0.027	0.018	0.012	0.027	0.018	0.013
				MLE using True Marginals			
τ_{12}	smean	0.50	0.50	0.50	0.50	0.50	0.50
	sse	0.023	0.015	0.010	0.023	0.014	0.011
τ	smean	0.40	0.40	0.40	0.39	0.39	0.39
	sse	0.021	0.013	0.010	0.020	0.014	0.010
				Naïve			
τ_{12}	smean	0.51	0.53	0.54	0.51	0.51	0.51
	sse	0.045	0.024	0.015	0.024	0.017	0.012
τ	smean	0.35	0.37	0.37	0.37	0.37	0.38
	sse	0.031	0.018	0.012	0.023	0.016	0.012

NOTE:

†sample mean,

‡sample standard error,

*mean of estimated standard error.

Table 2.

Summary Statistics of the Cancer Survivorship Study Data

	Total	$N(T_1^{\text{obs}})$ †	$\overline{T_1^{\text{obs}}}$ ‡	$N(T_2^{\text{obs}})$	$\overline{T_2^{\text{obs}}}$	$N(D^{\text{obs}})$	$\overline{D^{\text{obs}}}$
Overall	36735	11025	5.30	5468	6.48	11330	7.56
Diagnosis age groups							
<40	2617	1019	4.61	109	7.15	721	5.89
40+	34118	10006	5.37	5359	6.47	10609	7.68
Stage							
Early (I and II)	33032	9421	5.61	5044	6.65	9685	8.00
Late (III)	3703	1604	3.48	424	4.48	1645	4.96
Treatment							
Chemotherapy and Radiation Therapy	11159	3516	4.94	899	6.40	2752	6.20
Chemotherapy Only	2866	945	4.27	250	6.19	734	5.80
Radiation Therapy Only	14470	4033	6.13	2562	6.97	4295	8.71
No Chemotherapy or Radiation Therapy	8240	2531	4.86	1757	5.85	3549	7.59
Era§							
Era 1	7180	2772	5.27	2435	6.44	4876	8.42
Era 2	14166	4461	5.81	2285	6.73	3806	7.42
Era 3	15389	3792	4.71	748	5.86	2418	5.76

NOTES:

† $N(T_1^{\text{obs}})$, $N(T_2^{\text{obs}})$, $N(D^{\text{obs}})$: numbers of subjects with observed times to RSC, CVD, and death, respectively

‡ $\overline{T_1^{\text{obs}}}$, $\overline{T_2^{\text{obs}}}$, $\overline{D^{\text{obs}}}$: sample means of the observed times (in years) to RSC, CVD, and death, respectively.

§ Era 1: Born in 1900–1927; Era 2: Born in 1928–1945; Era 3: Born in 1946–1986