Harvard University
Harvard University Biostatistics Working Paper Series

Year 2004 *Paper* 15

Semiparametric Methods for Semi-competing Risks Problem with Censoring and Truncation

Hongyu Jiang* Jason Fine†
Richard J. Chappell‡

*Harvard University, hjiang@hsph.harvard.edu
†University of Wisconsin-Madison, jfine@bios.unc.edu
‡University of Wisconsin-Madison, chappell@stat.wisc.edu

http://biostats.bepress.com/harvardbiostat/paper15

Semiparametric Methods for Semi-competing Risks Problem with Censoring and Truncation

Hongyu Jiang, Jason Fine, and Richard J. Chappell

Abstract

Studies of chronic life-threatening diseases often involve both mortality and morbidity. In observational studies, the data may also be subject to administrative left truncation and right censoring. Since mortality and morbidity may be correlated and mortality may censor morbidity, the Lynden-Bell estimator for left truncated and right censored data may be biased for estimating the marginal survival function of the non-terminal event. We propose a semiparametric estimator for this survival function based on a joint model for the two time-to-event variables, which utilizes the gamma frailty specification in the region of the observable data. Firstly, we develop a novel estimator for the gamma frailty parameter under left truncation. Using this estimator, we then derive a closed form estimator for the marginal distribution of the non-terminal event. The large sample properties of the estimators are established via asymptotic theory. The methodology performs well with moderate sample sizes, both in simulations and in an analysis of data from a diabetes registry.

Semiparametric Methods for Semi-competing Risks Problem with Censoring and Truncation

Hongyu Jiang[1], Jason P. Fine[2], and Rick Chappell[3]

[1] Department of Biostatistics and Center for Biostatistics in AIDS Research, Harvard University, Boston, MA 02115 U.S.A.

[2,3] Department of Statistics and Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, WI 53706 U.S.A.

Abstract. Studies of chronic life-threatening diseases often involve both mortality and morbidity. In observational studies, the data may also be subject to administrative left truncation and right censoring. Since mortality and morbidity may be correlated and mortality may censor morbidity, the Lynden-Bell estimator for left truncated and right censored data may be biased for estimating the marginal survival function of the non-terminal event. We propose a semiparametric estimator for this survival function based on a joint model for the two time-to-event variables, which utilizes the gamma frailty specification in the region of the observable data. Firstly, we develop a novel estimator for the gamma frailty parameter under left truncation. Using this estimator, we then derive a closed form estimator for the marginal distribution of the non-terminal event. The large sample properties of the estimators are established via asymptotic theory. The methodology performs well with moderate sample sizes, both in simulations and in an analysis of data from a diabetes registry.
Key Words: Bivariate survival function; Concordance probability; Copula; Semi-competing risks; Truncation.

1. Introduction

Chronic life-threatening diseases usually involve multiple landmark events in the

progression of the disease. The landmarks are non-terminal and observation of such events occurs until death. For complex diseases, certain events may not be observed prior to death, in which case the censoring by death may be informative. Such data have been referred to as *semi-competing risks* data (Fine, Jiang and Chappell, 2001), since death may dependently censor landmark events, but not vice versa.

The motivating example for our methodologic research is registry data on diabetes patients, for whom landmark events might include the development of diabetic retinopathy, a major cause of blindness, and diabetic nephropathy, an indication of kidney failure (Andersen et al., 1983; Mogensen, 1984; Borch-Johnsen et al. 1985). When studying the natural history of diabetes, investigators may intend to report the marginal probabilities of the landmark events; see Bojestig et al., 1994; Remuzzi and Ruggenenti, 1998; and Dahlquist et al., 2001. However, because diabetic morbidities and death are associated, the Kaplan-Meier estimator for right censored data may be biased for the marginal distribution of the non-terminal event (Fine et al., 2001). For administratively right censored semi-competing risks data, Fine et al. (2001) showed that the dependence between morbidity and mortality can be estimated separately from their marginals under a gamma frailty copula in the region of the observable data (Day et al., 1997). They also provided a closed-form estimator for the marginal distribution of landmark events.

In the diabetes registry, there is another complication: administrative left truncation. The patients in this registry were treated at the Steno Memorial Hospital in Greater Copenhagen, a diabetes specialist hospital in Denmark (Borch-Johnsen et al., 1985; Ramlau-Hansen et al., 1987). Patients were referred to the Steno from general practitioners and/or other hospitals between 1933 and 1981. Those who died prior to the start of the study or failed to be referred were excluded from the registry. Thus, both time to landmark event and mortality are observed conditionally on death

occurring after study enrollment. To avoid bias, the observation of the event times on each enrolled patient was left-truncated at the time of first contact at the Steno. For an event time subject to independent left truncation and right censoring, the Lynden-Bell product limit estimator may be used to estimate the survival function (Lynden-Bell, 1971; Wang, et al., 1986; Tsai et al., 1987; Gu and Lai, 1990; Lai and Ying, 1991; Gijbels and Wang, 1993). In our set-up, the non-terminal event is subject to dependent censoring, in addition to administrative left truncation and right censoring, and the Lynden-Bell estimator may be biased.

Such left-truncated semi-competing risk data could be analyzed using a three-state illness-death transition intensity model (Andersen et al., 1993, Ch. IV.4.), However, methods for quantifying the association between illness and death and estimating the marginal distribution of the non-terminal illness event are not available. In the diabetes example, the marginal distribution is important in comparing cohorts from different calendar periods. The distribution corresponds to time to nephropathy in the absence of death prior to nephropathy. The overall management of diabetes has improved over calendar time, with recent medical care being more effective in prolonging times to nephropathy and to death following nephropathy, as well as time to death without nephropathy. The marginal distribution may be used to evaluate the net effect of care on nephropathy, independently of its effects on other aspects of diabetes which lead to dependent censoring via mortality prior to nephropathy.

To address these issues, we extend the methods of Fine et al. (2001) to the left-truncated semi-competing risk problem. In Section 2, we introduce our notation and the modified Clayton model on the upper wedge. In Section 3, new estimators are proposed for the association parameter in the Clayton model for left truncated semi-competing risks data. The closed-form estimator for the marginal distribution of the non-terminal event is given in Section 4. Technical details related to estimation

and inference are relegated to the Appendix. To illustrate the performance of our estimators, they are applied to simulated datasets and data from the aforementioned diabetic cohort study. The results are summarized in Sections 5 and 6 with some concluding discussion in Section 7.

2. Notation and model

To fix notation, let X and Y be two possibly dependent failure times observable from the same subject. The random variable Y can censor X, but not vice versa. Hence, data are only observable in the upper wedge of the support of (X, Y), where $X < Y$. Let A denote the left truncation variable and C be the right censoring variable. It is assumed that (A, C) are independent of (X, Y). This assumption is analogous to that made with univariate data and is reasonable when both truncation and censoring are administrative, as in the diabetes study. Because of censoring, we observe $Y' = Y \wedge C$, $\delta = I(Y < C)$, $X' = X \wedge Y \wedge C$, and $\eta = I\{X < (Y \wedge C)\}$, conditionally on $Y' > A$, where $x \wedge y$ is the minimum of x and y and $I(\cdot)$ is the indicator function. The observed data consists of n independent realizations of $(X', Y', A, \eta, \delta)$, denoted by $\{(X'_i, Y'_i, A_i, \eta_i, \delta_i), i = 1, ..., n\}$.

The traditional truncation problem only considers estimating the distribution of Y, the terminating endpoint, in the presence of independent left-truncation and right-censoring. With truncated data, it may not be possible estimate the unconditional survival curve $\Pr(Y > y)$ (Klein and Moeschberger, 1997, Ch. 4.6). Typically, one estimates the survival distribution given that the event occurs after a certain time point a. In practice, a could be $A_{(1)}$, the minimum of the A_i's or another time point greater than $A_{(1)}$ which is scientifically meaningful. The questions of interest are framed in terms of the conditional survival function $\Pr(Y > y | Y > a)$. If the truncation variable has a continuous distribution with support $[0, b]$ where $b \in [0, \infty)$,

then we can consistently estimate the unconditional survival function of Y since $A_{(1)} \to 0$ as $n \to \infty$ (Woodroofe, 1985). Applying the Lynden-Bell technique to $\{(Y_i', A_i, \delta_i), i = 1, ...n\}$ gives the estimator

$$\widehat{\Pr}(Y > y | Y > a) = \prod_{a < Y_i' \leq t} \left\{ 1 - \frac{d_y(Y_i')}{r_y(Y_i')} \right\},$$

where $d_y(t) = \sum_{i=1}^n I\{Y_i' = t, \delta_i = 1\}$ and $r_y(t) = \sum_{i=1}^n I\{A_i \leq t \leq Y_i'\}$. For the semi-competing risks problem, this is always a valid approach to estimate the distribution of Y since Y is only subject to independent censoring from C.

Estimation of the distribution of X is more complicated. Although X' is not necessarily $\geq$ the truncation variable A for Y', the observation of X' is conditional on a subject being sampled, that is, $Y' \geq A$. Hence, only the conditional distribution of X given the corresponding Y in the observable region $Y > a$ is identifiable. An exception is when X and Y are independent, in which case the unconditional distribution of X is identifiable and may be consistently estimated by the Kaplan-Meier estimator using data $\{(X_i', \eta_i), i = 1, \ldots, n\}$, regardless whether or not the unconditional distribution of Y is estimable. However, when Y can dependently censor X, neither the Kaplan-Meier estimator nor the Lynden-Bell estimator by artificially truncating X at A is valid for estimating either the unconditional or the conditional distribution of X. To recover the distribution of X with dependent censoring by Y, it is necessary to model the joint distribution of X and Y. In this paper, we propose a semi-parametric model which specifies a flexible parametric relationship between X and Y, while leaving the marginal survival functions completely unspecified. Using the approach in Day et al. (1997) and Clayton (1978), we posit a modified Clayton (1978) copula restricted to the observable region where $X < Y$ and $Y > a$.

Let $R_a(y)$ be the conditional survival function of Y, that is, $R(y)/R(a)$, where $R(y) = \Pr(Y > y)$. Define $F_a(x, y) = \Pr(X > x, Y > y | Y > a) = F(x, y)/R(a)$,

where $F(x,y) = Pr(X > x, Y > y)$. For $0 \leq \max(x,a) < y \leq \infty$, the model is

$$F_a(x,y) = \left\{ S_a(x)^{1-\theta} + R_a(y)^{1-\theta} - 1 \right\}^{\frac{1}{1-\theta}}, \tag{1}$$

where $S_a(x)$ meets the definition of a conditional survival function given $Y > a$. Interestingly, if $F_0(x,y) = F(x,y)$ satisfies (1) for $x < y$ with $S_0(x)$ and $R_0(y) = R(y)$ in place of $S_a(x)$ and $R_a(y)$, respectively, then $F_a(x,y)$ satisfies (1) with the same θ and $S_a(x) = F_0(x,a)/R_0(a)$ and $R_a(y) = R_0(y)/R_0(a)$ for all $a > 0$ (Hougaard, Section 7.3, 2002). Note that $S_0(x)$ may denote the unconditional distribution of X. We show later that this distribution is estimable under the assumed model (1) when $R_0(y)$, the unconditional distribution of Y, is estimable.

The parameter $\theta \geq 0$ quantifies the association between X and Y. When $\theta = 1$, $F_a(x,y) = S_0(x)R_a(y)$ and the event times are said to be independent on the upper wedge. For $\theta > 1$, there is positive association and θ is related to a gamma frailty model for X and Y (Day et al., 1997). For $\theta < 1$, negative dependence occurs (Oakes, 1989), but a frailty model representation does not exist.

To interpret S_a as the conditional marginal of X, we require that $S_a(t) = F_a(t,a) = F_0(t,a)/R(a)$ for all $t > a$. A mild additional assumption which guarantees this to be true is to require that the joint conditional survival function in the lower wedge $(x > y)$ can be written in a copula form in terms of the same functionals $S_a(.)$ and $R_a(.)$, and agrees with model (1) at $x = y > a$. The technical details of this model are given in Fine et al. (2001). In the sequel, when interpreting $S_a(\cdot)$ as the marginal, we implicitly assume that this additional assumption holds.

3. Inferences for the dependence structure

We develop an estimator for θ in (1) which does not require estimators for S_a and R_a. Let (X_i, Y_i) and (X_j, Y_j) denote independent pairs of event times from the ith and jth subjects. The estimator uses an indicator of the concordance or discordance

of these pairs,

$$\Delta_{ij} = \begin{cases} 1 & \text{if } (X_i - X_j)(Y_i - Y_j) > 0 \\ 0 & \text{if } (X_i - X_j)(Y_i - Y_j) < 0. \end{cases}$$

Under model (1), the expectation of Δ_{ij} given $Y_i \wedge Y_j > a$ is $\theta_0/(1+\theta_0)$, where θ_0 is the true value of θ (Hougaard, Section 7.3, 2002). In the left-truncated semi-competing risks setting, Δ_{ij} is computable from comparable pairs of event times when $(\tilde{A}_{ij} \vee \tilde{X}_{ij}) < \tilde{Y}_{ij} < \tilde{C}_{ij}$, where $\tilde{A}_{ij} = A_i \vee A_j, \tilde{X}_{ij} = X_i \wedge X_j, \tilde{Y}_{ij} = Y_i \wedge Y_j, \tilde{C}_{ij} = C_i \wedge C_j$, and $u \vee v$ stands for the maximum of u and v. Truncating $\tilde{Y}_{ij}$ by $\tilde{A}_{ij}$ ensures that Δ_{ij} has the desired conditional expectation.

The following extension of Oakes' (1982, 1986) estimator for θ exploits these properties of Δ_{ij}. Let $O_{ij} = I(\tilde{A}_{ij} \vee a \vee \tilde{X}_{ij} < \tilde{Y}_{ij} < \tilde{C}_{ij})$ and define $U(\theta) = \sum_{i<j} W(\tilde{X}'_{ij}, \tilde{Y}'_{ij}) O_{ij} \{\Delta_{ij} - \theta/(1+\theta)\}$, where $W(u,v)$ is a random weight function with a deterministic limit, $\tilde{W}(u,v)$, which is bounded for (u,v) in the support of $\{\tilde{X}'_{ij} = X'_i \wedge X'_j, \tilde{Y}'_{ij} = Y'_i \wedge Y'_j\}$. By solving $U(\theta) = 0$, we obtain the concordance estimator

$$\hat{\theta} = \frac{\sum_{i<j} W(\tilde{X}'_{ij}, \tilde{Y}'_{ij}) O_{ij} \Delta_{ij}}{\sum_{i<j} W(\tilde{X}'_{ij}, \tilde{Y}'_{ij}) O_{ij} (1 - \Delta_{ij})}. \tag{2}$$

As discussed in Fine et al. (2002), a useful weight function is

$$W^{-1}_{c,d}(x,y) = n^{-1} \sum_{i=1}^{n} I\{X'_i \geq x \wedge c, Y'_i \geq y \wedge d\}, \tag{3}$$

where c and d are some constants. With $c = d = 0$, $W_{0,0} \equiv 1$ and $\hat{\theta}$ reduces to unweighted concordance estimator. With $c = d = \infty$, $W_{c,d}(x,y)$ is the inverse of proportion of subjects in the risk set defined by (x,y) (Oakes, 1986). In practice, c and d in (3) may be selected so that excessive weight is not given to large x and y where the risk set may be small.

It is easy to show that as $n \to \infty$, $n^{-2}\{U(\theta) - \tilde{U}(\theta)\}$ vanishes uniformly for θ in a neighborhood of θ_0, where $\tilde{U}$ is U with W replaced by $\tilde{W}$. Thus, $\hat{\theta}$ has the same

limit as $\tilde{\theta}$, the root of $\tilde{U}(\theta) = 0$. Assuming model (1) holds, Δ_{ij} is independent of $(\tilde{X}_{ij}, \tilde{Y}_{ij})$. On the other hand, by the independence between $\{\tilde{A}_{ij}, \tilde{C}_{ij}\}$ and $\{\tilde{X}_{ij}, \tilde{Y}_{ij}\}$, we have Δ_{ij} is independent of $\tilde{W}(\tilde{X}'_{ij}, \tilde{Y}'_{ij})O_{ij}$ which implies that $E\{\tilde{U}(\theta_0)\} = 0$. The strong law of large numbers for U-statistics and a continuous mapping theorem give that $\tilde{\theta}$ is strongly consistent for θ_0. Hence, $\hat{\theta}$ is strongly consistent. In Appendix 1, we show that $n^{1/2}(\hat{\theta} - \theta_0)$ has a limiting normal distribution with variance Σ which is consistently estimated by $\hat{\Sigma} = \hat{I}^{-2}\hat{J}$, where

$$\hat{I} = n^{-2} \sum_{i<j} W(\tilde{X}'_{ij}, \tilde{Y}'_{ij}) O_{ij} (1 + \hat{\theta})^{-2},$$

$$\hat{J} = 2n^{-3} \sum_{k<l<m} (\hat{Q}_{kl}\hat{Q}_{km} + \hat{Q}_{kl}\hat{Q}_{lm} + \hat{Q}_{lm}\hat{Q}_{km}),$$

and $\hat{Q}_{kl} = W(\tilde{X}'_{kl}, \tilde{Y}'_{kl}) D_{kl}\{\Delta_{kl} - \hat{\theta}/(1 + \hat{\theta})\}$.

Since inferences about S_a rely on the copula (1), it would be helpful to assess this formulation. Similar to the model checking technique for evaluating the fitness of Clayton copula for semi-competing risks data (Fine et al. 2001), a goodness-of-fit statistic based on the distance between two estimators from $U(\theta)$ with different weights can be derived (Shih, 1998). Under misspecification, the estimators may converge to distinct values and the test rejects with probability one. Of course, the test has low power against certain alternatives. When $X > Y$, the (X, Y) pairs are unobservable and the relationship between X and Y is nonidentifiable.

Let $W_i = W_{c_i, d_i}, U_i$ be U with W_i in place of W, and $\hat{\theta}_i$ be the corresponding estimator, $i = 1, 2$. In Appendix 1, we show that when the copula is specified correctly $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2)$ is asymptotically normal with variance that is consistently estimated by

$$\hat{\Gamma} = 2n^{-3} \sum_{k<l<m} (\hat{Q}^*_{kl}\hat{Q}^*_{km} + \hat{Q}^*_{kl}\hat{Q}^*_{lm} + \hat{Q}^*_{lm}\hat{Q}^*_{km}),$$

where $\hat{Q}^*_{kl} = \hat{I}_1^{-1}\hat{Q}_{1kl} - \hat{I}_2^{-1}\hat{Q}_{2kl}$, and $\hat{I}_i$ and $\hat{Q}_{ikl}$ are $\hat{I}$ and $\hat{Q}_{kl}$ with W replaced by $W_i, i = 1, 2$. For a 2α level test, the critical region is $n^{1/2}|\hat{\theta}_1 - \hat{\theta}_2|\hat{\Gamma}^{-1/2} > \psi_{1-\alpha}$, where

ψ_q is the qth quantile of the standard normal distribution.

4. Estimating the marginal conditional distribution

To develop an estimator for $S_a(t)$ from model (1), we require estimators for both $R_a(t)$ and $F_a(t,t)$ in addition to θ. While it is straightforward to estimate $R_a(t)$ using the Lynden-Bell estimator, to our knowledge, the conditional distribution $F_a(t,t)$ cannot be nonparametrically estimated using available methods.

To circumvent this difficulty, we further condition on $X > a$ on both sides of model (1). Following Hougaard (Section 7.3, 2002), the relationship

$$F_{\tilde{a}}(x,y) = \Pr(X > x, Y > y | X > a, Y > a) = \{S_{\tilde{a}}(x)^{1-\theta} + R_{\tilde{a}}(y)^{1-\theta} - 1\}^{\frac{1}{1-\theta}}, \quad (4)$$

holds with the same θ as in (1), but where $R_{\tilde{a}}(y) = \Pr(Y > y | X > a, Y > a)$. As when conditioning only on $Y > a$, if $a = 0$, then the unconditional distribution of X is identifiable from the assumed model for $F_{\tilde{a}}$. Now, let $Z = X \wedge Y$ and $\gamma = I(Z \leq C)$. It is easy to see that $\gamma = \eta + (1-\eta)\delta$ and the minimum of Z and C is X'. Define $H_{\tilde{a}}(t) \equiv \Pr(X > t, Y > t | X > a, Y > a)$. Equality (4) implies $S_{\tilde{a}}(t) = g\{H_{\tilde{a}}(t), R_{\tilde{a}}(t), \theta\}$, where $g(a,b,c) = (a^{1-c} - b^{1-c} + 1)^{1/(1-c)}$. Both $H_{\tilde{a}}(t)$ and $R_{\tilde{a}}(t)$ can be consistently estimated by Lynden-Bell estimators using data $\{(Z_i, \gamma_i, A_i) : X_i > a, Y_i > a, i = 1, \ldots, n\}$ and $\{(Y_i, \delta_i, A_i) : X_i > a, Y_i > a, i = 1, \ldots, n\}$, respectively. Hence, using ideas from Fine et al. (2001), we propose an estimator $\hat{S}_{\tilde{a}}(t) = g\{\hat{H}_{\tilde{a}}(t), \hat{R}_{\tilde{a}}(t), \hat{\theta}\}$, where $\hat{H}_{\tilde{a}}(t)$ and $\hat{R}_{\tilde{a}}(t)$ are the Lynden-Bell estimators for $H_{\tilde{a}}(t)$ and $R_{\tilde{a}}(t)$, respectively, and $\hat{\theta}$ is the concordance estimator.

Note that $\hat{\theta}$ is strongly consistent for θ_0, and $\hat{H}_{\tilde{a}}(t)$ and $\hat{R}_{\tilde{a}}(t)$ are strongly consistent for $H_{\tilde{a}}(t)$ and $R_{\tilde{a}}(t)$, uniformly for $t \in [a, \tau]$, where $P(X' > \tau) > \epsilon > 0$, ϵ fixed (Tsai et al., 1987; Lai and Ying, 1991). This plus the fact that g has bounded derivatives gives the uniform strong convergence of $\hat{S}_{\tilde{a}}(t)$ to $\{S_{\tilde{a}}(t) : t > a\}$. The

convergence of $n^{1/2}\{\hat{S}_{\tilde{a}}(t) - S_{\tilde{a}}(t)\}$ to a Gaussian process for $t \in [a, \tau]$ can be easily established by extending the theoretical developments in Fine et al. (2001) to the left-truncated data. Hence, in Appendix 2, we present the variance estimator for $\hat{S}_{\tilde{a}}$ with the proof omitted. To obtain a confidence interval for $S_{\tilde{a}}$ bounded in [0,1], we can take monotone transformation of the estimator and apply the delta method.

Unlike the Lynden-Bell estimator for $S_{\tilde{a}}(t)$, which decreases at each $t > a$ with $\sum_i \eta_i I(X_i' = t) > 1$, the estimator $\hat{S}_{\tilde{a}}$ is a step-function which changes value at both the observed values of X and Y whenever $\hat{H}_{\tilde{a}}(t)^{1-\hat{\theta}} - \hat{R}_{\tilde{a}}(t)^{1-\hat{\theta}}$ jumps. In finite samples, $\hat{S}_{\tilde{a}}(t)$ may not be monotone or may not be well-defined for some t since $\hat{H}_{\tilde{a}}(t)$ may be greater than $\hat{R}_{\tilde{a}}(t)$, in contrast to the theoretical stochastic ordering of $H_{\tilde{a}}(t)$ and $R_{\tilde{a}}(t)$. The difficulties occur most often when estimating probabilities in the tail of $S_{\tilde{a}}$, especially when censoring of X by Y is heavy. To address tail instability, we restrict attention to the interval $[a, t^*]$, where $t^* \leq \max\{s : \hat{H}_{\tilde{a}}(u)^{1-\hat{\theta}} - \hat{R}_{\tilde{a}}(u)^{1-\hat{\theta}} > -1, 0 \leq \hat{S}_{\tilde{a}}(u) \leq 1, u \leq s\}$. The monotone estimator $\hat{S}^*_{\tilde{a}}(t) = \min_{a<s\leq t}\{\hat{S}_{\tilde{a}}(s)\}$ is asymptotically equivalent to $\hat{S}_{\tilde{a}}$ for $t \in [a, t^*]$.

5. Numerical Studies

To evaluate the performance of our proposed estimators, we generated (X, Y) pairs from model (1) with both X and Y following a Weibull marginal distribution with scale parameter 1 and shape parameter 0.5 and we chose a to be 0.5, $\theta = 1, 2,$ or 3 and sample size $n = 100$ or 200. The truncation variable A was generated from an independent Weibull(1, 0.5) distribution with a shift of 0.5. Conditional on A, a censoring variable C independent of (X, Y) follows a uniform distribution on $[A, A + 25]$, giving 15% independent censoring on Y. Censoring by either C or Y on X is 42%. 1000 datasets were simulated for each combination of θ and n.

In studying the estimator of the association parameter, $\hat{\theta}$, we compared two

weights of the form (3), one with $c = d = 0$ ($W = 1$) and the other with c and d equal to $x_{.95}$ and $y_{.95}$, the 95th percentiles of the uncensored X and Y values, respectively. In Table 1, we report the average of the following quantities from 1000 simulations: $\hat{\theta}$ (Ave), the empirical variance of $\hat{\theta}$ (EmpVar), the estimated variance (AveVar), the coverage probability of the nominal 95% confidence interval for θ (Cov95). In all cases, the bias of θ is small, decreasing as n increases. The estimated asymptotic variance $n^{-1}\hat{\Sigma}$ and the empirical variance agree well and the confidence intervals have the right coverage probability. In most cases, the weighted estimator is more efficient than the unweighted estimator and has slightly higher coverage probabilities.

The performance of $\hat{S}^*_{\tilde{a}}(t)$ in simulated datasets is summarized in Table 2. We estimated $S_{\tilde{a}}(t)$ with the weighted estimator of θ and constructed a confidence interval using $\log(-\log(.))$ transformation. At various quantiles of $S_{\tilde{a}}(t)$, we computed the mean estimates of the conditional probabilities (Ave), the empirical variance (EmpVar), the average of estimated variance (AveVar), the percentage of valid estimates (%Val) and the coverage probability of 95% confidence interval (Cov95). The average and the empirical variance of the naive Lynden-Bell estimator are also presented in Table 2.

In all cases, the proposed estimator for $S_{\tilde{a}}(t)$ is unbiased. The estimated variance is larger than the empirical variance on average, but the difference diminishes as sample size increases. The coverage is generally close to 0.95 and improves with larger samples. Under independence, the Lynden-Bell estimator is somewhat more efficient since θ is estimated and not fixed at 1 when computing $\hat{S}^*_{\tilde{a}}$. However, when there is moderate dependence between X and Y ($\theta = 2, 3$), the naive estimator is noticeably biased upwards, with bias increasing as the dependence strengthens.

6. Application to the Denmark Diabetes Registry

We applied our methods to the aforementioned prospective cohort study on insulin-dependent diabetic patients conducted in the Steno Memorial Hospital in Greater Copenhagen. From 1933 to 1981, the study accrued roughly 2700 patients, who were diagnosed with insulin-dependent diabetes mellitus prior to age 31 and between 1933 and 1972. At entry, patients' age, age of diabetes diagnosis and the presence of diabetic nephropathy (DN) were recorded. The patients were then followed to death, emigration, or December 31, 1984 and the incidence of DN if not present at entry. A detailed description can be found in Andersen et al. (1993, Ch.I.3.).

Our focus is to quantify the association between time to DN and time to diabetes-related death and to estimate the probability of developing DN after being diagnosed with insulin-dependent diabetes. The relevant time origin is the time of diabetes diagnosis. As discussed in Section 1, an analysis of the marginal distribution of DN quantifies the net effect of care on prolonging time to DN, independently of its effect on death prior to nephropathy. The association between times of DN onset and death and the marginal distributions of the event times may vary across birth cohorts, and it is of interest to study the changes in these quantities over calendar time. In this paper, we illustrate our methods with the [1935, 1940) and [1945, 1950) birth cohorts.

Table 3 summarizes the observed data in the cohorts. Not surprisingly, fewer deaths were observed in the 1945-50 birth cohort. A crude analysis showed a rate of 9.7 deaths per 1000 person-year in the 1935-40 birth cohort versus a rate of 6.0 deaths per 1000 person-year in the 1945-50 birth cohort. Two possible reasons are that the latter cohort is younger and have access to more recent treatments for diabetes. It is also noted that 7% and 5% of subjects in the 1935-40 and 1945-50 birth cohorts, respectively, had already developed DN at the time of admission to registry.

Since 24.4% and 27.7% of patients were truncated at the diagnosis time (i.e., $A = 0$) in the 1935-40 and 1945-50 birth cohorts, respectively, we assumed unconditional

Clayton copula model with $a = 0$ for both cohorts. The bottom part of Table 3 shows that model (1) fitted the data in each cohort reasonably well using the goodness-of-fit test described in Section 3. The estimates of θ indicate strong association between time to the development of DN and time to death in both cohorts, with the dependence in 1935-40 being noticeably stronger. For any $x < y$, the hazard of death at time y for a patient who has developed DN at time x is about 8.8 times the hazard of death at time y for a patient who has not developed DN by time x in the 1935-40 cohort. The ratio is about 5.4 in the 1945-50 cohort. This finding supports the well known fact that diabetic nephropathy is a strong prognostic factor for death. It also suggests that the increased risk of mortality following DN has decreased over calendar time. The difference in association is marginally significant, with p-values of 0.06 and 0.08 obtained with tests using the weighted and unweighted estimators, respectively.

To estimate the marginal distribution of DN in insulin-dependent diabetic patients, we used the weighted estimator for the association parameter in Table 3 for each cohort. Figure 1 plots the estimated DN-free probabilities curves (thick solid lines) and their point-wise confidence intervals (thin dashed lines). Recall that these curves refer to scenarios in which death prior to DN is not possible. Hence, when comparing the two birth cohorts, these probabilities may be interpreted as the net effects of changes in care over calendar time, after adjusting for the indirect effects of changes in care on death from other aspects of diabetes prior to the occurrence of DN. To indicate the potential bias from the Kaplan-Meier estimator which assumes independence between time to DN and time to diabetes-related death, these estimators (thick dotted lines) are also plotted in Figure 1. For both cohorts, the naive estimator markedly overestimates the probabilities obtained with $\hat{S}^*_{\tilde{a}}$. For the 1935-40 cohort, the dependent censoring rate for DN onset is moderate, and the Kaplan-Meier estimator lies almost entirely outside the point-wise 95% confidence interval based on

our proposed estimator. For the 1945-50 cohort, the dependent censoring is lighter and differences between the Kaplan-Meier estimator and our estimator are smaller.

Figure 2 simultaneously plots the estimated event-free probabilities side by side for the two birth cohorts. In both panels, the solid line and the dotted line correspond to the 1935-40 and the 1945-50 birth cohorts. The lower panel confirms that survival of diabetic patients is improving in 1945-50 birth cohort, with a clear 4-5% advantage at most time points. In the upper panel, it is interesting to notice that after adjusting for the dependent censoring from death prior to DN, the DN-free probability curves of the two birth cohorts do not differ much in the first 23 years post-diagnosis. The noticeable divergence after 23 years post-diagnosis should be interpreted with care since for those who did not develop DN by 23 years post-diagnosis in the 1935-40 birth cohort, 94% had their incidence of DN censored at a later time. The estimated tail probabilities may not be reliable. Still, bearing in mind this tail instability, the stochastic ordering of the DN and death distributions is different in the two cohorts. Based on the point estimates, it is more likely to observe the development of DN prior to death in the 1945-50 cohort, even with this cohort's improved survival. A possible explanation is that increased survival post diagnosis has enabled diabetics to live long enough to develop severe complications, like nephropathy, at a higher rate. The increase in DN and the lengthening of survival might occur if the improvements in survival result from the prevention of death prior to DN and individuals experiencing DN are able to live longer with this condition. This scenario does not seem out of line with our results, which showed that the risk of death following DN is lower in the 1945-50 cohort.

The validity of the above interpretation involves several layers of assumptions. The first is that the underlying model (1) is correctly specified. The gamma-frailty copula on the upper wedge was not rejected using the numerical tests, so there is at

least some evidence that the model fits well in the observable region. Jiang et al. (2003) have shown that when the dependence structure is misspecified, the estimates of the marginal of X tend to be robust to this misspecification. The second and more critical assumption is that $S_{\tilde{a}}$ defined on the upper wedge corresponds to the marginal distribution of X defined in the lower wedge, where no data is observable. Even when the adjustment for dependent censoring is correctly specified by (1), this additional assumption is still needed to interpret $S_{\tilde{a}}$ as the marginal of X and cannot be verified empirically.

7. Discussion

To analyze left-truncated semi-competing risk data, we proposed a conditional version of the Clayton copula, with completely unspecified conditional marginals. This specification respects that the marginal and joint distributions for X and Y may not always be identifiable from left-truncated data. The modeling approach has three merits. Firstly, it retains the same flexibility as the standard Lynden-Bell estimator for left-truncated data with independent right censoring. The analyst may choose the parameter a based on the observed pattern of truncation in the data. Secondly, the conditional Clayton model preserves the nice interpretation of the association parameter θ as the predictive hazard ratio (Day et al., 1997) in the observable region. The value of θ is the same in both the conditional model and the unconditional model with $a = 0$. In other words, the concordance probability is not affected by conditioning on event times larger than a. Thirdly, unlike previous methods of estimating event-free probabilities of non-terminal event, our methods do not discard data of subjects whose non-terminal events occur prior to the corresponding truncation times for the terminating event. This extra information is useful in the estimation of θ and $R_{\tilde{a}}(t)$.

In extending Fine et al. (2001) to left-truncation, we used data pairs satisfying

$$(A_i \vee A_j) \vee (X_i \wedge X_j) < Y_i \wedge X_j < C_i \wedge C_j$$

to estimate θ. This ensured the independence between Δ_{ij} and $\tilde{W}(\tilde{X}'_{ij}, \tilde{Y}'_{ij})O_{ij}$ in the concordance estimator $\hat{\theta}$. While this estimator performed well in simulations and the diabetes registry analysis, there may be loss of efficiency by excluding some observed concordance-discordance information using $A_i \vee A_j$. Methods for recovering this information are a topic for future research.

Following Fine et al. (2001), it can be shown that the closed form estimator for $S_{\tilde{a}}$ is robust to the misspecification of the copula model in the lower wedge where $X > Y$. In other simulation studies (not reported), we found that when the two time-to-event variables are associated but follow some bivariate distribution other than the Clayton model in the upper wedge, $\hat{S}_{\tilde{a}}$ is still less biased than the Lynden-Bell estimator. Intuitively, the association parameter in a misspecified model can, at least to some extent, capture the true dependence between the two time-to-event variables, whereas assuming independence completely ignores the relationship between X and Y.

APPENDIX 1: Asymptotic normality of $n^{1/2}(\hat{\theta} - \theta_0)$

A Taylor expansion of $U(\hat{\theta})$ in $\hat{\theta}$ around θ_0 and the consistency of $\hat{\theta}$ give $n^{1/2}(\hat{\theta} - \theta_0) = I^{-1}\{n^{-3/2}U(\theta_0)\} + o_p(1)$, where I is the probability limit of $\hat{I}$. Straightforward calculations show $n^{-3/2}U(\theta_0) = n^{-3/2}\sum_{i<j} Q_{ij} + o_p(1)$, where $Q_{ij} = \tilde{W}(\tilde{X}'_{ij}, \tilde{Y}'_{ij})O_{ij}$ $\{\Delta_{ij} - \theta_0(1+\theta_0)^{-1}\}$. A central limit theorem for U-statistics and Slutsky's law yield the normal distribution for $n^{1/2}(\hat{\theta} - \theta_0)$ with variance $I^{-2}J$, where J is the limit of $\hat{J}$.

Under the null hypothesis, the distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ are centered around the same θ_0. Using the previous results, $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2) = n^{-3/2}\sum_{i<j} Q^*_{ij} + o_p(1)$,

where $Q^*_{ij} = I_1^{-1}Q_{1ij} - I_2^{-1}Q_{2ij}$, $I_k = \lim_{n\to\infty}\hat{I}_k$ and Q_{kij} is Q_{ij} with $\tilde{W}$ replaced by $\tilde{W}_k = \lim_{n\to\infty} W_k$, $k = 1, 2$. A limit theorem for U-statistics gives asymptotic normality, with variance $\Gamma = \lim_{n\to\infty} 2n^{-3}\sum_{k<l<m}(Q^*_{kl}Q^*_{km} + Q^*_{kl}Q^*_{lm} + Q^*_{lm}Q^*_{km})$. A consistent estimator $\hat{\Gamma}$ is computed with $\hat{Q}^*_{ij}$ in place of Q^*_{ij} in Γ.

APPENDIX 2: Variance estimator of $n^{1/2}\{\hat{S}_{\tilde{a}}(t) - S_{\tilde{a}}(t)\}$

Similarly to Fine et al. (2001), it can be shown that $n^{1/2}\{\hat{S}_{\tilde{a}}(t) - S_{\tilde{a}}(t)\}$ is asymptotically equivalent to $n^{-3/2}\sum_{i<j} V_{ij}(t)$, where

$$V_{ij}(t) = -g_1\{H_{\tilde{a}}(t), R_{\tilde{a}}(t), \theta_0\}H_{\tilde{a}}(t)\int_a^t \pi_z(u)^{-1}\{dM_{zi}(u) + dM_{zj}(u)\}$$

$$-g_2\{H_{\tilde{a}}(t), R_{\tilde{a}}(t), \theta_0\}R_{\tilde{a}}(t)\int_a^t \pi_y(u)^{-1}\{dM_{yi}(u)+dM_{yj}(u)\}+g_3\{H_{\tilde{a}}(t), R_{\tilde{a}}(t), \theta_0\}I^{-1}Q_{ij},$$

and

$$g_1(a,b,c) = \partial g(a,b,c)/\partial a = a^{-c}(a^{1-c} - b^{1-c} + 1)^{c/(1-c)},$$

$$g_2(a,b,c) = \partial g(a,b,c)/\partial b = -b^{-c}(a^{1-c} - b^{1-c} + 1)^{c/(1-c)},$$

$g_3(a,b,c) = \partial g(a,b,c)/\partial c =$

$$g(a,b,c)\left[\frac{\log_e(a^{1-c} - b^{1-c} + 1)}{(1-c)^2} + \frac{-a^{1-c}\log_e(a) + b^{1-c}\log_e(b)}{(a^{1-c} - b^{1-c} + 1)(1-c)}\right],$$

$$M_{zi}(t) = I(A_i \vee a < X_i' \le t, \eta_{zi} = 1) - \int_0^t I(X_i' \ge u > A_i \vee a)d\Lambda_z(u) \text{ and}$$

$$M_{yi}(t) = I(a < X_i', A_i < Y_i' \le t, \eta_{yi} = 1) - \int_0^t I(Y_i' \ge u > A_i, X_i' > a)d\Lambda_y(u)$$

are martingales, and $\Lambda_z(u)$ and $\Lambda_y(u)$ are cumulative hazard functions for Z and Y, respectively, given $X > a$ and $Y > a$; π_z and π_y are the limits of

$$\hat{\pi}_z(t) = n^{-1}\sum_{i=1}^n I(X_i' \ge t > A_i \vee a),$$

$$\hat{\pi}_y(t) = n^{-1}\sum_{i=1}^n I(Y_i' \ge t > A_i, X_i' > a).$$

A consistent estimator for the covariance function $\sigma(s,t) = \mathrm{cov}[n^{1/2}\{\hat{S}_{\tilde{a}}(s) - S_{\tilde{a}}(s)\}, n^{1/2}\{\hat{S}_{\tilde{a}}(t) - S_{\tilde{a}}(t)\}]$ is given by

$$\begin{aligned}\hat{\sigma}(s,t) = n^{-3}\textstyle\sum_{k<l<m}\Big\{&\hat{V}_{kl}(s)\hat{V}_{km}(t) + \hat{V}_{lm}(s)\hat{V}_{km}(t) \\ &+\hat{V}_{kl}(s)\hat{V}_{lm}(t) + \hat{V}_{lm}(s)\hat{V}_{kl}(t) + \hat{V}_{km}(s)\hat{V}_{kl}(t) + \hat{V}_{km}(s)\hat{V}_{lm}(t)\Big\},\end{aligned} \tag{5}$$

where

$$\begin{aligned}\hat{V}_{ij}(t) = -g_1(\hat{H}_{\tilde{a}}(t), \hat{R}_{\tilde{a}}(t), \hat{\theta})\hat{H}_{\tilde{a}}(t)\int_a^t \hat{\pi}_z(u)^{-1}\Big\{d\hat{M}_{zi}(u) + d\hat{M}_{zj}(u)\Big\} \\ -g_2(\hat{H}_{\tilde{a}}(t), \hat{R}_{\tilde{a}}(t), \hat{\theta})\hat{R}_{\tilde{a}}(t)\int_a^t \hat{\pi}_y(u)^{-1}\Big\{d\hat{M}_{yi}(u) + d\hat{M}_{yj}(u)\Big\} + g_3(\hat{H}_{\tilde{a}}(t), \hat{R}_{\tilde{a}}(t), \hat{\theta})\hat{I}^{-1}\hat{Q}_{ij},\end{aligned}$$

$$\hat{M}_{zi}(t) = I(A_i \vee a < X_i' \le t, \gamma_i = 1) - \int_0^t I(X_i' \ge u > A_i \vee a)d\hat{\Lambda}_z(u),$$

$$\hat{M}_{yi}(t) = I(a < X_i', A_i < Y_i' \le t, \delta_i = 1) - \int_0^t I(Y_i' \ge u > A_i, X_i' > a)d\hat{\Lambda}_y(u),$$

and $\hat{\Lambda}_z$ and $\hat{\Lambda}_y$ are modified Nelson-Aalen estimators for Λ_z and Λ_y respectively:

$$\hat{\Lambda}_z(t) = \sum_{a<X_i'\le t}\frac{\gamma_i}{n\hat{\pi}_z(X_i')} \quad \text{and} \quad \hat{\Lambda}_y(t) = \sum_{\substack{a<Y_i'\le t \\ a<X_i'}}\frac{\delta_i}{n\hat{\pi}_y(Y_i')}.$$

ACKNOWLEDGEMENTS

This research was supported in part by grant AI24643 from the National Institute of Health. The authors thank Dr. A.R. Andersen for providing the data.

REFERENCES

Andersen A. R., Christiansen J. S., Andersen J. K., Kreiner S., and Deckert T. (1983). Diabetic nephropathy in Type 1 (insulin-dependent) diabetes: an epidemiological study. *Diabetologia* **25**, 496-501.

Andersen, P. K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**, 661-670.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes* Berlin, New York: Springer-Verlag Inc.

Bojestig M., Arnqvist H. J., Hermansson G., Karlberg B. E., and Ludvigsson J. (1994). Declining Incidence of Nephropathy in Insulin-Dependent Diabetes Mellitus. *New England Journal of Medicine* **330**, 15-18.

Borch-Johnsen, K., Kreiner, S., and Deckert, T. (1985). The effect of proteinuria on relative mortality in Type 1 (insulin-dependent) diabetes mellitus. *Diabetologia* **28**, 590-596.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease epidemiology. *Biometrika* **65**, 141-51.

Dahlquist, G., Stattin, E. L., and Rudberg, S. (2001). Urinary albumin excretion rate and glomerular filtration rate in the prediction of diabetic nephropathy; a long-term follow-up study of childhood onset type-1 diabetic patients. *Nephrology Dialysis Transplantation* **16(7)**, 1382-1386.

Day R., Bryant J., and Lefkopoulou M. (1997). Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika* **84**, 45-56.

Fine J. P., Jiang H., and Chappell R. (2001). On semi-competing risks data. *Biometrika* **88**, 907-919.

Gijbels, I., and Wang, J. L. (1993). Strong representations of the survival function estimator for truncated and censored data with applications. *Journal of Multivariate Analysis* **47**, 210-229.

Gu, M. G., and Lai, T. L. (1990). Functional laws of the iterated logarithm for the product-limit estimator of a distribution function under random censorship or truncation. *The Annals of Probability* **18**, 160-189.

Hougaard, P. (2000). *Analysis of multivariate survival data.* New York: Springer-Verlag Inc.

Jiang, H., Chappell, R., Fine, J.P. (2003). Estimating the distribution of nonterminating event times in the presence of mortality or informative dropout. *Controlled Clinical Trial* **24**, 135-146.

Klein, J. P., and Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data.* New York: Springer-Verlag Inc.

Kofoed-Enevoldsen, A., Borch-Johnsen, K., Kreiner, S., Nerup, J., and Deckert, T. (1987). Declining incidence of persistent proteinuria in Type 1 (insulin-dependent) diabetic patients in Denmark. *Diabetes* **36**, 205-209.

Lai, T. L., and Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *The Annals of Statistics* **19**, 417-442.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Roy. Astronom. Soc.* **155**, 95-188.

Mogensen C. E. (1984). Microalbuminuria predicts clinical proteinuria and early mortality in maturity-onset diabetes. *New England Journal of Medicine* **310**, 356-360.

Oakes, D. (1982). A model for association in bivariate survival data. *J. R. Statist. Soc.* B **44**, 414-422.

Oakes, D. (1986). Semiparametric inference in bivariate survival data. *Biometrika* **73**, 353-361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487-493.

Ramlau-Hansen, H., Jespersen, N. C. B., Andersen, P. K., Borch-Johnsen, K., and Deckert, T. (1987). Life insurance for insulin-dependent diabetics. *Scand. Actuar. J.* 19-36.

Remuzzi, G., and Ruggenenti, P. (1998). Prognosis of diabetic nephropathy: how to improve the outcome. *Diabetes Research and Clinical Practice* **39** S49-S53.

Tsai, W. Y., Jewell, N. P., and Wang, M. C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**, 883-886.

Wang, M. C., Jewell, N. P., and Tsai, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics* **14**, 1597-1605.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics* **13**, 163-177.

Table 1. *Comparison of the weighted and unweighted estimators for θ.*

θ	(c, d)	Ave	EmpVar	AveVar	Cov95
		$n = 100$			
1	$(0, 0)$	1.01	0.022	0.029	96.8
	$(x_{.95}, y_{.95})$	1.01	0.018	0.028	98.1
2	$(0, 0)$	2.01	0.108	0.125	96.1
	$(x_{.95}, y_{.95})$	2.02	0.100	0.126	96.6
3	$(0, 0)$	3.04	0.274	0.302	95.4
	$(x_{.95}, y_{.95})$	3.05	0.265	0.302	96.1
		$n = 200$			
1	$(0, 0)$	1.01	0.012	0.015	96.4
	$(x_{.95}, y_{.95})$	1.00	0.009	0.015	98.6
2	$(0, 0)$	2.02	0.060	0.064	96.3
	$(x_{.95}, y_{.95})$	2.02	0.056	0.065	96.9
3	$(0, 0)$	3.04	0.150	0.155	94.9
	$(x_{.95}, y_{.95})$	3.04	0.144	0.155	95.5

Table 2. *Comparison of $\hat{S}^*_{\tilde{a}}$ and the Lynden-Bell estimator.*

				$\hat{S}^*_{\tilde{a}}(t)$			Lynden-Bell	
$S_{\tilde{a}}(t)$	n	Ave	EmpVar	AveVar	%Val	Cov95	Ave	EmpVar
				$\theta = 1$				
0.9	100	0.90	0.790	3.363	100	91.5	0.90	0.770
	200	0.91	0.412	1.291	100	96.0	0.90	0.402
0.7	100	0.71	1.235	1.813	100	96.0	0.70	1.116
	200	0.70	0.622	1.201	100	97.2	0.70	0.568
0.5	100	0.50	1.574	1.906	100	97.4	0.50	1.163
	200	0.50	0.703	1.141	100	97.8	0.50	0.566
0.3	100	0.31	1.660	2.521	100	97.4	0.30	1.071
	200	0.30	0.777	1.481	100	99.0	0.30	0.542
0.1	100	0.11	2.625	3.236	99.6	83.5	0.12	1.062
	200	0.10	1.121	1.784	100	88.2	0.11	0.549
				$\theta = 2$				
0.9	100	0.91	0.489	1.216	100	93.4	0.91	0.418
	200	0.90	0.253	0.618	100	98.2	0.90	0.201
0.7	100	0.71	0.940	1.434	100	97.2	0.74	0.620
	200	0.70	0.532	0.820	100	97.8	0.74	0.305
0.5	100	0.50	1.146	1.363	100	95.7	0.58	0.672
	200	0.50	0.541	0.727	100	96.8	0.58	0.311
0.3	100	0.31	1.032	1.226	100	95.9	0.42	0.671
	200	0.31	0.531	0.584	100	96.0	0.42	0.364
0.1	100	0.13	1.589	6.431	96.3	86.4	0.23	0.894
	200	0.11	0.510	0.737	99.6	92.2	0.23	0.389
				$\theta = 3$				
0.9	100	0.91	0.458	1.213	100	94.1	0.91	0.344
	200	0.90	0.223	0.540	100	97.9	0.91	0.171
0.7	100	0.71	0.970	1.448	100	96.8	0.76	0.508
	200	0.70	0.484	0.775	100	97.6	0.75	0.253
0.5	100	0.51	1.138	1.396	100	96.3	0.61	0.550
	200	0.50	0.529	0.662	100	96.8	0.62	0.261
0.3	100	0.31	0.877	0.978	99.4	95.2	0.47	0.610
	200	0.31	0.452	0.466	100	95.1	0.47	0.293
0.1	100	0.12	0.803	22.70	92.3	88.2	0.27	0.698
	200	0.11	0.302	1.554	98.1	93.4	0.27	0.334

Ave, empirical mean; EmpVar, empirical variance; AveVar, model-based variance; %Val, percentage of valid estimators; Cov95, empirical coverage probability in percentage. Empvar and AveVar are multiplied by 100.

Table 3. *Association between time to DN and time to death in insulin-dependent diabetic patients by birth cohorts.*

	$[1935, 1940)$		$[1945, 1950)$	
n	349		394	
$(\eta, \delta) = (0, 0)$	235	(67.3%)	280	(71.1%)
$(\eta, \delta) = (1, 0)$	24	(6.9%)	62	(15.7%)
$(\eta, \delta) = (0, 1)$	35	(10.0%)	22	(5.6%)
$(\eta, \delta) = (1, 1)$	55	(15.8%)	30	(7.6%)
$\hat{\theta}$ unweighted[1]	8.03	(5.5, 10.6)	5.51	(3.1, 7.8)
$\hat{\theta}$ weighted[1]	8.77	(6.2, 11.4)	5.44	(3.1, 7.8)
Lack-of-fit test	1.4	(p = 0.16)	0.12	(p = 0.9)

[1] The figures in the parentheses are 95% confidence intervals for θ.

Fig. 1. Estimated curves of DN-free probabilities by birth cohorts.

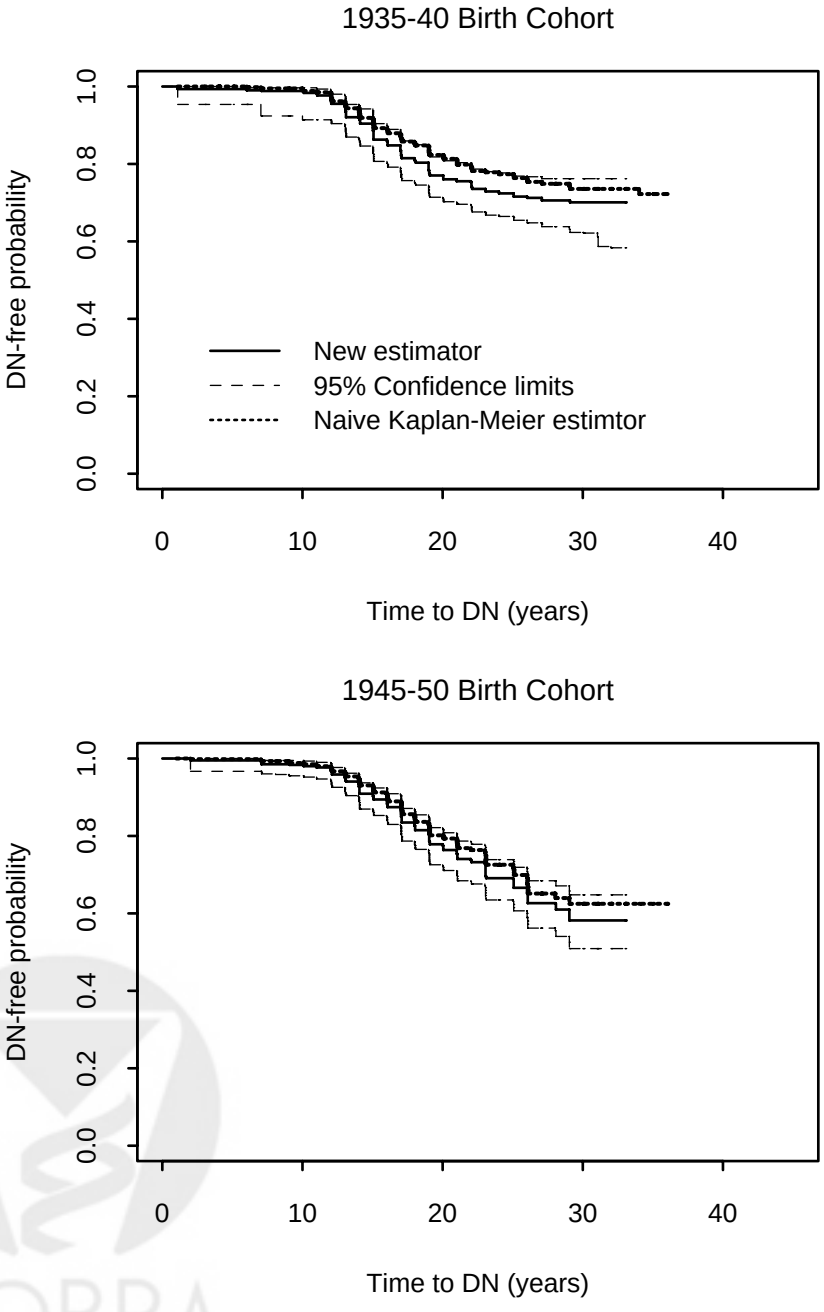

Fig. 2. Comparison of event-free probabilities by birth cohorts.