Nonparametric estimation with left truncated semi-competing risks data

Limin Peng and Jason P. Fine
Department of Statistics, University of Wisconsin, Madison, WI
Nonparametric estimation with left truncated semi-competing risks data

Department of Statistics, University of Wisconsin, Madison, WI, 53706 U.S.A.

SUMMARY

Cause-specific hazard and cumulative incidence function are of practical importance in competing risks studies. Inferential procedures for these quantities are well developed and can be applied to semi-competing risks data, where a terminating event censors a non-terminating event, after coercing the data into the competing risks format. Complications arise when there is left truncation of the terminating event, as often occurs in observational studies. The competing risks analysis naively truncates the non-terminating event using the left truncation time for the terminating event, which may lead to large efficiency losses. We propose simple nonparametric estimators which use all semi-competing risks information and do not require artificial truncation. The uniform consistency and weak convergence of the estimators are established and variance estimators are provided. Simulation studies and an analysis of a diabetes registry demonstrate large efficiency gains over the naive estimators.

Key Words: Counting process; Cumulative incidence; Dependent censoring; Left truncation; Uniform consistency; Weak convergence.
1. INTRODUCTION

Semi-competing risks data (Fine, Jiang & Chappell, 2001) occur when an event time $T_1$ can be dependently censored by another event time $T_2$ but not vice versa. The data is encountered with chronic diseases, where morbidity may be censored by mortality. As in the classic competing risks set-up, inference about $T_1$ is complicated by dependence between $T_1$ and $T_2$. Analyses of the marginal distribution of $T_1$ are problematic. The distribution is nonparametrically nonidentifiable (Tsiatis, 1975) and its interpretation hypothesizes the elimination of $T_2$, e.g., death. Nonparametric analyses of identifiable quantities, including cause-specific hazard and cumulative incidence functions, have been widely adopted (Prentice et al, 1978). To employ such analyses with semi-competing risks data, the convention is to ignore $T_2$ occuring subsequent to $T_1$ and to use only the competing risks information, $T = T_1 \wedge T_2$ and $\epsilon = (i : T = T_i)$, where $\wedge$ denotes minimum.

In observational studies, complications may arise when $T_2$ is left truncated at a time $L$ and $(T, \epsilon)$ is observed conditionally on $T_2 > L$. An example is a Denmark diabetes registry (Andersen et al, 1993), consisting of insulin-dependent diabetes patients referred to Steno Memorial Hospital in Greater Copenhagen. The cumulative incidence of diabetic nephropathy, a common morbidity, is helpful in characterising the disease process. The analysis must account for the facts that time to death, $T_2$, may dependently censor time to nephropathy, $T_1$, and that only patients living long enough to enter the registry provide data, that is, their data is available when $T_2$ is larger than the entry time $L$. Since some patients had nephropathy at first admission to the registry, where $T_1 < L$, application of the competing risks analysis is not straightforward. One must naively left truncate $T_1$ using $L$ in order that the left truncation time for $T$ is clearly defined. From the discussion in Andersen et al (Ch. III.3, 1993), it seems this ad hoc truncation is the standard approach to such data. Of course, much information on $T_1$ may be lost. In the analysis in §3.2, 20-30% of observed nephropathy events are excluded. The goal of this paper is to use all semi-competing risks information in estimation of the cause-specific hazard and cumulative incidence functions.

In §2.1, the estimators from the artificially truncated competing risks framework are reviewed. It is not obvious that these techniques can be modified to semi-competing risks data without imposing artificial truncation. In §2.2, we introduce nonparametric estimators for cause-specific hazard and cumulative incidence functions which employ alternative data formulations not requiring such truncation. The estimators’ asymptotic properties are given in §2.3, with proofs in the Appendix. Identifiability of the cumulative incidence function posits assumptions on the supports of the failure, censoring, and truncation variables,
analogous to those for univariate left truncated right censored data (Tsai et al, 1987). Conditional estimators are presented in §2.4 that are valid under weaker conditions. A simulation study reported in §3.1 shows that the new estimators may achieve two-three fold variance reductions over the naive estimators in scenarios akin to those in the diabetes registry. The registry data is analysed in §3.2 and some remarks conclude in §4.

2. NONPARAMETRIC ESTIMATION

2.1 Naive competing risks estimators

In practice, there is typically an independent censoring time, C. Without truncation, the observed semi-competing risks data is \( X = T \wedge C, Y = T_2 \wedge C, \delta = I(T_1 \leq Y), \) and \( \eta = I(T_2 \leq C), \) where \( I(\cdot) \) is the indicator function. With truncation, the observed data consists of \( n \) independent and identically distributed replicates of \( (X, Y, \delta, \eta, L) \), denoted by \( \{(X_i, Y_i, \delta_i, \eta_i, L_i)\}_{i=1}^n \), sampled conditionally on \( Y_i \geq L_i, i = 1, \ldots, n \). It is assumed that \( (L, C) \) is independent of \( (T_1, T_2) \).

Inference about the marginal survivor function of \( T_2, S_{T_2}(t) = \Pr(T_2 > t) \), can be based on the Lynden-Bell (1971) product limit estimator for left truncated right censored data using \( \{(Y_i, \eta_i, L_i), i = 1, \ldots, n\} \), sampled conditionally on \( Y_i \geq L_i, i = 1, \ldots, n \). This estimator has been thoroughly studied (Tsai et al., 1987; Gu & Lai, 1990; Lai & Ying, 1991; Gijbels & Wang, 1993). One may also use the Nelson-Aalen estimator with appropriately defined risk sets to estimate the marginal hazard function for \( T_2 \) (Andersen et al, 1993).

It is tempting to apply similar techniques to estimate the cause-specific hazard and the cumulative incidence function for the nonterminating event, corresponding to

\[
\lambda_1(t) = \lim_{h \to 0} \Pr(t \leq T < t + h, \epsilon = 1 | T \geq t) = \lim_{h \to 0} \Pr(t \leq T_1 < t + h, T_2 > t | T_1 \geq t, T_2 \geq t)
\]

and

\[
F_1(t) = \Pr(T \leq t, \epsilon = 1) = \Pr(T_1 \leq t, T_2 > T_1),
\]

respectively. To do so, the semi-competing risks data must be “coerced” into a competing risks set-up (Andersen et al, Ch. III.3, 1993).

To define the estimators, we use the counting process notation \( R_i(t) = I(L_i \leq t \leq X_i) \), \( N_{1i}(t) = I(L_i \leq X_i \leq t, \delta_i = 1) \), \( N_{2i}(t) = I(L_i \leq X_i \leq t, \delta_i = 0, \eta_i = 1) \), \( R(t) = \sum_{i=1}^n R_i(t) \), \( \tilde{N}_j(t) = \sum_{i=1}^n N_{ji}(t) \), \( j = 1, 2 \). The Nelson-Aalen type estimator for \( \Lambda_1(t) = \int_0^t \lambda_1(s)ds \) is

\[
\hat{\Lambda}_1(t) = \int_0^t \frac{d\tilde{N}_1(s)}{R(s)} = \sum_{i=1}^n \frac{I(X_i \leq t, L_i \leq X_i, \delta_i = 1)}{\sum_{j=1}^n I(L_j \leq X_i \leq X_j)}.
\]
This estimator may be combined with the right censored left truncated product limit estimator for \( S_T(t) = \Pr(T > t) \) using \( \{(X_i, I(T_i < C_i), L_i), i = 1, \ldots, n\} \),
\[
\hat{S}_T(t) = \prod_{X_i \leq t} \left[ 1 - \{d\hat{N}_1(X_i) + d\hat{N}_2(X_i)\}/\hat{R}(X_i) \right],
\]
to give an estimator for \( F_1(t) = \int_0^t S_T(u^-) \lambda_1(u) du, \hat{F}_1(t) = \int_0^t \hat{S}_T(u^-) d\hat{\Lambda}_1(u) \). Note that these estimators only employ data with \( X_i \geq L_i \). The removal of \( X_i \)'s which are smaller than the left truncation times for \( Y_i \)'s may incur considerable information loss. It is unclear that \( \hat{\Lambda}_1 \) and \( \hat{F}_1 \) can be modified to recover this information.

2.2 Proposed estimators

Our strategy is to use all available data instead of reducing it to a competing risks structure. Precursors to this work include earlier investigations of independently censored univariate and bivariate truncated data (Woodroofe, 1985; Gurler, 1996; Gijbels & Gurler, 1998). These papers considered nonparametric estimation of the univariate and bivariate distribution functions, while our focus is the estimation of \( \Lambda_1 \) and \( F_1 \). This involves a careful examination of the connections between the true joint distribution of \( (T_1, T_2) \), the conditional joint distribution under the left truncation mechanism, and \( \Lambda_1 \) and \( F_1 \). These relationships are rather subtle in the presence of both dependent censoring of \( T_1 \) by \( T_2 \) and independent censoring of \( T_1 \) and \( T_2 \) by \( C \). The resulting estimators of the cumulative cause-specific hazard and cumulative incidence functions must account for these complex censoring patterns.

Let \( H(y, x) = \Pr(T_2 > y, T_1 > x) \) and \( \tilde{H}(y, x) = \Pr(T_2 > y, T > x) \) and note that since \( \tilde{H}(y, x) = H(y, x) \) for \( y \geq x \), \( F_1(t) = \int f_y \int f_y H(du, dv) = \int f_y \int f_y \tilde{H}(du, dv) \). As shown in Fine et al (2001), \( \tilde{H} \) may be estimated directly when there is only right censoring by \( C \). However, it is not directly estimable in the presence of left truncation because all observables are now sampled from the conditional distribution given \( L \leq Y \). To proceed further, we consider the transformed bivariate survival function of \( (Y, X) \), \( F^*(y, x) = \Pr(Y > y, X > x, \eta = 1|L \leq Y) \). It is easy to show that \( F^*(y, x) = \alpha^{-1} \int_x \int_y G(u) \tilde{H}(du, dv), \) where \( \alpha = \Pr(L \leq Y) \) and \( G(y) = \Pr(L \leq y \leq C) \), which implies \( \tilde{H}(dy, dx) = \alpha F^*(dy, dx)/G(y) \). An estimator of \( \alpha/G(y) \) is obtained by substituting empirical quantities in \( \Pr(L \leq y \leq Y|L \leq Y) = \alpha^{-1} G(y) \Pr(T_2 \geq y) \). This suggests the following estimator of \( F_1(t) \),
\[
\int_0^t \int \frac{\hat{S}_{T_2}(u)}{C_n(u)} F_n^*(du, dv) = n^{-1} \sum_{i=1}^n \frac{\hat{S}_{T_2}(Y_i^-)}{C_n(Y_i)} I(X_i \leq t, X_i < Y_i, \eta_i = 1),
\]
where \( C_n(y) = n^{-1} \sum_{i=1}^n I(L_i \leq y \leq Y_i) \), \( F_n^*(y, x) \) is the empirical bivariate distribution
function of observed \((Y, X)\) with \(\eta = 1\), and

\[
\hat{S}_{T_2}(y) = \prod_{y_i \leq y} \left\{ 1 - \frac{\sum_{j=1}^n I(Y_j = Y_i, \eta_j = 1)}{\sum_{j=1}^n I(L_j \leq Y_i \leq Y_j)} \right\}.
\]

Technical issues arise because the estimator requires integration over an infinite region. Estimation is not possible when the upper bound on the support of \(C\) is less than that of \(T\). The problem can be seen more clearly by noting that as \(n \to \infty\), the numerator and denominator of \(\hat{S}_{T_2}(u)/\hat{C}_n(u)\) may converge to 0 for suitably large values of \(u\). Hence, a valid estimator cannot be derived without imposing stronger conditions on the failure, truncation, and censoring times. To circumvent this issue, we estimate \(F_1(t)\) in two parts: \(\Pr(T_1 \leq t, T_1 < T_2 \leq \tau)\) and \(\Pr(T_1 \leq t, T_2 > \tau)\), where \(\tau > t\) is a predetermined constant; precise conditions on \(\tau\) are deferred until the statement of the asymptotic results. These terms can be estimated consistently and can be added to give an estimator for \(F_1(t)\).

Reasoning as above shows that \(\Pr(T_1 \leq t, T_1 < T_2 \leq \tau)\) can be estimated by

\[
n^{-1} \sum_{i=1}^n \frac{\hat{S}_{T_2}(Y_i^-)}{C_n(Y_i)} I(X_i \leq t, X_i < Y_i \leq \tau, \eta_i = 1).
\]

This yields an integral over a compact region and does not involve tail estimation. In practice, \(\tau\) may be chosen to be slightly smaller than \(Y_{(n)} = \max_i Y_i\), which will tend towards the upper bound on the support of \(Y\) as \(n \to \infty\).

To estimate \(\Pr(T_1 \leq t, T_2 > \tau)\), notice that \(\Pr(T_1 \leq t, T_2 > \tau) = \Pr(T_2 > \tau) - \Pr(T_1 > t, T_2 > \tau)\) and for \(x > y\), \(\Pr(T_1 > x | T_2 > y) = \Pr(L \leq y < Y, X > x | L \leq Y) / \Pr(L \leq y < Y | L \leq Y)\). This suggests an estimator of \(\Pr(T_1 \leq t, T_2 > \tau)\) that takes the form

\[
\hat{S}_{T_2}(\tau) \left\{ 1 - \frac{C_{2,n}(\tau^+, t)}{C_n(\tau^+)} \right\}
\]

where \(C_{2,n}(y, x) = n^{-1} \sum_{i=1}^n I(L_i \leq y < Y_i, X_i > x)\).

The proposed estimator of the cumulative incidence function is then

\[
\hat{F}_1(t) = n^{-1} \sum_{i=1}^n \frac{\hat{S}_{T_2}(Y_i^-)}{C_n(Y_i)} I(X_i \leq t, X_i < Y_i \leq \tau, \eta_i = 1) + \hat{S}_{T_2}(\tau) \left\{ 1 - \frac{C_{2,n}(\tau^+, t)}{C_n(\tau^+)} \right\}.
\]

The new estimator \(\hat{F}_1\) does not require artificial truncation on \(X\), unlike \(\hat{F}_1\), and may be viewed as a weighted sum of the data over the feasible region, where \(\{T_1 \leq t, T_1 < T_2\}\). An estimator for the cumulative cause-specific hazard function is given by

\[
\hat{A}_1(t) = \sum_{i=1}^n \left[ \frac{\hat{S}_{T_2}(Y_i^-)}{C_n(Y_i)} I(X_i < Y_i \leq \tau, X_i \leq t, \eta_i = 1) + \frac{\hat{S}_{T_2}(\tau)}{C_n(\tau)} I(L_i \leq \tau < Y_i, X_i \leq t) \right] \left[ \sum_{j=1}^n \left\{ \frac{S_{T_2}(Y_j^-)}{C_n(Y_j)} I(X_j \leq Y_j \leq \tau, X_j \geq X_i, \eta_j = 1) + \frac{S_{T_2}(\tau)}{C_n(\tau)} I(L_j \leq \tau < Y_j, X_j \geq X_i) \right\} \right].
\]
It is easy to check that $\hat{F}_n(t) = \sum_{i=1}^{n} \hat{S}_{T_i}(Y_i) I(t \leq Y_i \leq \tau, X_i \geq t, \eta_i = 1) + \frac{\hat{S}_{T_i}(\tau)}{C_n(\tau)} I(L_i \leq \tau < Y_i, X_i \geq t)$.

Note that the restriction $a_L \leq a_X$ and $b_L \leq b_X$ is needed for estimating $S_T, A_1$, and $F_1$ using the naive competing risks approach, where for a nonnegative random variable $Z$, $a_Z$ and $b_Z$ are defined as $a_Z = \inf\{z \geq 0 : \Pr(Z \leq z) > 0\}$ and $b_Z = \sup\{z \geq 0 : \Pr(Z \leq z) < 1\}$. Modified versions of $\hat{S}_T, \hat{A}_1$ and $\hat{F}_1$ may be computed by only using observations with $X_i > (L_i \wedge K)$, where $K$ is some constant $> a_L$, and may be interpreted conditionally on $T > K$. Interestingly, it can be proved that when using all the semi-competing risks data, $A_1(t)$ and $F_1(t)$ are identifiable under the weaker condition $a_L \leq a_Y$ and $b_L \leq b_Y$. This condition is identical to that in Tsai et al. (1987) for univariate data under independent left truncation and right censoring and ensures the identifiability of $S_T$. In §2.4, we discuss conditional versions of the estimators $A_1$ and $\hat{F}_1$ which do not require these assumptions.

2.3 Large sample properties

In this subsection, we present asymptotic results for $\hat{F}_n(t)$. A strong representation for $\hat{F}_n(t)$ is established in Theorem 1. The uniform consistency and weak convergence follow immediately and are stated in Corollary 1. Detailed proofs are given in the Appendix.

To state the results, the following notation is needed. Let $C(y) = \Pr(L \leq y \leq Y|L \leq Y)$ and $C_2(y, x) = \Pr(L \leq y \leq Y, X > x|L \leq Y)$. Define $F^*_Y(y) = \Pr(Y \leq y, \eta = 1|L \leq Y)$,

$$\bar{L}_i(y) = \frac{I(Y_i \leq y, \eta_i = 1)}{C(Y_i)} - \int_y^\tau \frac{I(L_i \leq u \leq Y_i)}{\{C(u)\}^2} F^*_Y(du)$$

and $\bar{L}_n(y) = n^{-1} \sum_{i=1}^{n} \bar{L}_i(y)$. Let $F^*_Y, F^*$ be the empirical counterparts of $F^*_Y$ and $F^*$. Without loss of generality, we assume that $C(\cdot), C_2(\cdot, t)$ are continuous at $\tau$.

Theorem 1. Suppose $H$ is continuous in both components. Under condition C1-3, the following representation holds:

$$\hat{F}_n(t) - F_1(t) = \int_0^\tau \int_u^\tau \frac{S_{T_2}(u)}{C(u)} \left\{ F^*_n(du, dv) - F^*(du, dv) \right\} - \int_0^\tau \int_u^\tau \frac{S_{T_2}(u)\bar{L}_n(u)}{C(u)} F^*(du, dv)$$

$$- \int_0^\tau \int_u^\tau \frac{S_{T_2}(u)\{C_n(u) - C(u)\}}{\{C(u)\}^2} F^*(du, dv) - S_{T_2}(\tau) \left\{ 1 - \frac{C_2(\tau, t)}{C(\tau)} \right\} \bar{L}_n(\tau)$$

$$- \frac{S_{T_2}(\tau)}{C(\tau)} \{C_{2,n}(\tau, t) - C_2(\tau, t)\} + \frac{S_{T_2}(\tau)C_2(\tau, t)}{\{C(\tau)\}^2} \{C_n(\tau) - C(\tau)\} + R_n(t)$$

where $\sup_{0 \leq \tau \leq \tau} |R_n(t)| = o(n^{-1/2})$, almost surely.
The following corollary states the uniform consistency of \( \tilde{F}_1(t) \) as well as its weak convergence in \( D\{0, \tau\} \), which is immediate from Theorem 1 and the Glivenko-Cantelli theorem.

**Corollary 1.** Under the conditions of Theorem 1, \( \sup_{0 \leq t \leq \tau} |\tilde{F}_1(t) - F_1(t)| \rightarrow 0 \), a.s. In addition, for \( 0 \leq t < \tau \), \( \sqrt{n} |\tilde{F}_1(t) - F_1(t)| \) converges weakly to a tight Gaussian process with mean 0 and covariance \( \Sigma(s, t) = E\{i_j(s) \tilde{i}_j(t)\} \) where

\[
\begin{align*}
t_j(t) &= \left\{ \frac{S_{T_2}(Y_j)}{C(Y_j)} I(X_j \leq t, X_j < Y_j \leq \tau, \eta_j = 1) - \int_t^\tau \int_v \frac{S_{T_2}(u)}{C(u)} F^*(du, dv) \right\} \\
&\quad - \int_t^\tau \int_v \frac{S_{T_2}(u) \hat{L}_j(u)}{C(u)} F^*(du, dv) - \int_t^\tau \int_v \frac{S_{T_2}(u)}{C(u)^2} \{I(L_j \leq u \leq Y_j) - C(u)\} F^*(du, dv) \\
&\quad - \frac{S_{T_2}(\tau)}{C(\tau)^2} \{1 - \frac{C_2(\tau, t)}{C(\tau)}\} \hat{L}_j(\tau) - \frac{S_{T_2}(\tau)}{C(\tau)} \{I(X_j > t, L_j \leq \tau \leq Y_j) - C_2(\tau, t)\} \\
&\quad + \frac{S_{T_2}(\tau) C_2(\tau, t)}{C(\tau)^2} \{I(L_j \leq \tau \leq Y_j) - C(\tau)\}.
\end{align*}
\]

A consistent estimator of \( \Sigma(s, t) \) is \( \tilde{\Sigma}(s, t) = n^{-1} \sum_{j=1}^n i_j(s) \tilde{i}_j(t) \), where \( \tilde{i}_j(t) \) may be obtained by plugging \( \tilde{T}_2, C_n, F_n, C_{2,n} \), and other empirical quantities in \( t_j(t) \).

Proofs similar to those for Theorem 1 give that \( \sup_{0 \leq t \leq b_X \wedge \tau} |\hat{\Lambda}_1(t) - \Lambda_1(t)| \rightarrow 0 \), almost surely, and for \( 0 \leq t < b_X \wedge \tau \), \( \sqrt{n} |\hat{\Lambda}_1(t) - \Lambda_1(t)| \) converges weakly to a tight Gaussian process with covariance function \( \hat{\Sigma}(s, t) = E\{\hat{i}_j(s) \hat{i}_j(t)\} \), where

\[
\begin{align*}
\hat{i}_j(t) &= \left\{ \frac{S_{T_2}(Y_j)}{C(Y_j)} I(X_j < Y_j \leq \tau, X_j \leq t) - \int_t^\tau \int_v \frac{S_{T_2}(u)}{C(u)H(v, v)} F^*(du, dv) \right\} \\
&\quad - \int_t^\tau \int_v \frac{S_{T_2}(u) \hat{L}_j(u)}{C(u)H(v, v)} F^*(du, dv) - \int_t^\tau \int_v \frac{S_{T_2}(u)}{C(u)^2} \{I(L_j \leq u \leq Y_j) - C(u)\} F^*(du, dv) \\
&\quad - \int_t^\tau \int_v \frac{S_{T_2}(u) \hat{i}_j(u)}{C(u)^2 H(v, v)} F^*(du, dv) \\
&\quad + \left\{ \frac{S_{T_2}(\tau)}{C(\tau)} I(L_j \leq \tau < Y_j, X_j \leq t) - \int_t^\tau \frac{S_{T_2}(\tau)}{C(\tau)H(v, v)} F^{**}(dv) \right\} \\
&\quad - \int_t^\tau \frac{S_{T_2}(\tau) \hat{L}_j(\tau)}{C(\tau)H(v, v)} F^{**}(dv) - \int_t^\tau \frac{S_{T_2}(\tau)}{C(\tau)^2} \{I(L_j \leq \tau \leq Y_j) - C(\tau)\} F^{**}(dv) \\
&\quad - \int_t^\tau \frac{S_{T_2}(\tau) \hat{i}_j(u)}{C(\tau)^2 H(v, v)} F^{**}(dv),
\end{align*}
\]

\[
\begin{align*}
t_j, H(t) &= \left\{ \frac{S_{Y}(Y_j)}{C(Y_j)} I(t < Y_j \leq \tau, X_j \geq t, \eta_j = 1) - \int_t^\tau \frac{S_{Y}(u)}{C(u)} F^*(du, dv) \right\} \\
&\quad - \int_t^\tau \frac{S_{Y}(u) L_{j,n}(u)}{C(u)} F^*(du, dv) - \int_t^\tau \frac{S_{Y}(u)}{C(u)^2} \{I(L_j \leq u \leq Y_j) - C(u)\} F^*(du, dv) \\
&\quad + \left\{ \frac{S_{T_2}(\tau)}{C(\tau)} I(L_j \leq \tau \leq Y_j, X_j \geq t) - \int_t^\tau \frac{S_{T_2}(\tau)}{C(\tau)} F^{**}(dv) \right\}.
\end{align*}
\]
\[- \int_t^\infty \frac{S_{T_2}(\tau)L_{j,a}(\tau)}{C(\tau)} F^{**}(dv) - \int_t^\infty \frac{S_{T_2}(\tau)}{C(\tau)} \{I(L_j \leq \tau \leq Y_j) - C(\tau)\} F^{**}(dv) \]

and \(F^{**}(t) = \Pr(L \leq \tau < Y, X \leq t|L \leq Y)\).

2.4 Conditional cumulative incidence estimators

Without restricting the support of \((T_1, T_2, L, C)\), nonparametric identification of \(F_1(t)\) is not possible, similarly to the survival function with univariate left truncated right censored data (Tsai et al., 1987). We now propose conditional estimators which do not require these assumptions. Because of the complex truncation mechanism, the developments are not as straightforward as those in Tsai et al. (1987). The issue is that two meaningful definitions for the conditional cumulative incidence function are possible: \(F_1^{cl}(t|a) = \Pr(T_1 \leq t, T_2 > T_1|T \geq a)\) and \(F_1^{c2}(t|a) = \Pr(T_1 \leq t, T_2 > T_1|T_2 \geq a)\), where the choice of \(a\) is discussed below. Both definitions lead to consistent estimators in the region \(t > a\).

Estimation of \(F_1^{cl}\) can be accomplished by modifying either \(\hat{F}_1(t)\) or \(\hat{F}_1(t)\). Simply replacing \(\hat{S}_T(\cdot|a)\) in \(F_1^{cl}(t|a) = \int_a^t \hat{S}_T(u|a)d\hat{\Lambda}_1(t)\) gives a valid estimator \(\hat{F}_1^{cl}\), where \(\hat{S}_T(\cdot|a)\) is a modified Lynden-Bell estimator for \(S_T(\cdot|a) = \Pr(T > t|T > a)\) (Tsai et al., 1987). Note that artificial truncation on \(T_1\) is still enforced. As an alternative, simple algebra shows that for \(a_L < a < b_X \wedge \tau\) satisfying \(C_2(a, a) > 0\) and \(t > a\),

\[
F_1^{cl}(t|a) = \frac{C(a)}{C_2(a, a)} \left[ \int_a^t \int_r^\infty \frac{S_{T_2}(u|a)}{C(u)} F^{**}(du, dv) + \frac{S_{T_2}(\tau|a)}{C(\tau)} \{C_2(\tau, a) - C_2(\tau, t)\} \right].
\]

This formula suggests an estimator \(\hat{F}_1^{cl}(t)\), in which \(C, C_2,\) and \(F^{**}\) are replaced by their empirical counterparts and \(S_{T_2}(\cdot|a)\) by \(\hat{S}_{T_2}(\cdot|a)\), where \(\hat{S}_{T_2}(\cdot|a)\) is a modified Lynden-Bell estimator for \(S_{T_2}(\cdot|a) = \Pr(T_2 > t|T_2 > a)\). Similarly to \(\hat{F}_1\), all available data is utilised.

It is easy to see that an estimator of \(F_1^{c2}(t|a)\) cannot be obtained using only the competing risks aspects of the semi-competing risks data, since the conditioning is only on \(T_2\). However, it is straightforward to modify \(\hat{F}_1\). The estimator takes the form

\[
\hat{F}_1^{c2}(t|a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{S}_{T_2}(Y_i|a)}{C_2(Y_i)} I(X_i < Y_i \leq \tau, a \leq Y_i, X_i \leq t, \eta_i = 1) + \hat{S}_{T_2}(\tau|a) \left\{1 - \frac{C_2(a, \tau, t)}{C_2(\tau)} \right\}.
\]

A proof like that for Theorem 1 establishes the uniform consistency and weak convergence of \(\hat{F}_1^{cl}(t)\) and \(\hat{F}_1^{c2}(t)\). The results are stated in Corollary 2. The covariance functions of \(\hat{F}_1^{cl}(t)\) and \(\hat{F}_1^{c2}(t)\) can be consistently estimated with plug-in estimators.

Corollary 2. Suppose \(H\) is continuous in both components and conditions C2-3 hold. For \(a_L < a < b_X \wedge \tau\) and \(C(a) > 0\), \(\sup_{a \leq \tau} |\hat{F}_1^{cl}(t) - F_1^{cl}(t)| \to 0\) almost surely, and \(\sup_{a \leq \tau} |\hat{F}_1^{c2}(t) - F_1^{c2}(t)| \to 0\) almost surely. Furthermore, \(\sqrt{n} \{\hat{F}_1^{cl}(t) - F_1^{cl}(t)\}\) and
\[ \sqrt{n} \{ \hat{F}_1^2(t) - F_1^2(t) \} \text{ converge weakly in } D([a, \tau]) \text{ to tight Gaussian processes with zero mean and covariance } \Sigma^{s1}(s, t) = E\{ \hat{t}_j^{s1}(s) \hat{t}_j^{s1}(t) \} \text{ and } \Sigma^{s2}(s, t) = E\{ \hat{t}_j^{s2}(s) \hat{t}_j^{s2}(t) \}, \] respectively, where

\[ \hat{t}_j^{s1}(t) = \frac{C(a)}{C_2(a, a)} \left\{ \frac{S_{T_2}(Y_j|a)}{C(Y_j)} I(X_j < Y_j \leq \tau, a \leq X_j \leq t, \eta_j = 1) - \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} F^*(du, dv) \right\} 
- \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} L_{j,n}(u) F^*(du, dv) - \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} \{ I(L_j \leq u \leq Y_j) - C(u) \} F^*(du, dv) \]
- \frac{S_{T_2}(\tau|a)}{C(\tau)} \left\{ 1 - \frac{C_2(\tau, t)}{C(\tau)} \right\} \left[ \hat{L}_j(\tau) - \frac{S_{T_2}(\tau|a)}{C(\tau)} \{ I(X_j > t, L_j \leq \tau \leq Y_j) - C_2(\tau, t) \} \right] + \frac{F_1^{s1}(t)}{C(\tau)^2} \{ I(L_j \leq \tau \leq Y_j) - C(\tau) \} \]
- \frac{F_1^{s1}(t)}{C_2(a, a)} \{ I(L_j \leq \alpha \leq Y_j, X_j > a) - C_2(a, a) \}

and

\[ \hat{t}_j^{s2}(t) = \left\{ \frac{S_{T_2}(Y_j|a)}{C(Y_j)} I(X_j < Y_j \leq \tau, a \leq Y_j, X_j \leq t, \eta_j = 1) - \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} F^*(du, dv) \right\} 
- \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} \hat{L}_j(u) F^*(du, dv) - \int_a^t \int_v^\tau \frac{S_{T_2}(u|a)}{C(u)} \{ I(L_j \leq u \leq Y_j) - C(u) \} F^*(du, dv) \]
- \frac{S_{T_2}(\tau|a)}{C(\tau)} \left\{ 1 - \frac{C_2(\tau, t)}{C(\tau)} \right\} \left[ L_{j,n}(\tau) - \frac{S_{T_2}(\tau|a)}{C(\tau)} \{ I(X_j > t, L_j \leq \tau \leq Y_j) - C_2(\tau, t) \} \right] + \frac{S_{T_2}(\tau|a)C_2(\tau, t)}{C(\tau)^2} \{ I(L_j \leq \tau \leq Y_j) - C(\tau) \}.

3. NUMERICAL STUDIES

3.1 Simulations

Simulations are performed to compare the small sample behavior of \( \hat{F}_1(t) \) and \( \hat{F}_1(t) \). The failure times \((T_1, T_2)\) are generated from a gamma frailty model, where

\[ \Pr(T_1 > x, T_2 > y) = \left[ \left\{ \Pr(T_1 > x) \right\}^{(1-\theta)} + \left\{ \Pr(T_2 > y) \right\}^{(1-\theta)} - 1 \right]^{1/(1-\theta)} \]

with \( T_i \) following a Weibull(\( \gamma_i, \alpha_i \)) distribution with \( \Pr(T_i > x) = \exp(-\gamma_i x^{\alpha_i}), i = 1, 2 \) and \( \theta = 1.5 \). The truncation time \( L \) is generated from a mixture of a point mass at zero and a positive-valued random variable, mimicking the registry analysis in §3.2. The proportion of zero truncation times is either 0% or 20%. Conditional on \( L \), the right censoring variable \( C \) is \( \text{Unif}[L, L+c] \), where \( c \) is chosen to give 15% independent censoring of \( T_2 \). The simulations are conducted under two set-ups: (A) \( T_1 \sim \text{Weibull}(4, 1), T_2 \sim \text{Weibull}(4, 0.28), \hat{L} \sim \text{Unif}(0, 2) \)
and $c = 4.5$; (B) $T_1 \sim \text{Weibull}(3, 3.2)$, $T_2 \sim \text{Weibull}(3, 2)$, $\bar{L} \sim \text{Weibull}(0.9, 0.25)$ and $c = 3$; where $\bar{L}$ refers to the nonzero component of $L$. The sample size $n = 100$ or 200.

The summary statistics for the censoring and truncation patterns based on 1000 simulated datasets are presented in Table 1. Under (A), there is a low truncation level ($Pr(Y \leq L) = 1 - \alpha$), a moderate dependent censoring proportion ($Pr(\delta = 0, \eta = 1)$) and a high proportion of artificial truncation $Pr(X \leq L, Y \geq L)$. For set-up (B), there is a high truncation level, a high dependent censoring proportion, and a moderate artificial truncation proportion.

Tables 2–3 present the averages of $\hat{F}_1(t)$ (Ave), the empirical variance of $\hat{F}_1(t)$ (EmpVar), the average of the estimated variances of $\hat{F}_1(t)$ (AveVar) and the coverage probabilities of the nominal 95% confidence interval for $F_1(t)$ (Cov95) at time points corresponding to $F_1(t) = 0.1, 0.2, 0.3, 0.4$. For comparison, the averages of $\hat{F}_1(t)$ and its empirical variances are also reported. In all cases $\hat{F}_1(t)$ and $\hat{F}_1(t)$ are virtually unbiased. The variance estimates for $\hat{F}_1(t)$ are in good agreement with the empirical variances. The coverage probabilities are close to the nominal value, with their performance improving as $n$ increases.

The efficiency loss associated with artificial truncation is of substantive interest. Not surprisingly, $\hat{F}_1(t)$ generally has smaller variance than $\hat{F}_1(t)$. It is interesting that the magnitude of the variance reductions may be quite large. At early time points, two-three fold variance reductions are observed, with reductions ranging from 30-60%. At later time points, smaller but still substantial reductions of 20-30% are evident. These differences may be because artificial truncation tends to remove small $X$ which contribute more information to the estimation of the cumulative incidence function in the lower tail. The improvements with $\hat{F}_1$ seem to increase as the artificial truncation proportion increases.

### 3.2 Real data example

The focus of this analysis is to estimate the cumulative incidence of time to DN using the registry data from Andersen et al (1993). The registry includes all patients referred to Steno Memorial Hospital in Greater Copenhagen between 1931 and 1988 that were diagnosed before age 31 and between 1933 and 1972. The time origin is the age at diabetes diagnosis, with event times recorded in years since diagnosis. We restrict attention to patients born between 1915 and 1925 with disease onset age between 10 and 25 years. This cohort includes 196 patients and is summarized in Table 4. Observe that 26% (17/65) of DN events among male patients took place before admission to Steno, while the corresponding percentage in females is 21% (6/29). In view of §3.1, it is expected that naively truncating time to DN with the left truncation time for death will waste much information in estimation of $F_1(t)$. 

11
In Fig. 1, we plot $\hat{F}_1(t)$, $\hat{F}_1(t)$ and 95% pointwise confidence intervals for $F_1(t)$ based on $\hat{F}_1(t)$, separately for male and female patients. In calculating $\hat{F}_1(t)$, $\tau = 48.5$, which is slightly smaller than the largest event time, 49. Note that in this example, roughly 10% of patients have $L = 0$ so the regularity conditions for nonparametric identification of $F_1$ are satisfied and the conditional estimators in §3.4 are unnecessary. It appears that female diabetic patients have lower long term risk of nephropathy than do male patients, with cumulative incidence estimates of 0.3 versus 0.5 after 40 years, respectively. In both panels, $\hat{F}_1(t)$ and $\hat{F}_1(t)$ are rather similar, except for differences in the lower and upper tails.

We also calculated variance estimates for $\hat{F}_1(t)$ using modifications to the variance estimator in Lin (1997). These variance estimates are always larger than those for $\hat{F}_1(t)$, with the variance reductions ranging from 24% and 67% for males and from 19% to 38% for females. This confirms the simulation results which indicated that the higher the artificial truncation proportion, the larger the efficiency loss associated with the naive competing risks analysis.

4. REMARKS

The proposed nonparametric estimators utilise all semi-competing risks information and achieve large variance reductions over the naive competing risks estimators (Andersen et al., 1993). Given that $\hat{F}_1$ has a simple closed form and plug-in variance estimators are available, it would seem reasonable to recommend this procedure for practical usage. It would be of theoretical interest to explore whether these estimators are fully efficient and whether further efficiency gains are possible without additional computational burden.

Recall that the assumptions for nonparametric estimation of $F_1$ are weaker for $\hat{F}_1$ than for $\hat{F}_1$. Interestingly, when using all semi-competing risks data, there are two ways to define conditional estimators for the cumulative incidence function, $T > t$ and $T_2 > t$. However, when using the competing risks analysis, the latter cannot be estimated.

Determining $\tau$ is an important practical issue. While in theory $\tau$ should be deterministic, the simulations and data analysis suggest that in practice the proposed estimators work well when taking $\tau$ close to the largest value satisfying $C_n(\tau) > 0$.

ACKNOWLEDGMENT

The authors are grateful to P.K. Andersen for sharing the Copenhagen registry data.

APPENDIX

Proofs of lemmas and theorem 1
The following regularity conditions are needed in the proof of Theorem 1: (C1) $a_L \leq a_Y$, $b_L \leq b_Y$; (C2) $C(u) > 0$ for $a_{T_1} < u \leq \tau < b_{T_1}$; (C3) $\int \{G(u)\}^{-3} S_{T_2} (du) < \infty$.

The following lemmas are utilized in the proof of theorem 1.

Lemma 1. Under condition C2-3, (i) $\sup_t \int_t^{t+\tau} \frac{1}{C^2} F^*(d_u, d_v) < \infty$; (ii) $\sup_t \int_t^{t+\tau} \frac{S_{T^2}}{C^2} F^*(d_u, d_v) < \infty$; (iii) $\sup_t \int_t^{t+\tau} \frac{S_{T^2}}{C^2} F^*(d_u, d_v) < \infty$.

Proof. The proofs of (i) and (ii) are straightforward and are omitted. We can show (iii) by the following inequality:

$$\sup_t \int_t^{t+\tau} \frac{S_{T^2}}{C^2} F^*(d_u, d_v) \leq \int_0^{\infty} \frac{S_{T^2}}{C^2} \frac{C}{S_{T_2}} \tilde{H}(d_u, d_v) \leq \{S_{T_2}(\tau)\}^{-2} \int_0^{\tau} G^{-3} S_{T_2}(du) < \infty.$$ 

Lemma 2. For $\tau < b_Y$, (i) assuming the independence of $C$ and $L$,

$$\sup_{0 \leq \tau} \frac{C(Y_i)}{C_n(Y_i)} = O(\log n), \text{ with probability } 1;$$

and, (ii) without assuming the independence of $C$ and $L$,

$$\sup_{0 \leq \tau} \frac{C(Y_i)}{C_n(Y_i)} = O(n \frac{1-p+\epsilon}{2}), \text{ with probability } 1;$$

for any $0 \leq p < \frac{1}{2}$ and $\epsilon > 0$.

Proof. Part (i) directly follows from corollary 1.3 of Stute (1993). By lemma 4 of Lai and Ying (1991) and the fact that $nC_n(Y_i) \geq 1$,

$$\sup_{0 \leq \tau} \left| \frac{C(Y_i)}{C_n(Y_i)} \right| \leq \frac{\sup_{0 \leq \tau} \left| nC_n(Y_i) - nC(Y_i) \right|}{nC_n(Y_i)} + 1 \leq \frac{\sup_{0 \leq \tau} \left| nC_n(s) - nC(s) \right|}{nC_n(Y_i)} + 1 = O(n \frac{1-p+\epsilon}{2}), \text{ almost surely.}$$

On the other hand,

$$\sup_{0 \leq \tau} \left| \frac{C(Y_i)}{C_n(Y_i)} - 1 \right| \leq \frac{\sup_{s \geq 0} \left| C(s) - C_n(s) \right|}{Bn^{-p} - \sup_{s \geq 0} \left| C_n(s) - C(s) \right|}.$$ 

Using the $(\log \log n/n)^{1/2}$ bound for $\|C_n - C\|_{\infty}$ provided by LIL for empirical distribution function, it follows that

$$\sup_{0 \leq \tau} \left| \frac{C(Y_i)}{C_n(Y_i)} - 1 \right| = O\left( \frac{(\log \log n)^{1/2}}{n^{1/2-p}} \right), \text{ almost surely.}$$

Therefore, $\sup_{0 \leq \tau} \left| \frac{C(Y_i)}{C_n(Y_i)} \right| = O(1)$, almost surely, and the proof of (ii) is completed.
Proof of Theorem 1: To simplify the notation, the arguments of \(S_{T_2}(u), \hat{S}_{T_2}(u), C_n(u)\) and \(C(u)\) are suppressed inside the integral sign. Since \(\hat{H}(du, dv) = \{S_{T_2}(u)/C(u)\}F^*(du, dv)\) and \(C_2(\tau, t)/C(\tau) = \Pr(T_1 > t|T_2 > \tau)\), it follows that for \(t < \tau\),

\[
F_1^*(t) - F_1^*(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{S_{T_2}(Y_i^-)}{C_n(Y_i^-)} I(X_i \leq t, X_i < Y_i \leq \tau, \eta_i = 1) - \int_t^\tau \int_v \frac{S_{T_2}(u)}{C(u)} F^*(du, dv) + \hat{S}_{T_2}(\tau) \left\{ 1 - \frac{C_{2,n}(\tau, t)}{C_n(\tau)} \right\} - S_{T_2}(\tau) \left\{ 1 - \frac{C_2(\tau, t)}{C(\tau)} \right\} = I + II.
\]

Simple algebra shows that

\[
I = \int_t^\tau \int_v \frac{S_{T_2}(u)}{C} \left\{ F_n^*(du, dv) - F^*(du, dv) \right\} + \int_t^\tau \int_v \frac{\hat{S}_{T_2} - S_{T_2}}{C} F^*(du, dv) \\
- \int_t^\tau \int_v \frac{S_{T_2}(C_n - C)}{C^2} F^*(du, dv) + \int_t^\tau \int_v \frac{\hat{S}_{T_2} - S_{T_2}}{C} \left\{ F_n^*(du, dv) - F^*(du, dv) \right\} \\
- \int_t^\tau \int_v \frac{(C_n - C)S_{T_2}}{C^2} \left\{ F_n^*(du, dv) - F^*(du, dv) \right\} - \int_t^\tau \int_v \frac{(\hat{S}_{T_2} - S_{T_2})(C_n - C)}{C_nC} F_n^*(du, dv) \\
+ \int_t^\tau \int_v \frac{(C_n - C)^2S_{T_2}}{C^2} F_n^*(du, dv) \\
= I.a + I.b + I.c + R_{1n}(t) + R_{2n}(t) + R_{3n}(t) + R_{4n}(t), \text{ and }
\]

\[
II = \left\{ 1 - \frac{C_2(\tau, t)}{C(\tau)} \right\} \left\{ \hat{S}_{T_2}(\tau) - S_{T_2}(\tau) \right\} - \frac{S_{T_2}(\tau)}{C(\tau)} \left\{ C_{2,n}(\tau, t) - C_2(\tau, t) \right\} \\
+ \frac{S_{T_2}(\tau)C_2(\tau, t)}{C(\tau)} \left\{ C_n(\tau) - C(\tau) \right\} + \frac{S_{T_2}(\tau)}{C(\tau)C_n(\tau)} \left\{ C_n(\tau) - C(\tau) \right\} \left\{ C_{2,n}(\tau, t) - C_2(\tau, t) \right\} \\
- \frac{S_{T_2}(\tau)C_2(\tau, t)}{C(\tau)C_n(\tau)} \left\{ C_n(\tau) - C(\tau) \right\}^2 - \frac{1}{C(\tau)} \left\{ \hat{S}_{T_2}(\tau) - S_{T_2}(\tau) \right\} \left\{ C_{2,n}(\tau, t) - C_2(\tau, t) \right\} \\
+ \frac{C_{2,n}(\tau, t)}{C(\tau)C_n(\tau)} \left\{ \hat{S}_{T_2}(\tau) - S_{T_2}(\tau) \right\} \left\{ C_n(\tau) - C(\tau) \right\} \\
= II.a + II.b + II.c + R_{5n}(t) + R_{6n}(t) + R_{7n}(t) + R_{8n}(t).
\]

Next, we show that \(\sup_{0 \leq t \leq \tau} |r_n(t)| = o(n^{-1/2}), i = 1, 2, \ldots, 8\), with probability 1.

\[
\sup_{0 \leq t \leq \tau} |r_n(t)| \leq \sup_{\nu \leq \tau} |\hat{S}_{T_2}(\nu) - S_{T_2}(\nu)| \cdot \sup_{0 \leq t \leq \tau} \left| \int_t^\tau \int_v \frac{1}{C} \left\{ F_n^*(du, dv) - F^*(du, dv) \right\} \right|.
\]

By Lemma 1, the functional LIL (Goodman et al., 1981) gives

\[
\sup_{0 \leq t \leq \tau} \left| \int_t^\tau \int_v \frac{1}{C} \left\{ F_n^*(du, dv) - F^*(du, dv) \right\} \right| = O(n^{-1/2} \log^{1/2} n).
\]

In addition, following the lines of Stute (1993), we can derive the same strong representation for \(\hat{S}_{T_2}\) but the error term may have higher order since we may need to relax the
asymptotic order of \( \sup_i C(Y_i)/C_n(Y_i), O(\log n), \) to \( O(n^{-\frac{1}{2}+\epsilon}) \), as indicated by lemma 2. That is, \( \hat{S}_{T_2}(y) - S_{T_2}(y) = -S_{T_2}(y)\hat{L}_n(y) + \epsilon_n(y) \). Choosing \( p/2 - \epsilon = 1/5 \), we still get \( \sup_{y \leq \tau} |\epsilon(y)| = o(n^{-1/2}) \), and thus \( \sup_{y \leq \tau} |\hat{S}_{T_2}(y) - S_{T_2}(y)| = O(n^{-1/2} \log^{1/2} n) \) follows from functional LIL. Therefore, we have \( \sup_{0 \leq t \leq \tau} |R_{1n}(t)| = o(n^{-1/2}) \).

Since \( C_n(y) \) is the difference of two empirical distribution functions, it follows that \( \sup_{y \geq 0} |C_n(y) - C(y)| = O(n^{-1/2} \log^{1/2} n) \), and hence \( \sup_{0 \leq t \leq \tau} |R_{2n}(t)| = o(n^{-1/2}) \).

The result \( \sup_{0 \leq t \leq \tau} |R_{3n}(t)| = o(n^{-1/2}) \) can be obtained from Lemma 1 and Lemma 2, setting \( p/2 - \epsilon = 1/5 \), and

\[
\sup_{0 \leq t \leq \tau} |R_{3n}(t)| \leq \sup_{y \leq \tau} \left| \hat{S}_{T_2}(y) - S_{T_2}(y) \right| \cdot \sup_{y \geq 0} \left| C_n(y) - C(y) \right| \cdot \sup_i \left| \frac{C_n(Y_i)}{C(Y_i)} \right| \\
\cdot \sup \left| \int_0^t \int_0^\tau \frac{1}{C^2} F^*(du, dv) \right| \\
= O \left( \frac{\log^{1/2} n}{n^{1/2}} \right) \cdot O \left( \frac{\log^{1/2} n}{n^{1/2}} \right) \cdot O(1) = O(\log n / n^{1/5}) = o(n^{-1/2}).
\]

A similar argument gives \( \sup_{0 \leq t \leq \tau} |R_{4n}(t)| = o(n^{-1/2}) \). Now, since \( C(\tau) > 0 \), the same logic can be used to show that the asymptotic orders of \( \sup |R_{in}|(i = 5, 6, 7, 8) \) are also \( o(n^{-1/2}) \).

The proof is completed by replacing \( \hat{S}_{T_2} - S_{T_2} \) with \( -S_{T_2} \hat{L}_n \) in I and II.

**REFERENCES**


Table 1: Summary statistics under different simulation configurations

<table>
<thead>
<tr>
<th></th>
<th>%($L = 0$)</th>
<th>$1 - \alpha$</th>
<th>%($\delta = 0$)</th>
<th>%($\eta = 0$)</th>
<th>%($\delta = 0, \eta = 1$)</th>
<th>%($X &lt; L \mid L \leq Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.38</td>
<td>18</td>
<td>15</td>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.37</td>
<td>21</td>
<td>18</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0.81</td>
<td>33</td>
<td>14</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.80</td>
<td>36</td>
<td>16</td>
<td>25</td>
<td>16</td>
</tr>
</tbody>
</table>

Figure 1: Cumulative incidence estimates for male patients (left panel) and female patients (right panel) in birth cohort [1915, 1925] and age cohort [10, 25].
Table 2: Comparison of $\hat{F}_1(t)$ and $\tilde{F}_1(t)$ under set-up A.

<table>
<thead>
<tr>
<th>$F_1(t)$</th>
<th>n</th>
<th>Ave</th>
<th>EmpVar</th>
<th>AveVar</th>
<th>Cov95</th>
<th>Ave</th>
<th>EmpVar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
<td>0.098</td>
<td>1.2</td>
<td>1.1</td>
<td>91</td>
<td>0.099</td>
<td>2.4</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.100</td>
<td>0.6</td>
<td>0.6</td>
<td>92</td>
<td>0.100</td>
<td>1.4</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.198</td>
<td>2.0</td>
<td>1.9</td>
<td>93</td>
<td>0.198</td>
<td>3.5</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.198</td>
<td>1.0</td>
<td>1.0</td>
<td>93</td>
<td>0.200</td>
<td>1.8</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.302</td>
<td>2.7</td>
<td>2.5</td>
<td>94</td>
<td>0.303</td>
<td>4.3</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.301</td>
<td>1.3</td>
<td>1.3</td>
<td>95</td>
<td>0.299</td>
<td>2.0</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>0.400</td>
<td>3.0</td>
<td>2.9</td>
<td>94</td>
<td>0.403</td>
<td>4.6</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.399</td>
<td>1.6</td>
<td>1.5</td>
<td>94</td>
<td>0.398</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
<td>0.100</td>
<td>1.1</td>
<td>1.1</td>
<td>93</td>
<td>0.100</td>
<td>1.7</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.100</td>
<td>0.5</td>
<td>0.6</td>
<td>95</td>
<td>0.100</td>
<td>0.9</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.201</td>
<td>2.0</td>
<td>1.9</td>
<td>94</td>
<td>0.200</td>
<td>2.9</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.198</td>
<td>1.0</td>
<td>1.0</td>
<td>94</td>
<td>0.196</td>
<td>1.5</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.298</td>
<td>2.5</td>
<td>2.5</td>
<td>94</td>
<td>0.298</td>
<td>3.4</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.302</td>
<td>1.3</td>
<td>1.3</td>
<td>93</td>
<td>0.303</td>
<td>1.8</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>0.400</td>
<td>2.9</td>
<td>2.9</td>
<td>94</td>
<td>0.399</td>
<td>3.7</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.398</td>
<td>1.5</td>
<td>1.5</td>
<td>95</td>
<td>0.401</td>
<td>1.8</td>
</tr>
</tbody>
</table>

EmpVar: empirical variance($\times10^3$); AveVar: average of estimated variances($\times10^3$).
Table 3: Comparison of $\hat{F}_1(t)$ and $\hat{F}_1(t)$ under set-up B.

<table>
<thead>
<tr>
<th>$F_1(t)$</th>
<th>(\hat{F}_1(t)) Ave</th>
<th>(\hat{F}_1(t)) AveVar</th>
<th>(\hat{F}_1(t)) Cov95</th>
<th>(\hat{F}_1(t)) Ave</th>
<th>(\hat{F}_1(t)) EmpVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>100</td>
<td>0.100</td>
<td>1.2</td>
<td>1.2</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.102</td>
<td>0.6</td>
<td>0.6</td>
<td>94</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.198</td>
<td>2.0</td>
<td>2.0</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.202</td>
<td>1.0</td>
<td>1.0</td>
<td>94</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.301</td>
<td>2.6</td>
<td>2.6</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.300</td>
<td>1.3</td>
<td>1.3</td>
<td>94</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>0.402</td>
<td>3.0</td>
<td>3.0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.402</td>
<td>1.5</td>
<td>1.5</td>
<td>95</td>
</tr>
</tbody>
</table>

\(\%(L=0)=0\%\)

<table>
<thead>
<tr>
<th>$F_1(t)$</th>
<th>(\hat{F}_1(t)) Ave</th>
<th>(\hat{F}_1(t)) AveVar</th>
<th>(\hat{F}_1(t)) Cov95</th>
<th>(\hat{F}_1(t)) Ave</th>
<th>(\hat{F}_1(t)) EmpVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>100</td>
<td>0.101</td>
<td>1.1</td>
<td>1.1</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.099</td>
<td>0.6</td>
<td>0.6</td>
<td>94</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.198</td>
<td>2.2</td>
<td>2.2</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.198</td>
<td>1.1</td>
<td>1.1</td>
<td>93</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.301</td>
<td>2.6</td>
<td>2.6</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.300</td>
<td>1.4</td>
<td>1.4</td>
<td>95</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>0.402</td>
<td>3.0</td>
<td>3.0</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.400</td>
<td>1.4</td>
<td>1.5</td>
<td>95</td>
</tr>
</tbody>
</table>

\(\%(L=0)=20\%\)

EmpVar: empirical variance(\(\times 10^3\)); AveVar: average of estimated variances(\(\times 10^3\)).
Table 4: Summary statistics for birth×diagnosis age cohort= [1915, 1925] × [10, 25].

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>114</td>
<td>82</td>
</tr>
<tr>
<td>$(\delta, \eta) = (0, 0)$</td>
<td>22(19.3%)</td>
<td>31(37.8%)</td>
</tr>
<tr>
<td>$(\delta, \eta) = (0, 1)$</td>
<td>27(23.7%)</td>
<td>22(26.8%)</td>
</tr>
<tr>
<td>$(\delta, \eta) = (1, 0)$</td>
<td>5(4.4%)</td>
<td>3(3.7%)</td>
</tr>
<tr>
<td>$(\delta, \eta) = (1, 1)$</td>
<td>60(52.5%)</td>
<td>26(31.7%)</td>
</tr>
<tr>
<td>$L = 0$</td>
<td>9(7.9%)</td>
<td>10(12.2%)</td>
</tr>
<tr>
<td>$X &lt; L$</td>
<td>17(14.9%)</td>
<td>6(7.3%)</td>
</tr>
<tr>
<td>$X = L$</td>
<td>8(7.0%)</td>
<td>3(3.7%)</td>
</tr>
</tbody>
</table>