Published in final edited form as:
Biometrics. 2010 December ; 66(4): 1145–1152. doi:10.1111/j.1541-0420.2010.01404.x.

Inverse Probability of Censoring Weighted Estimates of Kendall's *τ* for Gap Time Analyses

Lajmi Lakhal-Chaieb[1,*], **Richard J. Cook**[2,**], and **Xihong Lin**[3,***]

[1]Département de mathématiques et statistique, Université Laval, Québec, Qc G1V 0A6, Canada

[2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[3]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A

Summary

In life history studies interest often lies in the analysis of the inter-event, or gap times and the association between event times. Gap time analyses are challenging however, even when the length of follow-up is determined independently of the event process, since associations between gap times induce dependent censoring for second and subsequent gap times. This paper discusses nonparametric estimation of the association between consecutive gap times based on Kendall's τ in the presence of this type of dependent censoring. A nonparametric estimator which uses inverse probability of censoring weights is provided. Estimates of conditional gap time distributions can be obtained following specification of a particular copula function. Simulation studies show the estimator performs well and compares favourably with an alternative estimator. Generalizations to a piecewise constant Clayton copula are given. Several simulation studies and illustrations with real data sets are also provided.

Keywords

copula; dependent censoring; gap times; Kendall's tau

1. Introduction

In many studies of life history processes, interest lies in the occurrence of two or more consecutive events. Bebchuk and Betensky (2002), for example, consider a three-state model in which the initial state is infection-free, the intermediate state represents HIV infection via blood transfusion, and the terminal state corresponds to AIDS diagnosis. Lin et al. (1999) discuss the analysis of follow-up data from a randomized trial of patients with colon cancer (Moertel et al., 1990). Recurrence of disease and death are the intermediate and terminal events respectively in this setting.

We are concerned with the analysis of data from the TREMIN Trust Study, which is part of the TREMIN Research Program for Women's Health. The particular study we consider is the Menstrual and Reproductive Health Study in which 1,997 women were recruited from

*lakhal@mat.ulaval.ca
**rjcook@uwaterloo.ca
***lin@hsph.harvard.edu

Supplementary Materials

Web Appendices referenced in Sections 2.3 and 3.2 ara available under the Paper information link at the *Biometrics* website: http://www.biometrics.tibs.org

1935-1939 and followed prospectively for up to 40 years. During this time participants were asked to keep a detailed diary of menstrual bleeding. As in Nan et al. (2006), we aim to examine the association between the time of the first menstrual cycle of at least 45 days in duration and menopause to better the understanding of the transitional phase to menopause. Multi-state models provide a natural representation for processes with an initial, intermediate and terminal state in which transitions between these states represent a progression. See Hougaard (2000) for a recent review of methods for the analysis of multi-state data and how these can be used to help understanding the time course of processes. Markov models are often adopted for progressive and degenerative processes, in which the operational time scale is the time since process initiation or some other common origin. However, in many settings interest lies in the sojourn time distributions for particular states. In such cases, the canonical models are semi-Markov but it may not be plausible to assume that successive gap times are independent. When the process is subject to type I right-censoring and there is an association between gap times, the second sojourn time is subject to dependent censoring (Lin et al., 1999), and analysis must take this dependent censoring into account.

There have been considerable advances in the analysis of consecutive sojourn, or gap times in settings such as these in the recent years. Kessing et al. (2004) consider models which assume conditional independence across gap time distributions given shared frailties and Kvist et al. (2007) develop a goodness-of-fit test for the latter model when the frailty is Gamma distributed. Lin et al. (1999) and Schaubel and Cai (2004), among others, derived nonparametric estimators for the joint and conditional survival functions. None of these methods yield simple summary measures of the association between sojourn times, and there is considerable appeal in developing methods with simple measures of association which are robust to dependent censoring.

Kendall's tau is among the most popular measures of association between two positive random variables. Kendall and Gibson (1990) gave an empirical estimate of τ from a sample of uncensored bivariate positive random variables. Several authors proposed estimators for τ with bivariate right-censored parallel observations; see Lakhal et al., (2009) for a review. Betensky and Finkelstein (1999) extended the estimation of τ to bivariate interval-censored observations and Wang and Wells (2000) derived an estimator for τ valid for any censoring scheme as long as a nonparametric estimator for the joint survivor function exists. Little work has been done, however, for nonparametric estimation of τ under more complicated censoring schemes, such as the dependent censoring scheme discussed above. In this paper we propose nonparametric estimators for Kendall's τ measuring the association between two consecutive gap times. The proposed estimators use inverse probability of censoring weights to address the impact of dependent censoring on the second gap times.

In the recent years, copulas had become an attractive framework for modeling the joint distribution of multiple failure times (He & Lawless 2003, Nan et al. 2006). Indeed, under a copula model, association parameters and complex joint probabilities can be easily expressed and estimated. He and Lawless (2003) consider the case of serial events, assume a Clayton copula to model the dependence between two successive gap times, and derive maximum likelihood estimates for the copula parameter and the survival function of the second gap time under weak marginal assumptions. Oakes (1986) proposed a more general copula model under which the cross-ratio is constant within particular regions of the plane. Nan et al. (2006) adapt this model for a joint analysis of the occurence times of an intermediate and a terminal event, and derive estimators of the copula parameters using the maximum pseudo-likelihood procedure of Shih and Louis (1995).

The second purpose of this paper is to develop methods for consecutive gap times analysis based on a standard copula formulation, and then to generalize this in the spirit of Nan et al. (2006), assuming a piecewise-constant cross-ratio model for consecutive gap times. The parameters of the latter model will be estimated by inverting the estimator of Kendall's tau derived in the first part of the paper.

The remainder of this paper is organized as follows. In the next section we define Kendall's τ and propose nonparametric estimators for this association measure using censored serial gap time data. An estimate of the conditional distribution of second gap time based on an assumed Clayton copula function is described in Section 3. These estimates are assessed through simulation studies and illustrated with data from a randomized trial of patients with colon cancer (Moertel et al., 1990). A goodness-of-fit procedure for the assumed copula is derived. This method is generalized in Section 4 to accommodate piecewise-constant cross-ratios as in Oakes (1986) and Nan et al. (2006). Application to the data from the Tremin Trust study of women's reproductive health illustrates the application of the proposed methods and facilitates comparisons with the findings of Nan et al. (2006). Section 5 contains general remarks and topics for further research.

2. Nonparametric Estimation of Kendall's *τ*

Suppose X and Y two time-to-event random variables, with joint survivor function $\pi(x, y) = P(X > x, Y > y)$ and (X_1, Y_1) and (X_2, Y_2) two independent replications of (X, Y). These points are said to be concordant if $(X_1 - X_2)(Y_1 - Y_2) > 0$, i.e. if the marginal rankings of individuals with respect to X and Y agree. They are otherwise discordant. Kendall's τ is defined as the probability of concordance among two pairs of points minus the probability of discordance and given as

$$\tau = \Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

This association measure is independent of the marginal distributions of X and Y and is equal to zero under independence. Moreover, if $\psi_{12} = I\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - I\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$ is the concordance/discordance status, equal to 1 if the pairs (X_1, Y_1) and (X_2, Y_2) are concordant and −1 otherwise, then

$$\tau = E(\psi_{12}) = 4\int_0^\infty \int_0^\infty \pi(x, y) \frac{\partial^2 \pi(x, y)}{\partial x \partial y} dxdy - 1. \quad (2.1)$$

Kendall and Gibson (1990) estimated τ from uncensored bivariate data $\{(X_i, Y_i), i = 1, \ldots, n\}$ by its empirical version

$$\binom{n}{2}^{-1} \sum_{i<j} \psi_{ij}.$$

2.1 Estimation with Parallel Survival Times

In the presence of censoring, it may not be possible to compute ψ_{ij} for some pairs of points, making estimation of τ more difficult; such pairs are called *non-orderable*, while pairs that can be ordered are *orderable*. Let C and D denote the censoring variables associated to X and Y, respectively. One may only observe $(X, \tilde{Y}, \delta_X, \delta_Y)$, where $X = \min(X, C)$, $\tilde{Y} = \min(Y, D)$, $\delta_X = I(X < C)$, and $\delta_Y = I(Y < D)$. Oakes (1982) shows that the pair (i, j) is orderable if

$\{\tilde{X}_{ij} < \tilde{C}_{ij}, \tilde{Y}_{ij} < \tilde{D}_{ij}\}$ where $\tilde{X}_{ij} = \min(X_i, X_j)$, $\tilde{Y}_{ij} = \min(Y_i, Y_j)$, $\tilde{C}_{ij} = \min(C_i, C_j)$ and $\tilde{D}_{ij} = \min(D_i, D_j)$. Denote by L_{ij} the indicator of this event. Oakes (1982) proposes to estimate τ by

$$\hat{\tau}_O = \binom{n}{2}^{-1} \sum_{i<j} L_{ij}\psi_{ij}.$$

This estimator is biased when τ is non-zero. Nevertheless, it is widely used to test independence of a pair of random variables based on censored data.

Recently, Lakhal et al. (2009) propose a method that greatly reduces the bias of $\hat{\tau}_O$ by incorporating use of inverse probability of censoring weights. Let $\hat{p}_{ij}$ be an estimator of the probability of being orderable $p_{ij} = \{\Pr(C > \tilde{X}_{ij}; D > \tilde{Y}_{ij} | \tilde{X}_{ij}, \tilde{Y}_{ij})\}^2$.

The weighted estimate is then

$$\hat{\tau}_{mo} = \binom{n}{2}^{-1} \sum_{i<j} \frac{L_{ij}\psi_{ij}}{\hat{p}_{ij}}. \quad (2.2)$$

This estimator may lie outside [−1, 1] for large values of $|\tau|$, so one may also consider

$$\hat{\tau}_{mo2} = \left(\sum_{i<j} \frac{L_{ij}}{\hat{p}_{ij}}\right)^{-1} \sum_{i<j} \frac{L_{ij}\psi_{ij}}{\hat{p}_{ij}}. \quad (2.3)$$

These estimators are shown to be consistent, asymptotically normally distributed and empirically to be superior to existing competitors in finite samples.

2.2 Estimation With Serial Gap Times and Dependent Censoring

In this section, we discuss the estimation of Kendall's tau between serial gap times. Let $T_1 = X$ and $T_2 = X+Y$ denote the times of occurrence of two successive events. Typically, the follow up process is subject to independent right-censoring by C. Denote by $G(\cdot)$ its survival function. Under this setting, one may only observe $\tilde{X} = \tilde{T}_1$, $\tilde{Y} = \tilde{T}_2 - \tilde{T}_1$, $\delta_X = I(X < C)$ and $\delta_Y = I(X + Y < C)$, where $\tilde{T}_1 = \min(T_1, C)$ and $\tilde{T}_2 = \min(T_2, C)$. Note that if T_1 is censored, T_2 is also censored and $\tilde{Y} = 0$. Hence, Y is censored by $D = \max(0, C - X)$. So, unless X and Y are independent, D is associated to Y.

Under these conditions, Lin et al. (1999), among others, derived a nonparametric estimator for $\pi(x; y)$ for each (x, y) such as $x + y \leq \bar{C}$, where $\bar{C} > 0$ satisfies $G(\bar{C}) > 0$. This estimator may be expressed as

$$\hat{\pi}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(\tilde{X}_i > x; \tilde{Y}_i > y)}{\hat{G}(\tilde{X}_i + y)}, \quad (2.4)$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimator of $G(\cdot)$ based on $\{(\tilde{X}_k + \tilde{Y}_k, 1 - \delta_{Y_k}), k = 1, \ldots, n\}$. One may incorporate (2.4) into (2.1) to estimate τ for serial gap times using the approach of Wang & Wells (2000) by

$$\hat{\tau}_W = 4\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\pi}(\tilde{X}_{(i)},\tilde{Y}_{(j)})\hat{\pi}(\Delta\tilde{X}_{(i)},\Delta\tilde{Y}_{(j)}) - 1, \quad (2.5)$$

where $\tilde{X}_{(0)} = 0 < \tilde{X}_{(1)} < \cdots < \tilde{X}_{(n)}$, and $\tilde{Y}_{(0)} = 0 < \tilde{Y}_{(1)} < \cdots < \tilde{Y}_{(m)}$ are the ordered samples of $\{\tilde{X}_k, k = 1, \ldots, n\}$ and $\{\tilde{Y}_k, k = 1, \ldots, n\}$, respectively, $m = \sum_{k=1}^{n}\delta_{X_i}$ and $\hat{\pi}(\Delta\tilde{X}_{(i)}, \Delta\tilde{Y}_{(j)}) = \hat{\pi}(\tilde{X}_{(i)}, \tilde{Y}_{(j)}) - \hat{\pi}(\tilde{X}_{(i-1)}, \tilde{Y}_{(j)}) - \hat{\pi}(\tilde{X}_{(i)}, \tilde{Y}_{(j-1)}) + \hat{\pi}(\tilde{X}_{(i-1)}, \tilde{Y}_{(j-1)})$ is the estimated density mass at $(\tilde{X}_{(i)}, \tilde{Y}_{(j)})$.

Here, we derive alternative estimators for τ based on an adaptation of (2.2) and (2.3) to deal with serial gap times. The main challenge in this adaptation is to identify orderable pairs and to express and estimate their associated p_{ij} terms.

The orderability condition L_{ij} is still expressed as $\{\tilde{X}_{ij} < \tilde{C}_{ij};\ \tilde{Y}_{ij} < \tilde{D}_{ij}\}$. By continuity, any orderable pair (i, j) satisfies $0 < \tilde{Y}_{ij} < \tilde{D}_{ij}$ and thus $D_i = \max(C_i - X_i, 0) > 0$ and $D_j = \max(C_j - X_j, 0) > 0$, which implies $C_i > X_i$ and $C_j > X_j$. Hence, X_i and X_j are uncensored and L_{ij} reduces to $\{\tilde{Y}_{ij} < \min(C_i - X_i, C_j - X_j)\}$, which can be re-written as $\{C_i > X_i + \tilde{Y}_{ij};\ C_j > X_j + \tilde{Y}_{ij}\}$. Note that for such pairs, $\tilde{Y}_{ij}$ is also observed, so the conditional probability of a pair being orderable is

$$\begin{aligned} p_{ij} &= \Pr\{C_i > X_i + \tilde{Y}_{ij}; C_j > X_j + \tilde{Y}_{ij} | X_i, X_j, \tilde{Y}_{ij}\} \\ &= \Pr\{C_i > X_i + \tilde{Y}_{ij} | X_i, X_j, \tilde{Y}_{ij}\} \times \Pr\{C_j > X_j + \tilde{Y}_{ij} | X_i, X_j, \tilde{Y}_{ij}\} \\ &= G(X_i + \tilde{Y}_{ij}) \times G(X_j + \tilde{Y}_{ij}). \end{aligned}$$

This probability is estimated by

$$\hat{p}_{ij} = \hat{G}(X_i + \tilde{Y}_{ij}) \times \hat{G}(X_j + \tilde{Y}_{ij}). \quad (2.6)$$

Kendall's tau is then estimated by (2.2) and (2.3), with $\hat{p}_{ij}$ given by (2.6) and we denote the resulting estimators by $\hat{\tau}_1$ and $\hat{\tau}_2$, respectively.

In many applications, as will be seen in Sections (3.2) and (4), interest lies on the conditional version of Kendall's tau $\tau_A = E[\psi_{12} | \nu_{12}(A)]$ where A is a subset of $[0; \infty)$ and $\nu_{12}(A) = I(X_1 \in A; X_2 \in A)$. One may compute $\nu_{ij}(A)$ for any orderable pair (i, j) since both X_i and X_j are uncensored for such a pair. Therefore, one may adapt $\hat{\tau}_2$ to estimate τ_A by

$$\hat{\tau}_A = \left\{\sum_{i<j}\frac{L_{ij}\nu_{ij}(A)}{\hat{p}_{ij}}\right\}^{-1}\sum_{i<j}\frac{\psi_{ij}L_{ij}\nu_{ij}(A)}{\hat{p}_{ij}}. \quad (2.7)$$

2.3 Asymptotic Properties

In web Appendix A, we prove the following result.

Theorem 1—*Let ξ_{ij} denote a random variable, measurable for any orderable pair (i, j) and such that $\xi_{ij} = \xi_{ji}$ and let $\gamma = E[\xi_{12}]$. Under some regularity conditions, the distribution of*

$$\sqrt{n}\{\hat{\gamma}-\gamma\}=\sqrt{n}\left\{\binom{n}{2}^{-1}\sum_{i<j}\frac{L_{ij}\xi_{ij}}{\hat{p}_{ij}}-\gamma\right\}$$

converges to a zero-mean normal with variance

$$\sigma_\gamma^2=I-\int_0^\infty\frac{q^2(u)}{\Pr(\tilde{X}+\tilde{Y}>u)}d\Lambda_c(u), \quad (2.8)$$

where I and q(·) *are given in* web Appendix A *and* $\Lambda_c(\cdot)=-\log[G(\cdot)]$ *is the cumulative hazard function of C. This variance can be estimated by plugging in consistent estimators for unknown quantities in* (2.8).

Applying this result with $\xi_{ij}=\psi_{ij}$ and $\gamma=\tau$ yields the asymptotic normality of $\sqrt{n}\{\hat{\tau}_1-\tau\}$ and an expression for its asymptotic variance. We show in web Appendix B that

$$\sqrt{n}\{\hat{\tau}_2-\tau\}=\sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}\frac{L_{ij}(\psi_{ij}-\tau)}{\hat{p}_{ij}}+o_p(1)$$

and thus applying Theorem 1 with $\xi_{ij}=\psi_{ij}-\tau$ and $\gamma=0$ yields the asymptotic normality of $\sqrt{n}\{\hat{\tau}_2-\tau\}$ as well as an expression for its asymptotic variance. Similary, one can show that

$$\sqrt{n}\{\hat{\tau}_A-\tau_A\}=\frac{\sqrt{n}}{E[\nu_{12}]}\binom{n}{2}^{-1}\sum_{i<j}\frac{L_{ij}\nu_{ij}(\psi_{ij}-\tau_A)}{\hat{p}_{ij}}+o_p(1).$$

Asymptotic normality and an expression for the asymptotic variance follow by applying Theorem 1 with $\xi_{ij}=\nu_{ij}(\psi_{ij}-\tau_A)$ and $\gamma=0$.

3. A Clayton Copula Model for serial gap times (*X*, *Y*)

Once the first event occurs, say at $X=x$, the conditional survival function $S_Y(\cdot|X=x)$ often becomes of interest. A convenient way to estimate this probability is to assume a Clayton copula for the pair (*X*, *Y*). In this section, we investigate such a model, derive related inference procedures, and discuss extensions.

3.1 Model and properties

Under a Clayton copula for (*X*, *Y*), the joint survival function is expressed as

$$\pi(x,y)=\max\left[\left\{S_X(x)^{-(\theta-1)}+S_Y(y)^{-(\theta-1)}-1\right\}^{-1/(\theta-1)},0\right], \quad (3.1)$$

where $\theta=\lambda_Y(y|X=x)/\lambda_Y(y|X>x)$ is the cross-ratio and $\lambda_Y(y|\cdot)=\lim_{dy\downarrow 0}\Pr(Y\le y+dy|Y>y;\cdot)/dy$ is the conditional hazard function of *Y*. Under this model, θ is constant and related to Kendall's τ through $\tau=(\theta-1)/(\theta+1)$. Moreover, one has

$$S_Y(y|X=x)=\{S_Y(y|X>x)\}^\theta \text{ for all } 0\le x,y. \quad (3.2)$$

Lin et al. (1999) note that, $S_Y(y|X > x)$ can be estimated by $\hat{\pi}(x; y)/\hat{\pi}(x; 0)$ for $x + y < C$, where $\hat{\pi}(\cdot, \cdot)$ is given by (2.4). We can also estimate θ by $\hat{\theta} = (\hat{\tau} + 1)/(\hat{\tau} - 1)$, where $\hat{\tau}$ is the nonparametric estimator of τ, derived in Section (2.2). A natural estimator for $S_Y(y|X = x)$ is then obtained by plugging in estimators for unknown quantities in (3.2).

Taylor series expansions of

$$\sqrt{n}\left\{\widehat{S}_Y(y|X=x) - S_Y(y|X=x)\right\} = \sqrt{n}\left[\left\{\widehat{S}_Y(y|X>x)\right\}^{\widehat{\theta}} - \{S_Y(y|X>x)\}^{\theta}\right],$$

along with the asymptotic presentations of $\sqrt{n}\{\widehat{S}_Y(y|X>x) - S_Y(y|X>x)\}$ given by Lin et al. (1999) and of $\sqrt{n}\{\widehat{\theta} - \theta\}$ given in Appendix A, prove that the distribution of $\sqrt{n}\{\widehat{S}_Y(y|X=x) - S_Y(y|X=x)\}$ converges to a zero mean normal. However, its variance involves complex formulas and we suggest variance estimation based on the Jackknife procedure.

3.2 A Goodness-of-Fit Test for the Clayton Copula Under Dependent Censoring

In the previous section, we derived an estimator for the conditional survival function $S_Y(y|X = x)$ under a Clayton copula model for the pair (X, Y). If this assumption does not hold, the estimator for $S_Y(y|X = x)$ may be biased and it is therefore desirable to check the adequacy of the Clayton copula. Several copula goodness-of-fit tests exist for complete data (Genest et al., 2009) and parallel censored observations (Lakhal-Chaieb 2010), but no procedure exists for successive gap times. We develop such a procedure here for the Clayton copula.

Let $M > 0$ and $A = [M; \infty)$. By Manatunga & Oakes (2006), the Clayton copula is the only family preserved under truncation; i.e. if (X, Y) follows a Clayton copula, then $(X, Y | X > M)$ follows also a Clayton copula, with the same association parameter. In particular, the conditionnal Kendall's tau τ_A is equal to τ. Our test is then based on $Q = \sqrt{n}\{\hat{\tau}_2 - \hat{\tau}_A\}$.

Under H_0, one may write the test statistics as $Q = \sqrt{n}\{(\hat{\tau}_2 - \tau) - (\hat{\tau}_A - \tau_A)\}$. It follows from web Appendix B that:

$$Q = \sqrt{n}\binom{2}{n}^{-1}\sum_{i<j}\frac{L_{ij}}{\hat{p}_{ij}}(\psi_{ij} - \tau)\left(1 - \frac{\nu_{ij}}{E[\nu_{12}]}\right) + o_p(1).$$

Applying the result of Theorem 1 with $\xi_{ij} = (\psi_{ij} - \tau)\left(1 - \frac{\nu_{ij}}{E[\nu_{12}]}\right)$ and $\gamma = 0$ yields the asymptotic normality and an expression for the asymptotic variance of Q under H_0. We reject H_0 at level α if $|Q/s.e.(Q)| > Z_{\alpha/2}$. The performance of this test will be investigated by simulations in the next Section.

Clearly, the choice of M dramatically affects the performance of our test. For small values of M, the difference between τ and τ_A is tiny for most copula families and the test may not be able to detect it. At the other extreme, for large values of M, the estimate of τ_A may be based on a relatively small number of points and thus may not be precise. Thus our test is most useful for moderate values of M. In our numerical investigations, we take M to be the estimated median from the Kaplan-Meier estimate based on $\{(\tilde{X}_k, \delta_{X_k}), k = 1, \cdots, n\}$.

3.3 Applications Involving the Clayton Copula

Colon cancer data—Moertel et al. (1990) discuss a clinical trial where patients treated for colon cancer are randomized into two groups: therapy and placebo, including 304 and 315 patients respectively. Patients are at risk of ordered events and the serial gap times in this example are: the time from randomization to cancer recurrence (X) and the time from cancer recurrence to death (Y). At the end of the study, 108 patients died among the 119 who had cancer recurrence in the therapy group and 155 died among 177 who had cancer recurrence in the placebo group. We computed $\hat{\tau}_1$ and $\hat{\tau}_W$ for both groups. We found $\hat{\tau}_1$ = 0.2725 (s.e.=0.062) and $\hat{\tau}_W = -0.796$ (s.e.=0.613) for the therapy group and $\hat{\tau}_1 = 0.2685$ (s.e.=0.058) and $\hat{\tau}_W = 0.012$ (s.e.=0.779) for the placebo group. Our estimator $\hat{\tau}_1$ detects a significant positive dependence between X and Y in both groups, as conjectured by Lin et al. (1999). Furthermore, $\hat{\tau}_1$ suggests that the magnitude of this dependence is not affected by the therapy. The variance of $\hat{\tau}_W$, estimated by the Jackknife procedure, is too large to make inference. A convenient way to illustrate this dependence is to investigate the conditional survival $S_Y(\cdot|X = x)$ under a Clayton copula for the pair (X, Y). In Figure 1, we report the median of $\hat{S}_Y(\cdot|X = x)$ versus x for both groups.

Figure 1 suggests that therapy decreases survival time following cancer recurrence. This is in agreement with the conclusions of Lin et al. (1999) and He and Lawless (2003). The Clayton copula assumption is tested for each group by the procedure presented in Section (3.2). It gives $Q = 1.983$ (s.e.=2.8115; p=0.48) and $Q = 1.906$ (s.e.=2.068; p=0.36) for the placebo and treatment group, respectively.

Simulations—The first set of simulations was conducted to assess and compare the finite sample performances of $\hat{\tau}_1$, $\hat{\tau}_W$ and the resulting estimators for $S_Y(y|X = x)$. Three real τ values (0.2, 0.5 and 0.8) and two sample sizes (100 and 200) were used to generate correlated pairs (X, Y) using a Clayton copula, with exponential margins with means equal to 1 and 1/2 respectively. The censoring variable C was generated from a Weibull distribution with parameters controlling the censoring fractions $cf_1 = \Pr(X > C)$ and $cf_2 = \Pr(X + Y > C)$. Two scenarios corresponding to $(cf_1, cf_2) = (0.15, 0.30)$ and $(0.20, 0.40)$ were considered. For each combination of the parameters above, 1000 samples were generated and for each simulated data set, we computed $\hat{\tau}_1$, $\text{var}(\hat{\tau}_1)$, $\hat{\tau}_W$ and the resulting $\hat{S}_Y(y_i|X = x_0)$; $i = 1, \ldots, 4$ at points $x_0, y_1, \ldots, y_4$ such that $S_X(x_0) = 1/2$ and $S_Y(y_i|X = x_0) = i/5$. The empirical means and mean square errors of $\hat{\tau}_1$, $\hat{\tau}_W$ and $\hat{S}_Y(y_i|X = x_0)$ are reported in Table 1.

The empirical means of the estimates of $\text{Var}(\hat{\tau}_1)$ along with their coverage rate of the 95% confidence interval are reported in Table 2.

As expected, the censoring fraction affects all estimators. Table 1 shows that $\hat{\tau}_1$ outperforms $\hat{\tau}_W$ under all simulation conditions, in terms of mean squared error. Note that $\hat{\tau}_1$ is virtually unbiased, except under the extreme conditions $cf_2 = 0.4$ and $\tau = 0.8$. On the other hand, the bias of $\hat{\tau}_W$ is non-negligible, even under light censoring and small values of τ. The same conclusions are made for estimators of $\hat{S}_Y(y_i|X = x_0)$ based on $\hat{\tau}_1$ and $\hat{\tau}_W$ respectively. Table 1 suggests that the estimator of $S_Y(y_i|X = x_0)$ based on τ_W is not particularly reliable, especially when $\tau = 0.8$. Table 1 also suggests that the performance $\hat{\tau}_1$ improves as τ increases while the opposite is observed for $\hat{S}_Y(y_i|X = x_0)$ based on τ_1. The results in Table 2 suggest that our estimator of $\text{Var}(\hat{\tau}_1)$ provides a reasonable measure of the variability of $\hat{\tau}_1$

A second set of simulations was conducted to investigate the performance of the goodness-of-fit test presented in Section 3.2. Samples of 200 pairs with $(cf_1, cf_2) = (0.2, 0.4)$ were generated from the Clayton, Frank and Gumbel families using the transformation method of Genest & Rivest (1993). Proportions of rejection of the null hypothesis corresponding to the Clayton copula are reported in Figure 2. These simulations show that the rejection rate of

this test is comparable with the nominal level. Furthermore, it indicates that the Gumbel copula is easier to distinguish from the Clayton copula than the Frank, in accordance with well known properties of these copulas.

4. A Piecewise Clayton Copula

4.1 Model and Method of Inference

Rejection of the null hypothesis by the test presented in the previous section implies that the assumption of a constant cross-ratio over the positive quadrant does not hold for the data at hand. Nan et al.(2006) discuss a copula model that relaxes this assumption and assumes that the cross-ratio depends on one of the time-to-events variables, say X. They considered a partition $0 = w_0 < w_1 < \cdots < w_K$ of the support of X such as the cross-ratio is constant inside each grid $(x, y) \in A_k \times [0, \infty]$ and equal to θ_k, where $A_k = [w_{k-1}, w_k)$. Under such conditions, (3.2) becomes

$$S_Y(y|X=x)=[S_Y(y|X>x)]^{\theta_k} \quad \text{for} \quad (x,y) \in A_k \times [0,\infty]. \quad (4.1)$$

and the resulting model is referred to as a piecewise Clayton copula. In web Appendix C we show that the conditional Kendall tau $\tau_{A_k} = E[\psi_{12}|\nu_{12}(A_k)]$ is related to θ_k through

$$\tau_{A_k}=g(a,b,\theta_k)=\frac{4}{(b-a)^2}\left[\frac{\theta_k}{2(\theta_k+1)}(b^2-a^2)+J(a,b,\theta_k)\right]-1, \quad (4.2)$$

where $a = S_X(w_k)$, $b = S_X(w_{k-1})$ and

$$J(a,b,\theta)=\begin{cases} \frac{1}{\theta-1}\int_0^{a^{-(\theta-1)}}(a^{-(\theta+1)}-t)^{-1/(\theta-1)}(b^{-(\theta+1)}-t)^{-\theta/(\theta-1)}\,dt-\frac{a^2}{2} & \text{if } \theta_k<1 \\ \frac{a^2}{2}-\frac{1}{\theta-1}\int_0^{\infty}(a^{-(\theta+1)}+t)^{-1/(\theta-1)}(b^{-(\theta+1)}+t)^{-\theta/(\theta-1)}\,dt & \text{if } \theta_k>1. \end{cases} \quad (4.3)$$

Nan et al. (2006) assume a piecewise Clayton copula for (T_1, T_2) and estimate the model parameters $\{\theta_k, k = 1, \ldots, K\}$ by adapting the methodology of Shih and Louis (1995). This approach ignores the ordered nature of (T_1, T_2) and as a consequence, the resulting estimator of $\Pr(T_1 > T_2)$ not identically equal to zero, as it should be. A more appealing approach is to assume a piecewise Clayton copula for (X, Y) since no order restrictions are required for these variables. Moreover, if the presence of a cycle of 45 days or more signals early changes prior to the onset of menopause, Y may well be a more natural quantity to focus on for some scientific questions.

For $k = 1, \cdots, K$, we estimate τ_{A_k} by (2.7). An estimate of θ_k is then obtained as the solution of $\hat{\tau}_{A_k} = g[\hat{S}_X(w_k), \hat{S}_X(w_{k-1}), \theta_k]$. The resulting estimate $\hat{\theta}_k$, along with (4.1), yields an estimator for $S_Y(y|X = x)$. The variance of $\hat{\theta}_k$ involves those of τ_{A_k}, $\hat{S}_X(w_k)$ and $\hat{S}_X(w_{k-1})$ and some covariance terms, and thus is complex to compute. We suggest using the Jackknife method to estimate $\mathrm{Var}(\hat{\theta}_k)$ and $\mathrm{Var}[\hat{S}_Y(y|X = x)]$.

4.2 Numerical Applications

Analysis of the Tremin Trust Data—We consider the data used by Nan et al. (2006) who analyzed follow up data from 562 women who were less than 25 years of age at the time of recruitment, provided data on the age of menarche, and were on study at 35 years of age. The purpose of this study is to characterize the association between several bleeding makers such as the age at the first 45-day cycle, and menopause.

Consider a reproductive life cycle where $T_1 = X$ represents the age at the first cycle of at least 45 days duration, and T_2 is the age at menopause; $Y = T_2 - X$ is then the time from the first 45 day cycle to menopause. At the end of the study, 193 women were observed to reach menopause among the 357 women who experienced the 45-day marker. We first test the assumption of a constant cross-ratio between X and Y by the procedure derived in Section (3.2). We find $Q = -4.507$, s.e.=1.0875 and p-value=3.4×10^{-5} leading us to reject the null hypothesis of a constant cross-ratio. We consider then a piecewise Clayton copula where the cross-ratio is assumed constant within the same intervals of $T_1 = X$ adopted by Nan et al. (2006), namely: 35 – 39, 40 – 45, 46 – 49 and 50+ years of age. These boundaries satisfy $\hat{S}_X(w_k) = \{0.8383, 0.562, 0.281\}$. We estimate $\hat{\tau}_{A_k}$, $k = 1, 2, 3, 4$ using equation (2.7) and obtain $\hat{\tau}_{A_1} = -0.049$ (s.e.=0.160), $\hat{\tau}_{A_2} = -0.374$ (s.e.=0.068), $\hat{\tau}_{A_3} = -0.235$ (s.e.=0.095) and $\hat{\tau}_{A_4} = -0.139$ (s.e.=0.112). Inverting these estimates yields $\hat{\theta}_1 = 0.411$; 95% C.I. [0.145;15.023], $\hat{\theta}_2 = 0.184$; 95% C.I. [0.129;0.260], $\hat{\theta}_3 = 0.327$; 95% C.I. [0.202;0.673] and $\hat{\theta}_4 = 0.756$; 95% C.I. [0.472;1.175].

We conclude that there is a significant association between the 45-day cycle and menopause only if the bleeding event occurs inside the age interval 40 – 49. This is in accordance with the results of Nan et al. (2006), who detect a significant association between T_1 and T_2 in the same region. Nan et al. (2006) present plots of $\Pr(T_2 > t_2 | T_1 = t_1)$ versus t_2 for different values of t_1. This probability is equal to $S_y(t_2 - t_1 | X = t_1)$ and thus can be estimated by our model. The results are presented in Figure 3 for $t_1 \in \{36, 39, 42, 45, 48, 51\}$.

Figure 3 is again in accordance with the one produced by Nan et al. (2006) for women who experience the 45 days cycle marker after 40 years of age. However, we obtain different results for women who experience the 45 days cycle before age 40. In particular, we estimate that such a woman has about a 20% chance of experiencing menopause before age 45 (see the $X = 36$ and 39 curves), while this probability is estimated to be approximately null by Nan et al. (2006). This suggests a lack of fit of one of the models and thus there is a need for appropriate goodness-of-fit tests for such copulas models. This is beyond the scope of this paper.

Simulations—A second set of simulations were conducted to assess the performances of the estimator of $\hat{\tau}_{A_k}$ given by (2.7). Censored samples were generated following the piecewise Clayton copula with conditions similar to those of the Tremin Trust dataset: $n = 562$, bound-aries such as $S_X(w_k) \in \{0.8383, 0.562, 0.281\}$ and $\tau_{A_k} \in \{-0.049, -0.374, -0.235, -0.139\}$. The marginal distributions of X and Y were exponential with means 1 and 1/2 respectively. The distribution of C was taken as Weibull with parameters such as $cf_1 = 0.35$ and $cf_2 = 0.65$. The empirical means of $\hat{\tau}_{A_k}$ from 1000 iterations are -0.049 (s.e.=0.10), -0.373 (s.e.=0.066), -0.229 (s.e.=0.054) and -0.136 (s.e.=0.053).

The squared roots of the means of the estimated variances are {0.09, 0.064, 0.057, 0.055}. These simulations show that the $\hat{\tau}_{A_k}$ are virtually unbiased and that their variances are consistently estimated.

5. Discussion

In this paper, we consider nonparametric estimation of the association between successive gap times when the second gap time is subject to dependent censoring. This dependent censoring is induced by the association between the first and second gap times and is present even when the process itself is independently right-censored. Inverse probability weights are used to obtain a nonparametic estimate of Kendall's tau. This estimator is then used to make inferences about the conditional survivor function $S_Y(y|X = x)$ under a Clayton copula. While this is possible by (3.2) for the Clayton copula, this particular relation does not hold

for other copula families. Alternative ways of estimating $S_Y(y|X = x)$ under arbitrary copula functions warrant investigation for settings where the Clayton copula is inappropriate.

Recurrent event processes are increasingly arising in health research and gap time models frequently offer an appealing framework for analysis. Generalizations of the proposed estimator of Kendall's tau would facilitate inferences about gap time models for recurrent event processes with copula formulations. Such models would allow the estimation of

$$\Pr(T_k > t_k | T_1 = t_1,\ T_2 = t_2, \ldots, T_{k-1} = t_{k-1}).$$

This is an area warranting development.

We employed a piecewise Clayton copula formulation in Section 4. Such models have received relatively little attention in the literature and methods to assist in specification of the regions with a constant cross-ratio would be helpful since at present they are based on *ad hoc* graphical methods. Derivation of formal goodness-of-fit tests for these models and objective procedures for specifying these regions would increase their practical utility.

In many settings event times are not observed precisely, but individuals are only assessed at periodic inspection times creating interval-censored data on gap times. In the context of a three-state progressive model, the intermediate event may be subject to interval-censoring and the terminal event right-censoring, or both events may be interval censored. In this case nonparametric estimation of the association between gap times is considerably more challenging, and parametric assumptions may be required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to two anonymous referees and to the associate editor for their constructive comments that led to substantial improvement of the initial version of this paper. This work was partially supported by individual grants from the Natural Sciences and Engineering Research Council of Canada to the two first authors and by grants R37 CA76404 and P01 CA134294 from the National Cancer Institute to the third author.

References

Bebchuk JD, Betensky RA. Local likelihood analysis of the latency distribution with interval censored intermediate events. Statist Med. 2002; 21:3475–3491.

Betensky R, Finkelstein D. An extension of Kendall's coefficient of concordance to bivariate interval-censored data. Statist Med. 1999; 18:3101–3109.

Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika. 1995; 82(4):835–845.

Genest C, Rémillard B, Beaudoin D. Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics. 2009; 44:199–213.

Genest C, Rivest L-P. Statistical inference procedures for bivariate Archimedean copulas. J Am Statist Assoc. 1993; 88:1034–1043.

Hougaard P. Multi-state models: A review. Lifetime Data Analysis. 2000; 5:239–264. [PubMed: 10518372]

He W, Lawless JF. Flexible maximum likelihood methods for bivariate proportional hazards models. Biometrics. 2003; 12:837–848. [PubMed: 14969462]

Kendall, M.; Gibbons, JD. Rank Correlation Methods. Fifth. London: Edward Arnold; 1990. A Charles Griffin Title

Kessing LV, Hansen MG, Andersen PK. Course of illness in depressive bipolar disorders: Naturalistic study, 1994–1999. Brit J Psych. 2004; 185:372–377.

Kvist K, Gerster M, Andersen PK, Kessing LV. Non-parameteric estimation and model checking procedures for marginal gap time distributions for recurrent events. Statist Med. 2007; 26:5394–5410.

Lakhal-Chaieb L. Copula inference under censoring. Biometrika. 2010 In Press.

Lakhal-Chaieb L, Rivest L-P, Beaudoin D. IPCW estimator for Kendall's tau under bivariate censoring. International Journal of Biostatistics. 2009; 5(1):8.

Lin DY, Sun W, Ying Z. Nonparametric estimation of the gap time distributions for serial events with censored data. Biometrika. 1999; 86:59–70.

Manatunga AK, Oakes D. A meaure of association for bivariate frailty distributions. Journal of Multivariate Analysis. 1996; 56:60–74.

Moertel CG, Fleming TR, McDonald JS, et al. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. New England Journal of Medicine. 1990; 322:352–358. [PubMed: 2300087]

Nan B, Lin X, Lisabeth LD, Harlow SD. Piece-wise constant cross-ratio estimates for the association between age at marker event and age at menopause. J Am Statist Assoc. 2006; 101:65–77.

Oakes D. A concordance test for independence in the presence of censoring. Bioemtrics. 1982; 38(2): 451–455.

Oakes, D. A model for bivariate survival data. In: Moolgavkar, SH.; Prentice, RL., editors. Modern Statistical Methods in Chronic Disease Epidemiology. Wiley; New York: 1986.

Schaubel DE, Cai J. Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. Biometrika. 2004; 91:291–303.

Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. Biometrics. 1995; 51:1384–1399. [PubMed: 8589230]

Wang W, Wells MT. Estimation of Kendall's tau under censoring. Statist Sinica. 2000; 10(4):1199–1215.

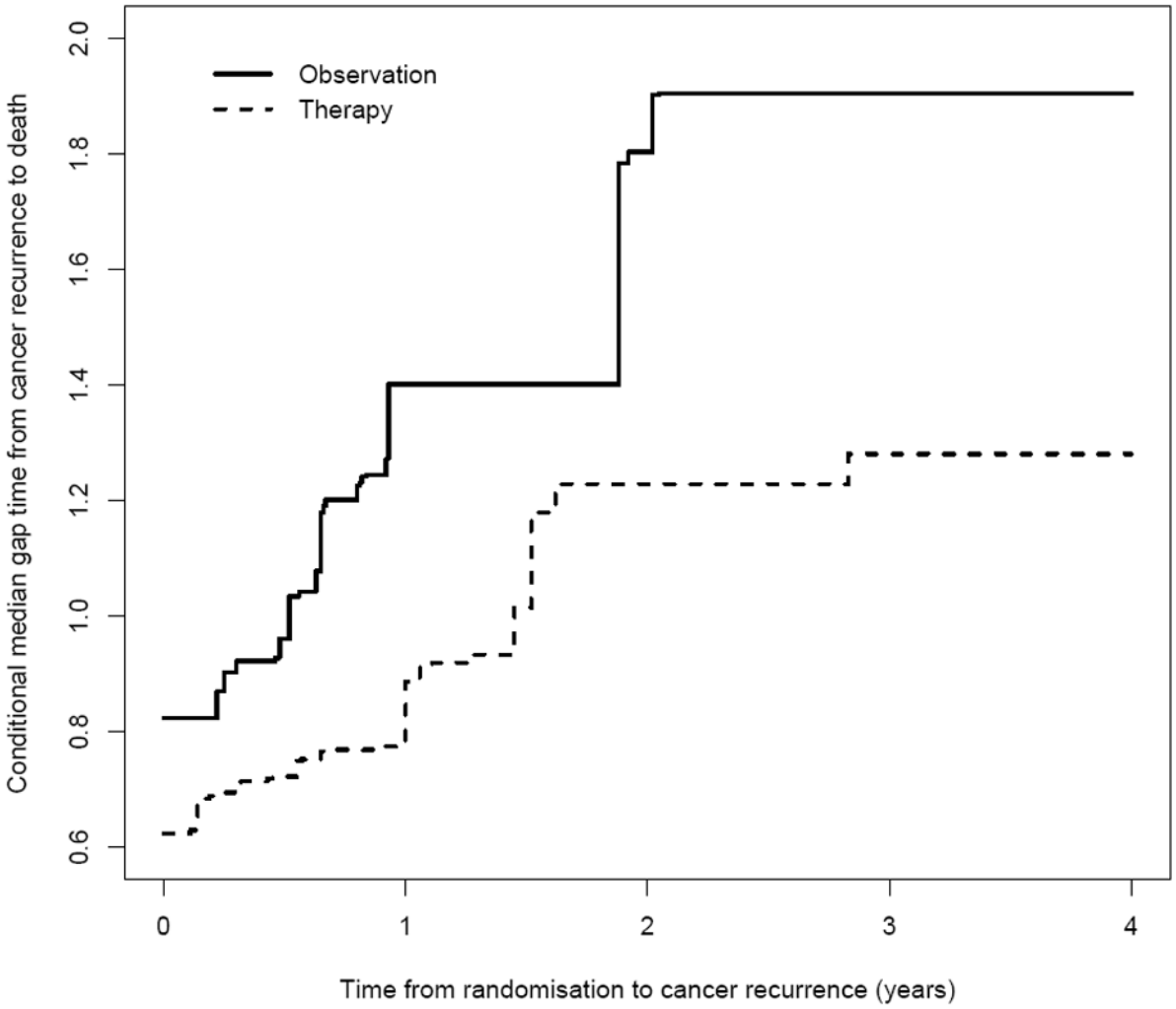

Figure 1.
Empirical estimates of conditional median time from cancer occurrence to death given cancer occurrence date.

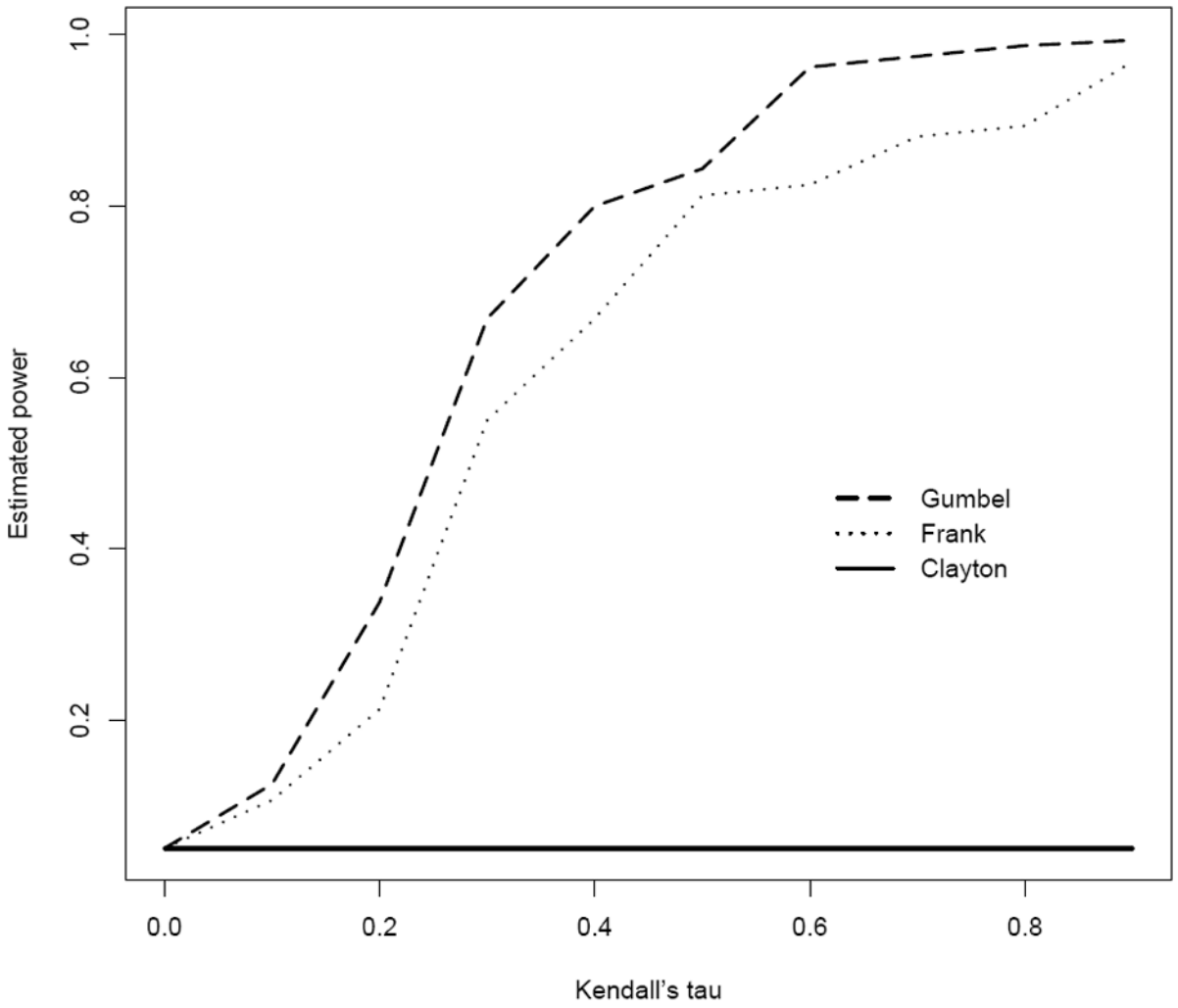

Figure 2.
Empirical power based on 1000 samples of size 200 from Gumbel, Frank and Clayton copulas.

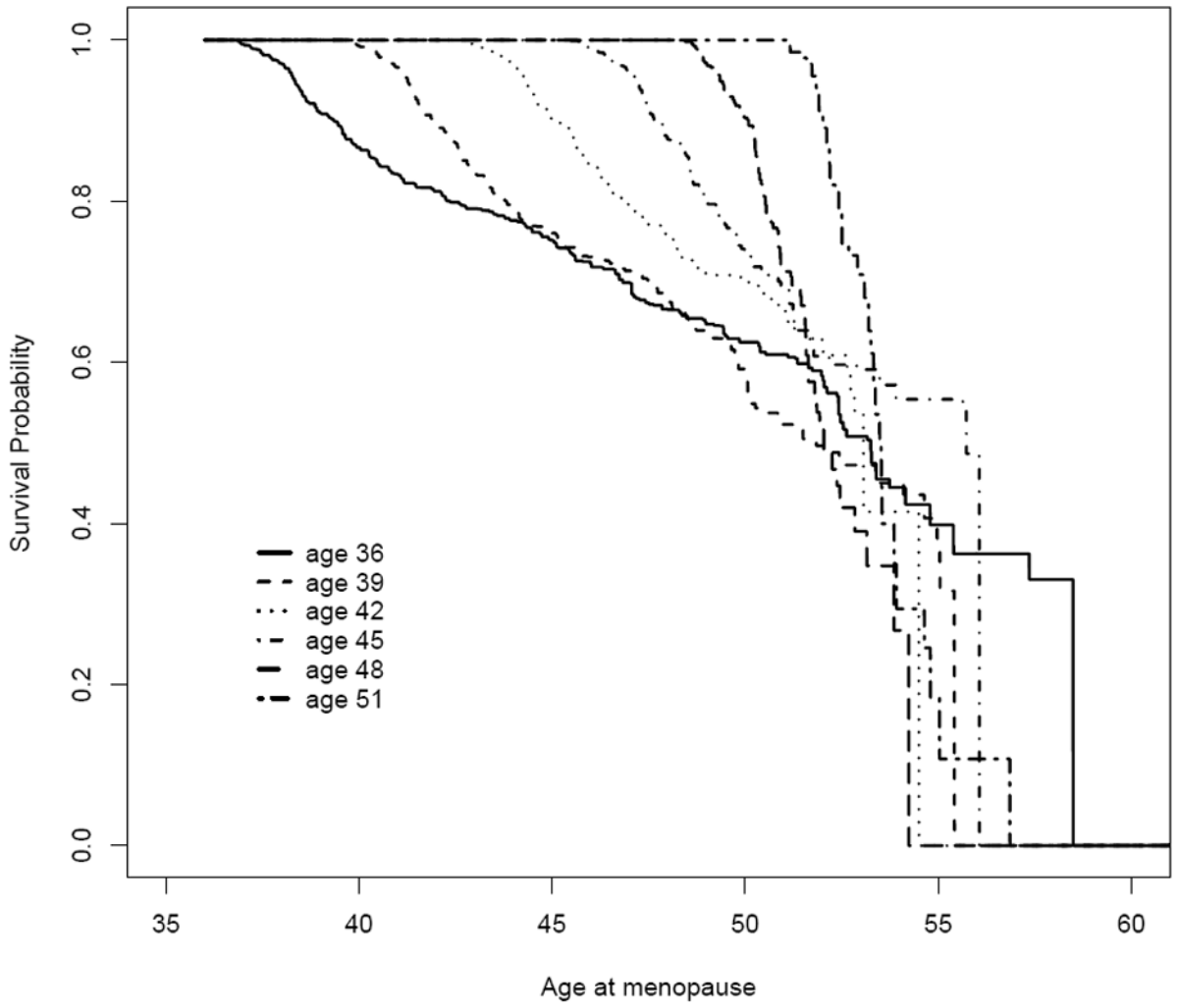

Figure 3.
Empirical estimates of conditional survivor functions for menopause given age at first cycle of ≥ 45 days duration.

Table 1

Means and Mean square errors (in parentheses; $\times 10^4$) of $\hat{\tau_1}$ (IPCW) and $\hat{\tau_W}$ (WW) and their resulting conditional survival estimates

			$\hat{\tau}$		$\hat{S}(y_1\|X=x_0)$		$\hat{S}(y_2\|X=x_0)$		$\hat{S}(y_3\|X=x_0)$		$\hat{S}(y_4\|X=x_0)$	
n	cf_1 cf_2	τ	**IPCW**	**WW**	**IPCW**	**WW**	**IPCW**	**WW**	**IPCW**	**WW**	**IPCW**	**WW**
100	15% 30%	0.2	0.187 (6.7)	0.229 (19.0)	0.201 (4.9)	0.179 (8.0)	0.405 (7.2)	0.374 (13.2)	0.601 (7.7)	0.570 (14.2)	0.801 (6.0)	0.780 (9.5)
		0.5	0.492 (4.1)	0.547 (25.7)	0.210 (8.6)	0.175 (13.2)	0.410 (14.9)	0.354 (27.3)	0.618 (18.7)	0.557 (34.8)	0.809 (16.0)	0.748 (29.9)
		0.8	0.797 (1.0)	0.849 (23.4)	0.232 (28.5)	0.203 (35.6)	0.401 (44.8)	0.352 (63.9)	0.547 (50.2)	0.481 (88.3)	0.670 (56.6)	0.592 (115.6)
	20% 40%	0.2	0.185 (9.1)	0.231 (37.0)	0.193 (7.1)	0.175 (11.1)	0.398 (10.3)	0.367 (19.6)	0.599 (10.9)	0.566 (21.9)	0.801 (7.7)	0.774 (15.3)
		0.5	0.484 (5.0)	0.540 (40.9)	0.201 (10.3)	0.174 (16.9)	0.409 (26.8)	0.360 (36.0)	0.615 (26.8)	0.548 (46.32)	0.808 (21.1)	0.736 (34.6)
		0.8	0.791 (1.3)	0.830 (36.6)	0.246 (42.8)	0.259 (57.4)	0.401 (54.9)	0.401 (76.6)	0.503 (64.9)	0.478 (95.5)	0.620 (84.7)	0.575 (135.4)
200	15% 30%	0.2	0.195 (3.3)	0.218 (8.3)	0.199 (2.5)	0.185 (4.0)	0.399 (3.3)	0.380 (5.9)	0.604 (3.2)	0.587 (5.3)	0.805 (2.7)	0.795 (3.7)
		0.5	0.495 (2.1)	0.522 (11.3)	0.204 (3.3)	0.183 (6.8)	0.411 (7.1)	0.379 (13.8)	0.608 (8.3)	0.577 (15.4)	0.804 (7.0)	0.780 (11.2)
		0.8	0.797 (0.4)	0.824 (9.8)	0.216 (14.0)	0.188 (21.3)	0.417 (32.0)	0.365 (48.9)	0.587 (35.7)	0.531 (62.1)	0.821 (34.7)	0.667 (65.6)
	20% 40%	0.2	0.188 (4.4)	0.205 (15.9)	0.194 (3.8)	0.185 (6.4)	0.397 (5.3)	0.383 (10.1)	0.599 (5.5)	0.584 (10.3)	0.800 (4.0)	0.789 (7.3)
		0.5	0.490 (2.5)	0.515 (18.3)	0.204 (5.9)	0.186 (10.8)	0.412 (11.7)	0.381 (15.7)	0.608 (13.9)	0.575 (24.7)	0.801 (9.2)	0.777 (17.9)
		0.8	0.795 (0.5)	0.821 (21.6)	0.228 (26.5)	0.227 (37.5)	0.415 (45.8)	0.394 (63.8)	0.548 (50.0)	0.519 (75.4)	0.664 (61.0)	0.623 (103.4)

Table 2

Empirical means of the estimates of Var($\hat{\tau_1}$) with their empirical coverage rate

	***n* = 100**		***n* = 200**	
τ	***cf*$_1$ = 20%**	***cf*$_1$ = 40%**	***cf*$_1$ = 20%**	***cf*$_1$ = 40%**
0.2	(6.9) 95.3	(9.0) 94.0	(3.1) 95.0	(4.5) 94.5
0.5	(4.0) 94.6	(4.8) 94.8	(2.2) 94.2	(2.6) 95.5
0.8	(1.0) 95.2	(1.4) 93.8	(0.6) 94.3	(0.6) 94.5