Published in final edited form as:

Biometrics. 2009 March ; 65(1): 125–134. doi:10.1111/j.1541-0420.2008.01054.x.

Measuring agreement of multivariate discrete survival times using a modified weighted kappa coefficient

Ying Guo^{*} and Amita K. Manatunga

Department of Biostatistics, Rollins School of Public Health of Emory University, 1518 Clifton RD NE, Atlanta, GA, 30322, USA

SUMMARY

Assessing agreement is often of interest in clinical studies to evaluate the similarity of measurements produced by different raters or methods on the same subjects. We present a modified weighted kappa coefficient to measure agreement between bivariate discrete survival times. The proposed kappa coefficient accommodates censoring by redistributing the mass of censored observations within the grid where the unobserved events may potentially happen. A generalized modified weighted kappa is proposed for multivariate discrete survival times. We estimate the modified kappa coefficients nonparametrically through a multivariate survival function estimator. The asymptotic properties of the kappa estimators are established and the performance of the estimators are examined through simulation studies of bivariate and trivariate survival times. We illustrate the application of the modified kappa coefficient in the presence of censored observations with data from a prostate cancer study.

Keywords

Agreement; Censoring; Multivariate discrete survival times; Weighted kappa

1. Introduction

In many scenarios in biomedical sciences, the same outcome may be measured by different methods or raters. For example, the disease status of patients may be evaluated by different raters; an event may be measured by a gold standard or a relatively simple method. Evaluation of agreement between these measurements is needed to determine whether different raters generate similar values on same patients; or whether the simple method reproduces the results of the gold standard.

Statistical methods for measuring agreement between categorical outcomes are well established. Cohen (1960) developed the kappa statistic as an agreement index for two binary variables. It has an appealing interpretation as a measure of chance-corrected agreement. Later, Cohen (1968) generalized the original kappa to the weighted kappa coefficient for ordinal discrete outcomes. Since its development, kappa with its extensions (Cohen, 1960 and 1968; Fleiss 1971; Kraemer, 1980) have been well studied in the literature and broadly applied in many areas (Maclure and Willett, 1987; Korten et al., 1992; Klar et

*yguo2@sph.emory.edu.

Supplementary Materials

Web Appendices and Tables referenced in Sections 2, 3 and 4 are available under the Paper Information link at the Biometrics web site <http://www.tibs.org/biometrics/>.

al., 2000; Williamson et al., 2000). However, applications of kappa in survival studies have been quite limited due to the presence of censoring.

In biomedical sciences, researchers are often interested in assessing agreement between two survival times that are measured on the same subjects. For example, in depression studies, the time of onset of clinical depression is measured using both clinician-administered and patient self-reported scales. Evaluating agreement between the disease onset times is useful in assessing the reliability of patient self-report and identifying an appropriate instrument for diagnosing depression. Later in this paper, we will describe another study where the time to the recurrence of prostate cancer is assessed by two different techniques.

Due to the presence of censored observations in survival studies, most existing agreement measures such as kappa cannot be directly applied. To assess the agreement between discrete survival times, Guo and Manatunga (2005) developed a local kappa coefficient that measures local agreement between bivariate survival times on each point of the two-dimensional time grid. The agreement between the survival times is reflected by the pattern of the local kappas on the two-dimensional time grid. This method can be used for investigating covariate effects on the agreement through modeling the local kappa coefficient.

In many studies, a single agreement index is needed to represent the overall strength of agreement between two survival times. Since discrete survival times can be viewed as ordinal outcomes, Cohen's weighted kappa (Cohen, 1968) seems to be a natural choice for our purpose. However, the presence of censoring makes it infeasible to estimate the weighted kappa with survival times. In this paper, we propose a modified weighted kappa coefficient that can incorporate censored observations. A major difference between the modified weighted kappa and Cohen's weighted kappa is that they are defined based on two types of cell probabilities in the contingency table of two ordinal outcomes. Cohen's weighted kappa is based on unconditional cell probabilities. The modified weighted kappa is based on redistributed cell probabilities which are the conditional probabilities for an event to occur in a particular cell given the observed survival times and censoring indicators. In calculating the redistributed cell probabilities, the mass of a censored observation is redistributed to the cells where the unobserved event may possibly occur. We also extend the method for multivariate cases by proposing a generalized modified weighted kappa for multivariate survival times. Estimates of the modified weighted kappa coefficient can be obtained through a nonparametric estimator of the bivariate survival function. Asymptotic properties such as strong consistency and asymptotic normality are established for the proposed estimator in the paper.

The paper is organized as follows. In Section 2, we present the methodology for the modified weighted kappa coefficient and its extension with multiple raters. In Section 3, we evaluate the performance of the proposed method through simulations. In Section 4, we present an example from a prostate cancer study. Finally, we conclude with a discussion.

2. Methods

2.1 Cohen's weighted kappa coefficient

Cohen (1960) proposed the kappa index as a measure of agreement for a binary test. It is interpreted as the proportion of observed agreement to its maximum possible value after chance agreement is removed from consideration. Cohen (1968) considered the extension of the original kappa to ordinal variables: measurements that incorporate natural orders such as the degree of severity of a disease that can be categorized as normal, mild, moderate and severe. For ordinal data, the disagreement in adjacent categories is less serious than the

disagreement in more disparate categories. Hence Cohen (1968) proposed the weighted kappa to allow various degrees of disagreement to be differentially weighted in evaluating the overall agreement.

Let Y_1 and Y_2 denote ordinal ratings with R ordered categories by two raters. Without loss of generality, we assume $Y_j = 1, \dots, R, j = 1, 2$. The joint distribution of Y_1 and Y_2 can be summarized by an $R \times R$ contingency table. The cell probability of the table is defined as $p_{ij} = \Pr(Y_1 = i, Y_2 = j)$ where $i, j = 1, \dots, R$. The sample estimates are $\hat{p}_{ij} = n_{ij}/n$ where n_{ij} is the observed cell frequency and n is the total number of observations. Denote w_{ij} as the weight assigned to the cell (i, j) to represent the degree of agreement of this cell. The weight function is restricted such that $0 \leq w_{ij} \leq 1$ with $w_{ij} \rightarrow 1$ indicating stronger agreement. The weighted kappa coefficient is defined as

$$\kappa = \frac{P_o^w - P_e^w}{1 - P_e^w}, \quad (1)$$

where $P_o^w = \sum_{i=1}^R \sum_{j=1}^R w_{ij} p_{ij}$ denotes weighted observed agreement probabilities and $P_e^w = \sum_{i=1}^R \sum_{j=1}^R w_{ij} p_{i.} p_{.j}$ represents the expected agreement when raters are independent. Estimates $\hat{\kappa}$ are calculated by replacing the true probabilities p_{ij} in (1) with the sample proportions $\hat{p}_{ij}$.

Various weighted kappa coefficients can be obtained by choosing different weight functions.

The two most common sets of weights are the quadratic weights with $w_{ij} = 1 - \frac{(i-j)^2}{(R-1)^2}$ and the Cicchetti weights (Cicchetti and Allison, 1971) with $w_{ij} = 1 - \frac{|i-j|}{|R-1|}$. When the quadratic weights are applied, the weighted kappa coefficient has been shown to be closely related to various agreement measures for continuous outcomes. Cohen (1968) proved the weighted kappa coefficient with quadratic weights is equivalent to the product-moment correlation coefficient under the assumption that the two marginal distributions are the same. Fleiss and Cohen (1973) showed the weighted kappa with quadratic weights is asymptotically equivalent to the intraclass correlation coefficient for ordinal ratings from a Gaussian general linear model. More recently, King and Chinchilli (2001) proved that the weighted kappa with quadratic weights is equivalent to Lin's concordance correlation coefficient (Lin, 1989), a popular measure of agreement for continuous measurements.

2.2 The modified weighted kappa coefficient

We now consider the assessment of agreement between survival outcomes. Let T_1 and T_2 denote two correlated discrete survival times and assume that the joint distribution of (T_1, T_2) is concentrated on an integral grid $\{(l_1, l_2), l_1 = 1, \dots, m_1, l_2 = 1, \dots, m_2\}$. The survival times (T_1, T_2) has the joint survival function of $S(l_1, l_2) = \Pr(T_1 > l_1, T_2 > l_2)$ and the density function of $p(l_1, l_2) = \Pr(T_1 = l_1, T_2 = l_2)$. Following Oakes (1989), the distribution of the censoring times (C_1, C_2) is assumed to be concentrated on the grid

$\left\{ \left(l_1 + \frac{1}{2}, l_2 + \frac{1}{2} \right), l_1 = 1, \dots, m_1, l_2 = 1, \dots, m_2 \right\}$ to avoid ties between censored and uncensored observations. The joint survival function for (C_1, C_2) is

$G(l_1 + \frac{1}{2}, l_2 + \frac{1}{2}) = \Pr(C_1 > l_1 + \frac{1}{2}, C_2 > l_2 + \frac{1}{2})$ and the density function is

$g(l_1 + \frac{1}{2}, l_2 + \frac{1}{2}) = \Pr(C_1 = l_1 + \frac{1}{2}, C_2 = l_2 + \frac{1}{2})$ It is assumed that the censoring times are independent of the survival times. The observed times and censoring indicators are $\tilde{T}_j = T_j \wedge C_j$ and $\delta_j = I(T_j \leq C_j)$ for $j = 1, 2$. The observed data consist of $(T_1, T_2, \delta_1, \delta_2)$.

We define the redistributed cell probability as

$$O_{l_1 l_2} = \Pr(T_1 = l_1, T_2 = l_2 | \tilde{T}_1, \tilde{T}_2, \delta_1, \delta_2), l_j = 1, \dots, m_j.$$

$O_{l_1 l_2}$ is the probability for the event to happen at (l_1, l_2) given a random observation $(\tilde{T}_1, \tilde{T}_2, \delta_1, \delta_2)$. Depending on the censoring status of (δ_1, δ_2) , $O_{l_1 l_2}$ can be written in terms of the joint survival function S in the following ways,

- if $(\delta_1, \delta_2) = (1, 1)$,

$$O_{l_1 l_2} = \begin{cases} 1 & \text{for } (l_1, l_2) = (\tilde{T}_1, \tilde{T}_2), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

- if $(\delta_1, \delta_2) = (0, 1)$,

$$O_{l_1 l_2} = \begin{cases} \frac{S(l_1 - 1, l_2 - 1) - S(l_1 - 1, l_2) - S(l_1, l_2 - 1) + S(l_1, l_2)}{S(\tilde{T}_1 - 1, \tilde{T}_2 - 1) - S(\tilde{T}_1 - 1, \tilde{T}_2)} & \text{for } (l_1, l_2) = (\tilde{T}_1 + a_1, \tilde{T}_2) \\ & \text{where } a_1 = 1, \dots, m_1 - \tilde{T}_1, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

- if $(\delta_1, \delta_2) = (1, 0)$,

$$O_{l_1 l_2} = \begin{cases} \frac{S(l_1 - 1, l_2 - 1) - S(l_1 - 1, l_2) - S(l_1, l_2 - 1) + S(l_1, l_2)}{S(\tilde{T}_1 - 1, \tilde{T}_2) - S(\tilde{T}_1, \tilde{T}_2)} & \text{for } (l_1, l_2) = (\tilde{T}_1, \tilde{T}_2 + a_2) \\ & \text{where } a_2 = 1, \dots, m_2 - \tilde{T}_2, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

- if $(\delta_1, \delta_2) = (0, 0)$,

$$O_{l_1 l_2} = \begin{cases} \frac{S(l_1 - 1, l_2 - 1) - S(l_1 - 1, l_2) - S(l_1, l_2 - 1) + S(l_1, l_2)}{S(\tilde{T}_1, \tilde{T}_2)} & \text{for } (l_1, l_2) = (\tilde{T}_1 + a_1, \tilde{T}_2 + a_2) \\ & \text{where } a_j = 1, \dots, m_j - \tilde{T}_j, \\ & j = 1, 2, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Therefore, the probability $O_{l_1 l_2}$ represents all the mass of an uncensored observation at (l_1, l_2) and redistributed mass at (l_1, l_2) from an earlier censored observation. In the next theorem, we show that the expectation of the redistributed cell probability $O_{l_1 l_2}$ is equal to the unconditional cell probability $p(l_1, l_2)$.

Theorem 1— $E(O_{l_1 l_2}) = \Pr(T_1 = l_1, T_2 = l_2) = p(l_1, l_2)$, for $l_j = 1, \dots, m_j$ and $j = 1, 2$.

Proof: See Web Appendix A.

We propose the following modified weighted kappa coefficient to measure the overall agreement between (T_1, T_2) ,

$$\kappa^w = \frac{\sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} O_{l_1 l_2} - \sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} O_{l_1} \cdot O_{l_2}}{1 - \sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} O_{l_1} \cdot O_{l_2}}, \quad (6)$$

where $O_{l_1} = \sum_{l_2=1}^{m_2} O_{l_1 l_2}$ and $O_{l_2} = \sum_{l_1=1}^{m_1} O_{l_1 l_2}$. For the weight function, we recommend using the quadratic weights because the quadratic weights lead to close connection between the weighted kappa coefficient and agreement measures for continuous outcomes such as Lin's concordance correlation coefficient (Lin, 1989). Other weight functions can be applied under the same framework.

Let $(\tilde{T}_{k1}, \tilde{T}_{k2}, \delta_{k1}, \delta_{k2})$, $k = 1, \dots, n$ be n random observations. The sample estimator of the proposed modified weighted kappa is

$$\hat{\kappa}^w = \frac{\sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} \hat{O}_{l_1 l_2} - \sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} \hat{O}_{l_1} \cdot \hat{O}_{l_2}}{1 - \sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} w_{l_1, l_2} \hat{O}_{l_1} \cdot \hat{O}_{l_2}}, \quad (7)$$

where $\hat{O}_{l_1 l_2} = \sum_{k=1}^n \hat{O}_{l_1 l_2}^k / n$ is the sample estimator of the redistributed cell probability with $\hat{O}_{l_1 l_2}^k = \Pr(T_1 = l_1, T_2 = l_2 | \tilde{T}_{k1}, \tilde{T}_{k2}, \delta_{k1}, \delta_{k2})$. Without censoring, $\hat{\kappa}^w$ is equivalent to the sample estimator of Cohen's weighted kappa. In the presence of censoring, a nonparametric estimator of the bivariate survival function $\hat{S}$ can be used to calculate $\hat{O}_{l_1 l_2}$ through equations (2)–(5). In this paper, we estimate the modified weighted kappa through Prentice-Cai estimator $\hat{S}$ (Prentice and Cai, 1992) because it is shown to be adequate for most practical uses (Kalbfleisch and Prentice, 2002).

Theorem 2—The estimator $\hat{\kappa}^w$ has the following asymptotic statistical properties as $n \rightarrow \infty$:

- i. $\hat{\kappa}^w$ is strongly consistent. That is, $|\hat{\kappa}^w - \kappa^w| \rightarrow 0$ with probability 1.
- ii. $\sqrt{n}(\hat{\kappa}^w - \kappa^w)$ converges weakly to a zero-mean normal distribution.
- iii. Let $(\tilde{T}_{k1}, \tilde{T}_{k2}, \delta_{k1}, \delta_{k2})$, $k = 1, \dots, n$ represent the observed data. A bootstrap estimator $\kappa^\#$ can be obtained based on a bootstrap sample that is drawn randomly with replacement from $(\tilde{T}_{k1}, \tilde{T}_{k2}, \delta_{k1}, \delta_{k2})$, $k = 1, \dots, n$. Then $\sqrt{n}(\kappa^\# - \hat{\kappa}^w)$, given the observed data, weakly converges to the same limit distribution as $\sqrt{n}(\hat{\kappa}^w - \kappa^w)$ in probability.

Proof: See Web Appendix B

The analytical expression for the variance of $\hat{\kappa}^w$ is technically challenging since it involves the covariance of $\hat{S}$ through a complicated function. We propose to use the bootstrap procedure for consistent estimation of the variance of $\hat{\kappa}^w$ and for the calculation of asymptotic confidence intervals. Specifically, we randomly sample B times with replacement from $(\tilde{T}_{k1}, \tilde{T}_{k2}, \delta_{k1}, \delta_{k2})$ ($k = 1, \dots, n$). From each bootstrap sample, a bootstrap estimate $\kappa^\#$ is obtained. The bootstrap variance estimator is the sample variance of the bootstrap estimates $\kappa_b^\#$, $b = 1, \dots, B$. For the confidence interval of κ^w , we use the bootstrap

percentile confidence interval which defines confidence limits as percentiles of bootstrap sample estimates.

2.3. Extension to multiple raters

In many studies, the time to events are measured by more than two raters or methods. We now extend the proposed modified weighted kappa to measure agreement among multivariate discrete survival times. Suppose the survival times of the same set of subjects are measured by R raters (or methods) as $T_1, \dots, T_R$. Let $\{(l_1, \dots, l_R), l_j = 1, \dots, m_j, j = 1, \dots, R\}$ denote the grid for the R dimensional discrete survival time. The survival times $(T_1, \dots, T_R)$ have a joint survival function of $S(l_1, \dots, l_R) = \Pr(T_1 > l_1, \dots, T_R > l_R)$. The distribution of censoring times $(C_1, \dots, C_R)$ is assumed to be concentrated on the grid

$\{(l_1 + \frac{1}{2}, \dots, l_R + \frac{1}{2}), l_j = 1, \dots, m_j, j = 1, \dots, R\}$ with a survival function of $G(l_1 + \frac{1}{2}, \dots, l_R + \frac{1}{2})$. It is assumed that the censoring times are independent of the survival times. The observed data are $(\tilde{T}_1, \dots, \tilde{T}_R, \delta_1, \dots, \delta_R)$ with $\tilde{T}_j = T_j \wedge C_j$ and $\delta_j = I(T_j \leq C_j)$ for $j = 1, \dots, R$.

Define the multivariate redistributed cell probabilities as

$$O_{l_1, \dots, l_R} = \Pr(T_1 = l_1, \dots, T_R = l_R | \tilde{T}_1, \dots, \tilde{T}_R, \delta_1, \dots, \delta_R), l_j = 1, \dots, m_j,$$

To measure the agreement among $(T_1, \dots, T_R)$, we propose the generalized modified weighted kappa,

$$\kappa_g^w = \frac{P_o - P_e}{1 - P_e}.$$

The observed agreement P_o and expected agreement P_e are defined as

$$P_o = \sum_{l_1=1}^{m_1} \dots \sum_{l_R=1}^{m_R} \frac{\sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \{1, \dots, R\}}} w(l_{i_1}, l_{i_2})}{\binom{R}{2}} O_{l_1, \dots, l_R} \quad (8)$$

$$P_e = \frac{1}{\binom{R}{2}} \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \{1, \dots, R\}}} \sum_{l_1=1}^{m_{i_1}} \sum_{l_2=1}^{m_{i_2}} w(l_{i_1}, l_{i_2}) O_{l_{i_1}} O_{l_{i_2}}, \quad (9)$$

where

$$\begin{aligned} O_{l_{i_1}} &= \Pr(T_{i_1} = l_{i_1} | \tilde{T}_1, \dots, \tilde{T}_R, \delta_1, \dots, \delta_R), \\ O_{l_{i_2}} &= \Pr(T_{i_2} = l_{i_2} | \tilde{T}_1, \dots, \tilde{T}_R, \delta_1, \dots, \delta_R). \end{aligned}$$

Note that when calculating the observed overall agreement P_o , the weight assigned to $O_{l_1, \dots, l_R}$ is the average of the pairwise weights for all the bivariate subsets from $(l_1, \dots, l_R)$.

The generalized modified weighted kappa κ_g^w is closely related to the pairwise modified weighted kappa κ^w defined in (6). When $R = 2$, the generalized kappa reduces to the pairwise kappa. For $R > 2$, the generalized kappa can be calculated from pairwise kappas in the following way,

$$\kappa_g^w = \frac{\sum_{i_1=1}^{R-1} \sum_{i_2=i_1+1}^R P_{e_{i_1 i_2}} \kappa_{i_1 i_2}^w}{\sum_{i_1=1}^{R-1} \sum_{i_2=i_1+1}^R P_{e_{i_1 i_2}}},$$

where $P_{e_{i_1 i_2}} = \sum_{l_{i_1}=1}^{m_{i_1}} \sum_{l_{i_2}=1}^{m_{i_2}} w(l_{i_1}, l_{i_2}) O_{l_{i_1}} O_{l_{i_2}}$ is agreement between raters i_1 and i_2 that is expected by chance and $\kappa_{i_1 i_2}^w$ is the pairwise modified kappa for raters i_1 and i_2 . Therefore, the generalized kappa is a weighted average of all the pairwise kappas among the multiple raters. Heavier weights are given to pairwise kappas between raters that have higher expected-by-chance agreement.

The redistributed cell probabilities $O_{11}, \dots, O_{RR}$ and $O_{ij\cdot}$ can be written in terms of the joint survival function $S(t_1, \dots, t_R)$ and its lower-dimensional marginal survival functions. Therefore, a nonparametric estimator $\hat{S}$ such as Prentice-Cai estimator (Prentice and Cai, 1992) can be used for calculating $\hat{O}_{11}, \dots, \hat{O}_{RR}$. The generalized modified weighted kappa can then be estimated by substituting $\hat{O}_{11}, \dots, \hat{O}_{RR}$ and $\hat{O}_{ij\cdot}$ into (8) and (9). Using similar arguments as those in the Web Appendix B, we can show that the estimator $\hat{\kappa}_g^w$ has the asymptotic properties that are presented in Theorem 2 for the estimator $\hat{\kappa}^w$.

3. Simulation Studies

3.1. Bivariate survival times

There are two types of mechanisms through which discrete survival times arise in practice (Kalbfleisch and Prentice, 2002). Often, discrete survival data arise when underlying continuous survival time is subject to interval censoring. In other instance, time may be truly discrete, as, for example, time to pregnancy measured as the number of menstrual cycles before a woman conceives.

The performance of the proposed estimator $\hat{\kappa}^w$ was evaluated using discrete bivariate survival times generated from both mechanisms. For grouped times, we first generated continuous times from the Clayton model (Clayton, 1978). We assumed that T_1 and T_2 had marginal exponential distributions with mean of 1. The bivariate survival function for Clayton model was therefore

$$S(t_1, t_2) = (e^{t_1/\theta} + e^{t_2/\theta} - 1)^{-\theta}, \quad (10)$$

with a constant odds ratio of $1 + \frac{1}{\theta}$, where $\theta \rightarrow \infty$ indicates independence and $\theta \rightarrow 0$ indicates maximal positive dependence. The continuous random survival times T_j^* , $j=1, 2$ were then grouped into 5 intervals: $[a_1, a_2)$, $[a_2, a_3)$, $[a_3, a_4)$, $[a_4, a_5)$, $[a_5, \infty)$ and the discrete survival times were defined as

$$T_j = 1I\{T_j^* \in [a_1, a_2)\} + 2I\{T_j^* \in [a_2, a_3)\} + 3I\{T_j^* \in [a_3, a_4)\} + 4I\{T_j^* \in [a_4, a_5)\} + 5I\{T_j^* \in [a_5, \infty)\}, \quad j=1, 2.$$

Here I is an indicator variable. The cut points a_k ($k = 1, \dots, 5$) were chosen so that $\Pr(T_j = 1) = 0.15$, $\Pr(T_j = 2) = 0.2$, $\Pr(T_j = 3) = 0.3$, $\Pr(T_j = 4) = 0.2$ and $\Pr(T_j = 5) = 0.15$.

We considered three sets of simulations by specifying three different θ parameters ($\theta = 0.95, 0.5$ and 0.25) in the Clayton model. The true weighted kappa coefficients of the three Clayton models were 0.472, 0.651 and 0.804, representing moderate, substantial and almost perfect agreement respectively (Landis and Koch, 1977). For each set of simulation, we generated 50, 100 or 200 pairs of discrete survival times. The censoring variables (C_1, C_2) were independent from the survival times and from each other. Each censoring variable was

assumed to follow a multinomial distribution that took values in $(\frac{1}{2}, 2\frac{1}{2}, 3\frac{1}{2}, 4\frac{1}{2}, 5\frac{1}{2})$ with probability $\underline{p} = (p_1, p_2, p_3, p_4, p_5)'$. Three scenarios were considered: (1) $\underline{p} = (0.05, 0.05, 0.05, 0.05, 0.80)$ which induced a censoring proportion of 10%; (2) $\underline{p} = (0.1, 0.15, 0.25, 0.2, 0.3)$ which induced a censoring proportion of 30%; and (3) $\underline{p} = (0.2, 0.3, 0.3, 0.17, 0.03)$ with a censoring proportion of 50%.

We also performed the simulation study using truly discrete survival data generated through the discrete Clayton model (Shih, 1989). The discrete survival times T_j , $j = 1, 2$ took values in 1, ..., 5 with marginal probabilities of 0.15, 0.2, 0.3, 0.2 and 0.15, respectively. Three values for the constant local odds ratio were considered such that true weighted kappa coefficients of the three discrete Clayton models were 0.472, 0.651 and 0.804. Censoring variables were generated from multinomial distributions to induce censoring proportion of 10%, 30% and 50%.

The variance of the kappa estimator was evaluated using the bootstrap procedure. As a comparison to the proposed method, we considered the currently available alternate method which does not take into account of censored observations and estimates sample kappa based on empirical bivariate distribution calculated based on complete observations in the simulated data sets.

The simulation results based on grouped times and truly discrete times are presented Table 1 and 2, respectively. The results are based on 500 simulation runs. Sample means and sample standard deviations for $\hat{\kappa}^w$ are presented, along with the mean of bootstrap standard error estimates based on 200 bootstrap samples. From Table 1, we can see the modified weighted kappa estimates show small bias under low and medium censoring percentages but are biased downwards with heavy censoring. However, The bias is reduced as the sample size increases. From Table 2, the estimator is biased downwards with the increase of censoring but the bias again decreases with larger sample size. In both tables, the average of the bootstrap standard error of the proposed estimator is very close to the standard deviation of the estimates. For grouped times (Table 1), the coverage probabilities of the proposed estimator are close to the nominal level with low and medium censoring but are lower with heavy censoring. For truly discrete times (Table 2), the coverage probabilities are close to the nominal level in all scenarios except when the data has a small sample size and heavy censoring. In both tables, our proposed kappa estimator has significantly smaller bias and higher coverage probabilities than the sample estimator of kappa based on complete observations.

Comparisons between simulation results from Table 1 and 2 show that the performance of the proposed kappa estimator is generally better for truly discrete survival times than for grouped times. For grouped times, the kappa estimator demonstrates a complicated pattern in small sample size, for example the estimator appears to be biased upwards for low censoring but downwards under heavy censoring. The explanation is that there are two sources of bias in estimating kappa coefficient based on grouped survival times. The first source of bias is due to censoring. For medium to heavy censoring, the proposed method underestimates kappa. The bias due to censoring is present for both grouped times and the truly discrete times. The second source of bias is present only with grouped times and is due to estimating the discrete bivariate distribution based on grouped data that are generated from continuous survival models. To demonstrate the bias due to grouping, we generated both grouped times and truly discrete survival times in the absence of censoring and evaluated the bias in the estimated discrete bivariate distribution and kappa estimator. Results in Table 3 show that the bias is negligible for truly discrete times but is noticeable for grouped survival times with small to moderate sample sizes (Table 3). More specifically, we found for grouped times the estimated discrete distribution tends to be over-estimated on the diagonal line, i.e. (T_1, T_2) s.t. $T_1 = T_2$, and negatively biased on the off-diagonal line (similar trend is also observed when grouping continuous times from other bivariate survival functions such as Gumbel model). Consequently, the estimated kappa is positively biased for grouped data without censoring. The additional source of variability due to the estimation of a discrete distribution based on grouped data causes the decrease in the accuracy in estimating kappa based on grouped survival times. Furthermore, the two sources of bias have opposite directions resulting in the complicated pattern in the kappa estimator with grouped times in small sample sizes. In the presence of light censoring (10%), the positive bias due to grouping is dominant and hence the kappa estimate is biased upwards. When the censoring proportion increases, the negative bias due to censoring becomes more prominent and hence the kappa estimates demonstrate downwards bias. However, both censoring and grouping bias diminish with the increase of sample size. Our simulation studies with more sample size scenarios (Web Table 1) show that the proposed estimator demonstrates desirable performance for grouped survival times with larger sample sizes.

3.2. multivariate survival times

In this section, we evaluate the performance of the generalized modified weighted kappa coefficient through a simulation study of discrete trivariate survival times that are obtained by grouping continuous times. Continuous trivariate survival times were generated from the continuous multivariate Clayton model with unit exponential marginals,

$$S(t_1, t_2, t_3) = (e^{t_1/\theta} + e^{t_2/\theta} + e^{t_3/\theta} - 2)^{-\theta}, \quad (11)$$

The discrete survival times were then created by grouping continuous times into intervals. Three sets of simulations were considered using the same θ parameters as in the bivariate simulations. We consider sample sizes of 100, 200 and 350. Table 4 provides a summary of the statistics for the trivariate simulation study. The estimates for the generalized modified weighted kappa coefficient demonstrate similar performance as the estimates of the bivariate modified kappa.

4. Example

Prostate cancer is the most common cancer in US men. Depending on the patients' demographic and disease characteristics, various kinds of treatments are available. One major difficulty in treating and monitoring prostate cancer is the lack of a standard definition for disease freedom after treatments. There is a general consensus that posttreatment disease

status is reflected in the prostate specific antigen (PSA) with high level PSA indicating relapse. However, there is no universal agreement regarding the exact pattern of the PSA level that defines disease recurrence. Various definitions had been proposed for different kinds of treatments and the disease-free survival rates based on these definitions are used as important guidance for physicians in treatment selection. Since disease-free survival rates depend heavily on the definition of disease freedom, potential discrepancies between definitions may lead to different conclusions regarding the treatments effectiveness. Therefore, it is important to assess the agreement between different definitions before comparing the disease-free survival rates derived from them.

Radical prostatectomy and irradiation are two commonly used treatments for prostate cancer (Critz et al., 1995). For radical prostatectomy, disease freedom is defined by reaching and maintaining an undetectable prostate specific antigen (PSA) nadir ranging from 0.2ng/ml to 0.5ng/ml (Critz et al., 1996). For irradiation, according to the American Society of Therapeutic Radiation Oncology (ASTRO) consensus criteria (1997), posttreatment disease freedom is represented by a non-rising PSA with a rising PSA defined as three consecutive PSA increases measured six months apart. For years, there is a controversy regarding the disease free rates between the two treatments. Some researchers claim irradiation cures fewer patients than radical prostatectomy while others argue the two treatments are equally effective (Critz et al., 1996). In order to establish the comparability of disease-free survival rates between the two treatments, researchers (Critz et al., 1996) are interested in studying the agreement between the two definitions of disease freedom. Additionally, potential difference in the strength of agreement among various covariate subgroups is also of interest.

In this study, 1305 men with prostate cancer received simultaneous irradiation by integrating iodine 125 prostate implant with a follow-up external beam radiation. The disease status of all subjects was evaluated every six months after the treatment of the external beam radiation. The survival time was defined as the time elapsed from the end of the irradiation to the recurrence of prostate cancer which was determined based on two different definitions. In specific, T_1^* is the time when patients' posttreatment PSA level exceeded the nadir of 0.2 ng/ml while T_2^* is based on the ASTRO definition and represents the midpoint between the time when the lowest PSA was achieved after irradiation and the time when the first of the three consecutive rises in the PSA level occurred. The survival time was measured in terms of months. The two cancer recurrence times were subject to censoring due to end of the follow-up on a patient. During the study period, 156 subjects experienced prostate cancer recurrence based on both definitions and 64 had cancer recurrence based on one of the definitions, representing approximately 80% of censoring. The absolute difference between the observed times based on the two definitions ranges between 0 and 108 months with the mean of 1.9 months.

In the prostate cancer study, cancer recurrence times were collected in a discrete manner since the subjects were evaluated only every six months. Therefore, the proposed modified kappa could be used to evaluate agreement between cancer recurrence times measured by the two definitions. Due to the sparsity of events, we grouped survival times T_1^* and T_2^* into five intervals: no more than 30 months, 31–60 months, 61–90 months, 91–162 months and >162 months, and the resulting discrete survival times are denoted as T_1 and T_2 . Without loss of generality, $T_j = 1, 2, \dots, 5$ for $j = 1, 2$, corresponding to the five time intervals. The estimated modified weighted kappa with the quadratic weight function was 0.842 with the bootstrap SE of 0.021 (based on 200 bootstrap samples). The 95% confidence interval for the modified weighted kappa coefficient was (0.798, 0.882) based on the 2.5% and 97.5% empirical percentiles of bootstrap sample estimates. Therefore, there was rather strong

agreement between the recurrence times measured by the two definitions in the prostate cancer data. Since there was heavy censoring with the prostate data, we performed a simulations study to confirm the applicability of our proposed method under this scenario. We evaluated the performance of the kappa estimator with sample size of 1300 and censoring proportion of 80%, which is the set up similar to the prostate data example (see Web Table 2). For all kappa levels, the proposed estimator performs reasonably well with the heavy censoring rate. More specifically, for $\kappa = 0.804$ which represents similar strength of agreement as our data, the bias of the kappa estimator is about 6% and coverage probability is close to 90% (Web Table 2).

As an alternative approach, we also measured the agreement between the two definitions using Lin's concordance correlation coefficient (CCC) (Lin, 1998), treating the ungrouped data T_j^* ($j=1, 2$) as continuous survival times. A nonparametric estimator (Guo and Manatunga, 2007) was used to accommodate censored observations. The estimated CCC for the ungrouped data is 0.792 and is quite close to the estimated modified kappa of the grouped data. In both cases, we concluded there is strong agreement between cancer recurrence times based on the two definitions in the prostate cancer data.

5. Discussion

In this paper, we propose an extension to Cohen's (1968) weighted kappa coefficient for measuring agreement between discrete survival times. To the best of our knowledge, there has been no previous work for adapting kappa coefficient to survival outcomes. To accommodate censored observations, we first estimate the joint survival function of two survival times and then redistribute the mass of a censored observation to those cells where the unobserved event may potentially happen. A key underlying assumption for the proposed method is that the censoring distribution is independent of the joint survival function. This assumption ensures the mass of censored observations can be appropriately redistributed based on the estimated survival function. Among various estimators of the survival function, we choose Prentice and Cai's (1992) estimator since it is adequate for most practical purposes and is more efficient than many alternatives (Kalbfleisch and Prentice, 2002). Additionally, Prentice-Cai estimator can incorporate both univariate and bivariate censoring, which is an advantage over estimators that are only applicable to univariate censoring (Lin and Ying, 1993; Tsai and Crowley, 1998).

The proposed modified weighted kappa is developed for measuring agreement between discrete survival times. However, it is also useful for continuous outcomes in certain situations: for example, when the survival times are actually measured in a discrete way or when the events are too sparse in the original continuous time scale. In these cases, one can discretize the continuous times and apply the proposed modified weighted kappa to measure agreement between the grouped survival times. We recommend several guidelines for the discretization. When it is possible, we suggest grouping the survival times into practically meaningful intervals which are associated with clinical interpretations. For example, time to onset of diabetes may be grouped into juvenile, adult and elderly diabetes. The second guideline is that discretization shall capture the empirical distribution of observed events so that the distribution of discrete survival times demonstrates a reasonable representation of the underlying continuous survival distribution. For example, it is undesirable to group in a way that most of the observed events fall into a few of the intervals while the rest of the intervals are empty. In that case, the estimated modified weighted kappa is calculated based on only a few nonempty cells in the contingency table and hence may not correctly reflect agreement between the original times. Finally, we recommend grouping times after the last observed event into one or two intervals where the last interval is either one sided or ends at the maximum time point where the event of interest may possibly happen. In practice,

different discretizations inevitably influence the magnitude of the estimated kappa because kappa is known to be dependent on the marginal distributions (Cook, 1998). The above guidelines can help the estimated kappa correctly reflect agreement between original survival times. Alternatively, one can use an agreement measure defined for continuous scales. One popular agreement index for continuous outcomes is the concordance correlation coefficient (CCC) (Lin, 1998). A nonparametric estimator for the CCC has been proposed for survival times that can accommodate censored observations (Guo and Manatunga, 2007).

The performance of the proposed estimates for the modified weighted kappa and the generalized modified weighted kappa are satisfactory with low to medium censoring. With heavy censoring, the estimates are more biased with relatively low coverage probabilities in small to medium sample sizes. With increased sample size, the performance of the estimator improves. It should be pointed out that estimation bias for a statistic constructed based on survival functions is unavoidable in the presence of heavy censoring due to the difficulty in estimating the right tail of the survival function. For example, Lin and Ying (1993) noted bias when estimating the correlation coefficient between two survival times in the presence of heavy censoring.

When there is poor agreement between two survival time measurements, we can evaluate the empirical distribution of the observed events in the contingency table to find out causes of disagreement. For example, if the proposed kappa is calculated based on discrete survival times that are grouped from continuous times. One may want to check whether current discretization is suitable and whether agreement improves when choosing more appropriate discretization by following the suggested guidelines we present above. Evaluation of the cause of disagreement can provide helpful information when investigators are interested in adjusting the two measurements to reduce discrepancy between them, especially when one measurement is consistently biased from the other.

As a reviewer pointed out, the proposed agreement method may be useful to the multiple time scale problem. In many survival studies, the time to event of interest may be measured using multiple plausible scales. Often, the first time scale is the original chronological time or age of a subject at failure. The second alternative time scale is based on time-varying covariates, such as usage or exposure measures, and is often viewed as the operational time. One classical example is that the lifetime of a car can be measured either by its age or by the mileage driven. In multiple time scale problems, the first chronological time scale is readily defined while the definition of the operational time scale is usually not obvious. Existing methods (Farewell and Cox, 1979; Oakes, 1995; Kordonsky and Gertsbakh, 1997; Duchesne and Lawless, 2000) aim to select a time scale that “captures” most of the variation in the failure times given the time-dependent covariates. Duchesne and Lawless (2000) have proven that the time scale with the smallest squared coefficient of variation is the ideal time scale within certain survival distribution family. Compared to the existing methods, our proposed agreement method is useful for identifying an ideal time scale that is equivalent to the original time scale in measuring the event of interest. We propose the following scheme: First, fix the partition on the chronological time scales where there exists a natural or meaningful grid. Second, partition the range of the operational time and measure the agreement between the two discretized time scales using the proposed modified weighted kappa. Repeat the second step for different partitions and select the time scale that maximizes the modified weighted kappa. The ideal scale that provides the strongest agreement to the original time scale can then be used in interpreting the risk associated with the operational time scale with respect to the risk associated with the chronological time scale.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the referees, Associate Editor and Editor for valuable comments resulting in substantial improvements to the manuscript. This work was supported by an NIH grant R01-MH079448-02.

REFERENCES

- American Society for Therapeutic Radiology and Oncology Consensus Panel. Consensus statement: guidelines for PSA following radiation therapy. *International Journal of Radiation Oncology, Biology, Physics* 1997;37:1035–1041.
- Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 1971;11:101–109.
- Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978;65:141–151.
- Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37–46.
- Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968;70:213–220. [PubMed: 19673146]
- Cook, RJ. Kappa and its dependence on marginal rates. In: Armitage, P.; Colton, T., editors. *The Encyclopedia of Bio-statistics*. New York: Wiley; 1998. p. 2166–2168.
- Critz FA, Tarlton RS, Holladay DA. Prostate specific antigen -monitored combination radiotherapy for patients with prostate cancer. *Cancer* 1995;75:2383–2391. [PubMed: 7536126]
- Critz FA, Levinson K, Williams WH, Holladay DA. Prostate-Specific Antigen Nadir: the optimum level after irradiation for prostate cancer. *Journal of Clinical Oncology* 1996;14:2893–2900. [PubMed: 8918485]
- Duchesne T, Lawless J. Alternative time scales and failure time models. *Lifetime Data Analysis* 2000;6:157–179. [PubMed: 10851840]
- Farewell VT, Cox DR. A note on multiple time scales in life testing. *Applied Statistics* 1979;28:73–75.
- Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971;76:378–382.
- Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973;33:613–619.
- Guo Y, Manatunga AK. Modeling the agreement of discrete bivariate survival times using kappa coefficient. *Lifetime Data Analysis* 2005;11:309–332. [PubMed: 16133882]
- Guo Y, Manatunga AK. Nonparametric estimation of the concordance correlation coefficient under univariate censoring. *Biometrics* 2007;63:164–172. [PubMed: 17447941]
- Kalbfleisch, JD.; Prentice, RL. *The statistical analysis of failure time data*. New Jersey: John Wiley & Sons; 2002.
- King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* 2001;20:2131–2147. [PubMed: 11439426]
- Klar N, Lipsitz SR, Ibrahim J. An estimating equations approach for modelling kappa. *Biometrical Journal* 2000;42:45–58.
- Korten AE, Jorm AF, Henderson AS, McCusker E, Creasey H. Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology* 1992;21:1121–1131. [PubMed: 1483817]
- Kraemer HC. Extension of the kappa coefficient. *Biometrics* 1980;36:207–216. [PubMed: 7190852]
- Kordonsky KB, Gertsbakh I. Multiple time scales and the lifetime coefficient of variation: Engineering applications. *Lifetime Data Analysis* 1997;3:139–156. [PubMed: 9384619]
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174. [PubMed: 843571]

- Lin DY, Ying Z. A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* 1993;80:573–581.
- Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–268. [PubMed: 2720055]
- Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* 1987;126:161–169. [PubMed: 3300279]
- Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 1989;84:487–493.
- Oakes D. Multiple time scales in survival analysis. *Lifetime Data Analysis* 1995;1:7–18. [PubMed: 9385093]
- Prentice RL, Cai J. Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* 1992;79:495–512.
- Shih JH. Modeling multivariate discrete failure time data. *Biometrics* 1998;54:1115–1128. [PubMed: 9750256]
- Tsai WY, Crowley J. A note on nonparametric estimators of the bivariate survival function under univariate censoring. *Biometrika* 1998;85:573–580.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag; 1996.
- Williamson JM, Manatunga A, Lipsitz SR. Modelling kappa for measuring dependent categorical agreement data. *Biostatistics* 2000;1:191–202. [PubMed: 12933519]

Table 1

Simulation results for the modified weighted kappa estimator with grouped survival times simulated from the Clayton model based on 500 simulation runs, with comparison to sample kappa estimator based on complete observations

κ_g^w	Sample size	Censoring %	Proposed Kappa Estimator			Sample estimator $\hat{\kappa}$ with complete obs. only		
			$\hat{\kappa}$ mean (SD)	Bootstrap SE mean (SD)	Coverage Prob.	sample estimator mean(SD)	Bootstrap SE (SD)	Coverage Prob.
0.472	50	10%	0.516(0.114)	0.109 (0.021)	92.1	0.512(0.123)	0.114(0.021)	87.3
		30%	0.492(0.125)	0.131 (0.058)	93.0	0.391(0.182)	0.144(0.024)	83.4
		50%	0.420(0.139)	0.134 (0.025)	91.1	0.315(0.234)	0.166(0.031)	72.8
		100	0.509(0.082)	0.079 (0.011)	90.2	0.499(0.088)	0.086(0.013)	89.3
	100	10%	0.497(0.089)	0.085 (0.012)	94.9	0.373(0.136)	0.115(0.015)	83.6
		30%	0.443(0.104)	0.101 (0.014)	92.2	0.261(0.162)	0.135(0.020)	65.2
		50%	0.483(0.058)	0.057 (0.006)	93.8	0.469(0.061)	0.062(0.007)	93.6
		100	0.477(0.064)	0.063 (0.007)	96.1	0.340(0.091)	0.089(0.009)	71.3
	200	10%	0.451(0.081)	0.077 (0.009)	92.6	0.206(0.115)	0.104(0.014)	35.5
		30%	0.677(0.084)	0.082 (0.020)	91.4	0.673(0.093)	0.089(0.024)	86.5
		50%	0.646(0.096)	0.108 (0.066)	95.9	0.539(0.161)	0.132(0.028)	83.8
		100	0.554(0.126)	0.124 (0.028)	89.9	0.414(0.218)	0.162(0.033)	68.9
0.651	50	10%	0.675(0.059)	0.058 (0.010)	91.8	0.665(0.065)	0.064(0.012)	90.6
		30%	0.656(0.068)	0.065 (0.011)	94.3	0.526(0.110)	0.103(0.017)	83.8
		50%	0.600(0.085)	0.088 (0.016)	91.0	0.370(0.151)	0.136(0.020)	52.5
		100	0.656(0.042)	0.043 (0.005)	95.3	0.642(0.046)	0.047(0.006)	94.5
	100	10%	0.647(0.048)	0.048 (0.006)	95.3	0.497(0.077)	0.078(0.010)	50.0
		30%	0.612(0.065)	0.065 (0.010)	89.8	0.329(0.111)	0.107(0.012)	15.0
		50%	0.812(0.054)	0.056 (0.016)	94.5	0.813(0.060)	0.059(0.021)	90.0
		100	0.779(0.066)	0.082 (0.055)	92.8	0.704(0.120)	0.106(0.032)	90.0
	200	10%	0.680(0.109)	0.107 (0.028)	85.8	0.566(0.185)	0.156(0.038)	73.6
		30%	0.812(0.039)	0.038 (0.008)	93.2	0.806(0.043)	0.042(0.010)	94.5
		50%	0.793(0.044)	0.046 (0.009)	93.4	0.692(0.081)	0.079(0.018)	81.8
		100	0.727(0.070)	0.075 (0.018)	86.0	0.536(0.126)	0.126(0.026)	45.7
0.804	50	10%	0.805(0.027)	0.027 (0.004)	95.3	0.795(0.030)	0.031(0.005)	95.7
		30%	0.779(0.066)	0.082 (0.055)	92.8	0.704(0.120)	0.106(0.032)	90.0
		50%	0.680(0.109)	0.107 (0.028)	85.8	0.566(0.185)	0.156(0.038)	73.6
		100	0.812(0.039)	0.038 (0.008)	93.2	0.806(0.043)	0.042(0.010)	94.5
	100	10%	0.793(0.044)	0.046 (0.009)	93.4	0.692(0.081)	0.079(0.018)	81.8
		30%	0.727(0.070)	0.075 (0.018)	86.0	0.536(0.126)	0.126(0.026)	45.7
		50%	0.805(0.027)	0.027 (0.004)	95.3	0.795(0.030)	0.031(0.005)	95.7
		100	0.779(0.066)	0.082 (0.055)	92.8	0.704(0.120)	0.106(0.032)	90.0
	200	10%	0.680(0.109)	0.107 (0.028)	85.8	0.566(0.185)	0.156(0.038)	73.6
		30%	0.812(0.039)	0.038 (0.008)	93.2	0.806(0.043)	0.042(0.010)	94.5
		50%	0.793(0.044)	0.046 (0.009)	93.4	0.692(0.081)	0.079(0.018)	81.8
		100	0.727(0.070)	0.075 (0.018)	86.0	0.536(0.126)	0.126(0.026)	45.7

κ_g^w	Sample size	Censoring %	Proposed Kappa Estimator			Sample estimator $\hat{\kappa}$ with complete obs. only		
			$\hat{\kappa}$ mean (SD)	Bootstrap SE mean (SD)	Coverage Prob.	sample estimator mean(SD)	Bootstrap SE SE (SD)	Coverage Prob.
		30%	0.793(0.031)	0.033 (0.001)	92.8	0.677(0.056)	0.059(0.010)	40.4
		50%	0.752(0.051)	0.051 (0.010)	81.6	0.516(0.092)	0.095(0.015)	12.0

Table 2

Simulation results for the modified weighted kappa estimator with discrete survival times simulated from the Discrete Clayton model based on 500 simulation runs, with comparison to sample kappa estimator based on complete observations

κ_g^w	Sample size	Censoring %	Proposed Kappa Estimator			Sample estimator of $\hat{\kappa}$ with complete obs. only		
			$\hat{\kappa}$ mean (SD)	Bootstrap SE mean (SD)	Coverage Prob.	sample estimator mean(SD)	Bootstrap SE (SD)	Coverage Prob.
0.472	50	10%	0.455(0.120)	0.112(0.016)	91.6	0.461(0.120)	0.115(0.017)	92.0
		30%	0.436(0.129)	0.116(0.016)	88.7	0.356(0.165)	0.137(0.022)	81.6
		50%	0.373(0.200)	0.127(0.020)	80.3	0.309(0.248)	0.154(0.031)	69.5
		10%	0.462(0.085)	0.082(0.011)	93.9	0.459(0.090)	0.086(0.011)	93.0
	100	30%	0.457(0.092)	0.086(0.011)	91.2	0.373(0.121)	0.109(0.012)	83.4
		50%	0.427(0.112)	0.103(0.014)	90.8	0.329(0.166)	0.133(0.022)	76.8
		10%	0.472(0.060)	0.058(0.006)	94.0	0.461(0.062)	0.061(0.006)	94.3
		30%	0.460(0.063)	0.062(0.007)	93.6	0.370(0.087)	0.082(0.009)	78.5
	200	50%	0.456(0.081)	0.077(0.009)	93.8	0.336(0.119)	0.106(0.013)	76.6
		10%	0.640(0.092)	0.089(0.018)	92.0	0.628(0.095)	0.095(0.021)	91.8
		30%	0.609(0.100)	0.100(0.019)	95.1	0.518(0.149)	0.130(0.025)	82.6
		50%	0.535(0.137)	0.123(0.022)	83.6	0.466(0.212)	0.154(0.034)	77.5
0.651	50	10%	0.646(0.062)	0.062(0.010)	93.9	0.631(0.067)	0.068(0.012)	96.1
		30%	0.634(0.067)	0.068(0.011)	95.9	0.536(0.099)	0.097(0.017)	85.0
		50%	0.600(0.092)	0.088(0.015)	91.2	0.505(0.139)	0.128(0.024)	81.3
		10%	0.650(0.044)	0.044(0.006)	95.1	0.635(0.046)	0.047(0.006)	94.7
	100	30%	0.640(0.047)	0.048(0.006)	94.9	0.543(0.070)	0.070(0.010)	71.9
		50%	0.630(0.064)	0.062(0.009)	93.8	0.513(0.099)	0.097(0.015)	75.8
		10%	0.790(0.059)	0.060(0.016)	93.6	0.784(0.066)	0.064(0.020)	94.9
		30%	0.765(0.066)	0.072(0.017)	95.1	0.704(0.112)	0.101(0.030)	90.0
	200	50%	0.688(0.130)	0.109(0.025)	86.7	0.663(0.159)	0.137(0.042)	86.7
		10%	0.799(0.039)	0.041(0.008)	95.7	0.795(0.041)	0.044(0.010)	95.5
		30%	0.780(0.044)	0.048(0.009)	95.7	0.720(0.069)	0.070(0.017)	87.3
		50%	0.746(0.065)	0.070(0.014)	91.6	0.683(0.105)	0.102(0.029)	86.5
0.804	50	10%	0.802(0.027)	0.028(0.004)	96.1	0.794(0.031)	0.030(0.005)	93.6

κ_g^w	Sample size	Censoring %	Proposed Kappa Estimator				Sample estimator of κ with complete obs. only			
			κ mean (SD)	Bootstrap SE mean (SD)	Coverage Prob.	SE (SD)	sample estimator mean(SD)	Bootstrap SE SE (SD)	Coverage Prob.	
		30%	0.792(0.031)	0.032(0.005)	95.9		0.721(0.050)	0.049(0.008)		68.8
		50%	0.767(0.046)	0.047(0.007)	92.6		0.698(0.070)	0.070(0.017)		76.2

Table 3
Comparison between discrete survival times arisen from two different mechanisms

κ_g^w	Sample Size	Grouped survival times ¹		Truly discrete survival times ²	
		$\hat{\kappa}_g^w$	bias(%) in the estimated discrete joint distribution *	$\hat{\kappa}_g^w$	bias(%) in the estimated discrete joint distribution *
0.472	50	0.518	19.4	0.469	3.8
	100	0.507	13.5	0.470	2.7
	200	0.483	4.8	0.471	1.4
	1000	0.475	2.6	0.473	0.4
0.651	50	0.686	17.8	0.651	4.2
	100	0.678	12.0	0.650	2.5
	200	0.659	5.6	0.651	2.0
	1000	0.654	2.7	0.651	0.9
0.804	50	0.825	26.8	0.801	6.3
	100	0.818	14.7	0.802	7.7
	200	0.808	9.5	0.804	3.9
	1000	0.806	8.1	0.804	2.3

¹ Grouping continuous survival times generated from the continuous Clayton model (Clayton, 1978)

² Discrete survival times generated from the Discrete Clayton model (Shih, 1998)

* bias (%) is defined as the
$$\frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \frac{|\hat{p}_{ij} - p_{ij}|}{p_{ij}}$$

Simulation results for the generalized modified weighted kappa estimator under trivariate survival models based on 500 simulation runs

Table 4

True κ_g^w	Sample Size	Censoring %	$\hat{\kappa}$ mean	$\hat{\kappa}$ SD	Bootstrap SE Est. mean (SD)	95% Coverage probability
0.472	100	10%	0.483	0.061	0.064 (0.012)	95.5
		30%	0.475	0.067	0.079 (0.022)	96.7
		50%	0.431	0.070	0.085 (0.016)	89.4
	200	10%	0.480	0.046	0.046 (0.009)	95.3
		30%	0.480	0.059	0.067 (0.026)	96.7
		50%	0.463	0.063	0.069 (0.020)	94.7
	350	10%	0.479	0.036	0.037 (0.011)	95.5
		30%	0.486	0.034	0.047 (0.015)	94.5
		50%	0.484	0.050	0.063 (0.021)	98.6
	100	10%	0.656	0.050	0.051 (0.010)	94.9
		30%	0.631	0.055	0.067 (0.023)	94.1
		50%	0.570	0.064	0.077 (0.014)	86.6
	200	10%	0.651	0.033	0.036 (0.006)	96.1
		30%	0.634	0.060	0.058 (0.028)	94.7
		50%	0.592	0.086	0.068 (0.024)	86.5
	350	10%	0.653	0.026	0.027 (0.008)	95.7
		30%	0.645	0.038	0.044 (0.021)	97.1
		50%	0.621	0.061	0.071 (0.031)	95.7
0.804	100	10%	0.802	0.033	0.036 (0.008)	96.5
		30%	0.775	0.043	0.052 (0.021)	87.3
		50%	0.704	0.060	0.068 (0.015)	80.9
	200	10%	0.801	0.024	0.025 (0.005)	96.3
		30%	0.782	0.045	0.047 (0.029)	89.5
		50%	0.738	0.047	0.059 (0.022)	80.2
	350	10%	0.805	0.017	0.019 (0.006)	95.7
		30%	0.789	0.062	0.044 (0.031)	93.0
		50%	0.757	0.071	0.065 (0.033)	90.4