11-16-2011

Assessing Association for Bivariate Survival Data with Interval Sampling: A Copula Model Approach with Application to AIDS Study

Hong Zhu
The Ohio State University, hongzhu@jhsph.edu

Mei-Cheng Wang
Johns Hopkins Bloomberg School of Public Health, mcwang@jhsph.edu

Biometrics 000, 1–24 DOI: 000
000 0000

Assessing Association for Bivariate Survival Data with Interval Sampling: A Copula Model Approach with Application to AIDS Study

Hong Zhu

Division of Biostatistics, College of Public Health, The Ohio State University,

1841 Neil Avenue, 248 Cunz Hall, Columbus, Ohio 43210, U.S.A.

email: hzhu@cph.osu.edu

and

Mei-Cheng Wang

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University,

615 N. Wolfe Street, Baltimore, Maryland 21205, U.S.A.

email: mcwang@jhsph.edu

Summary: In disease surveillance systems or 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 (e.g., HIV infection) within a calendar time interval, the time of the initiating event (e.g., birth) is retrospectively identified for all the cases in the registry, and subsequently the second failure event (e.g., death) is observed during the follow-up. Sampling bias is induced due to the selection process that the data are collected conditioning on the first failure event occurs within a time interval. Consequently, the first failure time is doubly truncated, and the second failure time is informatively right censored. A copula model under semi-stationary condition is considered to assess the association between the bivariate survival times with interval sampling. Estimation and inference are carried out by a two-stage procedure. We first obtain bias-corrected estimators of marginal survival functions, then a pseudo conditional likelihood method is developed to study the association parameter. Asymptotic properties of the proposed estimators are established, and finite sample performance is evaluated by simulation studies. The method is applied to a motivating community-based AIDS study in Rakai to investigate the effect of age at infection on survival time of HIV seroconverters.

Key words: Bivariate survival data; Copula model; Interval sampling; Semi-stationarity.

1. Introduction

In disease surveillance systems or registries, it is common to collect data with a certain failure event (e.g., diagnosis of disease) occurring within a calendar time interval and then obtain additional information retrospectively or/and prospectively. Such type of sampling is refereed to as interval sampling and we consider bivariate survival data with interval sampling in this paper. One example of such data is Acquired Immunodeficiency Syndrome (AIDS) blood transfusion data collected by the Centers for Disease Control (CDC), which is from a registry database, a common source of medical data (Bilker and Wang, 1996). 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 disease progression. The time of the initiating event of Human Immunodeficiency Virus (HIV) infection were retrospectively identified, and bivariate survival times of interest are the lag time from infection to AIDS diagnosis and the survival time after AIDS. Generally speaking, under interval sampling scheme, subjects experiencing the first failure event within a calendar time interval are identified as cases and enter into a registry. 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.

This paper is motivated by Rakai AIDS study in investigating the association between age at infection and survival time of HIV-seroconverters. This study in the rural Rakai district of southwestern Uganda conducted annual surveillance from Nov 1994 in an open cohort of individuals aged 15-49 years (Lutalo et al., 2007). 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 outmigration or other administrative censoring. The follow-up time was truncated on Dec 31st, 2003, before antiretrovival treatment (ART) became available to this population. The date of seroconversion was defined as the mid-interval between the last negative and the first positive HIV test. In this study, the initiating, the first failure and second failure events are, respectively, birth, incidence of HIV infection and death. Bivariate survival data refer to age at HIV infection and residual lifetime. Figure 1 provides a graphical presentation to illustrate how interval sampling design arises in Rakai AIDS study.

[Figure 1 about here.]

As shown by Figure 1, the sampling population consists of individuals who became HIV-infected between 1995 and 2003. Under interval sampling, the age at HIV-infection was observed subject to double truncation and residual lifetime was dependently right censored. Previous study (Lutalo et al., 2007) suggested that survival time decreased significantly with older age at infection. They compared Kaplan-Meier survival curves among different groups of age at infection by log-rank test and estimated hazard ratio of death associated with age at infection by Cox proportional hazards model. Age at infection was treated as a conditional variable and therefore their analytical results would be interpreted conditionally, if it appropriately adjusted for the bias on observed residual lifetime from dependent right censoring. As a contrast to conditional analysis with Cox regression, this paper focuses on unconditional analysis of the association between age at infection and residual lifetime. Moreover, the distribution of age at infection was typically estimated by empirical method, but this ignores the fact that age at infection of the sampling population is doubly truncated. Consequently, joint distribution of age at infection and residual lifetime is also sampling-biased. The purpose of this paper is to address the issue of interval sampling in assessing the association of bivariate survival data collected from disease registries. The scientific goal is to

quantitatively examine the association between age at infection and residual lifetime among HIV seroconverters, and study how the association varies with other important factors, such as 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 (Visser, 1996; Lin, Sun and Ying, 1999; Schaubel and Cai, 2004). When association of bivariate survival times is of interest, semiparametric copula model has been becoming an increasingly popular tool for modeling the dependence. Copula-based survival model has been proposed by Shih and Louis (1995) for bivariate data both subject to right censoring, Wang and Ding (2000) for bivariate current status data, and Lakhal-Chaieb, Cook and Lin (2010) for bivariate serial gap times. Copula family includes many useful bivariate survival models and enjoys flexibility in modeling. An appealing feature is that it allows separate modeling and estimation of margins and 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, 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 copula model have been used by Genest, Ghoudi and Rivest (1995) for complete data, Shih and Louis (1995) for right-censored data and Wang and Ding (2000) for current status data. The proposed estimators for association measure in these papers showed to have nice asymptotic properties.

In this paper, we consider a copula model for bivariate survival data with interval sampling, and the association parameter is estimated through a similar two-stage procedure based on pseudo conditional likelihood. Particularly, under reasonable model assumption, we study the data structure that the first failure time is doubly truncated and the second failure time

is informatively right censored. The rest of the paper is organized as follows. In Section 2, interval sampling design is discussed with more details, and copula model for bivariate survival data as well as the model assumption of semi-stationarity are introduced. In Section 3, marginal survival distribution for each failure time is studied and association parameter is estimated by a two-stage procedure. Asymptotic properties of the proposed estimators are established. Finite sample performance is examined by simulation studies in Section 4. In Section 5, for illustration, the proposed method is applied to Rakai AIDS study. Finally, concluding remarks and discussion are included in Section 6. Proofs of the results are provided in the Appendix and the Web Appendix.

2. Interval sampling, Copula Model and Semi-stationarity

In this section, we describe the data structure for bivariate survival data with interval sampling and some fundamental concepts of copula model, together with the model assumption of semi-stationarity. Statistical method and inference are developed for a target population of cases (e.g., HIV seroconverters). To begin, we define random variables for the target population. Let T denote the calendar time of the initiating event (e.g., birth), Y denote the time from the initiating event to the first failure event (e.g., HIV infection), Z denote the time from the first event to the second event (e.g., death), and C denote the calendar censoring time. The failure times Y and Z are possibly correlated and their dependent relationship is of primary interest. Under 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 $0 \leqslant Y + T \leqslant t_0$. Therefore, bivariate failure times are observed subject to sampling bias. Specifically, Y is doubly truncated. Denote the double truncation rate by $\beta = 1 - P(-T \leqslant Y \leqslant t_0 - T)$, and with this probability a person who experienced the first failure event (e.g., infected with HIV) will not be identified. If the second failure event occurs

before C ($C \leqslant t_0$), Z is uncensored which is described by a further constraint $Y+Z \leqslant C-T$. Otherwise, it is censored with censoring time $C-(T+Y)$.

Assume that the initiating event T occurs over the calendar time with a rate function $\lambda(t)$ for $t \leqslant t_0$. Let $f_{Y,Z}(y,z)$ denote the *population* joint density function of (Y,Z), and $F_Y(\cdot)$, $F_Z(\cdot)$ 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 discussion. Note that the constraint $y^+ < \infty$ is not an absolutely required assumption for the inferential results of (Y, Z), but it does make the likelihood discussion much easier. Therefore, the *population* density function of T, $g(t)$ could be defined as a normalized rate function in the interval $[-y^+, t_0 - y_-]$ as, $g(t) = \lambda(t) I(-y^+ \leqslant t \leqslant t_0 - y_-) / \int_{-y^+}^{t_0-y_-} \lambda(u)du$, and its *population* cumulative distribution function is denoted by $G(t)$.

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

$$S_{Y,Z}(y,z) = C_\alpha\{S_Y(y), S_Z(z)\}, \;\; y,\, z \geqslant 0$$

$$f_{Y,Z}(y,z) = c_\alpha\{S_Y(y), S_Z(z)\} f_Y(y) f_Z(z), \;\; y,\, z \geqslant 0$$

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

$$\tau = 4 \int_0^1 \int_0^1 C_\alpha(u,v) du dv - 1$$

We then introduce the following model assumption to facilitate the development of the proposed work.

S. The disease progression is independent of when the initiating event occurs. Or, equivalently, assume that (Y, Z) is independent of T.

The model is considered to be semi-stationary if (S) is satisfied. The time elapses between T and the end of the calendar time interval t_0 is mainly affected by experimental constraint so that this assumption may be often justified in biomedical study. In addition, if the initiating event occurs at a constant rate which implies T follows a uniform distribution, the model is considered to be stationary, which was studied in Zhu (2010) and an associated manuscript (Zhu and Wang, 2011). It is important, however, 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. Nevertheless, we focus on the semi-stationary condition in this paper, and the non-stationarity when (S) does not hold will be explored in the future.

3. Estimation and Inference for Semiparametric Copula Model

A semiparametric copula model for bivariate survival data with interval sampling is considered in this section under semi-stationary condition when (S) is satisfied. In some scenarios, 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 the distribution of T, where $\theta \in \Theta$ and Θ is an open set in R^k. Take Rakai AIDS study data for example, g describes the birth trend for HIV seroconverters. In the following, under semi-stationary condition that T is independent of (Y, Z), a conditional likelihood estimator of θ is obtained, and inverse probability weighting (IPW) method is employed to derive bias-corrected semiparametric consistent estimators of $S_Y(y)$ and $S_Z(z)$ at the first

stage. At the second stage, the association parameter α in copula model is estimated based on a pseudo conditional likelihood.

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

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

where $p_{T,Y}(t, y)$ and $p_Y(y)$ are the *sampling* joint density of (T, Y) and marginal density of Y respectively. Therefore, the conditional likelihood function of observed $\{t\}$ given observed $\{y\}$ is

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

in which the distribution of Y becomes a nuisance parameter and is eliminated by 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 (Serfling, 1980). Under regularity conditions, as $n \to \infty$, $\hat{\theta}$ 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

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

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

We then explore the probability structure of the bivariate data to obtain bias-corrected consistent estimators of marginal survival functions $S_Y(y)$ and $\mathrm{S}_Z(z)$. Due to 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, $\mathrm{p}_Y(y) = \frac{w(y,\theta)f_Y(y)}{\int w(u,\theta)f_Y(u)du}$, 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's are the observed first failure time, and $\hat{\theta}$ is the conditional likelihood estimator from $L_c(\theta)$. The weighted empirical survival function (WEMP) $\hat{S}_Y(y,\hat{\theta})$ is actually the semiparametric maximum likelihood estimator (SPMLE) of $S_Y(y)$.

For the second failure time Z, assume that observation of Z ends at calendar time C, we observe data $\{(y,x)\}$ where $x = min(z, c-t-y)$. Due to interval sampling and correlation between Y and Z, Z is dependently right censored and $S_Z(z)$ can not be simply estimated by the Kaplan-Meier estimator in general. Since the observation of Z is always coupled with the observation of Y, the same selection bias function $w(y,\theta)$ could be used to adjust for the sampling bias of observing Z induced by Y. Hence, a weighted Kaplan-Meier estimator (WKME) of $S_Z(z)$ is developed using the method of inverse probability weighting, and is given as

$$\begin{aligned}\hat{S}_Z(z,\hat{\theta}) &= \prod_{Z_{(j)}<z}\left(1 - \frac{\sum_{i\in d_j} w(Y_{(j)},\hat{\theta})^{-1}}{\sum_{i\in R_j} w(Y_i,\hat{\theta})^{-1}}\right) \\ &= \prod_{Z_{(j)}<z}\left(1 - \frac{\sum_{i\in d_j}\{G(t_0 - Y_{(j)},\hat{\theta}) - G(-Y_{(j)},\hat{\theta})\}^{-1}}{\sum_{i\in R_j}\{G(t_0 - Y_i,\hat{\theta}) - G(-Y_i,\hat{\theta})\}^{-1}}\right)\end{aligned}$$

where $\hat{\theta}$ is the conditional maximum likelihood estimator, $d_j = \{i : Z_i = Z_{(j)}\}$ and $R_j = \{i : Z_i \geqslant Z_{(j)}\}$ are the failure event set and risk set at $Z_{(j)}$ respectively, and $\{Z_{(1)},\ldots,Z_{(k)}\}$ are distinct ordered uncensored second failure time with their counterparts at the first failure time as $\{Y_{(1)},\ldots,Y_{(k)}\}$. The inverse probability weighting method has been widely used in literatures particularly to reduce selection bias in observational study. In our model setting,

the selection bias function is constructed according to the biased distribution of Y under 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 WKME $\hat{S}_Z(z,\hat{\theta})$ is a semiparametric consistent estimator of $S_Z(z)$.

Next, we present the estimation procedure for the association parameter α based on a pseudo conditional likelihood. Bivariate survival distribution of (Y,Z) is modeled by copula as $S_{Y,Z}(y,z) = C_\alpha\{S_y(y), S_z(z)\}$. 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 $\hat{S}_Y(z,\theta)$ respectively. For each subject i $(i = 1,\ldots,n)$, data $\{y_i, x_i, \delta_i, t_i\}$ are observed where $x_i = \min(z_i, c_i - t_i - y_i)$ and $\delta_i = I(z_i \leqslant c_i - t_i - y_i)$. 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

$$L_c(\alpha) = \prod_i \frac{f_{Y,Z}(y_i,x_i)^{\delta_i} \frac{\partial S_{Y,Z}(y_i,x_i)^{1-\delta_i}}{\partial y_i}}{S_Y(c_i - t_i) - S_Y(-t_i)} \propto \prod_i f_{Y,Z}(y_i,x_i)^{\delta_i} \frac{\partial S_{Y,Z}(y_i,x_i)^{1-\delta_i}}{\partial y_i}.$$

An interesting feature in the likelihood decomposition 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 copula model of bivariate survival data,

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

From previous discussion, the 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 estimated margins, and is given as

$$\begin{aligned} U(\alpha,\theta)^{(p)} &= \frac{\partial}{\partial\alpha}\left[\sum_{i=1}^n \delta_i \log[c_\alpha\{\hat{S}_Y(y_i,\theta), \hat{S}_Z(x_i,\theta)\}] + (1-\delta_i)\log\left[\frac{\partial C_\alpha\{\hat{S}_Y(y_i,\theta), \hat{S}_Z(x_i,\theta)\}}{\partial u_i}\right]\right] \\ &= 0 \end{aligned} \tag{2}$$

The estimator of the association parameter α, $\hat{\alpha}(\theta)$, is the solution to (2). The asymptotic theory for $\hat{\alpha}(\theta)$ is developed, and we list the required conditions and state the large sample results for $\hat{\alpha}(\theta)$ in Theorem 1. The details of the proof are provided in the Web Appendix.

The following conditions are assumed throughout the paper.

(1) Assume that standard regularity conditions for maximum likelihood estimate hold.

(2) Define functions

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

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

Assume that they are continuous and bounded for $(y, z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$. Based on these assumptions, $\hat{\alpha}(\theta)$ can be shown consistent and asymptotically normal.

THEOREM 1: *As $n \to \infty$, $\hat{\alpha}(\theta)$ 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 $\sigma^2 = (\rho_1^2 + \rho_2^2)/\rho_1^4$, where*

$$\rho_1^2 = E[-V_\alpha\{\alpha_0, S_Y(Y_i), S_Z(X_i)\}] = \int_{\mathcal{A}} -V_\alpha\{\alpha_0, S_Y(y), S_Z(z)\} dJ_{\alpha_0}(y, z, \delta),$$

$$\rho_2^2 = E[\{I_1(Y, \alpha_0) + I_2(X, \delta, \alpha_0)\}^2] = \int_{\mathcal{A}} \{I_1(y, \alpha_0) + I_2(z, \delta, \alpha_0)\}^2 dJ_{\alpha_0}(y, z, \delta),$$

with

$$I_1(Y, \alpha_0) = \int_{\mathcal{A}} V_{\alpha,1}\{\alpha_0, S_Y(y), S_Z(z)\} I_1^0(Y_i)(y) dJ_{\alpha_0}(y, z, \delta),$$

$$I_2(X, \delta, \alpha_0) = \int_{\mathcal{A}} V_{\alpha,2}\{\alpha_0, S_Y(y), S_Z(z)\} I_2^0(X_i, \delta_i)(z) dJ_{\alpha_0}(y, z, \delta),$$

$$I_1^0(Y_i)(y) = -S_Y(y)\Big\{ \int_0^y \frac{dN_{1i}(u)}{p(Y \geqslant u)} - \int_0^y \frac{I(Y_i \geqslant u) d\Lambda_1(u)}{p(Y \geqslant u)} \Big\},$$

$$I_2^0(X_i, \delta_i)(z) = -S_Z(z)\Big\{ \int_0^z \frac{dN_{2i}(u)}{p(Z \geqslant u, C_2 \geqslant u)} - \int_0^z \frac{I(X_i \geqslant u) d\Lambda_2(u)}{p(Z \geqslant u, C_2 \geqslant u)} \Big\},$$

J_{α_0} is the joint distribution of (Y, X, δ), $C_2 = C - T - Y$, $N_{1i}(u) = I(Y_i \leqslant u)$, and $N_{2i}(u) = I(Z_i \leqslant u, \delta_i = 1)$.

In the proof of Theorem 1, we show σ^2 can be consistently estimated by $\hat{\sigma}^2 = (\hat{\rho}_1^2 + \hat{\rho}_2^2)/\hat{\rho}_1^4$.

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 $U(\alpha, \hat{\theta})^{(p)} = 0$. Let the solution be denoted by $\hat{\alpha}(\hat{\theta})$. Note that the error 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 error in the first term has been explored in Theorem 1. The error 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 desired asymptotic properties.

THEOREM 2: *As $n \to \infty$, $\hat{\alpha}(\hat{\theta})$ 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 $\sigma_1^2 = \sigma^2 + \gamma I_C^{-1} \gamma^t$, where*

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

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

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

Furthermore, a natural estimator for bivariate survival function could be obtained by plugging in estimators for unknown quantities in copula model $S_{Y,Z}(y,z) = C_\alpha\{S_Y(y), S_Z(z)\}$. To be specific, the margins $S_Y(y)$ and $S_Z(z)$ are replaced by their semiparametric consistent estimators, the weighted empirical survival function for Y and the weighted Kaplan-Meier estimate for Z, 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 3 with the proof provided in the Appendix.

THEOREM 3: *As $n \to \infty$, $\hat{S}_{Y,Z}(y,z)$ converges to $S_{Y,Z}(y,z)$ in probability, 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* $[\frac{\partial C\{\alpha, S_Y(y), S_Z(z)\}}{\partial \alpha}]^2 \sigma_1^2 + \Sigma(y,z)$.

For semiparametric copula model under semi-stationary condition, 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 $\hat{G}(t,\hat{\theta})$. Since T is also doubly truncated subject to the constraint $-Y \leqslant T \leqslant t_0 - Y$, estimating G is essentially a dual problem as estimating S_Y. Shen (2008) provided an algorithm to jointly compute the nonparametric maximum likelihood estimators of both G and S_Y. In data analysis, the plot is used as a graphical tool to examine the adequate fit of the parametric distribution of T.

4. Numerical Studies

In this section, we present simulation studies to evaluate the performance of the proposed estimation and inference procedures under 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 $\{(t_1, y_1, z_1), \ldots, (t_n, y_n, z_n)\}$ is generated as follows. Define $T = -3W + 10$, where W follows $Exp(\theta)$ distribution with $\theta = 1.0$ and 2.0. Let bivariate failure times (Y, Z) be generated from the following three copula models $C_\alpha\{S_Y(y), S_Z(z)\}$.

Clayton's family: $C_\alpha(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}$ with positive association when $\alpha > 0$ and independence for $\alpha \to 0$.

Positive stable Copula: $C_\alpha(u,v) = \exp(-[\{-\log(u)\}^\alpha + \{-\log(v)\}^\alpha]^{1/\alpha})$ with positive association when $\alpha > 1$ and independence for $\alpha = 1$.

Frank's family: $C_\alpha(u,v) = -\frac{1}{\alpha}\log\{1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\}$ with positive association when $\alpha > 0$, negative association when $\alpha < 0$ and independence for $\alpha \to 0$.

We choose unit exponential margins and three different values for α in each of the copula model, in order to accommodate different levels of dependence of Y and Z. An observation (t, y, z) is included in interval sampled dataset if and only if $0 \leqslant t + y \leqslant 10$ and is censored if $t + y + z \geqslant 10$. For each choice of parameters (θ, α), 1000 simulated samples are generated with sample size $n = 400$.

[Figure 2 about here.]

The proposed estimation methods for marginal survival functions of Y and Z are evaluated in Figure 2. It demonstrates that the weighted empirical survival function (WEMP) outperforms the empirical one in estimating $S_Y(y)$, and the weighted Kaplan-Meier estimator (WKME) outperforms the Kaplan-Meier estimator in estimating $S_Z(z)$. The association parameter estimate is obtained as $\hat{\alpha}(\hat{\theta})$ by solving $U(\alpha, \hat{\theta})^{(p)} = 0$. Particularly, the joint survival function estimator $\hat{S}_{Y,Z}(y, z)$ is assessed at $(y, z) = (0.22, 0.51)$, denoted by $\hat{S}_1$ corresponding to marginal survival probabilities of 0.8 for Y and 0.6 for Z. Table 1 provides the simulation results about $\hat{\theta}$, $\hat{\alpha}(\hat{\theta})$ and $\hat{S}_1$, which include the empirical bias, average of model-based standard error estimates, empirical standard error and 95% nominal coverage probability. Wald confidence interval is constructed using estimated asymptotic variance, and empirical estimate of the 95% coverage probability is obtained based on Wald confidence interval over 1000 replications. It shows the proposed estimators $\hat{\theta}$, $\hat{\alpha}(\hat{\theta})$ and $\hat{S}_1$ work well with fairly small biases. For the association estimator $\hat{\alpha}(\hat{\theta})$, the average of model-based standard error estimates is very close to the empirical standard error, which implies satisfactory performance of the inferential result on 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.

[Table 1 about here.]

5. Application

The HIV seroconversion data from Rakai AIDS study provide an example of bivariate survival data with interval sampling. In this study, 837 subjects were ascertained with 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 administrative censoring at the end of 2003. 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 interval sampling. For the purposes of illustration, we apply the proposed semiparametric copula model method to analyze Rakai HIV seroconversion data, address the statistical issues of 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 HIV epidemic for treatment-naive 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 birth time of HIV seroconverters by T, age at infection by Y, and residual lifetime after infection by Z. These variables are all analyzed by a continuous scale in years. 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 $g(t) = c+\theta_1 t$, and a quadratic model $g(t) = c+\theta_1 t+\theta_2 t^2$, where c is a given positive-valued constant in both models. The choice of the parametric form of

the distribution of T can be and has been examined by comparing the parametric estimate of $G(t)$ with its nonparametric maximum likelihood estimate. Figure 3 (a) plots the empirical and model-based estimated density functions of T. It shows that the difference between linear and quadratic models is considerably small, and demonstrates unignorable bias in estimating birth density by the empirical method. A decreasing trend in birth rate of HIV seroconverters is found by both polynomial models. This may partly reflect the change in the population under surveillance, such as trends in HIV incidence and prevalence. Actually, 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 HIV prevalence declined from approximately 18% in 1995 to 13% in 2003 (Lutalo et al., 2007). In the analysis, we choose the quadratic model for birth density given the small difference between the two polynomial model fits. The parameter estimates together with their estimated standard errors are $(\hat{\theta}_1, \hat{\theta}_2) = (-1.869 \times 10^{-4}, 1.001 \times 10^{-5})$ $(s.e. = (1.420 \times 10^{-4}, 3.029 \times 10^{-6}))$. 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 $\hat{G}(t, \hat{\theta})$ and it shows the assumption of quadratic birth density is considerably reasonable.

[Figure 3 about here.]

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 (WEMP) and weighted Kaplan-Meier estimator (WKM) respectively, adjusting for the selection bias from interval sampling. In Figure 4 (a), it shows that the empirical method overestimates the marginal survival of age at infection comparing to the weighted one, However, as shown in Figure 4 (b), the difference between Kaplan-Meier estimator and the weighted one in estimating the residual lifetime is fairly small, perhaps due of a large percentage of censored observations.

[Figure 4 about here.]

Previous analysis (Lutalo et al., 2007) shows survival time decreased significantly with older age at infection ($p = 0.01$) based on a Cox proportional hazards model conditional on age at infection. However, the appropriateness of Cox model is under investigation since it does not take into account the fact that the data are collected under interval sampling. As discussed, due to interval sampling, age at infection is doubly truncated and residual lifetime is observed subject to dependent right censoring. Therefore, selection bias needs to be adjusted for in analyzing the data and studying the relationship between age at infection and residual lifetime. We consider a copula model where the dependency structure is fitted by Frank's family, and assess the association parameter α quantitatively through the two-stage estimation procedure. To estimate the standard error of the estimator, we adopt a nonparametric bootstrap method by sampling 837 subjects with replacement from the dataset. The resampling procedure is repeated 500 times. Wald 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 -0.195 with 95% Wald confidence interval being $[-1.226, 0.836]$. The corresponding Kendall's *tau* is estimated as -0.022 with Wald confidence interval being $[-0.126, 0.082]$. Different 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. Furthermore, 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 $\geqslant$ 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 $\geqslant$ 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 that the estimation of α shows.

Moreover, studies suggest that the progression of HIV infection is different by HIV subtype (Kaleebu et al., 2001). 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 Rakai HIV seroconversion data by 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 (PCR) 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 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 (Lutalo et al., 2007), so for the analysis purposes we compare A subtype with combined non-A virus subtypes. First, we consider marginally analyzing bivariate failure times of age at infection and residual lifetime by HIV subtype. As shown in Figure 5, 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 A infections, compared with 45 deaths among 349 non-A subtypes infections. It is consistent with the result from previous study in Uganda (Kaleebu et al., 2001) that subtype A has a slower disease progression rate, and it is thought to be less pathogenic than other subtypes. Next, we quantitatively examine the association between age at infection and residual lifetime by HIV subtype. The association parameter α in Frank's family copula model is estimated as 2.949 with 95% Wald confidence interval being $[-0.567, 6.465]$ for A subtype, -0.368 with $[-1.654, 0.918]$ for non-A subtypes, and

-0.349 with $[-1.174, 0.476]$ 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 Rakai HIV epidemic probably has a predominance of non-A subtypes infection and subtype A appears to be a very different virus subtype in HIV progression from other subtypes, which is consistent with the conclusions from other studies. However, the difference may be partly due to insufficient follow-up time given a large proportion of censored observations. While since Rakai community cohort study is still ongoing, additional death cases could be obtained by pushing forward the sampling window, which is expected to increase the power and precision of the analysis.

[Figure 5 about here.]

6. Discussion

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 conditioning on the first failure event (e.g., diagnosis of disease) occurs within a time interval. Under interval sampling scheme, the first failure time is subject to double truncation and the second failure time is subject to informative right censoring. We focus specifically on this data structure, and investigate the association between bivariate survival data with interval sampling by copula model under reasonable assumption of semi-stationarity. The association parameter is estimated through a two-stage estimation procedure. First, we obtain consistent estimators of marginal survival functions adjusting for bias from interval sampling, then assess the association based on a pseudo conditional likelihood. Our simulation results suggest that the proposed method works well for moderate sample size and asymptotic properties of the proposed estimators are established. Since the asymptotic variance has

rather complicated form and possibly involves censoring distribution, bootstrap method is applied as a direct and robust way to compute the estimated standard error in application.

To establish asymptotic properties of the proposed association parameter estimator, we assume some regularity conditions, that are, the score function and its partial derivatives are bounded. While for many popular copula functions, such as positive stable copula, this assumption may not be always valid especially on the boundary of the parameter space, and this problem was discussed by Chen et al. (2010). However, since it is not the major focus of this paper, we adopt the bounded assumption as in Shih and Louis (1995) and many others to derive the large sample properties of the two-stage estimator in copula model. Nevertheless, as the likelihood theory generally does not work for the copula model when the bounded assumption fails, a nonparametric test procedure to test the independence between bivariate survival data needs to be developed under the interval sampling mechanism considered in this paper and this is a subject of current research. In simulations and data application, some specific copula models are used to characterize the dependence structure of bivariate survival data given their modeling flexibility and computational convenience. While in fact, any copula model could be considered and this raises a closely related issue on 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 copulas needs to be addressed in future work. Potentially, model selection procedure and criteria could be developed for bivariate survival data with interval sampling.

The research is motivated by and applied to Rakai HIV seroconversion data to assess the association between age at HIV-infection and residual lifetime among treatment-naive HIV seroconverters, and study how the association varies with HIV subtype. Another interesting extension of this work would be incorporating covariates with regression coefficients in copula model. The covariates involved in regression model could be baseline variables or time-

dependent variables. For example, in Rakai Heath Science Program, ART became available in 2004 and this time-dependent treatment variable would further complicate the analysis. One scientific goal of the future research is to examine ART effect on HIV progression in Rakai cohort.

7. Supplementary Materials

Web Appendix referenced in Section 3 is available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

Acknowledgments

The authors wish to thank the Rakai Health Science Program at Johns Hopkins Bloomberg School of Public Health for providing data.

References

Bilker, W. B. and Wang M.-C. (1996). A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics* **52,** 10–20.

Chen, X., Fan, Y., Pouzo, D., and Ying, Z. (2010). Testing hypotheses in the functional linear model. Estimation and model selection of semiparametric multivariate survival functions under general censorship. *Journal of Econometrics* **157,** 129–142.

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.

Kaleebu P., Ross, A., Morgan, D., Yirrel, D., Oram, J., Rutebemberwa, A., et al. (2001). Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. *AIDS* **15,** 293–299.

Lakhal-Chaieb, L., Cook, R., and Lin, X. (2010). Inverse Probability of Censoring Weighted Estimates of Kendalls τ for Gap Time Analyses. *Biometrics* **66,** 1145–1152.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of gap time distributions for serial events with censored data. *Biometrika* **86,** 59-70.

Lutalo, T., Gray, R. H., Wawer, M., Sewankambo, N., Serwadda, D., Laeyendecker, O., et al. (2007). Survival of HIV-infected treatment-naive individuals with documented dates of seroconversion in Rakai, Uganda. *AIDS* **21,** (suppl 6) S15–S19.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Schaubel, D. E. and Cai, J. (2004). Nonparametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine* **23,** 1885–1900.

Shen, P.-S. (2008). Nonparametric analysis of doubly truncated data. *Annals of the institute of statistical mathematics* **62,** 835–53.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameters in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Visser, M. (1996). Nonparametric estimation on the bivariate survival function with application to vertically transmitted AIDS. *Biometrika* **83,** 507–518.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika* **87,** 879–893.

Zhu, H. (2010). *Statistical methods for bivariate survival data with interval sampling and application to biomedical studies*. PhD Dissertation, Johns Hopkins University.

Zhu, H. and Wang, M.-C. (2011). Analyzing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika* In revision.

Appendix

Proof of Theorem 2

Under the assumed conditions listed in Section 3, we study the asymptotic properties of $\hat{\alpha}(\hat{\theta})$. If θ is known, in the proof of Theorem 1 shown in the Web appendix, we prove that $\hat{\alpha}(\theta) - \alpha_0$ converges to 0 in probability.

Observe that

$$n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\} = n^{1/2}\{\hat{\alpha}(\theta) - \alpha_0\} + n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \hat{\alpha}(\theta)\} \tag{A.1}$$

As identified by Theorem 1 if θ is known, the first term in $(A.1)$ converges weakly to a normal distribution with mean 0 and variance σ^2. By counting process asymptotic techniques, it is asymptotically equivalent to a sum of n i.i.d. zero-mean random variables, expressed as

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

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

To develop the asymptotic results of the second term in $(A.1)$, the additional variation created by estimating θ by $\hat{\theta}$, the MLE of the conditional likelihood function $L_c(\theta)$, needs to be handled. Let $\gamma = E\{\partial\phi_i(\alpha,\theta)/\partial\theta\}$, under appropriate regularity conditions, the second term can be approximated by a sum of n i.i.d. zero-mean random variables, expressed as

$$n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \hat{\alpha}(\theta)\} = n^{-1/2}\gamma I_C^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i) + o_p(1) \tag{A.3}$$

which converges weakly to a normal distribution with mean 0 and variance $\gamma I_C^{-1}\gamma^t$. 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 $(A.2)$ and $(A.3)$, we get

$$n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\} \cong n^{-1/2}\sum_{i=1}^{n}\phi_i(\alpha,\theta) + n^{-1/2}\gamma I_C^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i) \tag{A.4}$$

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

$$E\Big\{\phi_i(\alpha,\theta)\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i)\Big\} = E\left[\phi_i(\alpha,\theta)E\{\frac{\partial}{\partial\theta}\log p_{T|Y}(T_i|Y_i)|Y_i\}\right] = 0 \tag{A.5}$$

$(A.4)$ and $(A.5)$ 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 $\sigma_1^2 = \sigma^2 + \gamma I_C^{-1}\gamma^t$. It is natural to estimate σ_1^2 by $\hat{\sigma}_1^2 = \hat{\sigma}^2 + \hat{\gamma}\hat{I}_C^{-1}\hat{\gamma}^t$, where $\hat{\sigma}^2$ is given in the proof of Theorem 1, and $\hat{\gamma}$ and $\hat{I}_C$ are the

corresponding moment-type empirical estimators. The consistency of $\hat{\sigma}^2$, $\hat{\gamma}$ and $\hat{I}_C$ implies that $\hat{\sigma}_1^2$ is a consistent estimator of σ_1^2.

Proof of Theorem 3

Consider bivariate survival estimate of $S_{Y,Z}(y,z)$ at any given time $(y,z) \in \mathcal{A} = [y_-, y_+] \times [z_-, z_+]$ is obtained under copula model, as $\hat{S}_{Y,Z}(y,z) = C\{\hat{\alpha}(\hat{\theta}), \hat{S}_Y(y), \hat{S}_Z(z)\}$, where $\hat{\alpha}(\hat{\theta})$ is the two-stage association parameter estimator, $\hat{S}_Y(y)$ and $\hat{S}_Z(z)$ are the weighted empirical survival function for Y and the weighted Kaplan-Meier estimate for Z. The asymptotic results of $\hat{S}_{Y,Z}(y,z)$ are proved under the assumed conditions listed in Section 3 and the following regularity conditions. Assume that copula function $C(\alpha,u,v)$ are continuous and differentiable at α, u, v 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 2, $\hat{\alpha}(\hat{\theta})$ converges in probability to α_0. Since copula function $C(\alpha,u,v)$ is a continuous function of α, u and v, then $C\{\hat{\alpha}(\hat{\theta}), \hat{S}_Y(y), \hat{S}_Z(z)\}$ converges in probability to $C\{\alpha_0, S_Y(y), S_Z(z)\}$ uniformly in $\mathcal{A} = [y_-, y_+] \times [z_-, z_+]$. Therefore, As $n \to \infty$, $\hat{S}_{Y,Z}(y,z)$ converges to $S_{Y,Z}(y,z)$ in probability.

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

$$\begin{aligned} n^{1/2}[C\{\hat{\alpha}(\hat{\theta}), \hat{S}_Y(y), \hat{S}_Z(z)\} - C\{\alpha_0, S_Y(y), S_Z(z)\}] &\cong n^{1/2}\frac{\partial C\{\alpha, S_Y(y), S_Z(z)\}}{\partial \alpha}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\} \\ &+ n^{1/2}\frac{\partial C\{\alpha_0, S_Y(y), S_Z(z)\}}{\partial u}(\hat{S}_Y - S_Y)(y) \\ &+ n^{1/2}\frac{\partial C\{\alpha_0, S_Y(y), S_Z(z)\}}{\partial v}(\hat{S}_Z - S_Z)(z) \end{aligned} \tag{A.6}$$

By Theorem 2, $n^{1/2}\{\hat{\alpha}(\hat{\theta}) - \alpha_0\}$ converges weakly to normal with mean zero and variance

σ_1^2. Therefore, the first term in (A.6) is asymptotically equivalent to

$$n^{-1/2}\frac{\partial C\{\alpha, S_Y(y), S_Z(z)\}}{\partial \alpha}\sigma_1 \sum_{i=1}^{n} \frac{\partial \log l\{\alpha, \hat{S}_Y(y), \hat{S}_Z(z)\}}{\partial \alpha} \tag{A.7}$$

which is a sum of n i.i.d. random variables. Applying counting process asymptotic techniques to $\hat{S}_Y$ and $\hat{S}_Z$, the sum of the second and the third terms in (A.6) is asymptotically equivalent to

$$n^{-1/2}\left[\sum_{i=1}^{n} \frac{\partial C\{\alpha_0, S_Y(y), S_Z(z)\}}{\partial u} I_1^0(Y_i)(y) + \frac{\partial C\{\alpha_0, S_Y(y), S_Z(z)\}}{\partial v} I_2^0(X_i, \delta_i)(z)\right] \tag{A.8}$$

where $I_1^0(Y_i)(y) = -S_Y(y)\{\int_0^y \frac{dN_{1i}(u)}{p(Y \geqslant u)} - \int_0^y \frac{I(Y_i \geqslant u) d\Lambda_1(u)}{p(Y \geqslant u)}\}$ and $I_2^0(X_i, \delta_i)(z) = -S_Z(z)\{\int_0^z \frac{dN_{2i}(u)}{p(Z \geqslant u, C_2 \geqslant u)} - \int_0^z \frac{I(X_i \geqslant u) d\Lambda_2(u)}{p(Z \geqslant u, C_2 \geqslant u)}\}$ with $N_{1i}(u) = I(Y_i \leqslant u)$, $N_{2i}(u) = I(Z_i \leqslant u, \delta_i = 1)$ and $C_2 = C - T - Y$. Note that (A.8) is a sum of n i.i.d. random variables, and the expectation of each term in (A.8) is zero. By the central limit theorem, (A.8) converges weakly to normal with zero mean and variance $\Sigma(y, z)$.

Moreover, we have

$$\begin{aligned} & E\left[\Big\{\frac{\partial \log l\{\alpha, \hat{S}_Y(y), \hat{S}_Z(z)\}}{\partial \alpha}\Big\}\{I_1^0(Y_i) + I_2^0(X_i, \delta_i)\}\right] \\ = \;& E\left[\{I_1^0(Y_i) + I_2^0(X_i, \delta_i)\} E\Big\{\frac{\partial \log l\{\alpha, \hat{S}_Y(y), \hat{S}_Z(z)\}}{\partial \alpha} \mid Y_i, X_i, \delta_i\Big\}\right] = 0 \end{aligned} \tag{A.9}$$

which means (A.7) and (A.8) are asymptotically orthogonal. Therefore, $(A.7)$, $(A.8)$ and $(A.9)$ imply that as $n \to \infty$, 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 covariance function $[\frac{\partial C\{\alpha, S_Y(y), S_Z(z)\}}{\partial \alpha}]^2 \sigma_1^2 + \Sigma(y, z)$.

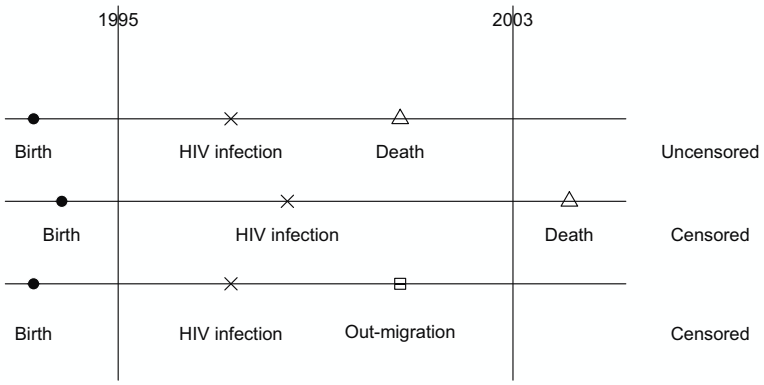

Figure 1. Exploratory plot of data for a cohort of Rakai HIV seroconverters.

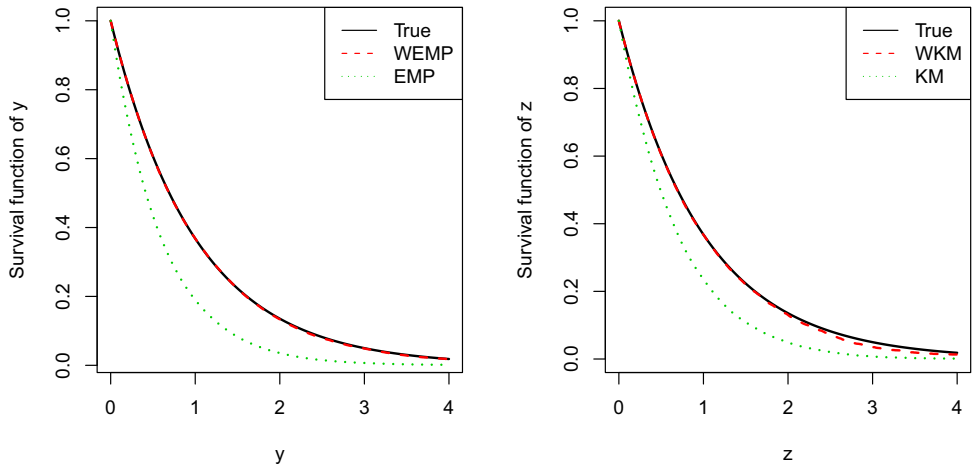

Figure 2. Simulation results of estimations of marginal survival functions of Y and Z: solid lines represent the true survival functions, dashed lines represent the weighted estimates by proposed methods, and dotted lines represent the estimates by conventional methods.

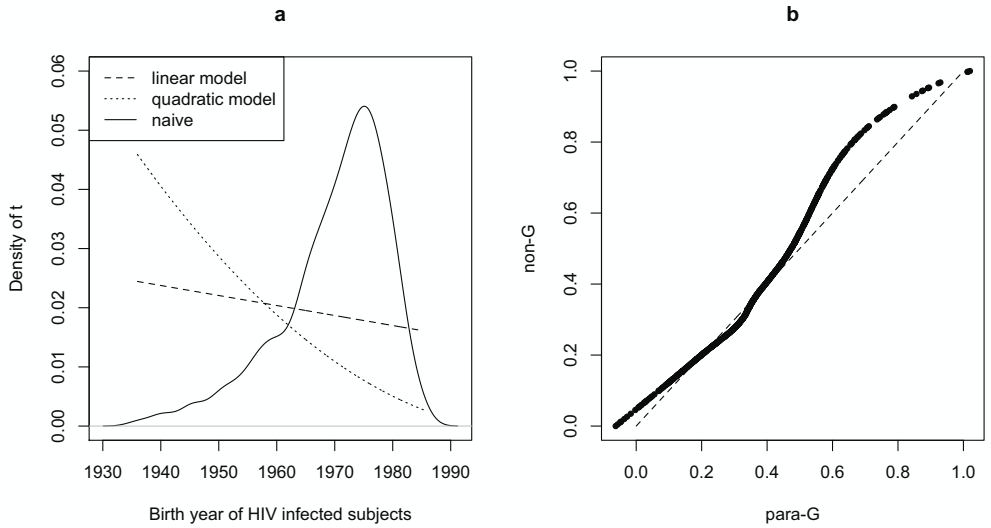

Figure 3. (a) Birth density plots: solid line represents the biased empirical estimate, dashed line represents the estimate from linear model fit, and dotted line represents the estimate from quadratic model fit. (b) Scatter plot of $\hat{G}_n(t)$ (non-G) against $\hat{G}(t, \hat{\theta})$ (para-G): dashed line represents $y = x$ as reference.

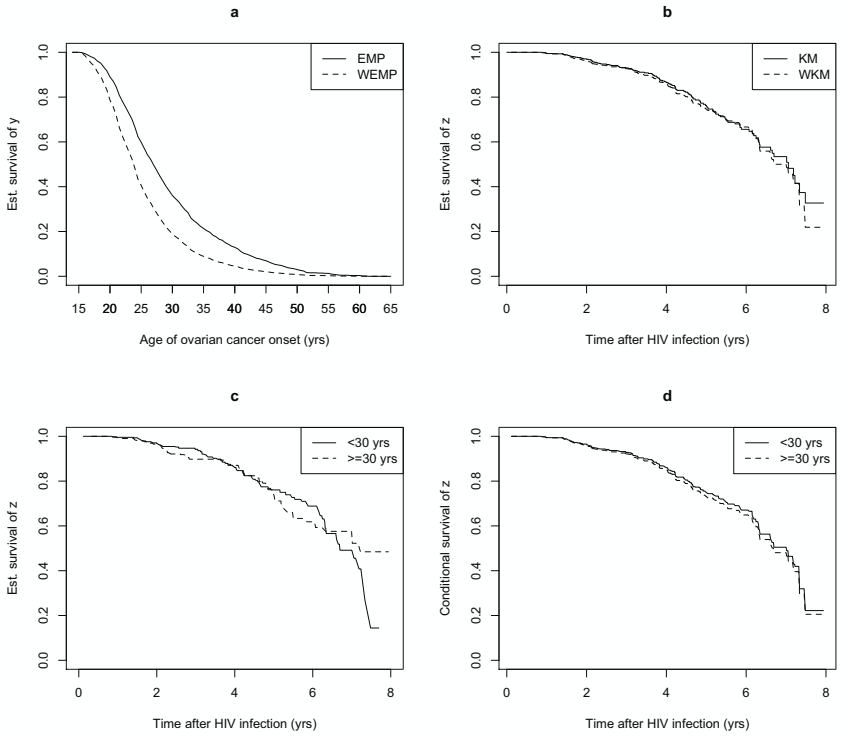

Figure 4. (a) Estimated marginal survival functions of age at infection (EMP vs WEMP). (b) Estimated marginal survival functions of residual lifetime (KM vs WKM). (c) Estimated marginal survival functions of residual lifetime for different categories of age at infection. (d) Estimated conditional survival functions of residual lifetime given different categories of age at infection.

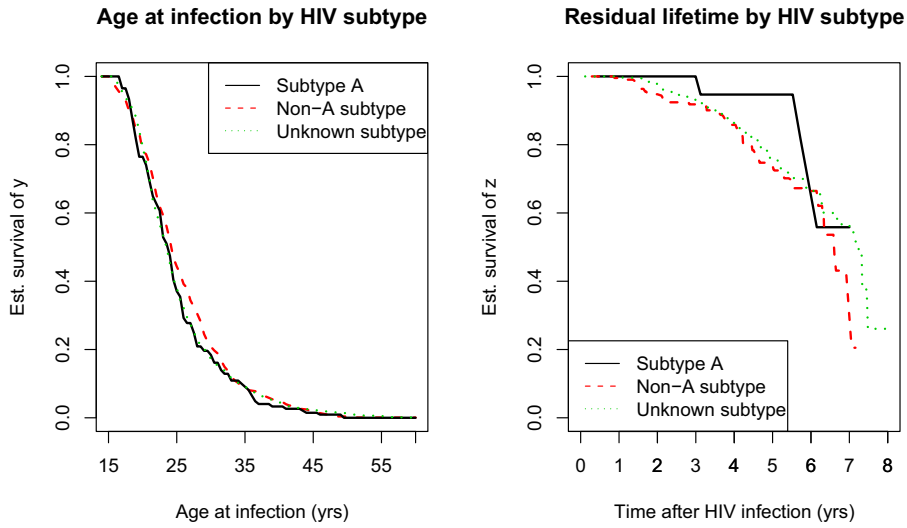

Figure 5. Estimated marginal survival functions of age at infection and residual lifetime by HIV subtype: solid lines represent curves for A subtype, dashed lines represent curves for non-A subtype, and dotted lines represent curves for unknown subtype.

Table 1
Simulation summary statistics of $(\hat{\theta}, \hat{\alpha}, \hat{S}_1)$ *under semi-stationary condition*

Model	θ	$Bias(\hat{\theta})$	$SE_m(\hat{\theta})$	α	$Bias(\hat{\alpha})$	$SE_e(\hat{\alpha})$	$SE_m(\hat{\alpha})$	$CP(\hat{\alpha})$	$Bias(\hat{S}_1)$	$SE_e(\hat{S}_1)$	$CP(\hat{S}_1)$
Clayton	1.0	0.003	0.084	0.50	0.011	0.154	0.132	95.3	-0.003	0.031	94.6
	1.0	0.001	0.072	1.33	0.028	0.243	0.186	95.6	-0.002	0.030	95.7
	1.0	0.001	0.076	3.00	0.085	0.455	0.421	95.8	-0.003	0.031	93.7
	2.0	0.002	0.128	0.50	0.005	0.242	0.230	95.5	-0.002	0.042	95.2
	2.0	0.007	0.137	1.33	0.037	0.368	0.346	95.7	-0.001	0.042	94.6
	2.0	0.009	0.131	3.00	0.131	0.580	0.562	96.0	-0.006	0.043	93.1
Pos. Stab.	1.0	0.001	0.072	1.25	0.018	0.066	0.051	93.2	0.002	0.032	95.1
	1.0	0.006	0.079	1.67	0.018	0.104	0.085	93.5	-0.001	0.031	94.5
	1.0	0.002	0.076	2.50	0.008	0.167	0.147	94.2	-0.003	0.031	94.8
	2.0	0.011	0.133	1.25	0.021	0.073	0.059	93.4	0.001	0.044	94.5
	2.0	0.019	0.143	1.67	0.010	0.116	0.102	93.8	-0.003	0.042	95.3
	2.0	0.012	0.133	2.50	0.015	0.210	0.193	94.5	-0.002	0.043	95.7
Frank	1.0	0.007	0.079	2.00	0.013	0.425	0.413	95.5	0.001	0.030	94.8
	1.0	0.002	0.075	-1.00	0.035	0.427	0.411	94.8	0.001	0.032	95.2
	1.0	0.001	0.078	-2.00	0.006	0.428	0.415	95.7	-0.001	0.031	94.1
	2.0	0.001	0.135	2.00	0.038	0.571	0.554	95.8	-0.002	0.042	96.0
	2.0	0.012	0.014	-1.00	0.030	0.526	0.512	95.1	-0.001	0.043	94.6
	2.0	0.008	0.136	-2.00	0.026	0.545	0.551	95.6	-0.002	0.043	96.0

Bias: empirical bias; SE_m : average of model-based standard error estimates; SE_e: empirical standard error; CP: 95% nominal coverage probability; $\hat{S}_1 = \hat{S}_{Y,Z}(0.22, 0.51)$.