Published in final edited form as:

Int J Biostat. 2015 May ; 11(1): 151–173. doi:10.1515/ijb-2013-0060.

A Semi-stationary Copula Model Approach for Bivariate Survival Data with Interval Sampling

Hong Zhu* and

Biostatistics, University of Texas Southwestern Medical Center – Dallas, Dallas, TX, USA

Mei-Cheng Wang

Biostatistics, Johns Hopkins University, Baltimore, MD, USA

Mei-Cheng Wang: mcwang@jhsph.edu

Abstract

In disease registries, bivariate survival data are typically collected under interval sampling. It refers to a situation when entry into a registry is at the time of the first failure event (i.e., HIV infection) within a calendar time window. For all the cases in the registry, time of the initiating event (i.e., birth) is retrospectively identified, and subsequently the second failure event (i.e., death) is observed during follow-up. In this paper we discuss how interval sampling introduces bias into the data. Given the sampling design that the first event occurs within a specific time interval, the first failure time is doubly truncated, and the second failure time is possibly informatively right censored. Consider semi-stationary condition that the disease progression is independent of when the initiating event occurs. Under this condition, this paper adopts copula models to assess association between the bivariate survival times with interval sampling. We first obtain bias-corrected estimators of marginal survival functions, and estimate association parameter of copula model by a two-stage procedure. In the second part of the work, covariates are incorporated into the survival distributions via the proportional hazards models. Inference of the association measure in copula model is established, where the association is allowed to depend on covariates. Asymptotic properties of proposed estimators are established, and finite sample performance is evaluated by simulation studies. The method is applied to a community-based AIDS study in Rakai to investigate dependence between age at infection and residual lifetime without and with adjustment for HIV subtype.

Keywords

bivariate survival data; Copula model; interval sampling; pseudo-conditional likelihood; semi-stationarity

1 Introduction

In disease surveillance systems or registries, it is common to collect data with a certain failure event, such as diagnosis of disease, occurring within a calendar time interval and then

*Corresponding author: hong.zhu@utsouthwestern.edu.

to obtain additional information retrospectively and/or prospectively. Such type of sampling is referred to as the interval sampling [1] and we consider bivariate survival data with interval sampling in this paper. One example of such data is AIDS blood transfusion data collected by the Centers for Disease Control, which is from a registry database, a common source of medical data [2]. Individuals who were diagnosed with AIDS during the course of the registry, July 1st, 1982, to June 30th, 1989, were recruited into the database and followed to study the progression of disease. In this example, the time of the initiating event of the HIV infection was retrospectively identified, and bivariate survival times of interest are the lag time from the infection to the AIDS diagnosis and the survival time after AIDS [2]. Generally speaking, under the interval sampling scheme, subjects experiencing the first failure event within a calendar time interval are identified as cases and entered into a registry [1]. For all the cases, the time of the initiating event is retrospectively confirmed and the occurrence of the second failure event is subsequently observed during the follow-up. Therefore, there is clearly a sampling bias due to the selection process, and subjects with the first failure events occurring before or after the course of the registry are unobservable and unaccountable. Any estimation and inference procedure done without consideration of this fact could possibly yield biased results. In the literature, sampling bias issues in disease surveillance data related to AIDS have been extensively studied; see the article of Brookmeyer [3] for an overview of research papers in the field. Methods were developed to handle, for example, various types of truncation, censoring and length-biased sampling.

This paper is partly motivated by a Rakai HIV study in investigating the relationship between age at infection and survival time of treatment-naïve HIV-infected individuals. This study in the rural Rakai district of southwestern Uganda conducted annual surveillance from November 1994 in an open cohort of individuals aged 15–49 years [4]. Interest is focused on a cohort of HIV seroconverters, who were initially HIV negative, then seroconverted between 1995 and 2003 and followed until they died or were censored by out-migration or the end of follow-up. With a wide range of age at infection, the risk of death may be positively or negatively associated with increasing age at infection. The scientific goal is to explore how the HIV progression differs by the age at infection and HIV subtype. Since antiretroviral treatment (ART) became available in 2004 in the Rakai Health Sciences Program, the follow-up time and survival analysis were truncated on Dec 31st, 2003, to assess the survival of ART-naïve HIV-infected individuals [4]. In this study, the initiating, first and second failure events correspond, respectively, to the birth, incidence of HIV infection and death. Bivariate survival times refer to the age at HIV infection and residual lifetime. Figure 1 provides a graphical presentation to illustrate how the interval sampling design arises in the Rakai HIV study.

As shown in Figure 1, the sampling population consists of individuals who became HIV infected between 1995 and 2003. Under the interval sampling, the age at HIV infection was observed subject to double truncation, that is, left truncation on Jan 1st, 1995, and right truncation on Dec 31st, 2003. The residual lifetime was possibly dependently right censored. A previous study [4] suggested that survival time decreased significantly with older age at infection. Their conclusion was based on comparing Kaplan–Meier survival curves among different groups of age at infection by the log-rank test and the estimated hazard ratio of death associated with age at infection by the Cox proportional hazards model. However, this

ignored the fact that data were collected with the interval sampling, which may introduce substantial bias in data analysis. Therefore, our first research problem is to correctly examine the association between age at infection and residual lifetime by removing the bias from the interval sampling. Second, the HIV subtype was found to have an impact on the progression of HIV infection [5]. Thus, we further study the effect of HIV subtype on the association between age at infection and residual lifetime and evaluate the association adjusting for the HIV subtype.

In statistical literature, bivariate and multivariate survival data have been extensively studied. Various statistical methods have been developed to nonparametrically analyze bivariate survival data with right censoring [6, 7, 8, 9]. When the association of bivariate survival times is of interest, the semiparametric copula model has been becoming an increasingly popular tool for modeling the dependence, and in particular, the copula-based survival model has been proposed [10] for bivariate data both subject to right censoring. Wang [11] for bivariate survival data under dependent censoring and Lakhal-Chaieb et al. [12] for bivariate serial gap times. It is noted that these papers consider bivariate survival data, where both failure times are subject to right censoring. They are different from the bivariate survival data with interval sampling of our interest, for which the first failure time is subject to double truncation and the second failure time is subject to right censoring. Nevertheless, the copula family includes many useful bivariate survival models and enjoys flexibility in modeling [10]. An appealing feature is that it allows separate modeling and estimation of margins and the dependency parameter. Estimation and inference could be carried out by a two-stage procedure. At the first stage, marginal survival functions of each failure time are consistently estimated. At the second stage, the association parameter is estimated by maximizing a pseudo-likelihood with marginal survival functions replaced by their consistent estimators. The ideas of two-stage estimation for the copula model have been used by Genest et al. [13] for complete data, Shih and Louis [10] for right-censored data and Wang and Ding [14] for bivariate current status data. The properties of the proposed association estimators depend on regularity conditions on the imposed copula model and the plugged-in estimators of marginal survival functions. In these works, the covariate effect is often modeled through marginal distributions, such as the marginal Cox regression model, assuming the association parameter in the copula model is constant.

In this paper, we consider the semi-stationary copula model for bivariate survival data with interval sampling, and the association parameter is estimated through a two-stage procedure based on a pseudo-conditional likelihood. One challenge is how to correctly estimate the marginal distributions using interval sampling data, which is crucial to ensure unbiased estimation of the association between bivariate survival times. Under the stationary condition, the model was studied in Zhu and Wang [1], where it showed interval sampling does not induce bias on each univariate failure time. This paper relaxes the stationarity assumption to the semi-stationarity, which is a much less restrictive condition, and investigates the data structure that the first failure time is doubly truncated and the second failure time is dependently right censored. Moreover, motivated by the Rakai HIV study where the HIV progression is likely to depend on the HIV subtype, it is of interest to quantify the covariate effect on the association. Therefore, we focus on two scenarios. First, in the absence of covariates, we propose bias-corrected estimators for marginal survival

functions and estimate the association parameter of the copula model by a two-stage procedure. Then we model the marginal distributions in a more flexible manner by the Cox proportional hazards models with covariates incorporated. The association is modeled in a parametric way to include covariates and we study the covariate-adjusted association. A novelty of the second model lies in the explicit linkage of the covariate effect to the association measure. Our approaches are not restricted to a particular copula but include the Clayton, positive stable and Frank copula models. The rest of the paper is organized as follows. In Section 2, the interval sampling design is discussed with more details, and the copula model for bivariate survival data as well as the semi-stationary model assumption are introduced. In Section 3, marginal survival distribution for each failure time and the association parameter in the copula are studied under the semi-stationary condition, without consideration of covariates. In Section 4, we incorporate covariates into the copula model and evaluate the association. Asymptotic properties of the proposed estimators are established. Finite sample performances are examined by simulation studies in Section 5. In Section 6, for illustration, the proposed method is applied to the Rakai HIV study. Finally, concluding remarks and discussion are included in Section 7. Technical details and proofs are provided in supplementary materials.

2 The semi-stationary Copula model

In this section, we describe the data structure for bivariate survival data with interval sampling and some fundamental concepts of the copula model, together with the semi-stationary model assumption. Statistical methods and inference are developed for a target population of individuals experiencing the first failure event of interest. To begin, we define random variables for the target population. Let T denote the calendar time of the initiating event, Y denote the time from the initiating event to the first failure event and Z denote the time from the first event to the second event. The failure times Y and Z are possibly correlated and their dependent relationship is of primary interest. Under the interval sampling, the sampling population is made up of subjects whose first failure events occur within a calendar time interval $[0, t_0]$, described by the constraint $0 \leq Y + T \leq t_0$. Therefore, bivariate failure times are observed subject to sampling bias. Specifically, Y is doubly truncated. Denote the double truncation rate by $\mu = 1 - \Pr(-T \leq Y \leq t_0 - T)$, and with this probability a person who experienced the first failure event will not be identified. Let C denote the censoring time measured from the first event, because censoring for the second event can only arise following the occurrence of the first event. Correspondingly, the observation of Z is censored by C .

Assume that the initiating event T occurs over the calendar time with a rate function $\lambda(t)$ for $t \in [0, t_0]$. Let $f_{Y,Z}(y, z)$ denote the population joint density function of (Y, Z) , and $F_Y(\cdot)$ and $F_Z(\cdot)$ denote the population marginal cumulative distribution functions of Y and Z , respectively. We set $y_- = \inf\{y : F_Y(y) > 0\}$, $y_+ = \sup\{y : F_Y(y) < 1\}$, $z_- = \inf\{z : F_Z(z) > 0\}$, $z_+ = \sup\{z : F_Z(z) < 1\}$, $t_- = \inf\{t : \lambda(t) > 0\}$, and assume that failure time Y has finite support with $y_+ < \infty$ to reduce mathematical complexity in the discussion. The population density function of T , $g(t)$, could be defined as a normalized rate function in the interval

$[-y_+, t_0 - y_-]$ as, $g(t) = \lambda(t)I(-y_+ \leq t \leq t_0 - y_-) / \int_{-y_+}^{t_0 - y_-} \lambda(u)du$, and its population cumulative distribution function is denoted by $G(t)$.

Suppose bivariate failure times (Y, Z) come from the C_a copula for some association parameter $a \in \mathcal{R}$, where C_a is a distribution function on $[0, 1]^2$ with density c_a , then the joint survival function and density function of (Y, Z) are given by

$$\begin{aligned} S_{Y,Z}(y, z) &= C\{S_Y(y), S_Z(z), a\}, y, z \geq 0 \\ f_{Y,Z}(y, z) &= c\{S_Y(y), S_Z(z), a\}f_Y(y)f_Z(z), y, z \geq 0 \end{aligned}$$

where $S_Y(y)$, $S_Z(z)$, $f_Y(y)$ and $f_Z(z)$ are the population marginal survival functions and densities of Y and Z , respectively. The association parameter a is closely related to Kendall's τ , the rank correlation coefficient, expressed as

$$\tau = 4 \int_0^1 \int_0^1 C(u, v, \alpha) dC(u, v, \alpha) - 1$$

Assume that $(T_1, Y_1, Z_1), \dots, (T_n, Y_n, Z_n)$ are independent and identically distributed (i.i.d.). We then introduce the following model assumption to facilitate the development of the proposed work.

Assume that bivariate failure times (Y, Z) are independent of the time of the initiating event T .

The model is considered to be semi-stationary if (S) is satisfied. It is noticed that the validity of the semi-stationarity assumption (S) may be questionable in the presence of systematic shifts in factors related to disease progression, such as availability of new treatments in the context of HIV infection. Nevertheless, the methods in this paper are proposed under the semi-stationary assumption for analyzing bivariate survival data with interval sampling. Actually, this assumption may be satisfied in some biomedical or epidemiological studies where no new treatment or diagnostic tool would have a significant impact on disease incidence or survival during a specific study period. In addition to the semi-stationarity, if the initiating events occur at a constant rate over the calendar time so that T follows a uniform distribution, the model is considered to be stationary as discussed in Zhu and Wang [1]. The stationary model forms a special case in which the bivariate survival data with interval sampling are unbiased. To be specific, under the stationarity, double truncation from the interval sampling does not result in bias on the first failure time Y , and the second failure time Z is independently right censored. In this paper, we focus on the model under the semi-stationary condition, which is less restrictive than the stationary condition, to address the issue of interval sampling bias.

Assume that (S) holds, the joint density of observed (t, y) can be written as

$$\begin{aligned}
 p_{T,Y}(t,y) &= \frac{g(t)f_Y(y)I(-y \leq t \leq t_0-y)}{pr(-T \leq Y \leq t_0-T)} \\
 &= \left\{ \frac{g(t)I(-y \leq t \leq t_0-y)}{G(t_0-y)-G(-y)} \right\} \times \left[\frac{\{G(t_0-y)-G(-y)\}f_Y(y)}{\int_0^\infty \{G(t_0-u)-G(-u)\}f_Y(y)du} \right] \\
 &= p_{T|Y}(t|y) \cdot p_Y(y),
 \end{aligned} \tag{1}$$

where $p_{T,Y}(t,y)$ is the sampling joint density of (T, Y) , $p_{T|Y}(t|y)$ refers to the first bracket of the above formula and $p_Y(y)$ refers to the second bracket and is the marginal density of Y .

3 The copula model without covariates

Under the semi-stationary condition when (S) is satisfied, we fit a copula model for bivariate survival data with interval sampling, without considering covariates in this section. The data structure is studied and bias-corrected consistent estimators for marginal survival functions are proposed. In some situations, there is sufficient information on the distribution of the initiating event time T to determine a well-fitted parametric form. In such cases, it is desirable to make use of this information and incorporate it into the analysis. Therefore, we assume a parametric density function $g(t; \theta)$ to model T with corresponding distribution function $G(t; \theta)$, where $\theta \in \Theta$ and Θ is an open set in R^k . Taking the Rakai HIV study data, for example, g describes the birth trend for HIV seroconverters. In the following, under the semi-stationary condition that T is independent of (Y, Z) , a conditional likelihood estimator of θ is obtained, and the inverse probability weighting method is employed to derive consistent estimators of marginal survival functions, $S_Y(y)$ and $S_Z(z)$, at the first stage. At the second stage, the association parameter α in the copula model is estimated based on a pseudo-conditional likelihood. The approach is general and can be applied to different classes of the copula model.

Under Assumption (S), using eq. (1), the conditional likelihood function of observed $\{t\}$ given observed $\{y\}$ can be derived as

$$L_c(\theta) = \prod_{i=1}^n P_{T|Y}(t_i|y_i, \theta) = \prod_{i=1}^n \left\{ \frac{g(t_i; \theta)}{G(t_0 - y_i; \theta) - G(-y_i; \theta)} \right\}$$

in which the distribution of Y becomes a nuisance parameter and is eliminated by the conditioning procedure. The conditional maximum likelihood estimator $\hat{\theta}$ is obtained by maximizing $L_c(\theta)$. The large sample properties of $\hat{\theta}$ can be obtained using techniques for M -estimators. Under regularity conditions, as $n \rightarrow \infty$, $\hat{\theta}$ is consistent and $n^{1/2}(\hat{\theta} - \theta)$ converges weakly to a mean zero multivariate normal distribution with variance-covariance matrix I_c^{-1} , where

$$I_c = E \left[\left\{ \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) \right\} \left\{ \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) \right\}^T \right]$$

is the Fisher information matrix of $L_c(\theta)$.

We then explore the probability structure of the bivariate data to estimate marginal survival functions $S_Y(y)$ and $S_Z(z)$. Due to the interval sampling, the sampling distributions of Y and Z are, in general, different from their population distributions. For the first failure time Y , as shown in formula (1), the sampling density $p_Y(y)$ is proportional to its population density $f_Y(y)$ as

$$p_Y(y) = \frac{w(y, \theta) f_Y(y)}{\int_0^\infty w(u, \theta) f_Y(u) du}$$

where $w(y, \theta) = G(t_0 - y, \theta) - G(-y, \theta)$ is called the selection bias function and it represents the probability that a subject in the target population will be observed during the calendar time interval. The correction for the bias from interval sampling will make use of this selection bias function. It is clear that weighting each observation of Y by a weight that is inversely proportional to the selection bias function at the value of that observation will adjust for the sampling bias. Thus a consistent estimator of $S_Y(y)$ can be derived as

$$\hat{S}_Y(y, \hat{\theta}) = \frac{\sum_{i=1}^n \{G(t_0 - y_i; \hat{\theta}) - G(-y_i; \hat{\theta})\}^{-1} I(y_i > y)}{\sum_{i=1}^n \{G(t_0 - y_i; \hat{\theta}) - G(-y_i; \hat{\theta})\}^{-1}}$$

where y_i is the observed first failure time and $\hat{\theta}$ is the conditional likelihood estimator from $L_c(\theta)$. The weighted empirical survival function $\hat{S}_Y(y, \hat{\theta})$ can be proved to be a semiparametric maximum likelihood estimator of $S_Y(y)$.

For the second failure time Z , we discuss different situations of censoring on Z according to the real data application. In the Rakai HIV study, following the previous work [4], we consider that censoring occurs either due to out-migration or the end of the study. It is noted that there may be dependent censoring caused by informative dropout in some HIV studies. Developing a method to handle this dependent censoring is an interesting problem and will be explored in future work. Let C_1 be the censoring time caused by out-migration and measured from the first event, where $C_1 = \infty$ if there is no out-migration. As for a prospective follow-up study, it is reasonable to assume the censoring time C_1 by out-migration is independent of the residual lifetime Z . The independent censoring of C_1 on Z may be satisfied by the study design and model assumption. Then, we only need to adjust for the interval sampling bias on Z . Since Y and Z are observed concurrently, the same selection bias function $w(y, \theta) = G(t_0 - y, \theta) - G(-y, \theta)$ could be used to adjust for the sampling bias of observing Z induced by Y , and a weighted Kaplan–Meier estimator of $S_Z(z)$ is developed

by the inverse probability weighting method. Let $\{(y, x)\}$ where $x = \min(z, c_1)$ denote the observed bivariate data, and the weighted Kaplan–Meier estimator of $S_Z(z)$ is given as

$$\begin{aligned}\hat{S}_Z(z, \hat{\theta}) &= \prod_{x_{(j)} < z} \left\{ 1 - \frac{\sum_{i \in d_{j2}} w(y_i, \hat{\theta})^{-1}}{\sum_{i \in d_{j2}} w(y_i, \hat{\theta})^{-1}} \right\} \\ &= \prod_{x_{(j)} < z} \left[1 - \frac{\sum_{i \in d_{j2}} \{G(t_0 - y_i, \hat{\theta}) - G(-y_i, \hat{\theta})\}^{-1}}{\sum_{i \in d_{j2}} \{G(t_0 - y_i, \hat{\theta}) - G(-y_i, \hat{\theta})\}^{-1}} \right],\end{aligned}$$

where $\hat{\theta}$ is the conditional maximum likelihood estimator, $d_{j2} = \{i : x_i = x_{(j)}\}$ and $R_{j2} = \{i : x_i < x_{(j)}\}$ are the failure event set and risk set at $x_{(j)}$, respectively, and $\{x_{(1)}, \dots, x_{(k)}\}$ are distinct ordered uncensored second failure times with their counterparts at the first failure time as $\{y_{(1)}, \dots, y_{(k)}\}$. The inverse probability weighting method has been widely used in literatures particularly to reduce selection bias in observational studies. In our model setting, the selection bias function is constructed according to the biased distribution of Y under the interval sampling, thus the inverse of this function plays the role of correcting for the induced sampling bias of observing the second failure time Z . The weighted Kaplan–Meier estimator $\hat{S}_Z(z, \hat{\theta})$ is a semiparametric consistent estimator of $S_Z(z)$. The weak convergence of $n^{1/2}\{\hat{S}_Z(\cdot) - S_Z(\cdot)\}$ can be constructed following the lines in Aalen [15].

Next, let C_2 denote the censoring time measured from the first event and caused by the end of the follow-up at time t_0 , that is, $T + Y + C_2 = t_0$. Let $M = T + Y$ denote the calendar time of the first event, then the censoring time of observing Z is $t_0 - M$. First of all, we consider the case without selection bias from the interval sampling or other random censoring. When $Z \perp M$, a standard estimate of $S_Z(z)$ is the Kaplan–Meier estimate. However, when Z and M are dependent, this estimator is inconsistent. One exception shown by Zhu and Wang [1] is that under the stationary condition when T follows a uniform distribution, M occurs at a constant rate and is independent of Z ; therefore, $S_Z(z)$ can still be estimated by the Kaplan–Meier estimate. For correlated (Z, M) , we can intentionally treat M as a covariate and fit a working Cox proportional hazards model of Z given M as

$$h_Z(z|m) = h_{0Z}(z) \exp(\mu m)$$

where $h_{0Z}(z)$ is an unspecified baseline hazard function and μ is the regression coefficient. In the Rakai HIV study, M is the calendar time of HIV infection and Z is residual lifetime. The “exact method” for estimating the survival function of Z is to estimate the conditional survival distribution of Z given M using the working Cox model, then $S_Z(z)$ can be simply estimated by the empirical average of the estimated conditional survival distribution of Z given M over all the observed covariate M [16]. Secondly, similar to the previous discussion, we observe $\{(x, \delta)\}$, where $x = \min(z, c_2) = \min(z, t_0 - t - y)$ and $\delta = I(z < t_0 - t - y)$ with the selection probability $w(y, \theta)$ due to the interval sampling. Therefore, an inverse

probability weighting method is applied for the Cox model estimation to adjust for the interval sampling bias. Define a weighted partial score function as

$$U_{j2}(\mu) = m_j - \frac{\sum_{i \in R_{j2}} w(y_i, \hat{\theta})^{-1} \exp(\mu m_i) m_i}{\sum_{i \in R_{j2}} w(y_i, \hat{\theta})^{-1} \exp(\mu m_i)}$$

Then a weighted estimating equation for μ is $U_2(\mu) = \sum_j w(y_j, \hat{\theta})^{-1} U_{j2}(\mu) = 0$ with the solution denoted by $\hat{\mu}$, and let $\hat{S}_0(z)$ denote the corresponding Breslow estimator for the baseline survival function of Z . The conditional survival function of Z given M is consistently estimated by $\hat{S}_Z(z|m) = \hat{S}_0(z)^{\exp(\hat{\mu}m)}$. Accordingly, a consistent estimator of $S_Z(z)$ is obtained as $\hat{S}_Z(z, \hat{\theta}) = \sum_{i=1}^n n^{-1} \hat{S}_{0Z}(z)^{\exp(\hat{\mu}m_i)}$. The asymptotic properties of $\hat{S}_Z(z, \hat{\theta})$ depend on the asymptotic properties of $\hat{\mu}$ and $\hat{S}_0(z)$, for which a detailed discussion is provided in Section 4. Following our notations, the common censoring time C can be expressed as $C = \min(C_1, C_2) = \min(C_1, t_0 - M)$, and the method based on the working Cox model would provide an appropriate estimator for $S_Z(z)$ in general.

We then present the estimation procedure for the association parameter α in the copula model based on a pseudo-conditional likelihood. Bivariate survival distribution of (Y, Z) is modeled by the copula as $S_{Y,Z}(y, z) = C\{S_Y(y), S_Z(z), \alpha\}$. First, we consider the situation when θ is known, which means the exact parametric distribution of T is available. The marginal survival functions are estimated by $\hat{S}_Y(y, \theta)$ and $\hat{S}_Z(z, \theta)$, respectively. For each subject i ($i = 1, \dots, n$), data $\{y_i, x_i, \delta_i, t_i\}$ are observed where $x_i = \min(z_i, c_i)$ and $\delta_i = I(z_i \leq c_i)$. The joint density function of (Y, X, δ, T) can be expressed as a product of the conditional density function of $(Y, X, \delta|T)$ and the marginal density function of T . The corresponding conditional likelihood function, $L_c(\alpha)$, could be derived as

$$\begin{aligned} L_c(\alpha) &= \prod_i \frac{f_{Y,Z}(y_i, x_i)^{\delta_i} \left\{ \frac{\partial S_{Y,Z}(y_i, x_i)}{\partial y_i} \right\}^{1-\delta_i}}{S_Y(t_0 - t_i) - S_Y(-t_i)} \\ &\propto \prod_i f_{Y,Z}(y_i, x_i)^{\delta_i} \left\{ \frac{\partial S_{Y,Z}(y_i, x_i)}{\partial y_i} \right\}^{1-\delta_i} \end{aligned}$$

An interesting feature in the decomposition of likelihood is that the marginal likelihood function does not involve the parameter of interest α . Therefore, it is appropriate to estimate and make inference on α solely based on the conditional likelihood function $L_c(\alpha)$. Denote $\{S_Y(y_i), S_Z(x_i)\}$ by (u_i, v_i) and by the copula model of bivariate survival data,

$$L_c(\alpha) \propto \prod_{i=1}^n l(u_i, v_i, \alpha) = \prod_{i=1}^n l(u_i, v_i, \alpha)^{\delta_i} \left\{ \frac{\partial C(u_i, v_i, \alpha)}{\partial u_i} \right\}^{1-\delta_i}$$

From the previous discussion, two margins $S_Y(y)$ and $S_Z(z)$ could be consistently estimated by $\hat{S}_Y(y, \theta)$ and $\hat{S}_Z(z, \theta)$, respectively. Therefore, a pseudo-conditional likelihood score equation is constructed by substituting $S_Y(y)$ and $S_Z(z)$ with their consistent estimators and is given as

$$U_\alpha\{\hat{S}_Y(\theta), \hat{S}_Z(\theta), \alpha\}^{(p)} = \left(\frac{\partial}{\partial \alpha} \sum_{i=1}^n \delta_i \log [c\{\hat{S}_Y(y_i, \theta), \hat{S}_Z(x_i, \theta), \alpha\}] + (1 - \delta_i) \log \left[\frac{\partial C\{\hat{S}_Y(y_i, \theta), \hat{S}_Z(x_i, \theta), \alpha\}}{\partial u_i} \right] \right) = 0 \quad (2)$$

The estimator of the association parameter α , $\hat{\alpha}(\theta)$, is the solution to eq. (2). Note that T follows a uniform distribution under the stationary condition, where $\hat{S}_Y(y, \theta)$ and $\hat{S}_Z(z, \theta)$ reduce to the empirical survival function and Kaplan–Meier estimator, respectively. As $n \rightarrow \infty$, $\hat{\alpha}(\theta)$ is a consistent estimator of α_0 , and $n^{1/2}\{\hat{\alpha}(\theta) - \alpha_0\}$ converges weakly to normal distribution with mean zero and variance σ^2 .

Now we consider the general case when θ is unknown. It is natural to replace θ by the conditional maximum likelihood estimator $\hat{\theta}$ and derive an estimator of α by solving the equation $U_\alpha\{\hat{S}_Y(\hat{\theta}), \hat{S}_Z(\hat{\theta}), \alpha\}^{(p)} = 0$. Let $\hat{\alpha}(\hat{\theta})$ denote its solution. The variability of $\hat{\alpha}(\hat{\theta})$ can be decomposed into two terms as, $\hat{\alpha}(\hat{\theta}) - \alpha_0 = \{\hat{\alpha}(\theta) - \alpha_0\} + \{\hat{\alpha}(\hat{\theta}) - \hat{\alpha}(\theta)\}$, where the variability in the first term is described by σ^2 , and that in the second term is generated by the use of $\hat{\theta}$ to estimate θ . The corresponding distributions of the two terms can be proven to be asymptotically orthogonal to each other because θ in the second term is estimated by a conditional likelihood. The proposed estimator $\hat{\alpha}(\hat{\theta})$ has the following asymptotic properties.

Theorem 1—As $n \rightarrow \infty$, $\hat{\alpha}(\hat{\theta})$ is a consistent estimator of α_0 , and $n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\}$ converges weakly to normal with mean zero and variance $\sigma_1^2 = \sigma^2 + \rho I_C^{-1} \rho^T$, where

$$\rho = E \left[\frac{\partial}{\partial \theta} \frac{-U_\alpha\{S_Y(y, \theta), S_Z(z, \theta), \alpha\}}{\sum_{i=1}^n V_\alpha\{S_Y(y, \theta), S_Z(z, \theta), \alpha\}} \right]$$

$$V_\alpha(u, v, \alpha) = \frac{\partial^2 \log l(u, v, \alpha)}{\partial \alpha^2}$$

and I_C is the Fisher information matrix of $L_C(\theta)$.

The details of the proof are given in supplementary materials, where we also show σ_1^2 can be consistently estimated by $\hat{\sigma}_1^2 = \hat{\sigma}^2 + \hat{\rho} \hat{I}_C^{-1} \hat{\rho}^T$.

An estimator for bivariate survival function could be obtained by plugging in consistent estimators for unknown quantities in $S_{Y,Z}(y, z) = C\{S_Y(y), S_Z(z), \alpha\}$. To be specific, the margins $S_Y(y)$ and $S_Z(z)$ are replaced by the proposed semiparametric consistent estimators, and α is replaced by the two-stage association estimator $\hat{\alpha}(\hat{\theta})$. The asymptotic properties of $\hat{S}_{Y,Z}(y, z)$ are summarized in Theorem 2, with the proof provided in supplementary materials.

Theorem 2—As $n \rightarrow \infty$, $\hat{S}_{Y,Z}(y, z)$ is a consistent estimator of $S_{Y,Z}(y, z)$, and the process $n^{1/2}\{\hat{S}_{Y,Z}(y, z) - S_{Y,Z}(y, z)\}$ converges weakly to a bivariate zero-mean Gaussian process with covariance function $[\partial C\{S_Y(y), S_Z(z), \alpha\}/\partial \alpha]^2 \sigma_1^2 + \sum (y, z)$.

It is often interesting to report Kendall's τ as a measure of association between bivariate survival data for convenient comparison across different copula models. With an estimator of the copula model association parameter α , a “plug-in” estimator of Kendall's τ can be obtained based on $\tau = 4 \int_0^1 \int_0^1 C(u, v, \alpha) dC(u, v, \alpha) - 1$. For certain copula models, there exists a further simplified one-to-one relationship between τ and α . While the asymptotic variance of τ may be obtained by the Delta method, the computation is rather complex and it is more convenient to use the bootstrap approach to compute the variance estimate.

For the semi-stationary copula model, we rely on a parametric specification of $G(t, \theta)$ to take advantage of the available information about the distribution of T . It is expected to be more efficient than the model when the distribution of T is totally unknown and nonparametrically estimated. Of course, it is important to check the validity of the assumption $H_0: T \sim G(t, \theta)$. This can be done by plotting the nonparametric maximum likelihood estimate $\hat{G}_n(t)$ against $G(t, \hat{\theta})$. Since T is also doubly truncated subject to the constraint $-Y \leq T \leq t_0 - Y$, estimating G is essentially a dual problem as estimating S_Y . Shen [17] provided an algorithm to jointly compute the nonparametric maximum likelihood estimators of both G and S_Y . In the data analysis, the plot is used as a graphical tool to examine the adequate fit of the parametric distribution of T .

4 The copula model with covariates

In this section, we propose to model the dependence of the copula association measure α on covariates by a transformation of γa , $\alpha = \eta(\gamma a)$, where $\eta(\cdot)$ is a known function, a is a $p \times 1$ vector of continuous or discrete covariates and γ is the corresponding $1 \times p$ vectors of regression coefficients. For example, $\eta(\cdot)$ corresponds to an exponential transformation, $\exp(\gamma a)$, for the Clayton copula. For the discrete covariate, we can either create a dummy variable for each level or assuming a linear trend across levels. Though the choice of $\eta(\cdot)$ depends on the form and specific constraints of the copula model, the inference procedure developed in this section is general for any type of the copula. The marginal distributions are modeled by proportional hazards models incorporating covariates in combination with the copula, under the semi-stationary condition. To be specific, bivariate survival distribution of (Y, Z) is modeled by the copula as $S_{Y,Z}(y, z) = C\{S_Y(y), S_Z(z), \alpha\}$, and marginally, we assume the failure times Y and Z satisfy the Cox proportional hazards models

$$h_Y(y|a) = h_{0Y}(y) \exp(\beta_Y a)$$

$$h_Z^*(z|a, m) = h_{0Z}^*(Z) \exp(\phi a + \mu * m),$$

where β_Y and ϕ are the corresponding $1 \times p$ vectors of regression coefficients, μ^* is the regression coefficient for the calendar time of the first event M , $h_{0Y}(y)$ and $h_{0Z}^*(z)$ are unspecified baseline hazard functions for Y and Z , respectively. Denote $\beta_Z = (\phi, \mu^*)$ and $b = (a, M)$. Note that the Cox model of Z is specified given both a and M to handle dependent censoring on Z . The marginal survival functions $S_Y(y)$ and $S_Z(z)$ will be estimated from these two proportional hazards models.

Similar to the discussion in Section 3 where there is no covariate, bivariate survival data (Y, Z) are collected with a selection probability equals $w(y, \theta) = G(t_0 - y, \theta) - G(-y, \theta)$ due to the interval sampling, so we adjust for the sampling bias by the inverse probability weighting method. First consider θ is known. For the first failure time Y , define $R_{j1} = \{i : y_i \leq y_j\}$ as the risk set at y_j . Define a weighted partial score function as

$$U_{j1}(\theta, \beta_Y) = a_j - \left\{ \frac{\sum_{i \in R_{j1}} w(y_i, \theta)^{-1} \exp(\beta_Y a_i) a_i}{\sum_{i \in R_{j1}} w(y_i, \theta)^{-1} \exp(\beta_Y b_i)} \right\} \quad (3)$$

Let $U_1(\theta, \beta_Y) = \sum_j w(y_j, \theta)^{-1} U_{j1}(\theta, \beta_Y)$, and an estimating equation for β_Y is obtained as $U_1(\theta, \beta_Y) = 0$, with its solution denoted by $\hat{\beta}_Y(\theta)$. The inverse of the selection probability $w(y, \theta)$ is used to remove the sampling bias, since the probability of observing y is proportional to $w(y, \theta)$. For the second failure time Z , we observe $\{(x, \delta)\}$ where $x = \min(z, c)$ and $\delta = I(z \leq c)$. A weighted estimating equation for β_Z , $U_2^*(\theta, \beta_Z)$, can be constructed, similar to the one for the working Cox model in Section 3. The solution is denoted by $\hat{\beta}_Z(\theta)$. Now suppose θ is unknown. We replace θ by the conditional maximum likelihood estimator $\hat{\theta}$ and derive estimators of β_Y and β_Z by solving the corresponding equations $U_1(\hat{\theta}, \beta_Y) = 0$ and $U_2^*(\hat{\theta}, \beta_Z) = 0$, respectively. The solutions are denoted by $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\beta}_Z(\hat{\theta})$, which enjoy nice asymptotic properties. Then from the proportional hazards models, marginal survival functions $S_Y(y)$ and $S_Z(z)$ can be consistently estimated by $\hat{S}_Y(y, \hat{\theta}) = \hat{S}_{0Y}(y) \exp\{\hat{\beta}_Y(\hat{\theta})a\}$ and $\hat{S}_Z(z, \hat{\theta}) = \hat{S}_{0Z}(z) \exp\{\hat{\beta}_Z(\hat{\theta})b\}$, where $\hat{S}_{0Y}(y)$ and $\hat{S}_{0Z}(z)$ are the Breslow estimators for baseline survival functions. The details of estimation procedures for $\hat{\beta}_Z(\hat{\theta})$, $\hat{\beta}_Y(\hat{\theta})$, $\hat{\beta}_Z(\hat{\theta})$, $\hat{S}_Y$ and $\hat{S}_Z$, the asymptotic properties and proof are provided in supplementary materials.

To estimate γ , using steps similar to those in Section 3, a pseudo-conditional likelihood score equation is constructed by replacing $S_Y(y)$ and $S_Z(z)$ with $\hat{S}_Y(y, \hat{\theta})$ and $\hat{S}_Z(z, \hat{\theta})$, a, m ,

$$U_{\gamma}\{\hat{S}_Y(\hat{\theta}), \hat{S}_Z(\hat{\theta}), \gamma\}^{(p)} = \frac{\partial}{\partial \gamma} \left(\sum_{i=1}^n \delta_i \log [c\{\hat{S}_Y(y_i, \hat{\theta}|a_i), \hat{S}_Z(x_i, \hat{\theta}|\alpha_i, m_i), \eta(\gamma a_i)\}] + (1-\delta_i) \log \left[\frac{\partial C\{\hat{S}_Y(y_i, \hat{\theta}|a_i), \hat{S}_Z(x_i, \hat{\theta}|a_i, m_i), \eta(\gamma a_i)\}}{\partial u_i} \right] \right) \\ = 0$$

Denote its solution by $\hat{\gamma}(\hat{\theta})$, for which the asymptotic properties can be developed using the approach similar to those steps in Section 3. Therefore, as $n \rightarrow \infty$, $\hat{\gamma}(\hat{\theta})$ is a consistent estimator of γ_0 , and $n^{1/2}\{\hat{\gamma}(\hat{\theta}) - \gamma_0\}$ converges weakly to a mean zero multivariate normal distribution with variance-covariance matrix Σ_{γ} . Given that asymptotic variance-covariance matrix involves second-order derivatives, the estimation of Σ_{γ} is not straightforward; therefore, we estimate it by a bootstrap procedure. The variance for the association measure α can be obtained by the Delta method, or alternatively, we can estimate its variance by the sample variance over the bootstrap samples. The bivariate survival function can also be estimated based on the copula model by substituting $S_Y(y)$, $S_Z(z)$, and γ with their corresponding estimators derived in this section.

5 Simulation studies

The first set of simulations is carried out to assess the performance of the proposed estimation and inference procedures for the copula model without covariates under a moderate sample size. Specifically, we examine finite sample properties of the proposed estimators for marginal survival functions, association parameter and joint survival function. A set of data $\{(t_1, y_1, z_1), \dots, (t_n, y_n, z_n)\}$ is generated as follows. Define $T = -3W + 10$, where $W \sim \exp(\theta)$ with $\theta = 1.0$ and 4.0 , corresponding to a decreasing density function similar to the setting in the real data application, and moderate-to-heavy censoring. Let bivariate failure times (Y, Z) be generated from three Archimedean copula models: the Clayton, positive stable and Frank copulas. We choose unit exponential margins and three different values for α in each of the copula models, in order to accommodate different levels of dependence of Y and Z . An observation (t, y, z) is included in the interval sampled dataset if and only if $0 \leq t + y \leq 10$, and is censored if $t + y + z \leq 10$ or $z \leq c_1$, where c_1 is generated randomly from $Uniform(0, 4)$. When $\theta = 1.0$, the truncation rate is about 25–30% and the censoring rate is about 30–35%. When $\theta = 4.0$, the truncation rate is about 55–60% and the censoring rate is about 45–50%. For each choice of parameters (θ, α) , 1,000 simulated samples are generated with sample size $n = 400$.

The proposed estimators for the marginal survival functions of Y and Z are evaluated in Figure 2, where bivariate failure times (Y, Z) are generated from a Clayton copula model with $\alpha = 3$. The plots are based on the means of 1,000 simulations, and demonstrate that the weighted empirical survival function outperforms the empirical one in estimating $S_Y(y)$, and the weighted estimator from a working Cox model outperforms the Kaplan–Meier estimator in estimating $S_Z(z)$. The copula association parameter estimate is obtained as $\hat{\alpha}(\hat{\theta})$ by solving eq. (2). Particularly, the bivariate survival function estimator $\hat{S}_{Y,Z}(y, z)$ is assessed at $(y, z) = (0.22, 0.51)$, denoted by $\hat{S}_1$. We also estimate Kendall's τ as another measure of association. Table 1 provides the simulation results about $\hat{\theta}$, $\hat{\alpha}(\hat{\theta})$, $\hat{S}_1$ and $\hat{\tau}$ for the copula model without covariates, including the empirical bias, average model-based

standard error, empirical standard error and 95% coverage probability. The confidence interval is constructed using the estimated asymptotic variance, and the empirical estimate of the 95% coverage probability is obtained based on the confidence interval over 1,000 replications. It shows the proposed estimators $\hat{\theta}$, $\hat{\alpha}(\hat{\theta})$, $\hat{S}_1$ and $\hat{\tau}$ work well with fairly small biases. For the association estimator $\hat{\alpha}(\hat{\theta})$, the average model-based standard error is very close to the empirical standard error, which implies satisfactory performance of the inferential result of it. The coverage probabilities are all quite close to 95%. Moreover, the estimated standard error of $\hat{\alpha}(\hat{\theta})$ increases in general with stronger dependence of bivariate data (Y, Z) indicated by a larger absolute value of α . This phenomenon is not very surprising since greater variations are usually expected for larger values.

The second set of simulations is conducted to examine the finite sample performance of parameters in the copula model with covariates. The data-generating procedure generally follows that in the first set of simulations. The differences are, first of all, the association measure α of each copula is allowed to depend on a covariate $A \sim \text{Bernoulli}(1/3)$. For the Clayton and positive stable copulas, α is modeled on the log scale as $\log(\alpha) = \gamma_0 + \gamma_1 a$. For the Frank copula, α is modeled on the original scale as $\alpha = \gamma_0 + \gamma_1 a$. We set slope $\gamma_1 = 1/8$ and choose different values of intercept γ_0 to achieve different levels of the association of Y and Z . Second, for marginal failure times Y and Z , we use the Weibull densities to construct the proportional hazards models

$$\begin{aligned} h_Y(y|a) &= h_{0Y}(y) \exp(\beta_Y a) \\ h_Z(z|a) &= h_{0Z}(z) \exp(\phi a) \end{aligned}$$

where $h_{0Y}(y) = 2y$, $h_{0Z}(z) = 2z$, and $\beta_Y = \phi = 0.2$. Finally, for the distribution of T , we set $\theta = 2.0$. An observation (a, t, y, z) is included in the interval sampled dataset if and only if $0 \leq t + y \leq 10$, and is censored if $t + y + z \leq 10$ or $z \leq c_1$, where c_1 is generated randomly from $\text{Uniform}(0, 4)$. The truncation rate is around 40% and the censoring rate is around 50%. For each setting, 1,000 simulated samples are generated with sample size $n = 400$. We report the association estimate in terms of Kendall's τ instead of the copula association parameter α for convenient comparison across different copula families, and study the bias and variance for $\hat{\tau}$ in different scenarios. For each dataset, the value of $\hat{\tau}$ is computed based on the estimated parameters. The variance estimation is obtained by the bootstrap approach, where 500 bootstrap samples are used. Table 2 summarizes the simulation results about $\hat{\theta}$, $\hat{\beta}_Y(\hat{\theta})$, $\hat{\phi}(\hat{\theta})$, and $\hat{\tau}$ for the copula model with covariates. Under all the simulation scenarios, $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\phi}(\hat{\theta})$ obtained from the weighted estimating equations perform well with small biases and variances. For the association estimate $\hat{\tau}$, the bias is generally small, the magnitudes of the empirical standard error and average bootstrap standard error are similar and the coverage probabilities are close to 95%. This demonstrates the proposed inference procedure for the copula model with covariates is reasonably good and suggests that the bootstrap variance estimator provides an appropriate measure of the variability of $\hat{\tau}$.

6 Application to the Rakai HIV study

6.1 Overall association

The HIV seroconversion data from the Rakai HIV study provide an example of bivariate survival data with interval sampling. In this study, 837 subjects were ascertained with a documented date of HIV seroconversion between 1995 and 2003, and followed until they died or by the end of 2003. Among them, 120 died and others were censored by out-migration or the end of the follow-up. The information on date of birth, date of death, sex, place of residence and HIV subtype is available. The bivariate survival times of interest are age at HIV infection and residual lifetime. Exclusion of subjects who were infected before 1995 or after 2003 results in selection bias of the interval sampling. For the purposes of illustration, we apply the proposed semi-stationary copula model methods to analyze the Rakai HIV seroconversion data, address statistical issues of the interval sampling and study the association between age at HIV infection and residual lifetime among HIV seroconverters. The data and analysis method allow one to model the HIV epidemic for treatment-naïve individuals, which would help provide guidance on the initiation of ART.

In the analysis we assume the semi-stationary condition holds, that is, the progression of HIV is independent of the birth time of the study cohort. Denote the birth time of HIV seroconverters by T with distribution function $G(t)$, age at infection by Y , residual lifetime after infection by Z and calendar time of HIV infection by $M = T + Y$. Given that the HIV seroconversion was identified between 1995 and 2003, and the cohort subjects' age ranges from 15 to 50 years old, we let the support of $G(t)$ start from -50 , corresponding to the birth year from 1945, to avoid non-identifiability problem regarding the estimation of $G(t)$. Recall that we assume the parametric distribution of T is known, and two polynomial functions are used to model the density of birth time T : a linear model $g(t) = k + \theta_1 t$, and a quadratic model $g(t) = k + \theta_1 t + \theta_2 t^2$, where k is a given positive-valued constant in both models and $t \in [-50, 50]$. The choice of the parametric form of the distribution of T is examined by comparing the parametric estimate of the distribution function $G(t)$ with its nonparametric maximum likelihood estimate [17]. Figure 3(a) plots the parametric estimators, the nonparametric maximum likelihood estimator and the empirical estimator of the density function for T . It demonstrates substantial discrepancies between the empirical estimator and other ones, while a similar decreasing trend in the birth rate of HIV seroconverters is found by the two polynomial models and the nonparametric method. As discussed in Section 3, T is doubly truncated with the sampling constraint $-Y \leq T \leq t_0 - Y$; therefore, the sampling density of T is generally biased and it is not appropriate to use the empirical method to estimate the birth rate. Further, the decreasing trend may partly reflect the change in the population under surveillance, such as trends in HIV incidence and prevalence. Actually, the HIV incidence in Rakai declined from approximately 2.0 per 100 person-years in 1995 to 1.3 per 100 person-years in 2003, and the HIV prevalence declined from approximately 18% in 1995 to 13% in 2003 [4]. Given the small difference between the two polynomial model fits as well as the closer pattern between the nonparametric and quadratic model estimates, we choose the quadratic model for birth density. The parameter estimates together with their estimated standard errors are $\hat{\theta}_1 = -2.00 \times 10^{-3} (4.35 \times 10^{-4})$ and $\hat{\theta}_2 = 8.30 \times 10^{-5} (2.17 \times 10^{-6})$. In addition, the parametric assumption of the distribution of T , $H_0: T \sim G(t, \theta)$, is checked in Figure 3(b)

by plotting the nonparametric maximum likelihood estimator $\hat{G}_n(t)$ against $\alpha(t, \hat{\theta})$ and it suggests the assumption of quadratic birth density is considerably reasonable.

Next, the marginal survival functions $S_Y(y)$ and $S_Z(z)$ for age at HIV infection and residual lifetime are estimated by the weighted empirical survival function and weighted estimator from a working Cox model, adjusting for the selection bias from interval sampling. Figure 4(a) shows that the empirical method overestimates the marginal survival of age at infection comparing to the weighted one, and Figure 4(b) shows that the Kaplan–Meier estimator overestimates the marginal survival of residual lifetime comparing to the weighted one from the working Cox model.

Previous analysis [4] showed survival time decreased significantly with older age at infection based on a Cox proportional hazards model of residual lifetime conditional on age at infection. However, the appropriateness of the Cox model is under investigation since it did not take into account the fact that the data are collected under the interval sampling. As discussed, due to the interval sampling, age at infection is doubly truncated and residual lifetime is observed subject to dependent right censoring. Therefore, selection bias needs to be corrected in analyzing the data and studying the association between age at infection and residual lifetime. We consider the proposed copula model without covariates in Section 3, where the dependency structure is fitted by the Frank copula, and assess the overall association. To estimate the standard error of the association estimator, we adopt a nonparametric bootstrap method by sampling 837 subjects with replacement from the dataset. The resampling procedure is repeated 500 times. The confidence interval is constructed based on the asymptotic normality, where the standard error is computed using bootstrap resamples. The association parameter α is estimated as -0.19 with the 95% confidence interval $(-1.21, 0.85)$. The corresponding Kendall's τ is estimated as -0.02 with the confidence interval $(-0.13, 0.09)$. Differing from the result of the previous study, our analysis suggests a non-significant negative association between age at infection and residual lifetime among HIV seroconverters after adjusting for the sampling bias. Further, the relationship between bivariate survival times is explored graphically in Figure 4(c) by plotting estimated marginal survival functions of residual lifetime for two categories of age at infection: < 30 years and ≥ 30 years, as well as in Figure 4(d) by plotting estimated conditional survival functions of residual lifetime given these two categories. Figure 4(c) demonstrates a slightly negative association and a trend towards lower survival probability with older age at infection. However, Figure 4(d) shows that the estimated survival probability conditional on age at infection ≥ 30 years is comparable to that conditional on age at infection < 30 years, which may explain the low degree and non-significance of negative association as the estimation of α shows.

6.2 Subtype-stratified association and subtype-adjusted association

Studies suggest that the progression of HIV infection is affected by the HIV subtype [5]. HIV subtypes differ in biological characteristics that may affect pathogenicity, such as viral fitness and plasma viral loads. These differences may theoretically influence virus infectivity and transmissibility. We investigate this issue by analyzing the Rakai HIV seroconversion data with information on HIV subtype. Among 837 HIV seroconverters, 413 individuals'

HIV subtypes could be identified because their blood serum samples had sufficient HIV RNA for reverse transcriptase polymerase chain reaction amplification. Subtypes were classified as A (15.4%), C (0.5%), D (58.3%), AD recombinants (20.2%) and multiple infections (5.6%). Earlier analysis of the Rakai data suggests that subtypes D, AD recombinants and multiple infections have similar disease progression rates and there is only one individual with subtype C infection in this data set, so for the analysis purposes we compare A subtype with combined non-A virus subtypes [4].

First of all, we stratify by the HIV subtype to assess the association between age at infection and residual lifetime. The estimated marginal survival functions of age at infection and residual lifetime by the HIV subtype are shown in Figure 5. It demonstrates that survival curves of age at infection for A subtype, non-A subtypes and unknown subtype are almost the same, but the survival probability of residual lifetime is substantially lower for non-A and unknown subtypes compared with A subtype. In fact, there are only 2 deaths among 64 subtypes of A infections, compared with 45 deaths among 349 non-A subtypes infections. It is consistent with the result from the previous study in Uganda [5] that A subtype has a slower disease progression rate, and it is thought to be less pathogenic than other subtypes. The association measure α in the Frank copula model is estimated as 2.96 with the 95% confidence interval $(-0.56, 6.48)$ for A subtype, -0.38 with $(-1.64, 0.93)$ for non-A subtypes and -0.36 with $(-1.16, 0.48)$ for unknown subtype. The corresponding Kendall's τ is estimated as 0.30 with the confidence interval $(-0.06, 0.54)$ for A subtype, -0.04 with $(-0.18, 0.10)$ for non-A subtypes and -0.04 with $(-0.13, 0.05)$ for unknown subtype. Very interestingly, it shows a comparable negative association for non-A and unknown subtypes and conversely a positive association for A subtype, though the associations are not significant. The result suggests that the Rakai HIV epidemic probably has a predominance of non-A subtypes infection and subtype A appears to be a very different virus subtype in the HIV progression from other subtypes, which is consistent with the conclusions from other studies.

Next, we examine the relationship between age at infection and residual lifetime adjusting for the HIV subtype by the copula model with covariates in Section 4. The dependence structure in the Frank copula is allowed to depend on the HIV subtype through $\alpha = \gamma_0 + \gamma_1 a$, where $a = (a_1, a_2)$, a_1 denotes the indicator for A subtype and a_2 denotes the indicator for non-A subtypes. An individual's subtype is unknown if $a = (0, 0)$. Marginally, age at infection and residual lifetime given the HIV subtype are modeled through the Cox proportional hazards models: $h_Y(y|a) = h_{0Y}(y) \exp(\beta_Y a)$ and

$h_z^*(z|a, m) = h_{0z}^*(z) \exp(\phi a + \mu^* m)$ where m is the calendar time of HIV infection. From weighted estimating equations, the proposed estimate vector for β_Y is $(0.05, -0.07)$ with P -values > 0.05 and that for ϕ is $(-0.41, 0.29)$ with P -values < 0.05 . The values of estimates for β_Y and ϕ (the estimates of the covariate effects on each failure time) suggest similar survival patterns among HIV subtype groups as shown in Figure 5, which indicates no significant difference in age at infection by the HIV subtype, significant lower risk of death comparing A subtype to unknown subtype, and significant higher risk of death comparing non-A subtypes to unknown subtype. After covariate adjustment, the copula association parameter α is estimated as 4.31 with the 95% confidence interval $(1.98, 6.64)$ for A subtype, -0.95

with $(-2.30, 0.52)$ for non-A subtypes and -0.44 with $(-1.76, 0.98)$ for unknown subtype. Correspondingly, Kendall's τ is estimated as 0.41 with the confidence interval $(0.21, 0.55)$ for A subtype, -0.10 with $(-0.24, 0.06)$ for non-A subtypes and -0.05 with $(-0.19, 0.11)$ for unknown subtype. The confidence interval is computed based on 500 realizations of the resampling method. For each subtype, the adjusted association estimate suggests the same direction of the association as its unadjusted counterpart but a larger magnitude. The patterns of the association are similar between non-A subtypes and unknown subtype, both of which are non-significantly negative. It shows A subtype presents a significant positive association, which again illustrates that A subtype is very different from other HIV subtypes in the progression of disease.

7 Concluding remarks

This paper considers statistical issues on bivariate survival data with interval sampling, which arises commonly in disease registries or surveillance systems where data are collected conditional on the first failure event occurring within a specific time interval. In this paper, the semi-stationary condition is assumed for statistical modeling and inference. It is important to indicate that the semi-stationary assumption could be violated when, for instance, improved diagnostic strategies over time lead to earlier detection, or an effective treatment becomes available and is given to the diseased individuals during the process of observation. The situation when (S) does not hold is beyond scope of this paper and will be explored in future research. Under the semi-stationary assumption, we investigated the association between bivariate survival data with interval sampling by the copula models without and with adjustment for covariates. Since the asymptotic variance for parameters in the copula model with covariates has a rather complicated form and possibly involves the censoring distribution, the bootstrap method is applied as a direct and robust way to compute standard errors.

The data structure considered in this paper assumes that the first failure time Y is observed exactly. However, as pointed out by one reviewer, there are situations in some HIV cohorts where Y is interval censored. The interval censoring problem of Y is worthy of further investigation and may be handled by extending the prior work on bivariate survival distribution estimation for interval-censored outcomes [18]. For convenience of discussion, the proportional hazards model is used to model the relationship between each failure time and covariates. In fact, any regression model for survival data, such as the semiparametric transformation model or accelerated failure time model, can be used. The proposed copula model framework is very general and can be modified to accommodate other regression model for the marginal distribution. Moreover, it is noted that the same covariates may not be relevant as causal factors to the two failure events of interest, as well as the association between them. Nevertheless, in development of the method, we consider a general case by including the same set of covariates in modeling marginal distributions and the association, which would allow one to test the significance and estimate the covariate effects. In real data applications, one can always start with a general model by including all factors of interest as covariates and only keep the significant ones in the final model.

In the simulations and data application, certain specific copula models are used to characterize the dependence structure of bivariate survival data given their popularity, modeling flexibility and computational convenience. The simulations show knowing the true copula, the estimation procedure provides good results, and we anticipate that the procedure will also perform well for other copulas. While in fact, any copula model could be considered and this raises closely related issues on how a wrong choice of the copula model would affect the estimated association measure and how to choose an appropriate copula model. Since different copula models may lead to different dependence properties of bivariate survival function, the problem of model selection of the copulas needs to be addressed in future work. Potentially, Goodness-of-fit procedures for the copulas could be developed for bivariate survival data with interval sampling. In absence of covariates, we suggest to compare the copula model fit with some nonparametric estimates of the association, such as cross-ratio function or Kendall's τ . Therefore, the nonparametric association estimation for such data may shed some light on it but is still under development.

The research is motivated by and applied to the Rakai HIV seroconversion data to evaluate the association between age at HIV infection and residual lifetime among treatment-naïve HIV seroconverters, and study how the association varies by the HIV subtype and changes after controlling for the HIV subtype. Another scientifically interesting problem for further research is to examine the ART effect on the HIV progression. In the Rakai Health Science Program, the ART became available in 2004 and this time-dependent treatment variable would further complicate the analysis.

Acknowledgments

This work was supported in part by the Cancer Center Support Grant from the National Cancer Institute awarded to the Harold C. Simmons Cancer Center at the University of Texas Southwestern Medical Center. 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. We also thank the Rakai Health Sciences Program at Johns Hopkins Bloomberg School of Public Health for providing the data.

References

1. Zhu H, Wang M-C. Analyzing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika*. 2012; 99:345–61. [PubMed: 23843662]
2. Bilker WB, Wang M-C. A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics*. 1996; 52:10–20. [PubMed: 8934583]
3. Brookmeyer R. AIDS, epidemics, and statistics. *Biometrics*. 1996; 52:781–96. [PubMed: 8805756]
4. Lutalo T, Gray RH, Wawer M, Sewankambo N, Serwadda D, Laeyendecker O, et al. Survival of HIV-infected treatment-naïve individuals with documented dates of seroconversion in Rakai, Uganda. *AIDS*. 2007; 21:S15–9. [PubMed: 18032934]
5. Kaleebu P, Ross A, Morgan D, Yirrel D, Oram J, Rutebemberwa A, et al. Relationship between HIV-1 ENV subtypes A and D and disease progression in a rural Ugandan cohort. *AIDS*. 2001; 15:293–9. [PubMed: 11273208]
6. Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variable. *Biometrika*. 1998; 85:785–96.
7. Lin D-Y, Sun W, Ying Z. Nonparametric estimation of gap time distributions for serial events with censored data. *Biometrika*. 1999; 86:59–70.
8. Schaubel DE, Cai J. Nonparametric estimation of gap time survival functions for ordered multivariate failure time data. *Stat Med*. 2004; 23:1885–900. [PubMed: 15195322]

9. Visser M. Nonparametric estimation on the bivariate survival function with application to vertically transmitted AIDS. *Biometrika*. 1996; 83:507–18.
10. Shih JH, Louis TA. Inferences on the association parameters in copula models for bivariate survival data. *Biometrics*. 1995; 51:1384–99. [PubMed: 8589230]
11. Wang W. Estimating the association parameter for copula models under dependent censoring. *J R Stat Soc B*. 2003; 65:257–73.
12. Lakhal-Chaieb L, Cook R, Lin X. Inverse probability of censoring weighted estimates of Kendall's τ for gap time analyses. *Biometrics*. 2010; 66:1145–52. [PubMed: 20337629]
13. Genest C, Ghoudi K, Rivest L-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*. 1995; 82:543–52.
14. Wang W, Ding AA. On assessing the association for bivariate current status data. *Biometrika*. 2000; 87:879–93.
15. Aalen OO. Weak convergence of stochastic integrals related to counting process. *Z Wahrsch Ver Geb*. 1977; 38:261–77.
16. Zeng D. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann Stat*. 2004; 32:1533–55.
17. Shen P-S. Nonparametric analysis of doubly truncated data. *Ann Inst Stat Math*. 2008; 62:835–53.
18. Betensky RA, Finkelstein DM. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Stat Med*. 1999; 18:3089–100. [PubMed: 10544308]
19. Van der Vaart, AW. Asymptotic statistics. Cambridge: Cambridge University Press; 1998.
20. Qi L, Wang CY, Prentice RL. Weighted estimator for proportional hazards regression with missing covariates. *J Am Stat Assoc*. 2005; 100:1250–63.

Appendix A: Asymptotic properties of $\hat{\alpha}(\theta)$ in the copula model without covariates

Proof of Theorem 1

Assume that the standard regularity conditions for the maximum likelihood estimator hold and the functions $W_{\alpha}\{S_Y(y), S_Z(z), \alpha\}$, $V_{\alpha}\{S_Y(y), S_Z(z), \alpha\}$, $V_{\alpha,1}\{S_Y(y), S_Z(z), \alpha\}$, and $V_{\alpha,2}\{S_Y(y), S_Z(z), \alpha\}$ are continuous and bounded for $(y, z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$, where

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

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

In the following, we study the asymptotic properties of $\hat{\alpha}(\hat{\theta})$.

If θ is known, as $n \rightarrow \infty$, $\hat{\alpha}(\theta) - \alpha_0$ converges to 0 in probability, and $n^{1/2}\{\hat{\alpha}(\theta) - \alpha_0\}$ converges weakly to normal distribution with mean zero and variance σ^2 . The large sample properties of $\hat{\alpha}(\theta)$ can be proved following the lines in Zhu and Wang [1], where marginal survival functions $S_Y(y)$ and $S_Z(z)$ are replaced by the corresponding consistent estimators involving θ . Observe that

$$n^{1/2}\{\hat{\alpha}(\hat{\theta})-\alpha_0\}=n^{1/2}\{\hat{\alpha}(\theta)-\alpha_0\}+n^{1/2}\{\hat{\alpha}(\hat{\theta})-\hat{\alpha}(\theta)\} \quad (4)$$

As previously discussed, if θ is known, the first term in eq. (4) converges weakly to a normal distribution with mean 0 and variance σ^2 . By the counting process methodology [19], it is asymptotically equivalent to a sum of n i.i.d. zero-mean random variables, expressed as

$$n^{1/2}\{\hat{\alpha}(\theta)-\alpha_0\}=n^{-1/2}\sum_{i=1}^n\phi_i(\alpha,\theta)+o_p(1) \quad (5)$$

where $\phi_i(\alpha,\theta)=-U_\alpha\{S_Y(y,\theta),S_Z(z,\theta),\alpha\}[\sum_{i=1}^nV_\alpha\{S_Y(y,\theta),S_Z(z,\theta),\alpha\}^{-1}$ and for each θ , $E\{\phi(\alpha,\theta)\}=0$.

To develop the asymptotic results of the second term in eq. (4), the additional variation created by estimating θ by $\hat{\theta}$, the conditional maximum likelihood estimator based on $L_\alpha(\theta)$, needs to be handled. Let $\rho=E\{\phi(\alpha,\theta)/\theta\}$, under appropriate regularity conditions, the second term can be approximated by a sum of n i.i.d. zero-mean random variables, expressed as

$$n^{1/2}\{\hat{\alpha}(\hat{\theta})-\hat{\alpha}(\theta)\}=n^{-1/2}\rho I_C^{-1}\sum_{i=1}^n\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i)+o_p(1) \quad (6)$$

which converges weakly to a normal distribution with mean 0 and variance $\rho I_C^{-1}\rho^T$. Thus, $\hat{\alpha}(\hat{\theta})-\hat{\alpha}(\theta)$ converges to 0 in probability. Therefore, $\hat{\alpha}(\hat{\theta})-\alpha_0=\{\hat{\alpha}(\theta)-\alpha_0\}+\{\hat{\alpha}(\hat{\theta})-\hat{\alpha}(\theta)\}$ converges to 0 in probability. This completes the proof of consistency of $\hat{\alpha}(\hat{\theta})$.

Combining the preceding results of eqs (5) and (6), we get

$$n^{1/2}\{\hat{\alpha}(\hat{\theta})-\alpha_0\}\cong n^{-1/2}\sum_{i=1}^n\phi_i(\alpha,\theta)+n^{-1/2}\rho I_C^{-1}\sum_{i=1}^n\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i) \quad (7)$$

Also the corresponding distributions of those two terms are asymptotically orthogonal to each other, since

$$E\{\phi_i(\alpha,\theta)\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i)\}=E\left[\phi_i(\alpha,0)E\left\{\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i)|Y_i\right\}\right]=0 \quad (8)$$

Equations (7) and (8) imply that $n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\}$ is asymptotically equivalent to a sum of n i.i.d. zero-mean random variables. By the central limit theorem, it converges weakly to a normal random variable with mean zero and variance $\sigma_1^2 = \sigma^2 + \rho I_C^{-1} \rho^T$. It is natural to estimate σ_1^2 by $\hat{\sigma}_1^2 = \hat{\sigma}^2 + \hat{\rho} \hat{I}_C^{-1} \hat{\rho}^T$, where $\hat{\sigma}^2$ is a consistent estimator of σ^2 , and $\hat{\rho}$ and $\hat{I}_C$ are the corresponding moment-type empirical estimators. The consistency of $\hat{\sigma}^2$, $\hat{\rho}$ and $\hat{I}_C$ implies that $\hat{\sigma}_1^2$ is a consistent estimator of σ_1^2 .

Appendix B: Asymptotic properties of $\hat{S}_{Y,Z}(y, z)$ in the copula model without covariates

Proof of Theorem 2

Consider an estimate of $S_{Y,Z}(y, z)$ at any given time $(y, z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$ is obtained under the copula model, as $\hat{S}_{Y,Z}(y, z) = C\{\hat{S}_Y(y), \hat{S}_Z(z), \hat{\alpha}(\hat{\theta})\}$, where $\hat{\alpha}(\hat{\theta})$ is the two-stage association parameter estimator, $\hat{S}_Y(y)$ and $\hat{S}_Z(z)$ are the proposed semiparametric consistent estimators. The asymptotic results of $\hat{S}_{Y,Z}(y, z)$ are proved under the assumed conditions stated in the proof of Theorem 1 and the following regularity conditions. Assume that the copula function $C(u, v, \alpha)$ is continuous and differentiable at u, v, α , respectively, and the parameter α lies in a compact set.

First of all, we show the consistency of $\hat{S}_{Y,Z}(y, z)$. We have that $\hat{S}_Y(\cdot)$ converges in probability to $S_Y(\cdot)$ uniformly in $[y_-, y_+]$, and $\hat{S}_Z(\cdot)$ converges in probability to $S_Z(\cdot)$ uniformly in $[z_-, z_+]$. Also by Theorem 1, $\hat{\alpha}(\hat{\theta})$ converges in probability to α_0 . Since the copula function $C(u, v, \alpha)$ is a continuous function of u, v and α , then $C\{\hat{S}_Y(y), \hat{S}_Z(z), \hat{\alpha}(\hat{\theta})\}$ converges in probability to $C\{S_Y(y), S_Z(z), \alpha_0\}$ uniformly in $\mathcal{A} = [y_-, y_+] \times [z_-, z_+]$. Therefore, As $n \rightarrow \infty$, $\hat{S}_{Y,Z}(y, z)$ converges to $S_{Y,Z}(y, z)$ in probability.

Next, we illustrate the asymptotic distribution of $\hat{S}_{Y,Z}(y, z)$. Using the functional delta method on $C\{\hat{S}_Y(y), \hat{S}_Z(z), \hat{\alpha}(\hat{\theta})\}$ around α_0, S_Y and S_Z , we get

$$\begin{aligned} & n^{1/2}[C\{\hat{S}_Y(y), \hat{S}_Z(z), \hat{\alpha}(\hat{\theta})\} - C\{S_Y(y), S_Z(z), \alpha_0\}] \\ & \cong n^{1/2} \frac{\partial C\{S_Y(y), S_Z(z), \alpha\}}{\partial \alpha} \{\hat{\alpha}(\hat{\theta}) - \alpha_0\} \\ & + n^{1/2} \frac{\partial C\{S_Y(y), S_Z(z), \alpha_0\}}{\partial u} (\hat{S}_Y - S_Y)(y) \\ & + n^{1/2} \frac{\partial C\{S_Y(y), S_Z(z), \alpha_0\}}{\partial v} (\hat{S}_Z - S_Z)(z) \end{aligned} \quad (9)$$

By Theorem 1, $n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\}$ converges weakly to normal with mean zero and variance σ_1^2 . Therefore, the first term in eq. (9) is asymptotically equivalent to

$$n^{-1/2} \frac{\partial C\{S_Y(y), S_Z(z), \alpha\}}{\partial \alpha} \sigma_1 \sum_{i=1}^n \frac{\partial \{\hat{S}_Y(y), \hat{S}_Z(z), \alpha\}}{\partial \alpha} \quad (10)$$

which is a sum of n i.i.d. random variables. Applying the counting process asymptotic techniques to $\hat{S}_Y$ and $\hat{S}_Z$, the sum of the second and the third terms in eq. (9) is asymptotically equivalent to

$$n^{-1/2} \left[\sum_{i=1}^n \frac{\partial C\{S_Y(y), S_Z(z), \alpha_0\}}{\partial u} I_1^0(Y_i)(y) + \frac{\partial C\{S_Y(y), S_Z(z), \alpha_0\}}{\partial v} I_2^0(X_i, \delta_i)(z) \right] \quad (11)$$

where

$I_1^0(Y_i)(y) = -S_Y(y) [\int_0^y \{dN_{1i}(u)\} \{p(Y \geq u)\}^{-1} - \int_0^y \{I(Y_i \geq u) d\Lambda_1(u)\} \{p(Y \geq u)\}^{-1}]$ and $I_2^0(X_i, \delta_i)(z) = -S_Z(z) [\int_0^z \{dN_{2i}(u)\} \{p(Z \geq u, C_2 \geq u)\}^{-1} - \int_0^z \{I(X_i \geq u) d\Lambda_2(u)\} \{p(Z \geq u, C_2 \geq u)\}^{-1}]$ with $N_{1i}(u) = I(Y_i \leq u)$, $N_{2i}(u) = I(Z_i \leq u, \delta_i = 1)$ and $C_2 = C - T - Y$. Note that eq. (11) is a sum of n i.i.d. random variables, and the expectation of each term in eq. (11) is zero. By the central limit theorem, eq. (11) converges weakly to normal with zero mean and variance $\Sigma(y, z)$.

Moreover, we have

$$\begin{aligned} & E \left(\left[\frac{\partial \log l\{\hat{S}_Y(y), \hat{S}_Z(z), \alpha\}}{\partial \alpha} \right] \{I_1^0(Y_i) + I_2^0(X_i, \delta_i)\} \right) \\ &= E \left(\{I_1^0(Y_i) + I_2^0(X_i, \delta_i)\} E \left[\frac{\partial \log l\{\hat{S}_Y(y), \hat{S}_Z(z), \alpha\}}{\partial \alpha} \middle| Y_i, X_i, \delta_i \right] \right) = 0 \end{aligned} \quad (12)$$

which means eqs (10) and (11) are asymptotically orthogonal. Therefore, eqs (10), (11) and (12) imply that as $n \rightarrow \infty$, the process $n^{1/2}\{\hat{S}_{Y,Z}(y, z) - S_{Y,Z}(y, z)\}$ converges weakly to a bivariate zero-mean Gaussian process with $[\partial C\{S_Y(y), S_Z(z), \alpha\} / \partial \alpha]^2 \sigma_1^2 + \Sigma(y, z)$ as the covariance function.

Appendix C: Estimations and asymptotic properties in the copula model with covariates

We provide detailed estimation procedures for $\hat{\beta}_Z(\theta)$, $\hat{\beta}_Y(\theta)$, $\hat{\beta}_Z(\hat{\theta})$, $\hat{S}_Y$ and $\hat{S}_Z$ in the copula model with covariates, as well as related asymptotic properties and the proof. Consider θ is known. Define a weighted partial score function as

$$U_{j2}^*(\theta, \beta_Z) = b_j - \left\{ \frac{\sum_{i \in R_{j2}} w(y_i, \theta)^{-1} \exp(\beta_Z b_i) b_i}{\sum_{i \in R_{j2}} w(y_i, \theta)^{-1} \exp(\beta_Z b_i)} \right\}$$

where $R_{j2} = \{i: x_i = x_{(j)}\}$ is the risk set at $x_{(j)}$, and $\{x_{(1)}, \dots, x_{(k)}\}$ are distinct ordered uncensored second failure time. Then the estimating equation for β_Z is

$U_2^*(\theta, \beta_Z) = \sum_j w(y_j, \theta)^{-1} U_{j2}^*(\theta, \beta_Z) = 0$ with the solution denoted by $\hat{\beta}_Z(\theta)$. The asymptotic properties of $\hat{\beta}_Y(\theta)$ and $\hat{\beta}_Z(\theta)$ can be established following the lines in Qi et al. [20]. As $n \rightarrow \infty$, the random vectors $n^{1/2}\{\hat{\beta}_Y(\theta) - \beta_Y\}$ and $n^{1/2}\{\hat{\beta}_Z(\theta) - \beta_Z\}$ converge weakly to multivariate normal distributions with mean zero and variance-covariance matrices A_Y and A_Z , respectively. When θ is unknown, we obtain estimators $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\beta}_Z(\hat{\theta})$ by replacing θ by the conditional maximum likelihood estimator $\hat{\theta}$ in the estimating equations $U_1(\theta, \beta_Y) = 0$ and $U_2^*(\theta, \beta_Z) = 0$. The errors of both $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\beta}_Z(\hat{\theta})$ can be decomposed into two terms as

$$\begin{aligned}\hat{\beta}_Y(\hat{\theta}) - \beta_Y &= \{\hat{\beta}_Y(\theta) - \beta_Y\} + \{\hat{\beta}_Y(\hat{\theta}) - \hat{\beta}_Y(\theta)\} \\ \hat{\beta}_Z(\hat{\theta}) - \beta_Z &= \{\hat{\beta}_Z(\theta) - \beta_Z\} + \{\hat{\beta}_Z(\hat{\theta}) - \hat{\beta}_Z(\theta)\}\end{aligned}$$

The first error terms in the above expressions have been discussed, and the second error terms are generated by the use of $\hat{\theta}$ to estimate θ . The distributions of the two terms are proved to be asymptotically orthogonal to each other. The asymptotic properties of $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\beta}_Z(\hat{\theta})$ as well as the proof are given as follows.

Theorem 3

As $n \rightarrow \infty$, the random vectors $\hat{\beta}_Y(\hat{\theta})$ and $\hat{\beta}_Z(\hat{\theta})$ are consistent estimators of β_Y and β_Z respectively, $n^{1/2}\{\hat{\beta}_Y(\hat{\theta}) - \beta_Y\}$ and $n^{1/2}\{\hat{\beta}_Z(\hat{\theta}) - \beta_Z\}$ converge weakly to multivariate normal distributions with mean zero and variance-covariance matrices $A_Y + A_Y B_Y I_C^{-1} B_Y^T A_Y$ and $A_Z + A_Z B_Z I_C^{-1} B_Z^T A_Z$, respectively.

Proof of Theorem 3

Observe that

$$n^{1/2}\{\hat{\beta}_Y(\hat{\theta}) - \beta_Y\} = n^{1/2}\{\hat{\beta}_Y(\theta) - \beta_Y\} + n^{1/2}\{\hat{\beta}_Y(\hat{\theta}) - \hat{\beta}_Y(\theta)\} \quad (13)$$

$$n^{1/2}\{\hat{\beta}_Z(\hat{\theta}) - \beta_Z\} = n^{1/2}\{\hat{\beta}_Z(\theta) - \beta_Z\} + n^{1/2}\{\hat{\beta}_Z(\hat{\theta}) - \hat{\beta}_Z(\theta)\} \quad (14)$$

Following the lines in Qi et al. [20], the first terms in eqs (13) and (14) can be proved to have the expressions

$$n^{1/2}\{\hat{\beta}_Y(\theta) - \beta_Y\} = n^{-1/2} A_Y \sum_{i=1}^n \psi_i(\theta, \beta_Y) + o_p(1)$$

$$n^{1/2}\{\hat{\beta}_Z(\theta) - \beta_Z\} = n^{-1/2} A_Z \sum_{i=1}^n \varphi_i(\theta, \beta_Z) + o_p(1)$$

which converge weakly to multivariate normal distributions with mean zero and variance-covariance matrices A_Y and A_Z , respectively. To be specific,

$$A_Y = \int \left\{ \frac{S_Y^{(2)}(\theta, \beta_Y, y)}{S_Y^{(0)}(\theta, \beta_Y, y)} - \frac{S_Y^{(1)}(\theta, \beta_Y, y)^{\otimes 2}}{S_Y^{(0)}(\theta, \beta_Y, y)^2} \right\} dF_{mY}(y)$$

$$A_Z = \int \left\{ \frac{S_Z^{(2)}(\theta, \beta_Z, z)}{S_Z^{(0)}(\theta, \beta_Z, z)} - \frac{S_Z^{(1)}(\theta, \beta_Z, z)^{\otimes 2}}{S_Z^{(0)}(\theta, \beta_Z, z)^2} \right\} dF_{mZ}(z)$$

$$S_Y^{(r)}(\theta, \beta_Y, y) = n^{-1} \sum_i \{w(y, \theta)/w(y_i, \theta)\} I(y \leq y_i) \exp\{\beta_Y a_i\} a_i^{\otimes r}$$

$$S_Z^{(r)}(\theta, \beta_Z, z) = n^{-1} \sum_i \{w(y, \theta)/w(y_i, \theta)\} I(z \leq z_i) \exp\{\beta_Z b_i\} b_i^{\otimes r}$$

$$\psi_i(\theta, \beta_Y) = \int \left\{ a_i - \frac{S_Y^{(1)}(\theta, \beta_Y, y)}{S_Y^{(0)}(\theta, \beta_Y, y)} \right\} dM_{iY}(y)$$

$$\varphi_i(\theta, \beta_Z) = \int \left\{ a_i - \frac{S_Z^{(1)}(\theta, \beta_Z, z)}{S_Z^{(0)}(\theta, \beta_Z, z)} \right\} dM_{iZ}(z)$$

$$M_{iY}(u) = I(y_i \leq u) - \int_0^u \{w(y, \theta)/w(y_i, \theta)\} I(y \leq y_i) h_Y(y|a_i) dy$$

$$M_{iZ}(u) = I(z_i \leq u) - \int_0^u \{w(y, \theta)/w(y_i, \theta)\} I(z \leq z_i) h_Z(z|b_i) dz$$

where for a $p \times 1$ vector $\mathbf{v}$, $\mathbf{v}^{\otimes r}$ refers to the scalar 1 for $r = 0$, $\mathbf{v}$ for $r = 1$ and $\mathbf{v}\mathbf{v}^T$ for $r = 2$, F_{mY} and F_{mZ} are the cumulative distribution functions of the observed Y_i or Z_i , respectively. The variability in each of the second term in eqs (13) and (14) essentially results from the use of $\hat{\theta}$ to estimate θ . Let $B_Y = E\{\psi_i(\theta, \beta_Y)/\theta\}$ and $B_Z = E\{\varphi_i(\theta, \beta_Z)/\theta\}$. Under appropriate regularity conditions, the second terms in eqs (13) and (14) can be expressed as

$$n^{1/2}\{\hat{\beta}_Y(\hat{\theta}) - \hat{\beta}_Y(\theta)\} = n^{-1/2} A_Y B_Y I_C^{-1} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) + o_p(1)$$

$$n^{1/2}\{\hat{\beta}_Z(\hat{\theta}) - \hat{\beta}_Z(\theta)\} = n^{-1/2} A_Z B_Z I_C^{-1} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) + o_p(1)$$

which converge weakly to multivariate normal distributions with mean zero and variance-covariance matrices $A_Y B_Y I_C^{-1} B_Y^T A_Y$ and $A_Z B_Z I_C^{-1} B_Z^T A_Z$, respectively. In both eqs (13) and (14), the two error terms are asymptotically orthogonal since

$$E\{\psi_i(\theta, \beta_Y) \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i)\} = E \left[\psi_i(\theta, \beta_Y) E\left\{ \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) | Y_i \right\} \right] = 0$$

$$E\{\varphi_i(\theta, \beta_Z) \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i)\} = E \left[\varphi_i(\theta, \beta_Z) E\left\{ \frac{\partial}{\partial \theta} \log p_{T|Y}(T_i|Y_i) | Y_i \right\} \right] = 0$$

These imply that eqs (13) and (14) converge weakly to multivariate normal distributions with mean zero and variance-covariance matrices $A_Y + A_Y B_Y I_C^{-1} B_Y^T A_Y$ and $A_Z + A_Z B_Z I_C^{-1} B_Z^T A_Z$, respectively.

From the proportional hazards models, marginal survival functions $S_Y(y)$ and $S_Z(z)$ can be consistently estimated by

$$\hat{S}_Y(y, \hat{\theta}|a) = \hat{S}_{0Y}(y) \exp\{\hat{\beta}_Y(\hat{\theta})a\}$$

and

$$\hat{S}_Z(z, \hat{\theta}|b) = \hat{S}_{0Z}(z) \exp\{\hat{\beta}_Z(\hat{\theta})b\}$$

where $\hat{S}_{0Y}(y) = \exp\{-\hat{H}_{0Y}(y)\}$ and $\hat{S}_{0Z}(z) = \exp\{-\hat{H}_{0Z}(z)\}$ are the Breslow estimators for baseline survival functions. The baseline cumulative hazard function estimators are

$$\hat{H}_{0Y}(y) = \sum_{y_j < y} \left[\sum_{i \in R_{j1}} \exp\{\hat{\beta}_Y(\hat{\theta})a_i\} \right]^{-1}$$

and

$$\hat{H}_{0Z}(z) = \sum_{x_{(j)} < z} d_{j2} \left[\sum_{i \in R_{j2}} \exp\{\hat{\beta}_Z(\hat{\theta})b_i\} \right]^{-1}$$

where R_{j1} is the risk set for Y at y_j , d_{j2} and R_{j2} are the failure event set and risk set for Z at $x_{(j)}$ as previously defined.

Figure 1.

An explanatory plot of data for a cohort of the Rakai HIV patients infected between 1995 (study initiation year) and 2003 (truncation year), T is the birth time, Y is the time from birth to HIV infection and Z is the time from HIV infection to death

Figure 2.

Simulation results of estimations of marginal survival functions of Y and Z : true survival functions (solid), the proposed weighted estimates (dash), and the estimates by conventional methods (dot). WEMP: weighted empirical survival function; EMP: empirical survival function; WCOX: weighted estimator by a working Cox model; KM: the Kaplan–Meier estimator.

Figure 3.

(a) Birth density plots: linear model estimate (dash), quadratic model estimate (dot), nonparametric estimate (dashdot), and the biased empirical estimate (solid). (b) Scatter plot of $\hat{G}_n(t)$ against $\alpha(t, \hat{\theta})$. The dashed diagonal line $y = x$ is shown as reference. Non-G, nonparametric maximum likelihood estimator $\hat{G}_n(t)$; Para-G, parametric estimator $\alpha(t, \hat{\theta})$

Figure 4.

(a) Estimated marginal survival functions of age at infection. EMP: empirical survival function; WEMP: weighted empirical survival function. (b) Estimated marginal survival functions of residual lifetime. KM: the Kaplan–Meier estimator; WCOX: weighted estimator by a working Cox model. (c) Estimated marginal survival functions of residual lifetime for different categories of age at infection. (d) Estimated conditional survival functions of residual lifetime given different categories of age at infection

Figure 5.

Estimated marginal survival functions of age at infection and residual lifetime by HIV subtype: curves for A subtype (solid), curves for non-A subtype (dash), and curves for unknown subtype (dot)

Table 1

Simulation summary statistics of $(\hat{\theta}, \hat{\alpha}, \hat{S}_1, \hat{\tau})$ for the copula model without covariates

θ	$b(\hat{\theta})$	$\hat{\sigma}(\hat{\theta})$	α	$b(\hat{\alpha})$	$\hat{\sigma}(\hat{\alpha})$	$\hat{\sigma}(\hat{\alpha})$	$cp(\hat{\alpha})$	$b(\hat{S}_1)$	$\hat{\sigma}(\hat{S}_1)$	$cp(\hat{S}_1)$	$b(\hat{\tau})$	$\hat{\sigma}(\hat{\tau})$
Clayton copula												
1.0	0.3	8.4	0.50	1.1	14.8	13.3	95.3	-0.3	3.2	94.6	0.3	4.9
	0.1	7.2	1.33	2.6	24.1	18.2	95.5	-0.2	3.0	95.7	1.0	4.7
	0.1	7.6	3.00	8.1	44.8	42.2	95.7	-0.2	3.1	93.5	1.3	3.6
4.0	0.4	16.2	0.50	2.7	38.3	36.8	95.3	-0.2	4.2	95.4	2.7	11.3
	1.0	17.5	1.33	13.4	65.9	64.2	95.9	-0.2	4.3	95.6	5.0	10.8
	1.7	16.4	3.00	19.5	97.1	94.7	96.4	-0.6	4.4	94.2	5.4	8.3
Positive stable copula												
1.0	0.1	7.2	1.25	1.5	6.2	5.0	93.1	0.2	3.2	95.1	0.7	3.9
	0.6	7.9	1.67	1.3	10.1	8.3	93.5	-0.2	3.1	94.6	0.6	3.4
	0.2	7.6	2.50	0.6	16.3	14.2	94.3	-0.3	3.2	94.8	0.3	2.6
4.0	1.3	15.8	1.25	1.8	8.5	6.7	94.3	0.2	4.6	95.1	0.8	5.6
	1.8	17.3	1.67	0.7	15.5	14.2	94.4	-0.1	4.4	95.7	0.4	5.4
	1.4	16.5	2.50	1.3	26.2	23.7	95.6	-0.3	4.5	96.1	0.7	4.4
Frank copula												
1.0	0.7	7.9	2.00	1.0	41.8	40.6	95.4	0.1	3.1	94.5	0.1	4.1
	0.2	7.5	-1.00	3.2	42.2	40.5	94.8	0.1	3.2	95.2	0.2	4.6
	0.1	7.8	-2.00	0.4	42.1	40.6	95.8	-0.1	3.1	94.2	0.2	4.3
4.0	1.0	17.5	2.00	3.1	82.7	81.0	96.6	-0.3	4.7	96.3	0.4	8.7
	1.7	21.7	-1.00	1.9	85.2	83.9	96.3	-0.2	4.7	95.7	1.3	9.7
	1.4	22.0	-2.00	6.5	92.7	91.1	96.5	-0.2	4.5	96.1	1.4	8.6

b , empirical bias ($\times 10^2$); $\hat{\sigma}$, average of model-based standard error ($\times 10^2$); $\hat{\sigma}_e$, empirical standard error ($\times 10^2$); cp , 95% nominal coverage probability; $\hat{S}_1 = \hat{S}_{Y,Z}(0.22, 0.51)$.

Table 2

Simulation summary statistics of $(\hat{\theta}, \hat{\beta}_Y, \hat{\phi}, \hat{\tau})$ for the copula model with covariates, where $\theta = 2.0$, $\gamma_1 = 1/8$, $\beta_Y = \phi = 0.2$

q	$\hat{\sigma}(\hat{\theta})$	b_Y	$\hat{\sigma}_Y$	b_Z	$\hat{\sigma}_Z$	A	τ	$h(\hat{\tau})$	$\hat{\sigma}_h(\hat{\tau})$	$dp(\hat{\tau})$
Clayton copula										
0.8	10.8	1.7	8.8	0.8	11.7	0	0.20	-0.1	3.5	95.6
						1	0.22	-0.4	3.5	95.6
0.9	10.2	0.4	8.8	1.1	11.8	0	0.40	-0.6	3.4	95.5
						1	0.43	-0.8	3.3	95.7
0.4	10.6	1.2	8.7	1.6	11.5	0	0.60	-1.0	2.7	96.1
						1	0.63	-0.9	2.5	96.2
Positive stable copula										
-0.4	10.9	0.6	8.8	0.4	11.5	0	0.20	0.4	3.9	93.6
						1	0.30	0.1	3.7	93.5
0.6	11.1	0.1	8.8	0.1	11.7	0	0.40	0.5	3.6	94.3
						1	0.47	0.2	3.3	94.5
-0.3	10.9	-0.4	8.9	-1.1	11.8	0	0.60	-0.2	3.0	94.7
						1	0.65	-0.3	2.7	95.1
Frank copula										
0.8	10.5	0.2	8.8	0.5	12.0	0	0.20	0.9	4.4	96.0
						1	0.23	0.2	4.0	96.2
0.5	10.7	-0.9	8.7	0.9	11.8	0	-0.10	0.3	6.5	95.5
						1	-0.09	-0.2	4.1	95.8
0.3	10.3	1.0	8.9	-0.6	12.2	0	-0.20	-0.5	4.2	95.7
						1	-0.21	1.2	4.2	95.8

b_Y empirical bias ($\times 10^2$); $b_Y = h(\hat{\beta}_Y)$; $b_Z = h(\hat{\beta}_Z)$; $qY = \hat{\sigma}(\hat{\beta}_Y)$; $qZ = \hat{\sigma}(\hat{\beta}_Z)$; $\hat{\sigma}_Y = \hat{\sigma}(\hat{\beta}_Y)$; $\hat{\sigma}_Z = \hat{\sigma}(\hat{\beta}_Z)$; A: covariate; $\hat{\sigma}_h$ empirical standard error ($\times 10^2$); $\hat{\sigma}_h$ average bootstrap standard error ($\times 10^2$); cp, 95% nominal coverage probability.