# Semiparametric estimation of marginal mark distribution

By YIJIAN HUANG

*Department of Biostatistics, Rollins School of Public Health, Emory University,*

*Atlanta, Georgia 30322, U.S.A.*

yhuang5@emory.edu

and KRISTIN BERRY

*Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.*

kberry@u.washington.edu

*Running title:* Marginal mark estimation

Summary

In many applications, the outcome of interest is a mark such that its observation is contingent upon occurrence of an event. With incomplete follow-up data, the marginal mark distribution is, however, nonparametrically nowhere identifiable in many practical situations. To address this problem, we suggest a semiparametric model that postulates a normal copula for the association between the mark and survival time, but leaves the marginals unspecified. We show identifiability of the marginal mark distribution under this model, and propose an inference procedure. The estimated marginal distribution function is consistent and asymptotically normal, and it provides a basis for estimating summaries of the mark. Furthermore, we propose graphical model-checking methods and Kolmogorov–Smirnov-type goodness-of-fit tests. Simulation studies demonstrate that the inference procedure performs well in practical settings. The method is applied to the estimation of lifetime medical cost in a lung cancer trial.

*Some key words:* Copula; Goodness-of-fit test; Identifiability; Induced dependent censoring; Kolmogorov–Smirnov statistic; Linear transformation model; Marked point process; Medical cost; Normal copula; Quality-adjusted survival time.

## 1. Introduction

A mark is a random variable associated with an event such that its observation is contingent upon occurrence of the event. Examples include lifetime medical cost and quality-adjusted survival time (Olschewski & Schumacher, 1990). Parametric estimation of the marginal mark distribution is not always feasible, particularly in the case of lifetime medical cost. The distribution of medical cost is highly skewed, and typical parametric distribution families may not provide a good fit. On the other hand, nonparametric estimation is plagued by thorny issues with incomplete follow-up data. First, if the mark is a cumulative measure at the event, such as lifetime medical cost, the induced censoring pattern on the mark scale is typically dependent (Glasziou et al., 1990; Lin et al., 1997). Secondly and even more troublesome, the marginal mark distribution is nowhere identifiable in many practical situations (Huang & Louis, 1998). To understand this, consider lifetime medical

cost with data collected from a study with a duration of, say, three years. Then no information on medical cost can be observed for those participants who accumulate zero cost within three years and survive beyond the study duration, so that the distribution function of the mark is not identifiable at any point away from zero. This explains why most previous investigations addressed time-restricted marks instead, including Glasziou et al. (1990), Lin et al. (1997), Zhao & Tsiatis (1997) and Bang & Tsiatis (2000). For instance, three-year-restricted cost is the accumulated cost at three years or death, whichever happens earlier, but the time limit is artificial and the time-restricted cost may not be a good approximation of lifetime cost. Huang & Louis (1998) did show that the joint distribution of the mark and survival time is identifiable except for the tail area on the time scale, but this may not be satisfactory when the marginal mark distribution is of interest.

These concerns motivate semiparametric estimation as an attempt to strike a balance between model flexibility and identifiability, and we suggest copula-based models that parameterise the association between the mark of interest and survival time but leave the marginals unspecified. Since any assumption is untestable in the tail area on the time scale, due to limited study duration, this proposal specifically targets the situations that such tail probability is practically small.

## 2. Normal copula model and its identifiability

Let $T$ be the survival time and let $U$ be the mark of interest. Suppose that they have a continuous joint distribution. The normal copula model postulates that

$$\left\{ \begin{array}{c} H_T(T) \\ H_U(U) \end{array} \right\} \sim \mathrm{BN}(\rho) \tag{2·1}$$

for unspecified increasing transformations $H_T(.)$ and $H_U(.)$, where $\mathrm{BN}(\rho)$ is the standard bivariate normal distribution with correlation coefficient $\rho$. Define marginal distribution functions $F_T(t) := \mathrm{pr}(T \leq t)$ and $F_U(u) := \mathrm{pr}(U \leq u)$. As implied from (2·1),

$$H_T(.) = \Phi^{-1}\{F_T(.)\}, \qquad H_U(.) = \Phi^{-1}\{F_U(.)\}, \tag{2·2}$$

where $\Phi(.)$ is the cumulative distribution function of the standard univariate normal distribution. This indicates that the marginal distributions of $T$ and $U$ are completely unspecified. Their association is parameterised by $\rho$.

Denote the censoring time by $C$. We work with the classical random censorship mechanism,

$$C \perp (T, U), \tag{2·3}$$

where $\perp$ represents independence. As a result of censoring, the set of underlying random variables $\{T, U, C\}$ are observed through

$$X := T \wedge C, \qquad Y := U\, I(T \leq C), \qquad \Delta := I(T \leq C),$$

where $\wedge$ is the minimisation operator and $I(.)$ is the indicator function. Originally identified in Huang & Louis (1998), this data structure is common to various data-collection schemes involving a mark of interest.

Denote the maximum support of $X$ by $\tau$.

THEOREM 1. *Under the normal copula model (2·1) and random censorship (2·3), $\rho$, $F_T(.)$ on $[0, \tau]$ and $F_U(.)$ on*

$$\begin{cases} (-\infty, \infty) & \text{if } \rho \in (-1, 1) \\ [F_U^{-1}\{1 - F_T(\tau)\}, \infty) & \text{if } \rho = -1 \\ (-\infty, F_U^{-1}\{F_T(\tau)\}] & \text{if } \rho = 1 \end{cases}$$

*are identifiable from the distribution of $\{X, Y, \Delta\}$.*

Thus, the marginal distribution of $U$ becomes identifiable over the whole line under the proposed semiparametric model, unless $T$ and $U$ are perfectly correlated. From now on, we restrict our attention to situations with $\rho \in (-1, 1)$; the cases of $\rho = \pm 1$ are of little practical interest.

Bivariate copula modelling has been previously employed for the inference of association between random variables, where both marginal distributions can be readily estimated in a nonparametric fashion; see Klaassen & Wellner (1997) as well as Shih & Louis (1995) and Genest et al. (1995). However, our current application is different.

## 3. PROPOSED INFERENCE PROCEDURE

### 3·1 *Estimation of $\rho$*

Suppose that the sample consists of $\{X_i, Y_i, \Delta_i\}$, $i = 1, \ldots, n$, as $n$ independent replicates of $\{X, Y, \Delta\}$. Among the three components in the normal copula model (2·1), $F_U(.)$ is of interest

whereas $F_T(.)$ and $\rho$ may be treated as nuisances. Nevertheless, the estimation of $F_U(.)$ requires that of $F_T(.)$ and $\rho$. The standard Kaplan–Meier approach applies to $F_T(.)$, and we focus now on $\rho$.

By normal distribution theory, model (2·1) can be equivalently specified as

$$H_U(U) \,\big|\, H_T(T) \;\sim\; N\{\rho H_T(T), 1-\rho^2\}, \tag{3·1}$$

with $H_T(.)$ and $H_U(.)$ given in (2·2). This leads to the following linear transformation model with standard normal error $\varepsilon$:

$$(1-\rho^2)^{-1/2} H_U(U) \,\big|\, H_T(T) \;\sim\; \theta H_T(T) + \varepsilon,$$

where $\theta = \rho(1-\rho^2)^{-1/2}$; that is, upon an unspecified increasing transformation, $U$ is linearly related to $H_T(T)$ with a standard normal error. Write $Z = H_T(X)$. The random censorship (2·3) then implies that

$$(1-\rho^2)^{-1/2} H_U(Y) \,\big|\, Z, \Delta = 1 \;\sim\; \theta Z + \varepsilon. \tag{3·2}$$

Following the idea of Cheng et al. (1995) for linear transformation models, we invoke pairwise comparisons of $Y_i$, $i = 1, \ldots, n$, to obtain the following identity that is free of the transformation $H_U(.)$:

$$\mathrm{pr}\left(Y_i \geq Y_j \,\big|\, Z_i, \Delta_i = 1, Z_j, \Delta_j = 1\right) = \Phi_2(\theta Z_{ij}) \qquad \text{for } i \neq j, \tag{3·3}$$

where $Z_{ij} = Z_i - Z_j$ and $\Phi_2(.)$ is the cumulative distribution function of a normal distribution with mean 0 and variance 2, i.e. $\Phi_2(z) = \Phi(2^{-1/2}z)$. This is a probit model for $I(Y_i \geq Y_j)$, but with the observations being dependent.

Let $\widehat{F}_T(.)$ be the Kaplan-Meier estimator of $F_T(.)$ based on data $\{X_i, \Delta_i\}$, $i = 1, \ldots, n$. Given (2·2), we estimate $H_T(.)$ by $\widehat{H}_T(.) = \Phi^{-1}\{\widehat{F}_T(.)\}$. To avoid technical difficulties arising from the unboundedness of $\Phi^{-1}(.)$ at 0 and 1 and from the tail instability of $\widehat{F}_T(.)$, we only consider individuals with $X_i \in [a, b]$ such that $0 < a < b < \tau$ for constants $a$ and $b$. Write $\Delta^\circ = \Delta\, I(X \in [a, b])$, $\widehat{Z} = \widehat{H}_T(X)$ and $\widehat{Z}_{ij} = \widehat{Z}_i - \widehat{Z}_j$. Identity (3·3) suggests the following estimating function for $\theta$:

$$\widehat{\Psi}(\vartheta) = n^{-2} \sum_{i,j=1}^{n} \Delta_i^\circ \Delta_j^\circ w(\vartheta \widehat{Z}_{ij}) \widehat{Z}_{ij} \{I(Y_i \geq Y_j) - \Phi_2(\vartheta \widehat{Z}_{ij})\}, \tag{3·4}$$

where $w(.)$ is a positive weight function. Following Cheng et al. (1995), one may adopt $w(.) = 1$ and $w(.) = \Phi_2'(.)/[\Phi_2(.)\{1 - \Phi_2(.)\}]$ to mimic ordinary linear regression and the quasilikelihood approach, respectively, for the probit model with independent observations. Write $\widehat{\theta}$ as any zero-crossing of $\widehat{\Psi}(\vartheta)$. Since $\rho = \theta(1+\theta^2)^{-1/2}$, an estimator of $\rho$ is obtained as $\widehat{\rho} = \widehat{\theta}(1+\widehat{\theta}^2)^{-1/2}$.

THEOREM 2. *Suppose that normal copula model (2·1) holds with $\rho \in (-1, 1)$. Under random censorship (2·3), $\widehat{\rho}$ converges to $\rho$ almost surely, and $n^{1/2}(\widehat{\rho} - \rho)$ is asymptotically normal with mean 0.*

Interval estimation of $\rho$ will be investigated along with that of $F_U(.)$, which is the focus of interest.

### 3·2 *Estimation of $F_U(.)$*

Identity (3·2) implies that,

$$\text{given } \Delta^\circ = 1, \quad H_U(Y) \;\sim\; \rho Z + (1-\rho^2)^{1/2}\varepsilon.$$

Thus,

$$G(v) := \text{pr}\{H_U(Y) \le v \,|\, \Delta^\circ = 1\} = E\left[\Phi\left\{(1-\rho^2)^{-1/2}(v - \rho Z)\right\} \,\middle|\, \Delta^\circ = 1\right],$$

which is naturally estimated by

$$\widehat{G}(v) := m^{-1}\sum_{i=1}^{n} \Delta_i^\circ \Phi\left\{(1-\widehat{\rho}^2)^{-1/2}(v - \widehat{\rho}\widehat{Z}_i)\right\}$$

with $m = \sum_{i=1}^n \Delta_i^\circ$. Write $F_{Y\,|\,\Delta^\circ=1}(u) = \text{pr}(Y \le u \,|\, \Delta^\circ = 1)$ and denote its empirical counterpart by $\widehat{F}_{Y\,|\,\Delta^\circ=1}(u)$. Note that

$$H_U(.) = G^{-1}\{F_{Y\,|\,\Delta^\circ=1}(.)\}.$$

Thus, we obtain an estimator of $H_U(.)$ in the form

$$\widehat{H}_U(.) = \widehat{G}^{-1}\{\widehat{F}_{Y\,|\,\Delta^\circ=1}(.)\}.$$

Furthermore, (2·2) suggests an estimator of $F_U(.)$ in the form

$$\widehat{F}_U(.) = \Phi\left[\widehat{G}^{-1}\{\widehat{F}_{Y\,|\,\Delta^\circ=1}(.)\}\right].$$

More explicitly, the estimated marginal distribution $\widehat{F}_U(.)$ assigns probability mass to, and only to, those $Y$ values in the subsample $\{i:\ \Delta_i^\circ = 1,\ i = 1, \ldots n\}$. Order them as $Y_{(1)} \leq \cdots \leq Y_{(m)}$. For $Y_{(k)}$, the assigned probability mass is

$$\widehat{p}_{(k)} = \begin{cases} \Phi\left\{\widehat{G}^{-1}\left(\frac{1}{m}\right)\right\} & k = 1 \\ \Phi\left\{\widehat{G}^{-1}\left(\frac{k}{m}\right)\right\} - \Phi\left\{\widehat{G}^{-1}\left(\frac{k-1}{m}\right)\right\} & k = 2, \ldots, m-1 \\ 1 - \Phi\left\{\widehat{G}^{-1}\left(\frac{m-1}{m}\right)\right\} & k = m \end{cases} .$$

THEOREM 3. *Suppose that normal copula model (2·1) holds with $\rho \in (-1, 1)$. Under random censorship (2·3), $\widehat{F}_U(.)$ is uniformly consistent for $F_U(.)$ almost surely. Furthermore, for any constants $c$ and $d$ such that $0 < F_U(c) < F_U(d) < 1$, $n^{1/2}\{\widehat{F}_U(.) - F_U(.)\}$ on $[c, d]$ converges weakly to a zero-mean Gaussian process.*

Estimator $\widehat{F}_U(.)$ provides a basis for estimating summaries of the mark $U$. Of particular interest are the mean $\mu := E(U)$ and the median $\nu := \inf\{u : F_U(u) \geq 0.5\}$. They are naturally estimated by

$$\widehat{\mu} := \int u \, d\widehat{F}_U(u) = \sum_{k=1}^{m} \widehat{p}_{(k)} Y_{(k)}$$

and

$$\widehat{\nu} := \inf\{u : \widehat{F}_U(u) \geq 0.5\}.$$

COROLLARY 1. *Suppose that the conditions in Theorem 3 hold. If $U$ is bounded, $\widehat{\mu}$ converges almost surely to $\mu$.*

COROLLARY 2. *Under the conditions in Theorem 3, $\widehat{\nu}$ converges almost surely to $\nu$, and $n^{1/2}(\widehat{\nu} - \nu)$ is asymptotically normal with mean 0.*

To complete the inference procedure, interval estimation for $\rho$, $F_U$ on $[c, d]$ and $\nu$ remains to be addressed. One approach is to derive the influence curves of their maps from the distribution of $\{X, Y, \Delta\}$; see the definition of these maps in the Appendix. However, the derivation is algebraically

complex. An alternative is empirical or nonparametric bootstrap, which can be justified by the Hadamard-differentiability of these maps as established in the Appendix (Gill, 1989; van der Vaart & Wellner, 1996, § 3.9).

## 4. Model checking

From Huang & Louis (1998), the continuous joint distribution of $\{T, U\}$ is nonparametrically identifiable on the support $[0, \tau] \times (-\infty, \infty)$ only, under censorship (2·3). Therefore, model checking is inevitably limited to the validity of the normal copula on $[0, F_T(\tau)] \times [0, 1]$, which would be satisfactory if $F_T(\tau)$ is sufficiently large. In this section, we devise graphical model-checking methods and propose Kolmogorov–Smirnov-type goodness-of-fit tests.

Recall that the normal copula model (2·2) implies the linear transformation model (3·2), under random censorship (2·3). Graphical methods may be developed to assess the linearity of $H_U(Y)$ versus $Z$ and the normality of $H_U(Y) - \rho Z$, conditioning on $\Delta = 1$. Upon the estimation of $F_T(.)$, $\rho$ and $F_U(.)$ as described in § 3, one may plot $\Phi^{-1}\{\widehat{F}_U(Y)\}$ versus $\widehat{Z}$ among uncensored individuals, i.e. with $\Delta = 1$. A deviation of the straight line through $(0, 0)$ of slope $\widehat{\rho}$ would suggest a lack of fit. In addition, one may construct a normal Q-Q plot for the residuals $\Phi^{-1}\{\widehat{F}_U(Y)\} - \widehat{\rho}\widehat{Z}$ of uncensored individuals. If the normal copula model holds, the Q-Q plot should be approximately a straight line through $(0, 0)$ with slope $(1 - \widehat{\rho}^2)^{1/2}$.

Huang & Louis (1998) derived the nonparametric maximum likelihood estimator for the joint distribution $F_{TU}(t, u) := \mathrm{pr}(T \leq t, U \leq u)$. This Huang–Louis estimator, denoted by $\widetilde{F}_{TU}(t, u)$, is uniformly consistent on $[0, \tau] \times (-\infty, \infty)$, regardless of whether the normal copula model holds or not. Meanwhile, our proposal provides a natural estimator under the normal copula model, given by

$$\widehat{F}_{TU}(t, u) := \Omega\left\{\widehat{F}_T(t), \widehat{F}_U(u); \widehat{\rho}\right\} \qquad t \leq \tau, \tag{4·1}$$

where $\Omega(., .; \varrho)$ is the bivariate normal copula function with correlation coefficient $\varrho$. Then, a goodness-of-fit test may be based on the Kolmogorov–Smirnov statistic

$$\xi_{\mathcal{B}} := \sup_{(t,u)\in\mathcal{B}} \left|\widetilde{F}_{TU}(t, u) - \widehat{F}_{TU}(t, u)\right|$$

for a selected region $\mathcal{B}$. One natural choice for $\mathcal{B}$ is $[0,\tau]\times(-\infty,\infty)$.

THEOREM 4. *Suppose that random censorship (2·3) holds. Under the normal copula model (2·1) with $\rho\in(-1,1)$, $\widehat{F}_{TU}(.,.)$ is uniformly consistent for $F_{TU}(.,.)$ on $[0,\tau]\times(-\infty,\infty)$ almost surely. Thus, the goodness-of-fit test based on $\xi_{[0,\tau]\times(-\infty,\infty)}$ is consistent against any alternative copula that differs from the normal copula family on $[0,F_T(\tau)]\times[0,1]$. Furthermore, for constants $b$, $c$ and $d$ such that $0<b<\tau$ and $0<F_U(c)<F_U(d)<1$, $n^{1/2}\{\widehat{F}_{TU}(.,.)-\widetilde{F}_{TU}(.,.)\}$ converges weakly to a zero-mean Gaussian process on $[0,b]\times[c,d]$ when the normal copula model (2·1) holds with $\rho\in(-1,1)$.*

For critical values of the test, the distribution of $\xi_{\mathcal{B}}$ needs to be determined under the normal copula model. Apparently, this distribution is governed by $F_T(.)$, $F_U(.)$, $\rho$ and $F_C(.)$, where $F_C(t):=\mathrm{pr}(C\le t)$ is the distribution function of $C$. If these quantities are given, the distribution function of $\xi_{\mathcal{B}}$ can be calculated by the Monte Carlo method. Of course, none of them is known. Nevertheless, we suggest employing the semiparametric bootstrap approach, i.e. using their consistent estimators instead; $F_C(.)$ can be estimated by the Kaplan–Meier approach. Note that this bootstrap is different from the nonparametric one in § 3.2, where the bootstrap samples are generated without the normal copula assumption. If $\mathcal{B}\subseteq[0,b]\times[c,d]$, this semiparametric bootstrap for $\xi_{\mathcal{B}}$ may be justified by the functional delta method, the same argument as used for the nonparametric bootstrap. In the case of $\xi_{[0,\tau]\times(-\infty,\infty)}$ which is invariant to increasing transformations of the mark scale, the computation can be simplified since $F_U(.)$ or its estimator would not be needed. However, since $[0,\tau]\times(-\infty,\infty)\not\subseteq[0,b]\times[c,d]$, there may be technical difficulties with the above asymptotic argument for the semiparametric bootstrap. Nevertheless, our simulation results reported in § 5 provide empirical justification.

## 5. NUMERICAL STUDIES

### 5·1 *Preamble*

Simulations were conducted to evaluate the proposal with small and moderate samples. We studied the estimation procedure both under the normal copula model and under nonnormal cop-

ulas, i.e. when the normal copula model is misspecified. In addition, the goodness-of-fit tests were investigated. Finally, we applied this proposal to the estimation of lifetime medical cost with data from a lung cancer clinical trial.

Instead of using $\widehat{Z} = \Phi^{-1}\{\widehat{F}_T(X)\}$, we used $\widehat{Z} = \Phi^{-1}[\{\widehat{F}_T(X-) + \widehat{F}_T(X)\}/2]$ in our numerical studies. The resulting estimator is asymptotically equivalent, but had slightly better small-sample performance. Additionally, for each dataset, we chose sufficiently small $a$ and sufficiently large $b$ such that $\Delta_i^\circ = \Delta_i$ for all $i = 1, \ldots, n$.

In our studies, the use of weight functions $w(.) = 1$ and $w(.) = \Phi_2'(.)/[\Phi_2(.)\{1 - \Phi_2(.)\}]$ in the estimating function (3·4) resulted in virtually the same estimates of $\rho$ and $F_U(.)$. Only results with the former weight are presented.

### 5·2 *Performance under the normal copula model*

A number of settings were investigated under moderate censoring. The mark and survival time were considered at various association levels with $\rho = 0.8$, 0.4, 0, $-0.4$ and $-0.8$. With marginal distributions $F_U(.)$ and $F_T(.)$ given, they were generated by marginally transforming the standard bivariate normal distribution; that is, $U = F_U^{-1}\{\Phi(V)\}$ and $T = F_T^{-1}\{\Phi(S)\}$ where $(V, S) \sim \mathrm{BN}(\rho)$. The marginal survival distribution was set to be exponential with unit rate. The censoring time followed the exponential distribution with rate 0.2 but curtailed at 2. As such, any event with $T > 2$, corresponding to the top 13.5% in the survival time, was surely censored. The overall censoring rate was about 25%. Three marginal mark distributions were studied, namely uniform on $[0, 12^{1/2}]$, exponential with unit rate and standard lognormal scaled by $(e^2 - e)^{-1/2}$. Their respective means are $3^{1/2}$, 1 and $(e - 1)^{-1/2}$. With the same unit variance, these distributions are in ascending order of long tails; a long-tailed distribution would be expected in the case of medical cost.

For comparison, we investigated two naive approaches to the marginal mark estimation. One treats the observed marks as a random sample from the marginal mark distribution, resulting in the complete-case estimator. The other applies the Kaplan–Meier approach to the mark scale, in the case that the mark is a cumulative measure over time. Let $U(.)$ be the accumulation process.

Unlike the proposed and complete-case estimators, this naive Kaplan–Meier estimator typically depends on the accumulation process by using data $\{U_i(X_i), \Delta_i\}$, $i = 1, \ldots, n$. For this purpose, we set $U(t) = F_U^{-1}\{F_U(U)(t \wedge T)/T\}$; that is, on the transformed scale $F_U(.)$, the accumulation rate of the mark is constant over time for each individual. Note that the naive Kaplan–Meier estimator of $1 - F_U(.)$ may not reach 0 eventually. In this case, its integral up to the largest uncensored $Y_i$ was used as the estimated mean of $U$.

Table 1 presents the simulation results for the estimation of $\rho$ and marginal mark distribution with sample size 50; results for a sample size of 100, not shown, displayed similar features. Each scenario was simulated with 1,000 replications. The 95% nonparametric bootstrap percentile confidence interval was constructed with a bootstrap size of 39. This size is adequate for the assessment of coverage probability in these simulation studies, although it needs to be much larger in constructing a reliable confidence interval for a specific dataset. As shown, the proposed estimator for $\rho$ is nearly unbiased and the coverage probability of its nonparametric bootstrap percentile confidence interval is fairly accurate. This seems remarkable with sample size as small as 50. Of more interest is the marginal mark distribution. Note that both the proposed and the complete-case estimators of $F_U(.)$ at a percentile are invariant to the marginal mark distribution. The naive Kaplan–Meier estimator is so only with special accumulation processes including the one specified. For $F_U(.)$ at the 25th, 50th and 75th percentiles, the performance of the proposed estimator is again remarkable. The proposed mean estimator has small bias and good coverage probability of its confidence interval when the marginal mark distribution is uniform. However, its performance deteriorates with a long-tailed marginal distribution such as lognormal especially in the case of a strong positive association between the mark and survival time. This might be expected because the mean is sensitive to a long tail, which is largely censored under a strong positive association. The proposed median estimator is presented only for the lognormal marginal mark distribution, with similar and satisfactory performance observed under the other two distributions. In contrast, the two naive estimators are generally biased except for the special circumstance of $\rho = 0$, when the complete-case estimator for the marginal distribution is consistent and efficient. In this case,

the efficiency of the proposed estimator appears to be reasonable.

We also evaluated the semiparametric estimator $\widehat{F}_{TU}(t,u)$ for the joint distribution as given in (4·1), along with the nonparametric Huang–Louis estimator $\widetilde{F}_{TU}(t,u)$. Table 2 summarises the results with sample size 50 given at pairs of $(t,u)$, where $t$ takes the 20th, 40th, 60th and 80th percentiles of $F_T(.)$, and $u$ takes the 25th, 50th and 75th percentiles of $F_U(.)$. Note that these results are invariant to the marginal distribution of $U$. Given $\tau = 2$, where the censoring time is curtailed, $F_{TU}(t,u)$ is nonparametrically identifiable for all these points under consideration. Both the proposed and the Huang–Louis estimator have negligible bias, but the former is more efficient, as expected.

### 5·3 *Behaviour under nonnormal copulas*

In the absence of censoring, the normal copula model turns out to play little role in the estimation; as can be verified, $\widehat{F}_U(.)$ is asymptotically equivalent to the empirical marginal mark distribution in this case. However, the presence of censoring is the factor of interest here.

Clayton's and Frank's families (Shih & Louis, 1995) were considered. Copulas in these two families generally differ from normal copulas except when $T$ and $U$ are independent. We adopted the same simulation scenarios as described in § 5.2 except for the copula. Like the normal copula, both of these two nonnormal copula families are governed by a single parameter. In addition, the parameter has a one-to-one mapping with the association measure Kendall's tau within each family. For comparison purposes, corresponding to each nonzero $\rho$ value considered in § 5.2, we chose a nonnormal copula with the same Kendall's tau. However, only positive association was considered for the Clayton's copula, given its limitation in accommodating negative association.

The results are presented in Table 3, with sample size 100. In the case of Clayton's family, moderate bias is observed with the marginal mark distribution estimator at the tail. The bias becomes more serious for the mean estimator, especially when the underlying distribution has a longer tail. In contrast, the performance of the median estimator is reasonable. On the other hand, the proposed inference procedure appears to be quite robust with Frank's family. These results suggest that model checking might be important in practice.

### 5·4 *Characteristics of the goodness-of-fit tests*

We investigated the Kolmogorov–Smirnov-type goodness-of-fit test based on $\xi_{[0,\tau]\times(-\infty,\infty)}$ with both normal and nonnormal copulas under the same scenarios as in §§ 5.2 and 5.3. Sample sizes of 50, 100 and 200 and nominal test levels at 0.10, 0.05 and 0.01 were considered. The results are presented in Table 4 based on 1,000 replications. The critical values were calculated using the semiparametric bootstrap with size 99. For the same reason as for the nonparametric bootstrap, this size is adequate here for assessing the significant levels and powers of the tests. Overall, the achieved significance levels are reasonably accurate, when the normal copula model holds. On the other hand, when the underlying copula belongs to Clayton's or Frank's family, the power of the test increases with sample size although it can be limited with small samples.

### 5·5 *Application to a lung cancer study*

Our investigation was largely motivated by the analysis of lifetime medical cost in a randomised trial conducted by the Southwest Oncology Group. This trial compared paclitaxel plus carboplatin versus vinorelbine plus cisplatin therapies in earlier untreated patients with advanced non-small cell lung cancer (Kelly et al., 2001). Study outcomes included survival time, as the primary endpoint, and utilised resources which consisted of supportive care medications, blood products, medical procedures, protocol and non-protocol related treatments, and medical care inpatient days or outpatient visits. Cost was assigned to each resource using national databases and was adjusted to 1998 U.S. dollars according to the medical care component of the Consumer Price Index. Previous economic analyses focused on the study-duration-restricted medical cost (Ramsey et al., 2002) or the joint distribution of lifetime medical cost and survival time (Huang & Lovato, 2002; Huang, 2002). None of these methods was capable of addressing the marginal distribution of lifetime medical cost directly for each therapy. The current proposal is an effort to fill the gap. As an illustration, we present its application to those participants randomised to the paclitaxel plus carboplatin therapy.

The cost data were collected every 3 months for the first 6 months, and every 6 months thereafter, up to 24 months. After exclusion of those with insufficient documentation or no follow-up

for cost data collection, 183 out of the 206 participants randomised to paclitaxel plus carboplatin remained in the current analysis. Approximately 30% of these 183 participants were censored, and the survival rate at 24 months was 19% as estimated by the Kaplan–Meier approach.

We applied the proposed procedure to this dataset. For interval estimation, 95% nonparametric bootstrap percentile confidence intervals were constructed with a size of 999. The estimated $\rho$ is 0.783, with confidence interval (0.695, 0.847), showing a fairly strong association between lifetime medical cost and survival time. Figure 1 displays the estimated marginal survival function of lifetime medical cost. The estimated mean lifetime medical cost is \$52,696, with confidence interval (47,557, 58,088), whereas the estimated median is \$46,207, with confidence interval (41,983, 54,019). The difference between the mean and median indicates heavy skewness of the distribution. These results contrast the estimated mean two-year-restricted medical cost of \$48,940 given in Ramsey et al. (2002).

For model checking, we employed the methods developed in § 4. Figure 2 shows the scatterplot of $\Phi^{-1}[\{\widehat{F}_U(Y-)+\widehat{F}_U(Y)\}/2]$ versus $\widehat{Z}$ and the normal Q-Q plot of $\Phi^{-1}[\{\widehat{F}_U(Y-)+\widehat{F}_U(Y)\}/2]-\widehat{\rho}\widehat{Z}$, among uncensored observations. These plots suggest a good fit. The Kolmogorov–Smirnov statistic $\xi_{[0,\tau]\times(-\infty,\infty)}$ is 0.0356, corresponding to a $p$-value of 0.251 as obtained by the semiparametric bootstrap with a size of 999.

## 6. Final remarks

Although the focus of this article is on the normal copula model, the marginal identifiability result may be extended to any copula family so long as its members can be discriminated from each other on the support $[0, F_T(\tau)] \times [0, 1]$. Estimation with families such as Clayton's and Frank's is under current investigation, so are selection strategies of copula models.

The marginal identifiability of the mark is achieved through parameterisation of the copula function. It is important to recognise that certain elements of the postulated copula are not testable. Indeed, at best model-checking is limited to the fit of the copula model on the partial support $[0, F_T(\tau)] \times [0, 1]$, rather than the whole support $[0, 1]^2$. Two copulas that are common

on the support $[0, F_T(\tau)] \times [0, 1]$ cannot be discriminated from each other with the data available. Given that statistical analysis should be mainly driven by data, we would recommend this proposal only when $F_T(\tau)$ is sufficiently large that the model assumption can be reasonably checked. If $F_T(\tau)$ is small, any approaches to the marginal mark estimation would be inevitably model-driven to a high degree.

The proposed inference procedure has a minimal data requirement. In practice, additional information could be available. For example, in the case of lifetime medical cost, one might also observe accumulated cost at censoring time for each censored individual, or the cost accumulation process at discrete time points as in the study discussed in § 5.5. In that case, more efficient estimation might be possible. However, a general approach to taking advantage of such additional information is not likely to be available and developments might need to be tailored to each specific situation.


## Acknowledgement

This research was partially supported by funds from the U.S. National Institutes of Health. The authors thank Dr. John Crowley and Ms. Laura Lovato for providing the lung cancer dataset. They also gratefully acknowledge that helpful comments from the reviewers have led to improvement in the presentation.


## Appendix

### *Proofs*

*Proof of Theorem 1.* Huang & Louis (1998) showed that, under censorship (2·3), the continuous joint distribution of $\{T, U\}$ is identifiable up to the support $[0, \tau] \times (-\infty, \infty)$. The identifiability of $F_T(.)$ on $[0, \tau]$ then follows. Given relationships (2·2) and in the light of identity (3·1), it becomes clear that $F_U(.)$ would be identifiable on the support specified in the theorem if $\rho$ is known. Note that, when $\rho = \pm 1$, $U$ and $T$ are perfectly correlated and the support on which $F_U(.)$ is identifiable is mapped from $[0, \tau]$ where $F_T(.)$ is identifiable. The remaining task is to show the identifiability of $\rho$.

Similarly to the development in § 3.1, we can show that

$$\mathrm{pr}(U_1 \geq U_2 \,|\, T_1, T_2) = \begin{cases} \Phi\left[2^{-1/2}\rho(1-\rho^2)^{-1/2}\{H_T(T_1) - H_T(T_2)\}\right] & \text{if } \rho \in (-1,1), \\ I(\rho T_1 \geq \rho T_2) & \text{if } \rho = \pm 1, \end{cases}$$

where $(T_1, U_1)$ and $(T_2, U_2)$ are independent replicates of $(T, U)$. The identifiability of $\rho$ then follows. □

*Proof of Theorem 2.* We first show that $\widehat{\rho}$ is a plug-in estimator in a map from the distribution of $\{X, Y, \Delta\}$ to $\rho$; that is, $\widehat{\rho}$ can be obtained by substituting the empirical distribution of $\{X, Y, \Delta\}$ into the aforementioned map. We then use the functional delta method (Gill, 1989; van der Vaart & Wellner, 1996, § 3.9) to establish the properties of $\widehat{\rho}$.

Define $F_{XY\Delta}(t, u, \delta) := \mathrm{pr}(X \leq t, Y \leq u, \Delta \leq \delta)$, $F_{X\Delta}(t, \delta) := F_{XY\Delta}(t, \infty, \delta)$ and $F_{XY,\Delta^\circ=1}(t, u) := \mathrm{pr}(X \leq t, Y \leq u, \Delta^\circ = 1)$. Write their respective empirical counterparts as $\widehat{F}_{XY\Delta}$, $\widehat{F}_{X\Delta}$ and $\widehat{F}_{XY,\Delta^\circ=1}$. In addition, let

$$\begin{aligned} F_{ZY,\Delta^\circ=1}(z, u) &:= \mathrm{pr}(Z \leq z, Y \leq u, \Delta^\circ = 1), \\ \widehat{F}_{ZY,\Delta^\circ=1}(z, u) &:= n^{-1} \sum_{i=1}^{n} I(\widehat{Z}_i \leq z, Y_i \leq u, \Delta_i^\circ = 1). \end{aligned}$$

Thus,

$$\begin{aligned} F_{ZY,\Delta^\circ=1}(z, u) &= F_{XY,\Delta^\circ=1}\{H_T^{-1}(z), u\}, \\ \widehat{F}_{ZY,\Delta^\circ=1}(z, u) &= \widehat{F}_{XY,\Delta^\circ=1}\{\widehat{H}_T^{-1}(z), u\}. \end{aligned}$$

Note that

$$\widehat{\Psi}(\vartheta) = \iint w(\vartheta z_{12}) z_{12} \{I(u_1 \geq u_2) - \Phi_2(\vartheta z_{12})\} \, d\widehat{F}_{ZY,\Delta^\circ=1}(z_1, u_1) \, d\widehat{F}_{ZY,\Delta^\circ=1}(z_2, u_2),$$

where $z_{12} = z_1 - z_2$. Obviously, its theoretical counterpart is

$$\begin{aligned} \Psi(\vartheta) &:= \iint w(\vartheta z_{12}) z_{12} \{I(u_1 \geq u_2) - \Phi_2(\vartheta z_{12})\} \, dF_{ZY,\Delta^\circ=1}(z_1, u_1) \, dF_{ZY,\Delta^\circ=1}(z_2, u_2) \quad \text{(A·1)} \\ &= E[\Delta_1^\circ \Delta_2^\circ w(\vartheta Z_{12}) Z_{12} \{I(Y_1 \geq Y_2) - \Phi_2(\vartheta Z_{12})\}]. \end{aligned}$$

Identity (3·3) implies that

$$\Psi(\vartheta) = E[\Delta_1^\circ \Delta_2^\circ w(\vartheta Z_{12}) Z_{12} \{\Phi_2(\theta Z_{12}) - \Phi_2(\vartheta Z_{12})\}],$$

which clearly has a zero-crossing at $\theta$. Since $w(.)$ is positive and $\Phi_2(.)$ is monotone, $(\theta - \vartheta)\Psi(\vartheta)$ is 0 only if $\vartheta = \theta$. Therefore, the zero-crossing is unique, and so we can write $\theta = \Psi^{-1}(0)$. It now becomes clear that $\widehat{\rho}$ is a plug-in estimator in the map $F_{XY\Delta} \mapsto \rho$, which is decomposed as

$$F_{XY\Delta} \mapsto \left\{ \begin{array}{l} F_{X\Delta} \mapsto F_T \text{ on } [0,b] \mapsto H_T \text{ on } [a,b] \mapsto H_T^{-1} \text{ on } [H_T(a), H_T(b)] \\ F_{XY,\Delta^\circ=1} \end{array} \right\} \mapsto F_{ZY,\Delta^\circ=1} \mapsto \Psi \mapsto \theta \mapsto \rho.$$

Next, we show that the map $F_{XY\Delta} \mapsto \rho$ is Hadamard-differentiable, working with appropriate spaces of univariate and bivariate cadlag functions endowed with supnorm; see Neuhaus (1971) and van der Vaart & Wellner (1996, § 3.9). The Hadamard-differentiability of the maps $F_{XY\Delta} \mapsto \{F_{X\Delta}, F_{XY,\Delta^\circ=1}\}$ and $\theta \mapsto \rho$ is obvious, and that of the maps $F_{X\Delta} \mapsto F_T$ on $[0,b]$, $F_T$ on $[0,b] \mapsto H_T$ on $[a,b]$, $H_T$ on $[a,b] \mapsto H_T^{-1}$ on $[H_T(a), H_T(b)]$, $\{H_T^{-1}$ on $[H_T(a), H_T(b)], F_{XY,\Delta^\circ=1}\} \mapsto F_{ZY,\Delta^\circ=1}$ and $\Psi \mapsto \theta$ directly follows the results in van der Vaart & Wellner (1996, § 3.9). The remaining decomposed map is $F_{ZY,\Delta^\circ=1} \mapsto \Psi$. From (A·1), this map is further decomposed into the inner and outer integrations. The former is linear and continuous, and hence Hadamard-differentiable, whereas Gill et al. (1995, Lemma 5.1) asserted the Hadamard-differentiability for the latter. The chain rule now yields the Hadamard-differentiability of $F_{XY\Delta} \mapsto \rho$.

The continuity of $F_{XY\Delta} \mapsto \rho$ is implied by the Hadamard-differentiability. Since $\widehat{F}_{XY\Delta}$ is strongly consistent for $F_{XY\Delta}$ by the Glivenko–Cantelli theorem, $\widehat{\rho}$ is strongly consistent for $\rho$. By the functional delta method, the asymptotic normality of $\widehat{\rho}$ follows that of $\widehat{F}_{XY\Delta}$. □

*Proof of Theorem 3.* Write $F_{X \,|\, \Delta^\circ=1}(t) := \mathrm{pr}(X \le t \,\big|\, \Delta^\circ = 1)$ and $F_{Z \,|\, \Delta^\circ=1}(z) := \mathrm{pr}(Z \le z \,\big|\, \Delta^\circ = 1) = F_{X \,|\, \Delta^\circ=1}\{H_T^{-1}(z)\}$. We decompose the map $F_{XY\Delta} \mapsto F_U$ on $[c,d]$ as follows:

$$F_{XY\Delta} \mapsto \left\{ \begin{array}{l} \rho \\ H_T^{-1} \text{ on } [H_T(a), H_T(b)] \\ F_{X \,|\, \Delta^\circ=1} \\ F_{Y \,|\, \Delta^\circ=1} \end{array} \right\} \mapsto F_{Z \,|\, \Delta^\circ=1} \Bigg\} \mapsto G \Bigg\} \mapsto H_U \text{ on } [c,d] \mapsto F_U \text{ on } [c,d].$$

It can be verified that $\widehat{F}_U$ is the plug-in estimator in the above map. We use the same approach as in the proof of Theorem 2 to establish the strong consistency and asymptotic normality of $\widehat{F}_U$ on $[c,d]$.

We now show the strong consistency of $\widehat{F}_U$ with the support extended to the whole line. Note that $\widehat{F}_U(.)$ is a proper cumulative distribution function with monotonicity, $\widehat{F}_U(-\infty) = 0$ and $\widehat{F}_U(\infty) = 1$. We then have

$$\sup_{u\in(-\infty,c)} \left|\widehat{F}_U(u) - F_U(u)\right| \leq \widehat{F}_U(c) + F_U(c) \leq 2F_U(c) + \left|\widehat{F}_U(c) - F_U(c)\right|.$$

Similarly,

$$\sup_{u\in(d,\infty)} \left|\widehat{F}_U(u) - F_U(u)\right| \leq 2\{1 - F_U(d)\} + \left|\widehat{F}_U(d) - F_U(d)\right|.$$

Given the strong consistency of $\widehat{F}_U$ on $[c, d]$ for any fixed $c$ and $d$ such that $0 < F_U(c) < F_U(d) < 1$, almost surely the right-hand sides of the above expressions can be made arbitrarily small by appropriate choice of $c$ and $d$. The strong consistency of $\widehat{F}_U$ on the whole line follows. □

*Proof of Corollary 1.* This is a direct result of Theorem 3. □

*Proof of Corollary 2.* The assertion follows Theorem 3 and the Hadamard-differentiability of the map $F_U \mapsto \nu$ (van der Vaart & Wellner, 1996, Lemma 3.9.20). □

*Proof of Theorem 4.* The normal copula function $\Omega(s, v; \varrho)$ is symmetric between $s$ and $v$, and it is continuous in $s$, $v$ and $\varrho$. We note that $\partial\Omega(s, v; \varrho)/\partial s$ and $\partial\Omega(s, v; \varrho)/\partial\varrho$ are uniformly continuous and bounded on $[0, 1] \times [0, 1] \times \{\rho\}$. Thus, the map $\{F_T, F_U, \rho\} \mapsto F_{TU}$ defined as $F_{TU}(t, u) = \Omega\{F_T(t), F_U(u); \rho\}$ is continuous and Hadamard-differentiable (van der Vaart & Wellner, 1996, § 3.9.4.3).

The consistency of $\widehat{F}_{TU}(.,.)$ on $[0, \tau] \times (-\infty, \infty)$ under the normal copula model then follows that of $\widehat{F}_T(.)$ on $[0, \tau]$, $\widehat{F}_U(.)$ on $(-\infty, \infty)$ by Theorem 3, and $\widehat{\rho}$ by Theorem 2. Given the consistency of $\widetilde{F}_{TU}(.,.)$ on $[0, \tau] \times (-\infty, \infty)$ (Huang & Louis, 1998, Theorem 4), the omnibus property of the test based on $\xi_{[0,\tau]\times(-\infty,\infty)}$ is subsequently obtained.

Using the chain rule, we obtain that the map from the empirical distribution of $\{X, Y, \Delta\}$ to $\widehat{F}_{TU}(.,.)$ on $[0, b] \times [c, d]$ is Hadamard-differentiable. Huang & Louis (1998) established the Hadamard-differentiability of the map to $\widetilde{F}_{TU}(.,.)$. The asymptotic normality result then follows by the functional delta method. □

## References


Bang, H. & Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika* **87**, 329–43.

Cheng, S. C., Wei, L. J. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–45.

Genest, C., Ghoudi, K. & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543–52.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises Method — I. *Scand. J. Statist.* **16**, 97–128.

Gill, R. D., van der Laan, M. J. & Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. *Annales de L'I.H.P. Prob. Stat.* **31**, 545–97.

Glasziou, P. P., Simes, R. J. & Gelber, R. D. (1990). Quality adjusted survival analysis. *Statist. Med.* **9**, 1259–76.

Huang, Y. (2002). Calibration regression of censored lifetime medical cost. *J. Am. Statist. Assoc.* **97**, 318–27. correction, 661.

Huang, Y. & Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* **85**, 785–98.

Huang, Y. & Lovato, L. (2002). Tests for lifetime utility or cost via calibrating survival time. *Statist. Sinica* **12**, 707–23.

Kelly, K., Crowley, J., Bunn, P. A., Jr., Presant, C. A., Grevstad, P. K., Moinpour, C. M., Ramsey, S. D., Wozniak, A. J., Weiss, G. R., Moore, D. F., Israel, V. K., Livingston, R. B. & Gandara, D. R. (2001). Randomized phase III trial of Paclitaxel plus Carboplatin versus Vinorelbine plus Cisplatin in the treatment of patients with advanced non-small-cell lung cancer: A Southwest Oncology Group trial. *J. Clin. Oncol.* **19**, 3210–8.

Klaassen, C. A. J. & Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3**, 55–77.

Lin, D. Y., Feuer, E. J., Etzioni, R. & Wax Y. (1997). Estimating medical costs from in-

complete follow-up data. *Biometrics* **53**, 419–34.

Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.* **42**, 1285–95.

Olschewski, M. & Schumacher, M. (1990). Statistical analysis of quality of life in cancer clinical trials. *Statist. Med.* **9**, 749–63.

Ramsey, S. D., Moinpour, C. M., Lovato, L. C., Crowley, J. J., Grevstad, P., Presant, C. A., Rivkin, S. E., Kelly, K. & Gandara, D. R. (2002). Economic analysis of Vinorelbine plus Cisplatin versus Paclitaxel plus Carboplatin for advanced non-small-cell lung cancer. *J. Natl. Cancer Inst.* **94**, 291–7.

Shih, J. H. & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–99.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Zhao, H. & Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**, 339–48.

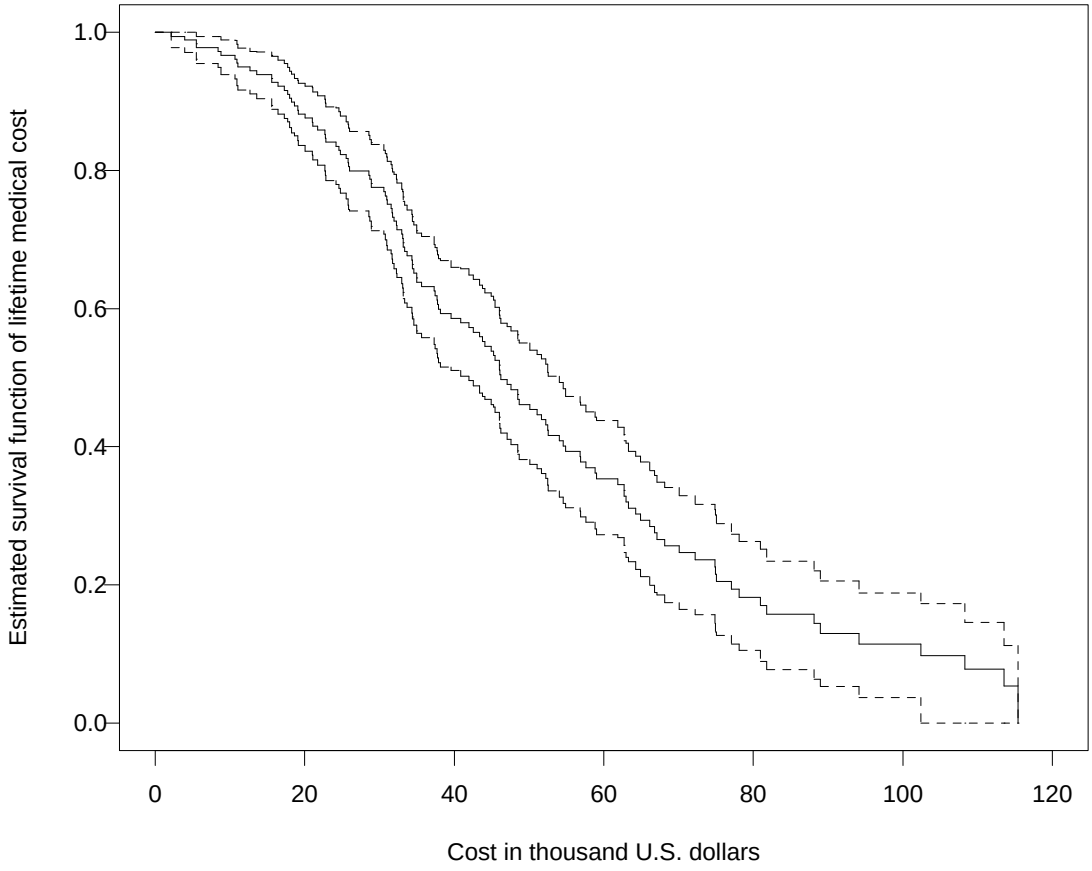


Figure 1: Southwest Oncology Group study. Estimated survival function of lifetime medical cost under the Paclitaxel plus Carboplatin therapy, along with pointwise 95% nonparametric bootstrap percentile confidence intervals.

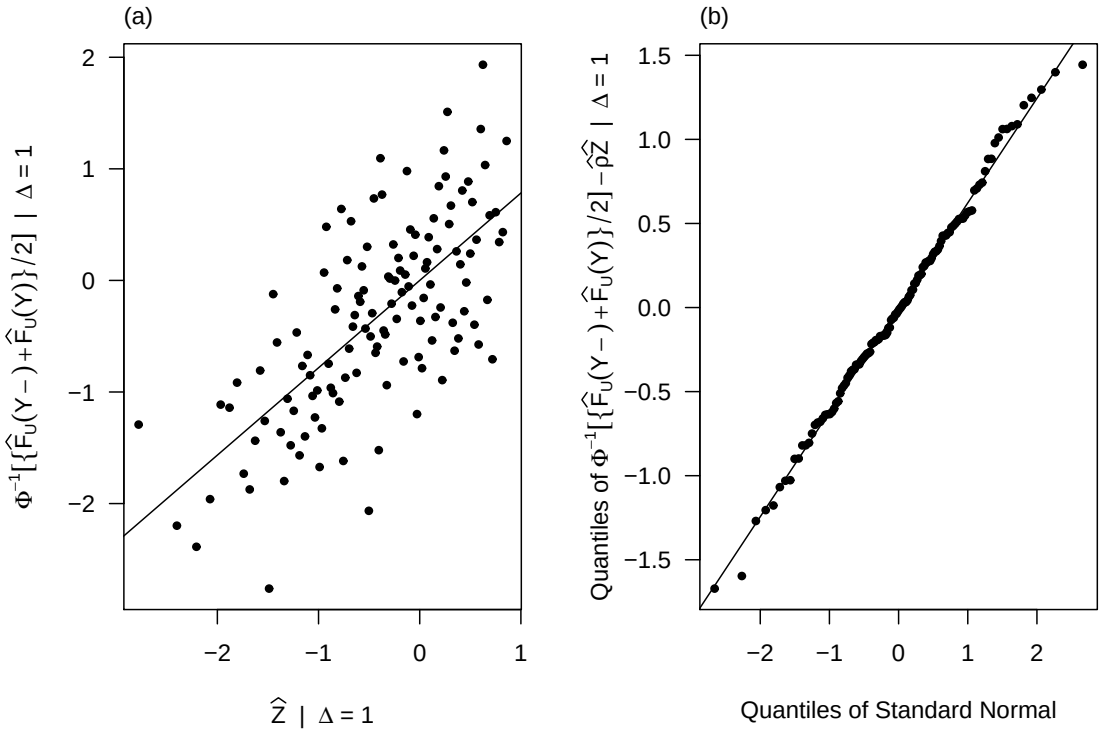


Figure 2: Southwest Oncology Group study. Graphical check of the normal copula model. (a) scatterplot of $\Phi^{-1}[\{\widehat{F}_U(Y-)+\widehat{F}_U(Y)\}/2]$ versus $\widehat{Z}$ given $\Delta = 1$; the straight line passes through (0,0) with slope $\widehat{\rho}$. (b) normal Q-Q plot of $\Phi^{-1}[\{\widehat{F}_U(Y-)+\widehat{F}_U(Y)\}/2]-\widehat{\rho}\widehat{Z}$ given $\Delta = 1$; the straight line passes through (0,0) with slope $(1-\widehat{\rho}^2)^{1/2}$.

Table 1: *Simulation summary statistics with sample size of 50: (a) bias ($\times 10^3$), standard deviation ($\times 10^3$); (b) coverage (%) of the 95% nonparametric bootstrap percentile confidence interval*

| Method | | $\rho$ | $F_U$ at percentile 25% | 50% | 75% | $\mu$ uni | $\mu$ exp | $\mu$ lgnm | $\nu$ lgnm |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\rho = 0.8$ | | | | | |
| Proposed | (a) | -16 72 | -2 61 | 3 75 | 8 77 | -16 152 | -61 161 | -81 141 | 8 90 |
| | (b) | 96.5 | 95.3 | 94.8 | 93.1 | 94.5 | 83.9 | 71.8 | 94.8 |
| CC | (a) | | 59 74 | 103 78 | 107 58 | -255 143 | -282 106 | -249 78 | -91 64 |
| NKM | (a) | | -4 61 | 6 74 | 30 70 | -43 140 | -116 125 | -132 101 | 1 82 |
| | | | | $\rho = 0.4$ | | | | | |
| Proposed | (a) | -14 166 | -2 65 | -2 83 | 2 79 | -2 170 | -9 188 | -12 187 | 12 100 |
| | (b) | 95.5 | 96.0 | 94.9 | 93.4 | 94.6 | 90.3 | 85.2 | 94.9 |
| CC | (a) | | 33 71 | 47 81 | 47 67 | -123 159 | -130 144 | -115 131 | -46 81 |
| NKM | (a) | | -21 59 | -29 77 | -11 73 | 50 148 | 17 157 | -2 157 | 42 96 |
| | | | | $\rho = 0$ | | | | | |
| Proposed | (a) | -2 202 | -3 71 | -5 87 | -3 76 | 8 180 | 7 180 | 6 179 | 18 107 |
| | (b) | 94.0 | 95.4 | 93.4 | 92.9 | 94.1 | 92.3 | 90.2 | 93.4 |
| CC | (a) | | -5 68 | -5 83 | -1 73 | 8 165 | 1 164 | -1 165 | 13 98 |
| NKM | (a) | | -47 58 | -61 78 | -43 76 | 137 152 | 112 169 | 86 183 | 87 108 |
| | | | | $\rho = -0.4$ | | | | | |
| Proposed | (a) | 14 171 | -4 78 | -7 85 | -6 69 | 16 178 | 12 162 | 9 158 | 16 105 |
| | (b) | 95.6 | 93.9 | 93.6 | 94.8 | 93.9 | 93.4 | 91.2 | 93.6 |
| CC | (a) | | -49 65 | -56 81 | -41 74 | 136 163 | 109 168 | 85 175 | 77 113 |
| NKM | (a) | | -78 57 | -93 77 | -68 76 | 222 152 | 181 169 | 141 183 | 131 116 |
| | | | | $\rho = -0.8$ | | | | | |
| Proposed | (a) | 18 74 | -9 80 | -7 78 | -4 65 | 25 165 | 14 146 | 10 142 | 14 95 |
| | (b) | 96.2 | 90.4 | 94.4 | 94.1 | 92.1 | 94.4 | 92.4 | 94.4 |
| CC | (a) | | -107 59 | -106 79 | -67 77 | 263 151 | 194 163 | 146 168 | 132 111 |
| NKM | (a) | | -121 54 | -125 77 | -79 78 | 303 145 | 227 163 | 171 171 | 157 112 |

CC, NKM: complete-case and naive Kaplan–Meier estimators, respectively.
uni, exp, lgnm: Marginal mark distribution $\sim$ uniform on $[0, 12^{1/2}]$, exponential with unit rate and standard log-normal scaled by $(e^2 - e)^{-1/2}$, respectively.

Table 2: *Simulation summary statistics for the joint distribution with sample size of 50: (a) bias ($\times 10^3$), standard deviation ($\times 10^3$) of the semiparametric estimator $\widehat{F}_{TU}(t,u)$; (b) bias ($\times 10^3$), standard deviation ($\times 10^3$) of the nonparametric Huang–Louis estimator $\widetilde{F}_{TU}(t,u)$*

| $u$ | $t = F_T^{-1}(0.2)$ | | $F_T^{-1}(0.4)$ | | $F_T^{-1}(0.6)$ | | $F_T^{-1}(0.8)$ | |
|---|---|---|---|---|---|---|---|---|
| | (a) | (b) | (a) | (b) | (a) | (b) | (a) | (b) |
| | | | | $\rho = 0.8$ | | | | |
| $F_U^{-1}(0.25)$ | -5 45 | -1 51 | -5 55 | -2 58 | -3 59 | -3 60 | -1 61 | -2 61 |
| $F_U^{-1}(0.50)$ | -4 53 | -1 56 | -6 64 | -2 69 | -5 69 | -2 72 | -1 73 | -1 74 |
| $F_U^{-1}(0.75)$ | -3 56 | -2 56 | -5 69 | -3 71 | -6 71 | -4 74 | -2 67 | -1 68 |
| | | | | $\rho = 0.4$ | | | | |
| $F_U^{-1}(0.25)$ | -3 34 | -3 40 | -4 44 | -3 50 | -4 52 | -4 54 | -3 58 | -3 59 |
| $F_U^{-1}(0.50)$ | -4 45 | -2 49 | -4 58 | -4 62 | -4 65 | -4 68 | -1 71 | -2 73 |
| $F_U^{-1}(0.75)$ | -3 52 | -1 55 | -3 66 | -1 70 | -3 70 | -1 74 | 1 70 | 1 71 |
| | | | | $\rho = 0$ | | | | |
| $F_U^{-1}(0.25)$ | -1 23 | -2 31 | -2 35 | -2 42 | -3 45 | -3 50 | -2 55 | -2 58 |
| $F_U^{-1}(0.50)$ | -3 35 | -3 41 | -4 50 | -5 56 | -5 59 | -5 63 | -3 67 | -3 70 |
| $F_U^{-1}(0.75)$ | -2 47 | -2 50 | -2 63 | -2 66 | -2 70 | -2 72 | 1 70 | 0 72 |
| | | | | $\rho = -0.4$ | | | | |
| $F_U^{-1}(0.25)$ | 1 12 | 0 19 | 0 24 | -1 31 | -1 37 | -1 43 | 0 52 | 0 55 |
| $F_U^{-1}(0.50)$ | 0 25 | -2 33 | -1 42 | -2 47 | -3 55 | -3 60 | -1 69 | -1 71 |
| $F_U^{-1}(0.75)$ | -1 38 | -2 42 | -2 57 | -2 62 | -2 68 | -2 71 | 0 70 | 0 72 |
| | | | | $\rho = -0.8$ | | | | |
| $F_U^{-1}(0.25)$ | 1 3 | 0 4 | 2 10 | 0 13 | 2 23 | -1 27 | 3 47 | 1 48 |
| $F_U^{-1}(0.50)$ | 2 8 | 0 14 | 3 26 | -1 33 | 1 48 | -2 54 | 1 69 | 1 71 |
| $F_U^{-1}(0.75)$ | 2 26 | -2 32 | 1 52 | -3 57 | -2 71 | -3 72 | -1 74 | 0 74 |

Table 3: *Simulation summary statistics for marginal mark distribution with the proposed method under nonnormal copulas, with sample size of 100: (a) bias ($\times 10^3$), standard deviation ($\times 10^3$); (b) coverage (%) of the 95% nonparametric bootstrap percentile confidence interval*

| | $F_U$ at percentile | | | $\mu$ | $\mu$ | $\mu$ | $\nu$ |
|---|---|---|---|---|---|---|---|
| | 25% | 50% | 75% | uni | exp | lgnm | lgnm |
| | Clayton's copula corresponding to $\rho = 0.8$ | | | | | | |
| (a) | 1 47 | 16 59 | 46 54 | -77 109 | -171 86 | -183 62 | -13 60 |
| (b) | 93.9 | 93.2 | 83.5 | 88.2 | 51.5 | 25.6 | 93.2 |
| | Clayton's copula corresponding to $\rho = 0.4$ | | | | | | |
| (a) | 9 48 | 25 60 | 36 53 | -69 114 | -114 105 | -120 87 | -20 65 |
| (b) | 95.9 | 93.0 | 89.5 | 92.1 | 80.8 | 68.8 | 93.0 |
| | Frank's copula corresponding to $\rho = 0.8$ | | | | | | |
| (a) | 1 45 | 2 55 | -3 53 | 6 113 | 39 157 | 52 205 | 2 63 |
| (b) | 94.4 | 95.1 | 94.7 | 94.9 | 92.5 | 88.6 | 95.1 |
| | Frank's copula corresponding to $\rho = 0.4$ | | | | | | |
| (a) | -1 46 | 3 58 | -4 55 | 4 116 | 20 134 | 25 144 | 5 72 |
| (b) | 95.8 | 94.6 | 95.2 | 95.7 | 94.4 | 93.5 | 94.6 |
| | Frank's copula corresponding to $\rho = -0.4$ | | | | | | |
| (a) | 2 48 | -2 44 | -2 44 | -1 106 | 0 106 | -1 105 | 5 52 |
| (b) | 96.5 | 98.3 | 96.6 | 96.7 | 95.1 | 93.2 | 98.3 |
| | Frank's copula corresponding to $\rho = -0.8$ | | | | | | |
| (a) | -1 51 | -4 48 | -1 43 | -2 108 | 2 104 | 1 103 | 7 58 |
| (b) | 96.3 | 96.0 | 95.0 | 96.1 | 94.5 | 91.8 | 96.0 |

uni, exp, lgnm: Marginal mark distribution $\sim$ uniform on $[0, 12^{1/2}]$, exponential with unit rate and standard log-normal scaled by $(e^2 - e)^{-1/2}$, respectively.

Table 4: *Empirical size and power ($\times 10^3$) of the Kolmogorov–Smirnov-type goodness-of-fit test based on* $\xi_{[0,\tau]\times(-\infty,\infty)}$

| Copula | | Normal | | | Clayton's | | | Frank's | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample size | | 50 | 100 | 200 | 50 | 100 | 200 | 50 | 100 | 200 |
| $\rho$ | Level | | | | | | | | | |
| 0.8 | 0.100 | 86 | 94 | 94 | 203 | 449 | 791 | 113 | 201 | 356 |
| | 0.050 | 38 | 41 | 57 | 116 | 312 | 649 | 54 | 109 | 223 |
| | 0.010 | 8 | 8 | 11 | 25 | 100 | 314 | 8 | 25 | 72 |
| 0.4 | 0.100 | 91 | 94 | 102 | 88 | 114 | 189 | 95 | 95 | 139 |
| | 0.050 | 46 | 48 | 57 | 51 | 60 | 100 | 47 | 47 | 75 |
| | 0.010 | 5 | 7 | 13 | 6 | 14 | 23 | 7 | 9 | 18 |
| 0.0 | 0.100 | 89 | 96 | 107 | | | | | | |
| | 0.050 | 50 | 45 | 48 | | | | | | |
| | 0.010 | 8 | 10 | 8 | | | | | | |
| -0.4 | 0.100 | 87 | 90 | 102 | | | | 103 | 135 | 179 |
| | 0.050 | 51 | 54 | 53 | | | | 40 | 70 | 100 |
| | 0.010 | 11 | 9 | 14 | | | | 7 | 19 | 23 |
| -0.8 | 0.100 | 90 | 89 | 89 | | | | 104 | 205 | 342 |
| | 0.050 | 35 | 46 | 47 | | | | 50 | 115 | 217 |
| | 0.010 | 5 | 10 | 7 | | | | 5 | 30 | 75 |