Biometrics 000, 1–23 DOI: 000
000 0000

Estimation of Kendall's *Tau* For Bivariate Survival Data with Truncation

Hong Zhu* and Mei-Cheng Wang**

*Division of Biostatistics, College of Public Health, The Ohio State University,
248 Cunz Hall, 1841 Neil Avenue, Columbus, Ohio 43210, U.S.A.
email: hzhu@cph.osu.edu

**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: Quantifying dependence between bivariate survival data has long been a major topic in many areas of risk analysis including environmental risk, financial risk, and health risk, etc. This leads to an intense interest in, and development of, estimation of Kendall's *tau*, τ, one of the most popular measures of association between two random variables. For bivariate right-censored data, several nonparametric estimators of τ have been developed by Oakes (1982), Wang and Wells (2000) and Lakhal et al. (2009). In many follow-up studies that involve cross-sectional sampling, truncated bivariate survival data arise when one or both components of failure times is observed only if it falls within a specific truncation set. Little work has been done, however, for the nonparametric estimation of τ under both truncation and censoring. Additional selection bias due to truncation effect must be correctly handled in comparing a pair of observed failure times. In this paper, the approach of inverse probability weighting is employed to account for both comparability (Martin and Betensky, 2005) and orderability of pairs. Nonparametric estimators of τ for measuring the association are proposed for bivariate left-truncated data, and bivariate survival data with interval sampling where one component is subject to double truncation and the other is subject to possibly dependent right censoring. The methods developed are applicable to many patterns of truncation for bivariate survival data, including left truncation, right truncation and double truncation for one or both components, where no other nonparametric estimator of τ is currently available. Simulation studies demonstrate that the estimators perform well with moderate sample sizes. The method is applied to the investigation of the association between ovarian cancer onset-age and survival.

Key words: Bivariate survival data; Inverse probability weighting; Kendall's tau; Nonparametric estimation; Truncation; U-statistic.

1. Introduction

In many areas of risk analysis including health risk, measuring the strength of dependence between bivariate failure times has long been a major topic. For example, in genetic epidemiologic research, assessment of the patterns of disease association among family members is often the first step towards identifying disease-related genes. In HIV-AIDS studies, the dependence between the time from HIV infection to AIDS and the time from AIDS to death reveals useful information about the disease progression. Bivariate survival data typically fall into one of the two categories, 'parallel' and 'serial' (Lin et al., 1999). In particular, parallel data arise when each observation corresponds to a pair of failure times, and serial data are encountered when two consecutive failure events occur on one individual. In many follow-up studies involving cross-sectional sampling, bivariate survival data are collected subject to truncation in addition to right censoring, which means one or both components of the bivariate failure times is observed only if it falls within a specific truncation set. This type of data represents a non-randomly screened subset of a population, thus any analysis that fails to account for the biased nature of the sample is inappropriate. The need to accommodate truncation makes the analysis of survival data, especially multivariate survival data, a challenge subject. Bivariate estimation with truncated survival data has received considerable attention recently. Gurler (1997) developed estimation method for the case when only one component of the bivariate is subject to truncation. Van der Laan (1996) proposed a nonparametric maximum likelihood estimator for the bivariate distribution when both components are randomly truncated. This paper focuses on the dependent relationship between bivariate failure times, and proposes methodology for estimating the association for bivariate survival data with truncation.

Kendall's *tau*, τ, the rank correlation coefficient, is among the most popular measures of association between two random variables. Different from the well-known Pearson correlation coefficient, τ does not require parametric information about the marginal distribution. Because of its rank invariant property, it is suitable for measuring dependence in non-Gaussian lifetime models. Copulas have become an attractive tool for semiparametrically modeling bivariate survival data. It has been shown that τ is closely related to the association parameter in the semiparametric copula models proposed by Gumbel (1960), Clayton (1978) and Frank (1979). Therefore, estimator of the parameter in copula model can be naturally obtained by estimator of τ. For complete bivariate data, an empirical estimate of τ is given by Kendall and Gibbons (1990), which possesses desirable asymptotic properties by U-statistic theory. Under censoring, several estimators of τ have been developed by Brown, Hollander and Korwar (1974), Weier and Basu (1980) and Oakes (1982), but none of them is consistent when the two margins are not independent. Wang and Wells (2000) derived a consistent estimator for τ expressed as an integral of an estimate of the joint survival function. Lakhal et al. (2009) introduced a modification of the estimator proposed by Oakes (1982) using a Horvitz-Thompson-type correction for the pair that are not orderable. Little work has been done, however, for the nonparametric estimation of τ under both truncation and censoring. For truncated bivariate survival data, selection bias due to truncation effect must be correctly handled and attention should be further restricted to comparable paris (Martin and Betensky, 2005), in addition to orderable pairs and uncertainty of pair ranking due to censoring. In this paper, nonparametric estimators of τ are proposed using inverse probability weighting method, where the contribution of each comparable and orderable pair is weighted by the inverse of the associated probability. Specifically, the paper discusses the estimation methods for bivariate left-truncated survival data, and bivariate survival data with interval sampling where one component is subject to double truncation and the other is subject to possibly dependent censoring. Under suitable regularity conditions, the estimators are shown to be consistent and asymptotically normally distributed, with consistent bootstrap variance estimator. The methods developed are applicable to many patterns of truncation for bivariate survival data, including left truncation, right truncation and double truncation for one or both components, where no other nonparametric estimator of τ is currently available. The methods also avoid the need to estimate the bivariate survival function. Series of simulation studies demonstrate the estimators perform well with moderate sample sizes, and the method is applied to the investigation of the association between ovarian cancer onset-age and survival.

2. Nonparametric estimation of Kendall's τ

2.1 *Preliminaries*

Let (X, Y) be bivariate random variables, and let (X_i, Y_i) and (X_j, Y_j) $(i \neq j)$ be two independent replications from (X, Y). This pair is called concordant if $(X_i - X_j)(Y_i - Y_j) > 0$, and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. Kendall's *tau* (Kendall and Gibbons, 1990) is defined as

$$\begin{aligned} \tau &= Pr\{(X_i - X_j)(Y_i - Y_j) > 0\} - Pr\{(X_i - X_j)(Y_i - Y_j) < 0\} \\ &= E[sgn\{(X_i - X_j)(Y_i - Y_j)\}] \\ &= E(a_{ij} b_{ij}), \end{aligned}$$

where $a_{ij} = 2I(X_i - X_j > 0) - 1$ and $b_{ij} = 2I(Y_i - Y_j > 0) - 1$. It is easy to see that $-1 \leqslant \tau \leqslant 1$ and $\tau = 0$ if (X, Y) are independent. In absence of censoring, τ can be estimated by

$$\hat{\tau} = \binom{n}{2}^{-1} \sum_{i<j} a_{ij} b_{ij},$$

and it has been shown that $\hat{\tau}$ is a U-statistic, an unbiased estimate of τ, and $n^{1/2}(\hat{\tau} - \tau)$ is asymptotically normal as $n \to \infty$.

Under censoring, the concordance/discordance status may be not clear for some pairs, making estimation of τ difficult. Let C_X and C_Y be the censoring variables associated with X and Y respectively. The observed variables are $(\tilde{X}, \tilde{Y}, \delta_X, \delta_Y)$ where $\tilde{X} = \min(X, C_X)$,

$\tilde{Y} = \min(Y, C_Y)$, $\delta_X = I(X < C_X)$ and $\delta_Y = I(X < C_Y)$. Oakes (1982) showed that the pair (i, j) is orderable if $\{\tilde{X}_{ij} < \tilde{C}_{X,ij}, \tilde{Y}_{ij} < \tilde{C}_{Y,ij}\}$ where $\tilde{X}_{ij} = \min(X_i, X_j)$, $\tilde{C}_{X,ij} = \min(C_{X,i}, C_{X,j})$, $\tilde{Y}_{ij} = \min(Y_i, Y_j)$ and $\tilde{C}_{Y,ij} = \min(C_{Y,i}, C_{Y,j})$, and proposed an estimator of τ by summing over only the orderable pairs as

$$\hat{\tau}_o = \binom{n}{2}^{-1} \sum_{i<j} \delta_{ij} a_{ij} b_{ij},$$

where $\delta_{ij} = I(\tilde{X}_{ij} < \tilde{C}_{X,ij}, \tilde{Y}_{ij} < \tilde{C}_{Y,ij})$ is the indicator of an orderable pair. However, this ignores partial information provided by censored data. Other alternatives were proposed by Brown et al. (1974) and Weier and Basu (1980) by modifying a_{ij} and b_{ij}. However, all of these estimators are inconsistent when $\tau \neq 0$ and have bias increasing in τ. Nevertheless, $\hat{\tau}_o$ is widely used to test independence of a pair of random variables based on censored data. To improve the performance of $\hat{\tau}_o$, Oakes (2008) suggested a renormalized estimator of τ as

$$\hat{\tau}_{rn} = \frac{C - D}{C + D},$$

where C and D are the numbers of definite concordant pairs and discordant pairs respectively, and proved that $\hat{\tau}_{rn}$ is consistent when (X, Y) follow Clayton copula model. Lakhal et al. (2009) used the inverse probability weighting method to modify and reduce the bias of $\hat{\tau}_o$. The weighted estimator is given as

$$\hat{\tau}_{l1} = \binom{n}{2}^{-1} \sum_{i<j} \frac{\delta_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}},$$

where $\hat{p}_{ij}$ is an estimator of $p_{ij} = \mathrm{pr}(\tilde{\mathrm{X}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{X,ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{Y,ij}} | \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}})$, the selection probability that the pair (i, j) is orderable. Specifically,

$$\begin{aligned} p_{ij} &= \mathrm{pr}(\tilde{\mathrm{X}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{X,ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{Y,ij}} | \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}}) \\ &= \mathrm{pr}(\tilde{\mathrm{X}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{X,i}}, \tilde{\mathrm{X}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{X,j}}, \tilde{\mathrm{Y}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{Y,i}}, \tilde{\mathrm{Y}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{Y,j}} | \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}}) \\ &= \mathrm{pr}(\tilde{\mathrm{X}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{X}}, \tilde{\mathrm{Y}}_{\mathrm{ij}} < \mathrm{C}_{\mathrm{Y}} | \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}})^2. \end{aligned}$$

The estimation of p_{ij} for orderable pair depends on the censoring mechanism. For large values of $|\tau|$, $\hat{\tau}_{l1}$ may lie outside $[-1, 1]$. To address this issue, an alternative is given as

$$\hat{\tau}_{l2} = \Big(\sum_{i<j} \frac{\delta_{ij}}{\hat{p}_{ij}}\Big)^{-1} \sum_{i<j} \frac{\delta_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}},$$

which is shown to be consistent and asymptotically normally distributed.

2.2 *Estimation for bivariate left-truncated survival data*

Bivariate left-truncated survival data arise when one observes only those pairs in which both failure events occur after the corresponding left truncation events, and meanwhile, the failure events are observed subject to right censoring. They typically fall in the category of parallel bivariate survival data. Consider a general setting, and let (T_X, T_Y) denote bivariate left truncation times, and (C_X, C_Y) denote bivariate censoring times. Each observation is of the form $(\tilde{X}, \tilde{Y}, T_X, T_Y, \delta_X, \delta_Y) \mid (\tilde{X} \geqslant T_X, \tilde{Y} \geqslant T_Y)$, with bivariate observed failure times $\tilde{X} = \min(X, C_X)$ and $\tilde{Y} = \min(Y, C_Y)$, and bivariate censoring indicators $\delta_X = I(X < C_X)$ and $\delta_Y = I(X < C_Y)$. Assume that (X, Y) and (T_X, T_Y, C_X, C_Y) are independent. It is important, however, to indicate that the independence assumption could be violated, as discussed in Wang (1991). For instance, T is the truncation time from disease onset to recruitment and X is the survival time from onset to death. The independence of T and X may not hold, when an effective drug or treatment was developed and given to the diseased individuals during the process of observation.

Two examples of bivariate left-truncated survival data are given in the following. In a pediatric AIDS cohort study (Shen and Yan, 2008), bivariate data consist of (X, Y), where X is mother's AIDS incubation time and Y is the time from birth to development of AIDS for a child. A mother-child pair is selected only when both are HIV-positive and have not yet developed AIDS at the time of recruitment. Let T_X be the time from infection to recruitment for mother, and T_Y be the time from birth to recruitment for child. Bivariate failure times (X, Y) are not left-truncated only if $(X \geqslant T_X, Y \geqslant T_Y)$, and are subject to right censoring due to loss to follow-up or end of the study with censoring times (C_X, C_Y). Figure 1 highlights all the different times for bivariate left-truncated survival data as described in this example. Ino et al. (2001) discussed a bivariate left-truncated and right-censored dataset comprising pairs of survival data for 50 brain tumor patients. Time from diagnosis to initiation of radiation therapy and time from diagnosis to tumor progression are left-truncated by time from diagnosis to chemotherapy and right-censored by the end of follow-up or death. The data in this study are of the form $(\tilde{X}, \tilde{Y}, T, \delta_X, \delta_Y) \mid (\tilde{X} \geqslant T, \tilde{Y} \geqslant T)$, a special case of the general form discussed earlier.

[Figure 1 about here.]

The estimator of τ developed for untruncated survival data can be extended to truncated survival data by considering comparability (Martin and Betensky, 2005) in addition to orderability. The bivariate left-truncated failure times are observed only when they exceed the truncation times, which results in selection bias. Further, the concordance/discordance status of some observed pairs may not be established due to censoring. The task is to adjust for selection bias from truncation and handle uncertainty in pair order due to censoring. The proposed estimation method for τ generalizes Lakahl's modified τ estimator, $\hat{\tau}_{l2}$, for censored data, to data that are both left-truncated and right-censored. For such data, a_{ij} and b_{ij} are computable for pairs of failure times that are comparable and orderable. A nonparametric estimator of τ is provided using inverse probability weighting method, where the contribution of each comparable and orderable pair is weighted by the inverse of the associated probability. Identify the comparable and orderable pair and estimate the associated probability p_{ij}. The pair (i, j) is comparable, if $\{\tilde{T}_{X,ij} \leqslant \tilde{X}_{ij}, \tilde{T}_{Y,ij} \leqslant \tilde{Y}_{ij}\}$ where $\tilde{T}_{X,ij} = \max(T_{X,i}, T_{X,j})$, $\tilde{T}_{Y,ij} = \max(T_{Y,i}, T_{Y,j})$, $\tilde{X}_{ij} = \min(X_i, X_j)$ and $\tilde{Y}_{ij} = \min(Y_i, Y_j)$. The pair (i, j) is orderable if $\{\tilde{X}_{ij} < \tilde{C}_{X,ij}, \tilde{Y}_{ij} < \tilde{C}_{Y,ij}\}$ where $\tilde{C}_{X,ij} = \min(C_{X,i}, C_{X,j})$ and $\tilde{C}_{Y,ij} = \min(C_{Y,i}, C_{Y,j})$. Let $\omega_{ij} = I(\tilde{T}_{X,ij} \leqslant \tilde{X}_{ij} < \tilde{C}_{X,ij}, \tilde{T}_{Y,ij} \leqslant \tilde{Y}_{ij} < \tilde{C}_{Y,ij})$ be the indicator that a pair is comparable and orderable, then the corresponding conditional probability p_{ij} is

$$\begin{aligned} p_{ij} &= \mathrm{pr}(\tilde{\mathrm{T}}_{\mathrm{X,ij}} \leqslant \tilde{\mathrm{X}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{X,ij}}, \tilde{\mathrm{T}}_{\mathrm{Y,ij}} \leqslant \tilde{\mathrm{Y}}_{\mathrm{ij}} < \tilde{\mathrm{C}}_{\mathrm{Y,ij}} \mid \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}}) \\ &= \mathrm{pr}(\mathrm{T_X} \leqslant \tilde{\mathrm{X}}_{\mathrm{ij}} < \mathrm{C_X}, \mathrm{T_Y} \leqslant \tilde{\mathrm{Y}}_{\mathrm{ij}} < \mathrm{C_Y} \mid \tilde{\mathrm{X}}_{\mathrm{ij}}, \tilde{\mathrm{Y}}_{\mathrm{ij}})^2. \end{aligned}$$

Suppose that (T_X, C_X) and (T_Y, C_Y) are independent, and denote the bivariate distributions of (T_X, C_X) by $H_X(\cdot,\cdot)$ and (T_Y, C_Y) by $H_Y(\cdot,\cdot)$, and the distributions of truncation time T_X by $G_X(\cdot)$ and T_Y by $G_Y(\cdot)$. Then, the probability of being comparable and orderable, p_{ij}, could be expressed as

$$p_{ij} = [\{G_X(\tilde{X}_{ij}) - H_X(\tilde{X}_{ij}, \tilde{X}_{ij})\}\{G_Y(\tilde{Y}_{ij}) - H_Y(\tilde{Y}_{ij}, \tilde{Y}_{ij})\}]^2.$$

An estimator of p_{ij}, $\hat{p}_{ij}$, is obtained by replacing H_X, H_Y, G_X and G_Y by the corresponding NPMLEs for left-truncated and right-censored data (Wang, 1991). Kendall's τ for bivariate left-truncated survival data is estimated by

$$\hat{\tau}_{lr} = \Big(\sum_{i<j} \frac{\omega_{ij}}{\hat{p}_{ij}}\Big)^{-1} \sum_{i<j} \frac{\omega_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}}.$$

The asymptotic theory of $\hat{\tau}_{lr}$ is facilitated by U-statistic techniques. Define an un-rescaled estimator $\hat{\tau}_u$ as

$$\hat{\tau}_u = \binom{n}{2}^{-1} \sum_{i<j} \frac{\omega_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}}.$$

Under some regularity conditions, in the Appendix it is shown that $n^{1/2}(\hat{\tau}_u - \tau)$ is asymptotically equivalent to a zero-mean U-statistic of order 2. Similar asymptotic result is available for $\hat{\tau}_{lr}$. Consistency and asymptotic normality follow the lines in Van der Vaart (1998), which are summarized in Theorem 1 with the proof provided in the Appendix.

Theorem 1: *The estimator $\hat{\tau}_{lr}$ is consistent, and as $n \to \infty$, $n^{1/2}(\hat{\tau}_{lr} - \tau)$ converges weakly to normal distribution with mean zero and variance σ^2_{lr}.*

The variance has rather complex form, therefore, for consistent estimation of the variance of $\hat{\tau}_{lr}$ and for asymptotic confidence interval calculations, a bootstrap procedure may be used.

2.3 *Estimation for bivariate survival data with interval sampling*

In many studies of disease natural history, interest often lies in consecutive failure events and the association between event times. For instance, considering the progression of cancer through successive stages, birth is the initiating event, cancer-onset and death are the serial bivariate failure events. Disease registry or surveillance systems commonly collect data with onset of disease occurring within a calendar time interval. This type of sampling is referred to as interval sampling, where the initiating event (i.e., birth) is retrospectively identified and the subsequent failure event (i.e., death) is observed during the follow-up. An example of such data is ovarian cancer registry data collected by SEER (Surveillance, Epidemiology, and End Results) program between 1973 and 2002, which is described in more details in data analysis. The fact that the data are collected conditioning on the first failure event (i.e., cancer-onset) occurring within a time interval results in sampling bias. The data fall in the category of serial bivariate survival data, and Figure 2 shows the schema of it. A nonparametric estimator of Kendall's τ is proposed for this type of data to measure the association, particularly, adjusting for the substantial sampling bias.

[Figure 2 about here.]

Suppose the study cohort under interval sampling is made up of subjects experiencing the first failure event within a calendar time interval $[0, t_0]$. Denote the time from initiating event to the beginning of data collection by T, the time from the initiating event to the first failure event by X, and the time from the first event to the second event by Y. Typically, the follow-up process is subject to independent right censoring, with the time from the initiating event to the time of censoring denoted by $C \leqslant T + t_0$. Bivariate failure times (X, Y) are the outcome of interest. Assume that T and (X, Y) are independent, as required by most of the methods for analyzing survival data under random truncation. Due to interval sampling, X is doubly truncated with the constraint $T \leqslant X \leqslant T + t_0$, and Y is dependently right censored with censoring time $C - X$. Under this setting, one may observe $(T, X, \tilde{Y}, \delta)$ $|\mathrm{T}\leqslant X \leqslant T + t_0$, where $\tilde{Y} = \min(Y, C - X)$ and $\delta = I(Y < C - X)$. Identify the comparable and orderable pair, and compute the associated probability p_{ij} for bivariate survival data with interval sampling . The pair (i, j) is comparable if $\{T_{ij}^{max} \leqslant X_{ij}^{min}, X_{ij}^{max} \leqslant t_0 + T_{ij}^{min}\}$, where $T_{ij}^{max} = \max(T_i, T_j)$, $T_{ij}^{min} = \min(T_i, T_j)$, $X_{ij}^{max} = \max(X_i, X_j)$, and $X_{ij}^{min} = \min(X_i, X_j)$. The pair is orderable if $\{Y_{ij}^{min} < \min(C_i - X_i, C_j - X_j)\}$ where $Y_{ij}^{min} = \min(Y_i, Y_j)$. The indicator of comparability and orderability is denoted by $\lambda_{ij} = I\{T_{ij}^{max} \leqslant X_{ij}^{min}, X_{ij}^{max} \leqslant t_0 + T_{ij}^{min}, Y_{ij}^{min} < \min(C_i - X_i, C_j - X_j)\}$. Assume that truncation time T and overall censoring time C are independent, the conditional probability that a pair being comparable and orderable is

$$
\begin{aligned}
p_{ij} &= \mathrm{pr}(\mathrm{T}_{\mathrm{ij}}^{\mathrm{max}} \leqslant \mathrm{X}_{\mathrm{ij}}^{\mathrm{min}}, \mathrm{X}_{\mathrm{ij}}^{\mathrm{max}} \leqslant \mathrm{t}_0 + \mathrm{T}_{\mathrm{ij}}^{\mathrm{min}}, \mathrm{Y}_{\mathrm{ij}}^{\mathrm{min}} < \min(\mathrm{C}_{\mathrm{i}} - \mathrm{X}_{\mathrm{i}}, \mathrm{C}_{\mathrm{j}} - \mathrm{X}_{\mathrm{j}}) \mid \mathrm{X}_{\mathrm{i}}, \mathrm{X}_{\mathrm{j}}, \mathrm{Y}_{\mathrm{ij}}^{\mathrm{min}}) \\
&= \mathrm{pr}(\mathrm{X}_{\mathrm{ij}}^{\mathrm{max}} - \mathrm{t}_0 \leqslant \mathrm{T}_{\mathrm{i}} \leqslant \mathrm{X}_{\mathrm{ij}}^{\mathrm{min}}, \mathrm{X}_{\mathrm{ij}}^{\mathrm{max}} - \mathrm{t}_0 \leqslant \mathrm{T}_{\mathrm{j}} \leqslant \mathrm{X}_{\mathrm{ij}}^{\mathrm{min}}, \\
&\quad C_i > X_i + Y_{ij}^{min}, C_j > X_j + Y_{ij}^{min} \mid X_i, X_j, Y_{ij}^{min}) \\
&= Pr(X_{ij}^{max} - t_0 \leqslant T \leqslant X_{ij}^{min} | X_i, X_j)^2 \times Pr(C_i > X_i + Y_{ij}^{min} | X_i, X_j, Y_{ij}^{min}) \\
&\quad \times Pr(C_j > X_j + Y_{ij}^{min} \mid X_i, X_j, Y_{ij}^{min}).
\end{aligned}
$$

Denote the distribution function of T by $G(\cdot)$, and the survival function of C by $K(\cdot)$. Then, p_{ij} could be expressed as

$$
p_{ij} = \{G(X_{ij}^{min}) - G(X_{ij}^{max} - t_0)\}^2 \times K(X_i + Y_{ij}^{min}) \times K(X_j + Y_{ij}^{min}).
$$

This probability could be estimated by replacing G and K by the corresponding appropriate estimators. As mentioned before, assume the overall follow-up process is subject to independent censoring. Thus, for the overall censoring time C, its survival function $K(\cdot)$, can be estimated by the Kaplan–Meier estimator $\hat{K}(\cdot)$ based on $\{(x_i + y_i, 1 - \delta_i)\}$. Then, discuss the estimation of the distribution function $G(\cdot)$ of truncation time T. Nonparametric method has been developed to analyze doubly-truncated data by Efron and Petrosian (1999). Since T is also doubly truncated with the constraint $X - t_0 \leqslant T \leqslant X$, estimating $G(\cdot)$ is essentially dual to estimating the distribution function of X. Shen (2008) provided an algorithm to jointly compute the NPMLEs for both $G(\cdot)$ and distribution function of X. Thus, $G(\cdot)$ can be nonparametrically estimated by $\hat{G}_n(\cdot)$ by this approach. Or alternatively, if some parametric information about the distribution of T is available, the following parametric method would provide a more efficient estimator of $G(\cdot)$. Consider a joint model of (T, X) and parameterize the distribution of T by $G(\cdot, \theta)$. The joint density of observed $\{t, x\}$ can be expressed as

$$\begin{aligned} p_{T,X}(t,x) &= \frac{g(t) f_X(x) I(x - t_0 \leqslant t \leqslant x)}{Pr(T \leqslant X \leqslant T + t_0)} \\ &= \frac{g(t) I(x - t_0 \leqslant t \leqslant x)}{G(x) - G(x - t_0)} \times \frac{\{G(x) - G(x - t_0)\} f_X(x)}{\int \{G(u) - G(u - t_0)\} f_X(u) du} \\ &= p_{T|X}(t \mid x) \times p_X(x), \end{aligned}$$

where g and f_X are the population densities of T and X respectively. Therefore, the conditional likelihood function of observed $\{t\}$ given $\{x\}$ is

$$L_c(\theta) = \prod_i p_{T|X}(t_i \mid x_i, \theta) = \prod_i \frac{g(t_i, \theta)}{G(x_i, \theta) - G(x_i - t_0, \theta)},$$

which could be used to estimate θ, and a parametric estimator of $G(\cdot)$ is obtained as $\hat{G}(\cdot, \hat{\theta})$. With $G(\cdot)$ either nonparametrically or parametrically estimated by $\hat{G}_n(\cdot)$ or $\hat{G}(\cdot, \hat{\theta})$, and $K(\cdot)$ estimated by the Kaplan–Meier estimator $\hat{K}(\cdot)$, an estimator of the conditional probability p_{ij} is

$$\hat{p}_{ij} = \{\hat{G}(X_{ij}^{min}) - \hat{G}(X_{ij}^{max} - t_0)\}^2 \times \hat{K}(X_i + Y_{ij}^{min}) \times \hat{K}(X_j + Y_{ij}^{min}).$$

Kendall's τ for bivariate survival data with interval sampling is estimated by

$$\hat{\tau}_{is} = \Big(\sum_{i<j} \frac{\lambda_{ij}}{\hat{p}_{ij}} \Big)^{-1} \sum_{i<j} \frac{\lambda_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}}.$$

Similar to the discussion in Section 2.2, define an un-rescaled estimator $\hat{\tau}_v$ as

$$\hat{\tau}_v = \binom{n}{2}^{-1} \sum_{i<j} \frac{\lambda_{ij} a_{ij} b_{ij}}{\hat{p}_{ij}}.$$

Under some regularity conditions, $n^{1/2}(\hat{\tau}_v - \tau)$ is asymptotically equivalent to a zero-mean U-statistic of order 2, and similar results hold for $n^{1/2}(\hat{\tau}_{ls} - \tau)$. Theorem 2 summarizes the asymptotic properties of $\hat{\tau}_{is}$ with the proof provided in the Appendix.

THEOREM 2: *The estimator $\hat{\tau}_{is}$ is consistent, and as $n \to \infty$, $n^{1/2}(\hat{\tau}_{is} - \tau)$ converges weakly to normal distribution with mean zero and variance σ^2_{is}.*

The computation of variance involves complex formulas and bootstrap method provides a direct and robust way to estimate variance in practice.

3. Simulations

The first set of simulations was carried out to assess the performance of the nonparametric estimator $\hat{\tau}_{lr}$ for bivariate left-truncated survival data and the inference procedure under moderate sample sizes. Bivariate failure times (X, Y) were generated from the Gumbel copula (Gumbel, 1960)

$$S_{X,Y}(x, y) = \exp\{-([-\log\{S_X(x)\}]^\alpha + [-\log\{S_Y(y)\}]^\alpha)^{1/\alpha}\},\ \alpha \geqslant 1,$$

where $\tau = (\alpha - 1)/\alpha$, with unit exponential margins and a Kendall's τ of 0.2, 0.5 and 0.8. Truncation vectors (T_X, T_Y) were generated from independent identical exponential distribution with $T = 3W - 2$ with $W \sim \exp(2)$, and censoring vectors were defined as $(C_X, C_Y) = (T_X + 3, T_Y + 3)$. The truncation rate and the censoring rate were both about 15% to 20%. For each value of parameter τ, 1000 simulated samples were generated with sample sizes $n = 200$ and $n = 400$ respectively. Table 1 summarizes the empirical biases, empirical standard errors, average of bootstrap standard errors, mean squared errors and 95% coverage probabilities for the nonparametric estimator $\hat{\tau}_{lr}$. Confidence interval was constructed based on asymptotic normality, in which the standard error of $\hat{\tau}_{lr}$ was computed using 500 bootstrap resamples. The empirical 95% coverage probability was obtained based on the 1000 confidence intervals. Under all simulation conditions, the estimators are approximately unbiased with small mean squared errors and the empirical coverage probabilities close to 95%. The empirical standard error and the average of bootstrap standard errors are very close, suggesting satisfactory performance of the inferential result on $\hat{\tau}_{lr}$. The bootstrap method provides good precision estimation of $\hat{\tau}_{lr}$. The standard error decreases as sample size increases or Kendall's τ increases.

[Table 1 about here.]

The second set of simulations was conducted to examine the finite sample performance of the nonparametric estimator $\hat{\tau}_{is}^{n}$ for bivariate survival data with interval sampling. For bivariate survival data, Shih and Louis (1995) proposed a semiparametric estimator for the association through a copula model-based two-stage procedure. This method was extended to bivariate survival data with interval sampling in Zhu (2010) and an associated manuscript (Zhu and Wang, 2011). The nonparametric estimator $\hat{\tau}_{is}^{n}$ of Kendall's τ and the copula model-based semiparametric estimator $\hat{\tau}_{is}^{s}$ were compared through simulations. A set of data $\{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\}$ was generated with interval sampling: $T = 13W - 9$ with $W \sim$ u(0,1), and correlated pairs (X, Y) were generated from the Clayton copula (Clayton, 1978)

$$S_{X,Y}(x, y) = \{S_X(x)^{-\alpha} + S_Y(y)^{-\alpha} - 1\}^{-1/\alpha},\ \alpha > 0,$$

where $\tau = \alpha/(2 + \alpha)$, with unit exponential margins and a Kendall's τ of 0.2, 0.5 and 0.8. An observation (t, x, y) was included in the data set if and only if $t \leqslant x \leqslant t + 10$ and was censored if $x + y \geqslant t + 10$. The censoring fraction was around 20%. For each value of τ, 1000 simulated samples were generated with $n = 200$ and $n = 400$ respectively. The simulation results are presented in Table 2, which includes the empirical biases, empirical standard errors, average of bootstrap standard errors and 95% coverage probabilities for $\hat{\tau}_{is}^{n}$ and $\hat{\tau}_{is}^{s}$. The nonparametric estimator $\hat{\tau}_{is}^{n}$ works well with small bias, and the bias of $\hat{\tau}_{is}^{n}$ is comparable to that of $\hat{\tau}_{is}^{s}$. However, the semiparametric method relies on a correct copula model specification, thus it is not as robust as the nonparametric method and may lose its efficiency when the assumed copula model is mis-specified. The empirical standard error of $\hat{\tau}_{is}^{n}$ is very close to the average of bootstrap standard errors and the empirical coverage probabilities of $\hat{\tau}_{is}^{n}$ are reasonably close to 95%, which may imply that the inference about τ based on the nonparametric method is reasonably good and suggest that the bootstrap estimator of the variance σ_{is}^{2} provides an appropriate measure of the variability of $\hat{\tau}_{is}^{n}$.

[Table 2 about here.]

4. Application to SEER ovarian cancer registry data

The proposed method was applied to ovarian cancer risk analysis, in the investigation of the association between age of cancer onset and residual lifetime based on the cancer registry data collected by Surveillance, Epidemiology and End-Results program. Knowledge of the strength of the association may provide important leads with respect to ovarian cancer etiology. This epidemiological surveillance system consists of population-based cancer registries designed to track cancer incidence and survival in the United States. Data collection began from January 1, 1973 (Ries et al., 2005) and the registries routinely collect information on newly diagnosed cancer patients. In the ovarian cancer registry dataset, there are 36,728 cancer patients diagnosed from 1973 to 2002, among whom 24,236 died before December 31, 2002. Clearly, the observed data are biased due to interval sampling. Any estimation or analysis method without consideration of this fact could possibly yield biased results. The ovarian cancer data were analyzed to estimate Kendall's τ as a measure of the association, adjusting for the bias from interval sampling.

The time from birth to the start of the data collection is denoted by T, and the bivariate failure times are age at cancer onset, X, and residual lifetime, Y. Under interval sampling, X is doubly truncated with the constraint $T \leqslant X \leqslant T+30$, and Y is dependently right censored. To illustrate the method, the analysis focuses on three groups of non-white ovarian cancer patients: 1,325 women born between 1920 and 1930, 911 women born between 1930 and 1940, and 251 women born between 1940 and 1950. A preliminary analysis by Cox regression model of cancer mortality conditional on age at onset suggests a significant negative association, but this does not account for interval sampling, nor gives an explicit measure of the association. Therefore, in the analysis, Kendall's τ is nonparametrically estimated by $\hat{\tau}_{is}^{n}$ and semiparametrically estimated by $\hat{\tau}_{is}^{s}$ for the three groups, where $\hat{\tau}_{is}^{s}$ is obtained by fitting the Frank model (Frank, 1979). Table 3 presents the point estimators, bootstrap standard errors and 95% confidence intervals for $\hat{\tau}_{is}^{n}$ and $\hat{\tau}_{is}^{s}$, respectively. Confidence interval is constructed based on the asymptotic normality, in which the standard error is computed using 500 bootstrap resamples. It is found that $\hat{\tau}_{is}^{n} = -0.510$ $(se = 0.048)$ and $\hat{\tau}_{is}^{s} = -0.340$ $(se = 0.037)$ for the 1920-1930 birth cohort, $\hat{\tau}_{is}^{n} = -0.549$ $(se = 0.017)$ and $\hat{\tau}_{is}^{s} = -0.276$ $(se = 0.069)$ for the 1930-1940 birth cohort, and $\hat{\tau}_{is}^{n} = -0.515$ $(se = 0.068)$ and $\hat{\tau}_{is}^{s} = -0.190$ $(se = 0.130)$ for the 1940-1950 birth cohort. Based on $\hat{\tau}_{is}^{n}$, significant negative associations are detected between X and Y in all of the three groups, and the magnitudes of the associations are comparable among these groups. The semiparametric estimator $\hat{\tau}_{is}^{s}$ suggests smaller negative associations compared with $\hat{\tau}_{is}^{n}$, and the association for the 1940-1950 birth cohort is no longer significant possibly due to small sample size. It is noticed that the semiparametric method relies on a specific copula, Frank model, thus it is less robust and possibly subject to model misspecification. The difference between the nonparametric and semiparametric methods may indicate a certain level of lack of fit of the Frank model and thus model selection procedure for copulas needs to be developed, though this is beyond the scope of the current paper.

[Table 3 about here.]

5. Discussion

Methodology for analyzing bivariate survival data in risk analysis must ensure the correctness of inferences, accounting for the dependence between bivariate failure times. This paper proposes the nonparametric estimation method of Kendall's *tau* for measuring the strength of dependence between bivariate survival data with truncation. Such data are important due to the common use of cross-sectional sampling and prevalence sampling designs in practice. There has been growing interest in assessing the relationship between bivariate failure times, which could be a pair of parallel failure times in family studies or two serial failure times in studies of disease natural history. Although many types of nonparametric estimator of τ have been proposed for bivariate survival data, until now few work on nonparametric estimation of τ for bivariate survival data with truncation has appeared in the literature. In estimation of τ, uncertainty in determining concordant or discordant status caused by censoring should be carefully handled. Furthermore, for truncated bivariate survival data, attention must be restricted to comparable pairs of bivariate failure times in addition to orderable pairs. Oakes (2008) pointed out this issue and briefly discussed the extension of the renomalized estimator to the case when bivariate data are subject to both left truncation and right censoring. In this article, the inverse probability weighting approach is employed to estimate τ for bivariate survival data with truncation, where the contribution of each comparable and orderable pair is weighted by the inverse of the associated probability. Under suitable regularity conditions, the proposed estimators are shown to possess desirable asymptotic properties. Simulations demonstrate the proposed methods perform well. For bivariate survival data with interval sampling, nonparametric estimation method and copula model-based semiparametric estimation method are compared in simulations and data analysis, which suggests semiparametric method can be less biased when a copula model is correctly assumed but it is not robust to model mis-specification.

Potentially, the nonparametric estimation of Kendall's τ may be used to develop model selection procedure or goodness-of-fit test of copulas, which could increase the practical utility of copula model.

Acknowledgments

References

Brown, B. W., Hollander, M., and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. *Reliability and Biometry* 327–354.

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

Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* **94,** 824–834.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19,** 194–226.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55,** 698–707.

Gurler, U. (1997). Bivariate distribution and hazard functions when a component is randomly truncated. *Journal of Multivariate Analysis* **60,** 20–47.

Ino, Y., Betensky, R. A., Zlatescu, M. C., Sasaki, H., Macdonald, D. R., Stemmer-Rachamimov, A. O., Ramsay, D. A., Cairncross, J. G. and Louis, D. N. (2001). Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clinical Cancer Research* **7,** 839–845.

Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*, 5th edition. A Charles Griffin Title. London: Edward Arnold.

Lakhal-Chaieb, L., Rivest, L.-P., and Beaudoin, D. (2009). IPCW estimator for Kendall's tau under bivariate censoring. *International Journal of Biostatistics* **5,** 1121–1141.

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

Martin, E. C. and Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *Journal of the American Statistical Association* **100,** 484–492.

Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics* **38(2),** 451–455.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73 (2),** 353–361.

Oakes, D. (2008). On consistency of Kendall's tau under censoring. *Biometrika* **95(4),** 997–1001.

Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hankey, B. F., Miller, B. A., Clegg, L., Mariotto, A., Feuer, E. J. and Edwards, B. K. (eds). (2005). *SEER Cancer Statistics Review, 1975-2002.* National Cancer Institute. Bethesda, MD.

Shen, P.-S. (2008). Nonparametric analysis of doubly truncated data. *Annals of the institute of statistical mathematics* **62,** 835–853.

Shen, P.-S. and Yan, Y.-F. (2008). Nonparametric estimation of the bivariate survival function with left-truncated and right-censored data. *Journal Statistical Planning and Inference* **138,** 4041–4054.

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

Van der Laan, M. J. (1996). Nonparametric estimation of the bivariate survival function with truncated data. *Journal of Multivariate Analysis* **58,** 107–131.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86,** 130–143.

Wang, W. and Wells, M. T. (2000). Estimation of Kendall's tau under censoring. *Statistica Sinica* **10,** 1199–1215.

Weier, D. R. and Basu, A. P. (1980). An investigation of Kendalls τ modified for censored data with applications. *Journal Statistical Planning and Inference* **4,** 381–390.

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 minor revision.

Appendix

Proof of Theorem 1

First show the asymptotic results of $\hat{\tau}_u$. One has

$$\begin{aligned} n^{1/2}(\hat{\tau}_u - \tau) &= n^{1/2}\left\{\binom{n}{2}^{-1}\sum_{i<j}\frac{\omega_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}} - \tau\right\} \\ &= n^{1/2}\binom{n}{2}^{-1}\sum_{i<j}\left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\right) \\ &\quad + n^{1/2}\binom{n}{2}^{-1}\sum_{i<j}\omega_{ij}a_{ij}b_{ij}\left(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}\right). \qquad (A.1) \end{aligned}$$

It is clear that the first term in (A.1) is a U-statistic of order 2, and

$$\begin{aligned} E\left\{\sum_{i<j}\left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\right)\right\} &= \sum_{i<j}E\left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\right) \\ &= \sum_{i<j}\left[E\left\{E\left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}}|\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right\} - \tau\right] \\ &= \sum_{i<j}\left[E\left\{\frac{1}{p_{ij}}E\left(\omega_{ij}a_{ij}b_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right\} - \tau\right], \end{aligned}$$

where $\tilde{X}_{ij} = \min(X_i, X_j)$ and $\tilde{Y}_{ij} = \min(Y_i, Y_j)$. Following the lines of Oakes (1986), $a_{ij}b_{ij}$ that indicates the concordance/discordance status is unconditionally independent of ω_{ij} that denotes the comparable and orderable event. Thus,

$$E(\omega_{ij}a_{ij}b_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}) = E(\omega_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij})E(a_{ij}b_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}) = p_{ij}E(a_{ij}b_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}).$$

Then, $E\Big\{\sum_{i<j}\Big(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Big)\Big\} = 0$. So the first term in (A.1) is a zero-mean U-statistic of order 2. By the asymptotic properties of U-statistic, the asymptotic variance of the first term is equal to

$$\begin{aligned} lim_{n\to\infty}\frac{4}{n^3} \quad \Bigg\{ & \sum_{i<j<k}\left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\right)\left(\frac{\omega_{ik}a_{ik}b_{ik}}{p_{ik}} - \tau\right) + \left(\frac{\omega_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\right)\left(\frac{\omega_{jk}a_{jk}b_{jk}}{p_{jk}} - \tau\right) \\ & + \left(\frac{\omega_{ik}a_{ik}b_{ik}}{p_{ik}} - \tau\right)\left(\frac{\omega_{jk}a_{jk}b_{jk}}{p_{jk}} - \tau\right)\Bigg\}. \end{aligned}$$

For the second term in (A.1), the additional variation created by estimating p_{ij} needs to be handled. From the discussion in Section 2.2,

$$\begin{aligned} \frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}} &= \frac{1}{[\{\hat{G}_X(\tilde{X}_{ij}) - \hat{H}_X(\tilde{X}_{ij}, \tilde{X}_{ij})\}\{\hat{G}_Y(\tilde{Y}_{ij}) - \hat{H}_Y(\tilde{Y}_{ij}, \tilde{Y}_{ij})\}]\}^2} \\ & \quad - \frac{1}{[\{G_X(\tilde{X}_{ij}) - H_X(\tilde{X}_{ij}, \tilde{X}_{ij})\}\{G_Y(\tilde{Y}_{ij}) - H_Y(\tilde{Y}_{ij}, \tilde{Y}_{ij})\}]^2}, \end{aligned}$$

where $\hat{G}$ and $\hat{H}$ are the NPMLEs, $[\{\hat{G}_X(t) - \hat{H}_X(t,t)\} - \{G_X(t) - H_X(t,t)\}]$ and $[\{\hat{G}_Y(s) - \hat{H}_Y(s,s)\} - \{G_Y(t) - H_Y(s,s)\}]$ can be approximated by a sum of n independent and identically distributed zero-mean terms (Wang, 1991), for simplicity denoted by $\hat{A}_1(t) - A_1(t)$ and $\hat{A}_2(s) - A_2(s)$ respectively. Then,

$$\begin{aligned} A_1(t) - \hat{A}_1(t) &= \frac{1}{n}\sum_{k=1}^{n}\phi_1(X_k, T_{X,k}, \delta_{X,k}, t) + o_p(n^{-1/2}), \\ A_2(s) - \hat{A}_2(s) &= \frac{1}{n}\sum_{k=1}^{n}\phi_2(Y_k, T_{Y,k}, \delta_{Y,k}, s) + o_p(n^{-1/2}), \end{aligned} \tag{A.2}$$

where $E\{\phi_1(X_k, T_{X,k}, \delta_{X,k}, t)\} = 0$ and $E\{\phi_2(Y_k, T_{Y,k}, \delta_{Y,k}, s)\} = 0$. Note that

$$n^{1/2}(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}) = 2n^{1/2}\frac{\{A_1(\tilde{X}_{ij})A_2(\tilde{Y}_{ij}) - \hat{A}_1(\tilde{X}_{ij})\hat{A}_2(\tilde{Y}_{ij})\}}{A_1(\tilde{X}_{ij})A_2(\tilde{Y}_{ij})\hat{A}_1^2(\tilde{X}_{ij})\hat{A}_2^2(\tilde{Y}_{ij})} + o_p(1).$$

Since $A_1(t)A_2(s) - \hat{A}_1(t)\hat{A}_2(s) = [A_1(t)\{A_2(s) - \hat{A}_2(s)\}] + [\hat{A}_2(s)\{A_1(t) - \hat{A}_1(t)\}]$, together with (A.2),

$$n^{1/2}(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}) \;=\; 2n^{-1/2}\frac{\{A_1(\tilde{X}_{ij})\sum_{k=1}^{n}\phi_2(\tilde{Y}_{ij}) + \hat{A}_2(\tilde{Y}_{ij})\sum_{k=1}^{n}\phi_1(\tilde{X}_{ij})\}}{A_1(\tilde{X}_{ij})A_2(\tilde{Y}_{ij})\hat{A}_1^2(\tilde{X}_{ij})\hat{A}_2^2(\tilde{Y}_{ij})} + o_p(1).$$

Therefore, the second term in (A.1) can be expressed as

$$n^{-1/2}\binom{n}{2}^{-1}\sum_{i<j}2\omega_{ij}a_{ij}b_{ij}\frac{\{A_1(\tilde{X}_{ij})\sum_{k=1}^{n}\phi_2(\tilde{Y}_{ij}) + \hat{A}_2(\tilde{Y}_{ij})\sum_{k=1}^{n}\phi_1(\tilde{X}_{ij})\}}{A_1(\tilde{X}_{ij})A_2(\tilde{Y}_{ij})\hat{A}_1^2(\tilde{X}_{ij})\hat{A}_2^2(\tilde{Y}_{ij})} + o_p(1)$$

$$= n^{-1/2}\sum_{k=1}^{n}\binom{n}{2}^{-1}\sum_{i<j}2\omega_{ij}a_{ij}b_{ij}\left\{\frac{\phi_2(\tilde{Y}_{ij})}{A_2(\tilde{Y}_{ij})\hat{A}_1^2(\tilde{X}_{ij})\hat{A}_2^2(\tilde{Y}_{ij})} + \frac{\phi_1(\tilde{X}_{ij})}{A_1(\tilde{X}_{ij})A_2(\tilde{Y}_{ij})\hat{A}_1^2(\tilde{X}_{ij})\hat{A}_2(\tilde{Y}_{ij})}\right\} + o_p(1)$$

$$= n^{-1/2}\sum_{k=1}^{n}2E\left[\omega_{12}a_{12}b_{12}\left\{\frac{\phi_2(\tilde{Y}_{12})}{A_2(\tilde{Y}_{12})\hat{A}_1^2(\tilde{X}_{12})\hat{A}_2^2(\tilde{Y}_{12})} + \frac{\phi_1(\tilde{X}_{12})}{A_1(\tilde{X}_{12})A_2(\tilde{Y}_{12})\hat{A}_1^2(\tilde{X}_{12})\hat{A}_2(\tilde{Y}_{12})}\right\}\right] + o_p(1),$$

which is a sum of independent and identically distributed zero-mean terms. Then, $n^{1/2}(\hat{\tau}_u - \tau)$ is asymptotically equivalent to a zero-mean U-statistic of order 2.

Next develop the asymptotic properties of $\hat{\tau}_{lr}$. One has

$$\begin{aligned}
n^{1/2}(\hat{\tau}_{lr} - \tau) &= n^{1/2}\left\{\frac{\binom{n}{2}^{-1}\sum_{i<j}\frac{\omega_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}}}{\binom{n}{2}^{-1}\sum_{i<j}\frac{\omega_{ij}}{\hat{p}_{ij}}} - \tau\right\} \\
&= n^{1/2}(\hat{\tau}_u - \tau) - n^{1/2}\tau\left\{\binom{n}{2}^{-1}\sum_{i<j}\frac{\omega_{ij}}{\hat{p}_{ij}} - 1\right\} + o_p(1) \\
&= n^{1/2}(\hat{\tau}_u - \tau) - n^{1/2}\tau\left\{\binom{n}{2}^{-1}\sum_{i<j}\frac{\omega_{ij}}{p_{ij}} - 1\right\} \\
&\quad - n^{1/2}\tau\left\{\binom{n}{2}^{-1}\left(\sum_{i<j}\frac{\omega_{ij}}{\hat{p}_{ij}} - \sum_{i<j}\frac{\omega_{ij}}{p_{ij}}\right)\right\} + o_p(1). \qquad (A.3)
\end{aligned}$$

It has been shown the first term in (A.3) is asymptotically equivalent to a zero-mean U-statistic of order 2. The second term in (A.3) is a sum of n independent and identically distributed zero-mean terms. Similar to the development of asymptotic results for the second term in (A.1), the third term in (A.3) is also asymptotically equivalent to a sum of independent and identically distributed zero-mean terms. Putting all of the three terms together, $n^{1/2}(\hat{\tau}_{lr} - \tau)$ is asymptotically equivalent to a zero-mean U-statistic of order 2. Following the lines in Van der Vaart (1998), $n^{1/2}(\hat{\tau}_{lr} - \tau)$ converges weakly to normal distribution with mean zero and variance σ^2_{lr} as $n \to \infty$.

Proof of Theorem 2

First develop the asymptotic results of $\hat{\tau}_v$. One has

$$\begin{aligned} n^{1/2}(\hat{\tau}_v - \tau) &= n^{1/2}\Bigg\{\binom{n}{2}^{-1}\sum_{i<j}\frac{\lambda_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}} - \tau\Bigg\} \\ &= n^{1/2}\binom{n}{2}^{-1}\sum_{i<j}\Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Bigg) \\ &\quad + n^{1/2}\binom{n}{2}^{-1}\sum_{i<j}\lambda_{ij}a_{ij}b_{ij}\Bigg(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}\Bigg). \end{aligned} \tag{A.4}$$

The first term in (A.4) is a U-statistic of order 2, and

$$\begin{aligned} E\Bigg\{\sum_{i<j}\Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Bigg)\Bigg\} &= \sum_{i<j}E\Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Bigg) \\ &= \sum_{i<j}\Bigg[E\Bigg\{E\Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}}|X_i, X_j, Y_{ij}^{min}\Bigg)\Bigg\} - \tau\Bigg] \\ &= \sum_{i<j}\Bigg[E\Bigg\{\frac{1}{p_{ij}}E\Big(\lambda_{ij}a_{ij}b_{ij}|X_i, X_j, Y_{ij}^{min}\Big)\Bigg\} - \tau\Bigg]. \end{aligned}$$

Since the concordance/discordance status is unconditionally independent of comparability and orderability of a pair (Oakes, 1986),

$$E(\lambda_{ij}a_{ij}b_{ij}|X_i, X_j, Y_{ij}^{min}) = E(\lambda_{ij}|X_i, X_j, Y_{ij}^{min})E(a_{ij}b_{ij}|X_i, X_j, Y_{ij}^{min}) = p_{ij}E(a_{ij}b_{ij}|X_i, X_j, Y_{ij}^{min}).$$

Then, $E\Big\{\sum_{i<j}\Big(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Big)\Big\} = 0$. So the first term in (A.4) is a zero-mean U-statistic of order 2. By the standard theory of U-statistic, the asymptotic variance of the first term is equal to

$$\begin{aligned} lim_{n\to\infty}\frac{4}{n^3} \quad \Bigg\{ & \sum_{i<j<k}\Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Bigg)\Bigg(\frac{\lambda_{ik}a_{ik}b_{ik}}{p_{ik}} - \tau\Bigg) + \Bigg(\frac{\lambda_{ij}a_{ij}b_{ij}}{p_{ij}} - \tau\Bigg)\Bigg(\frac{\lambda_{jk}a_{jk}b_{jk}}{p_{jk}} - \tau\Bigg) \\ & + \Bigg(\frac{\lambda_{ik}a_{ik}b_{ik}}{p_{ik}} - \tau\Bigg)\Bigg(\frac{\lambda_{jk}a_{jk}b_{jk}}{p_{jk}} - \tau\Bigg)\Bigg\}. \end{aligned}$$

The variation in the second term in (A.4) is due to the estimation of p_{ij}, the conditional probability that a pair is comparable and orderable.

$$\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}} = \frac{1}{(\Delta\hat{G})^2\hat{K}_i\hat{K}_j} - \frac{1}{(\Delta G)^2 K_i K_j},$$

where $\Delta G = G(X_{ij}^{min}) - G(X_{ij}^{max} - t_0)$, $K_i = K(X_i + Y_{ij}^{min})$, $K_j = K(X_j + Y_{ij}^{min})$, $\Delta\hat{G} = \hat{G}(X_{ij}^{min}) - \hat{G}(X_{ij}^{max} - t_0)$, $\hat{K}_i = \hat{K}(X_i + Y_{ij}^{min})$ and $\hat{K}_j = \hat{K}(X_j + Y_{ij}^{min})$. To be specific, $\hat{G}$ is either the NPMLE $\hat{G}_n(\cdot)$ (Shen, 2008) or the parametric estimator $\hat{G}(\cdot, \hat{\theta})$ from the conditional likelihood, and $\hat{K}$ is the Kaplan–Meier estimator. Note that $\Delta\hat{G} - \Delta G$ and $\hat{K} - K$ can be approximated by a sum of independent and identically distributed zero-mean terms.

$$\begin{aligned} G(t) - \hat{G}(t) &= \frac{1}{n}\sum_{k=1}^{n} \psi_1(T_k, X_k, t) + o_p(n^{-1/2}), \\ K(s) - \hat{K}(s) &= \frac{1}{n}\sum_{k=1}^{n} \psi_2(X_k, Y_k, \delta_k, s) + o_p(n^{-1/2}), \end{aligned} \tag{A.5}$$

where $E\{\psi_1(T_k, X_k, t)\} = 0$ and $E\{\psi_2(X_k, Y_k, \delta_k, s)\} = 0$. Then with (A.5), one has

$$\begin{aligned} n^{1/2}(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}) &= n^{1/2}\left\{\frac{(\Delta G)^2 K_i K_j - (\Delta\hat{G})^2\hat{K}_i\hat{K}_j}{(\Delta G)^2 K_i K_j (\Delta\hat{G})^2\hat{K}_i\hat{K}_j}\right\} \\ &= n^{1/2}\left[\frac{\Delta G\{K_i(K_j - \hat{K}_j) + \hat{K}_j(K_i - \hat{K}_i)\} + 2\hat{K}_i\hat{K}_j(\Delta G - \Delta\hat{G})}{\Delta G K_i K_j (\Delta\hat{G})^2\hat{K}_i\hat{K}_j}\right] + o_p(1) \\ &= n^{-1/2}\left[\frac{\Delta G\{K_i\sum_{k=1}^{n}\psi_{2j} + \hat{K}_j\sum_{k=1}^{n}\psi_{2i}\} + 2\hat{K}_i\hat{K}_j\sum_{k=1}^{n}(\psi_{1,min} - \psi_{1,max})}{\Delta G K_i K_j (\Delta\hat{G})^2\hat{K}_i\hat{K}_j}\right] + o_p(1), \end{aligned}$$

where $\psi_{1,min} = \psi_1(T_k, X_k, X_{ij}^{min})$, $\psi_{1,max} = \psi_1(T_k, X_k, X_{ij}^{max} - t_0)$, $\psi_{2i} = \psi_2(X_k, Y_k, \delta_k, X_i + Y_{ij}^{min})$, and $\psi_{2j} = \psi_2(X_k, Y_k, \delta_k, X_j + Y_{ij}^{min})$. Therefore, the second term in (A.4) can be expressed as

$$n^{-1/2}\sum_{k=1}^{n}\binom{n}{2}^{-1}\sum_{i<j}\lambda_{ij}a_{ij}b_{ij}\left\{\frac{\psi_{2j}}{K_j(\Delta\hat{G})^2\hat{K}_i\hat{K}_j} + \frac{\psi_{2i}}{K_iK_j(\Delta\hat{G})^2\hat{K}_i} + \frac{2(\psi_{1,min} - \psi_{1,max})}{\Delta G K_i K_j (\Delta\hat{G})^2}\right\} + o_p(1)$$

$$= n^{-1/2}\sum_{k=1}^{n} E\left[\lambda_{12}a_{12}b_{12}\left\{\frac{\psi_{22}}{K_2(\Delta\hat{G})^2\hat{K}_1\hat{K}_2} + \frac{\psi_{21}}{K_1K_2(\Delta\hat{G})^2\hat{K}_1} + \frac{2(\psi_{1,min}^{12} - \psi_{1,max}^{12})}{\Delta G K_1 K_2 (\Delta\hat{G})^2}\right\} + o_p(1),\right.$$

where $\psi_{1,min}^{12} = \psi_1(T_k, X_k, X_{12}^{min})$ and $\psi_{1,max}^{12} = \psi_1(T_k, X_k, X_{12}^{max} - t_0)$, and it is a sum of independent and identically distributed zero-mean terms. Then, $n^{1/2}(\hat{\tau}_v - \tau)$ is asymptotically equivalent to a zero-mean U-statistic of order 2.

Next derive the asymptotic properties of $\hat{\tau}_{is}$. One has

$$\begin{aligned} n^{1/2}(\hat{\tau}_{is} - \tau) &= n^{1/2}\left\{ \frac{\binom{n}{2}^{-1}\sum_{i<j}\frac{\lambda_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}}}{\binom{n}{2}^{-1}\sum_{i<j}\frac{\lambda_{ij}}{\hat{p}_{ij}}} - \tau \right\} \\ &= n^{1/2}(\hat{\tau}_v - \tau) - n^{1/2}\tau\left\{ \binom{n}{2}^{-1} \sum_{i<j} \frac{\lambda_{ij}}{\hat{p}_{ij}} - 1 \right\} + o_p(1) \\ &= n^{1/2}(\hat{\tau}_v - \tau) - n^{1/2}\tau\left\{ \binom{n}{2}^{-1} \sum_{i<j} \frac{\lambda_{ij}}{p_{ij}} - 1 \right\} \\ & \quad -n^{1/2}\tau\left\{ \binom{n}{2}^{-1} \left(\sum_{i<j} \frac{\lambda_{ij}}{\hat{p}_{ij}} - \sum_{i<j} \frac{\lambda_{ij}}{p_{ij}} \right) \right\} + o_p(1). \qquad \text{(A.6)} \end{aligned}$$

The first term in (A.6) has been shown to be asymptotically equivalent to a zero-mean U-statistic of order 2. The second term in (A.6) is a sum of n independent and identically distributed zero-mean terms. Similar to the development of asymptotic results for the second term in (A.4), the third term in (A.6) is also asymptotically equivalent to a sum of independent and identically distributed zero-mean terms. These three terms together imply $n^{1/2}(\hat{\tau}_{is} - \tau)$ converges weakly to normal distribution with mean zero and variance σ_{is}^2 as $n \to \infty$.

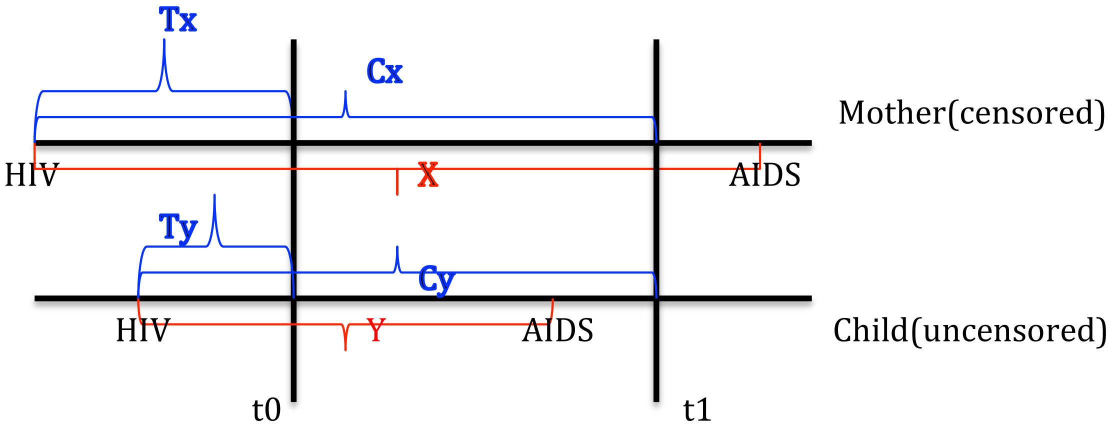

Figure 1. Schematic depiction of mother-child incubation times from pediatric AIDS cohort study : t_0 is the time of recruitment, t_1 is the time of the end of study.

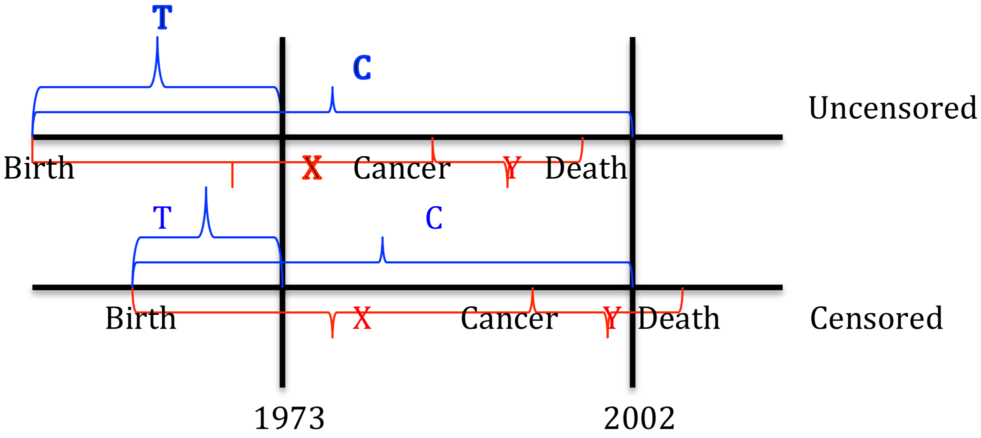

Figure 2. Schematic depiction of bivariate survival data with interval sampling collected by SEER cancer registries.

Table 1
Simulation results of estimated Kendall's τ for bivariate left-truncated survival data

n	τ	$bias(\hat{\tau}_{lr})$	$se_e(\hat{\tau}_{lr})$	$se_b(\hat{\tau}_{lr})$	$mse(\hat{\tau}_{lr})$	$cp(\hat{\tau}_{lr})$
200	0.2	0.001	0.065	0.063	0.0042	95.3
	0.5	0.008	0.051	0.054	0.0027	95.6
	0.8	0.008	0.026	0.023	0.0007	95.9
400	0.2	0.002	0.039	0.037	0.0015	96.2
	0.5	0.002	0.033	0.032	0.0011	96.5
	0.8	0.004	0.019	0.020	0.0004	96.7

$bias(\hat{\tau}_{lr})$: empirical bias of $\hat{\tau}_{lr}$; $se_e(\hat{\tau}_{lr})$: empirical standard error of $\hat{\tau}_{lr}$; $se_b(\hat{\tau}_{lr})$: average of bootstrap standard errors of $\hat{\tau}_{lr}$; $mse(\hat{\tau}_{lr})$: mean squared error of $\hat{\tau}_{lr}$; $cp(\hat{\tau}_{lr})$: 95% coverage probability of $\hat{\tau}_{lr}$.

Table 2

Simulation results of estimated Kendall's τ for bivariate survival data with interval sampling

n	τ	$bias(\hat{\tau}^n_{is})$	$se_e(\hat{\tau}^n_{is})$	$se_b(\hat{\tau}^n_{is})$	$cp(\hat{\tau}^n_{is})$	$bias(\hat{\tau}^s_{is})$	$se_e(\hat{\tau}^s_{is})$	$se_b(\hat{\tau}^s_{is})$	$cp(\hat{\tau}^s_{is})$
200	0.2	0.018	0.070	0.061	94.8	0.023	0.183	0.174	95.4
	0.5	0.025	0.053	0.050	93.9	0.025	0.129	0.110	95.3
	0.8	0.018	0.023	0.023	95.2	0.020	0.059	0.050	96.1
400	0.2	0.016	0.045	0.044	95.2	0.015	0.144	0.142	96.1
	0.5	0.028	0.036	0.036	94.4	0.017	0.099	0.097	96.3
	0.8	0.018	0.016	0.017	95.5	0.012	0.042	0.040	96.6

$\hat{\tau}^n_{is}$: nonparametric estimator of τ; $bias(\hat{\tau}^n_{is})$: empirical bias of $\hat{\tau}^n_{is}$; $se_e(\hat{\tau}^n_{is})$: empirical standard error of $\hat{\tau}^n_{is}$; $se_b(\hat{\tau}^n_{is})$: average of bootstrap standard errors of $\hat{\tau}^n_{is}$; $cp(\hat{\tau}^n_{is})$: 95% coverage probability of $\hat{\tau}^n_{is}$; $\hat{\tau}^s_{is}$: copula model-based semiparametric estimator of τ; $bias(\hat{\tau}^s_{is})$: empirical bias of $\hat{\tau}^s_{is}$; $se_e(\hat{\tau}^s_{is})$: empirical standard error of $\hat{\tau}^s_{is}$; $se_b(\hat{\tau}^s_{is})$: average of bootstrap standard errors of $\hat{\tau}^s_{is}$; $cp(\hat{\tau}^s_{is})$: 95% coverage probability of $\hat{\tau}^s_{is}$.

Table 3
Estimations of Kendall's τ for SEER ovarian cancer data (non-white patients)

birth cohort	group size	$\hat{\tau}_{is}^{n}$	$se_b(\hat{\tau}_{is}^{n})$	$ci(\hat{\tau}_{is}^{n})$	$\hat{\tau}_{is}^{s}$	$se_b(\hat{\tau}_{is}^{s})$	$ci(\hat{\tau}_{is}^{s})$
1920-1930	1325	-0.510	0.048	$(-0.604,-0.416)$	-0.340	0.037	$(-0.412,-0.267)$
1930-1940	911	-0.549	0.017	$(-0.582,-0.516)$	-0.276	0.069	$(-0.411,-0.141)$
1940-1950	251	-0.515	0.068	$(-0.643,-0.377)$	-0.190	0.130	$(-0.445,0.065)$

$\hat{\tau}_{is}^{n}$: nonparametric estimator of τ; $se_b(\hat{\tau}_{is}^{n})$: bootstrap standard error of $\hat{\tau}_{is}^{n}$ based on 500 replications; $ci(\hat{\tau}_{is}^{n})$: 95% confidence interval of $\hat{\tau}_{is}^{n}$; $\hat{\tau}_{is}^{s}$: copula model-based semiparametric estimator of τ; $se_b(\hat{\tau}_{is}^{s})$: bootstrap standard error of $\hat{\tau}_{is}^{s}$ based on 500 replications; $ci(\hat{\tau}_{is}^{s})$: 95% confidence interval of $\hat{\tau}_{is}^{s}$.