Published in final edited form as:

Biom J. 2016 May ; 58(3): 635–651. doi:10.1002/bimj.201400126.

Nonparametric association analysis of bivariate left-truncated competing risks data

Yu Cheng[*,1], **Pao-sheng Shen**[2], **Zhumin Zhang**[3], and **HuiChuan J. Lai**[4]

[1]Department of Statistics and Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260, USA

[2]Department of Statistics, Tunghai University, Taichung 40704, Taiwan

[3]Department of Nutritional Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

[4]Departments of Nutritional Science, Pediatrics, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, USA

Abstract

We develop time-varying association analyses for onset ages of two lung infections to address the statistical challenges in utilizing registry data where onset ages are left-truncated by ages of entry and competing-risk censored by deaths. Two types of association estimators are proposed based on conditional cause-specific hazard function and cumulative incidence function that are adapted from unconditional quantities to handle left truncation. Asymptotic properties of the estimators are established by using the empirical process techniques. Our simulation study shows that the estimators perform well with moderate sample sizes. We apply our methods to the Cystic Fibrosis Foundation Registry data to study the relationship between onset ages of *Pseudomonas aeruginosa* and *Staphylococcus aureus* infections.

Keywords

Cause-specific hazard function; Conditional quantity; Cumulative incidence function; Empirical process; Left truncation; Lung infection

[*]Corresponding author: yucheng@pitt.edu, Phone: +1 412 648 1851, Fax: +1 412 648 8814.

Software

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author (yucheng@pitt.edu).

Supplementary materials

Technical details were placed in a separate file for Supplementary Material.

Conflict of interest

The authors have declared no conflict of interest.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

Association analysis of bivariate correlated event times is often of interest in genetic family studies, demography, and medical investigations. For example, in the Cystic Fibrosis Foundation Registry (CFFR) data, the onset ages of *Pseudomonas aeruginosa* (*Pa*) and *Staphylococcus aureus* (*Sa*) are closely monitored, since these two bacterial infections are commonly observed in patients with cystic fibrosis (CF) and often lead to deterioration of lung functions (Kosorok et al., 2001; Flume et al., 2007, 2009). Bacteria have elaborate chemical signaling systems that enable them to communicate within and between species. The interplay between microorganisms in CF airway may influence disease prognosis and response to therapy (Rogers et al., 2010). To examine the possible interaction between *Pa* and *Sa*, we focus on the association between the onset ages of these two infections. There is extensive work in quantifying the association between paired event times when they are subject to independent censoring; see Hougaard (2000) for an overview. However, in the studies with composite endpoints, the event times of interest are often subject to competing-risk censoring. If a CF patient died before he/she has developed any infection, the event times of interest would be dependently censored by the competing event death, which complicates the quantification of the association between the onset ages of these two respiratory infections.

There are a few attempts in extending methods from bivariate survival data to bivariate competing risks settings. Some commonly used global dependent measures (e.g., Kendall's τ and Spearman's correlation) can be extended for bivariate competing risks data. However, as onset of lung infections during childhood occurs at various ages, the direction and strength of the association between the onset ages of *Pa* and *Sa* infections may vary with time.

Hence a time-varying association analysis is more appropriate. Some local-dependent measures have been adapted from bivariate survival data to bivariate competing risks data. Bandeen-Roche and Liang (2002) and Bandeen-Roche and Ning (2008) focused on a cause-specific cross-hazard ratio, and Cheng and Fine (2008) proposed an alternative representation of the association measure based on bivariate hazard functions. The estimation of the two equivalent measures requires binning the observed time region into finer grids and assuming constant association within each grid.

We are interested in the association between *Pa* and *Sa* infections over a life course between 1.5 and 20 years of age, a period when frequent infections occur that are also closely correlated with other key clinical outcomes of CF, such as growth and nutritional status. As a first attempt, we resort to time-varying association measures that do not require binning or the piecewise-constant assumption. Cheng et al. (2007) proposed two association measures for bivariate competing risks data based on bivariate cause-specific hazard (CSH) functions and bivariate cumulative incidence functions (CIFs), and developed nonparametric inference without any model assumptions. However, their methods cannot be readily applied to analyze the association between *Pa* and *Sa* infections as both event times are truncated at ages of entry (Lai et al., 2004). Left truncation is very common in registry data as subjects must be alive and have a certain disease to be included in a registry. Therefore, there is a need to develop association analysis for bivariate competing risks data with left truncation.

To incorporate left truncation, in this paper we consider conditional CSH functions and conditional CIFs. A Nelson–Aalen type estimator of the bivariate cumulative CSH function is proposed as is done in Cheng et al. (2007). To estimate the bivariate conditional CIF, we need an estimator for the bivariate conditional survival function of times to composite events. Nonparametric estimators of the bivariate survival function under independent censoring have been proposed by Campbell (1981), Dabrowska (1988), Burke (1988), Pruitt (1991), Prentice and Cai (1992), and van der Laan (1996a) among others. There has also been some work on estimating a bivariate distribution when observations are subject to truncation; see for example Gürler (1996, 1997) and Gijbels and Gürler (1998) for application when a single component of the bivariate data is subject to truncation, van der Laan (1996b) and Huang et al. (2001) for bivariate truncated data, and Shen (2006), Shen and Yan (2008), Shen (2010), and Dai and Fu (2012) for bivariate truncated and censored data. In our application, when we consider bivariate distribution of times to composite events, the event times are subject to bivariate truncation and right censoring. However, the methods proposed in Shen (2006), Shen and Yan (2008), and Shen (2010) require iteration algorithms to calculate the distribution estimates, which are computationally intensive and impractical for our data that involve 15,176 subjects. Recently, Dai and Fu (2012) proposed an estimator for the bivariate unconditional survival function based on a polar coordinate transformation, using the data with bivariate left-truncation and random censoring. They then constructed an inversely weighted estimator for the unconditional bivariate distribution function based on which an estimator of the truncation probability was obtained, coupled with the bivariate survival estimator for the truncation times. It is not clear how the estimator would perform if we use their bivariate survival estimator divided by the truncation probability estimator to obtain the bivariate conditional survival function. Hence, in Section 2, we extend the Dabrowska method for the bivariate conditional survival function, which does not require iteration or estimation of intermediate unconditional quantities and the truncation probability.

The rest of this paper is organized as the following. To quantify the association between the onset age of Pa and that of Sa among CF patients that are subject to competing-risk censoring and left truncation, we define two association measures as functions of the conditional cumulative CSH functions and the conditional CIFs in Section 2.2. For the time-varying association measure based on the conditional cumulative CSH functions, we propose a nonparametric estimation procedure based on the Nelson–Aalen type of estimators. For the time-varying association measure defined by the conditional CIFs, we develop a nonparametric estimator using the generalization of the Dabrowska (1988, 1989) estimator. Details on the estimators and test procedures are given in Section 2.3. The asymptotic properties of the proposed estimators and test statistics are established in Section 2.4. We conduct simulation studies to evaluate the finite-sample properties of our estimators and to examine the size and the power of our proposed tests in Section 3. The practical utility of our methods is illustrated in an analysis of the CFFR data in Section 4. We conclude with some remarks in Section 5.

2 Method

2.1 Data and notation

The CFFR contains information on majority of CF patients who have been treated by accredited CF centers in United States (FitzSimmons, 1993). Time to first *Pa* infection and time to first *Sa* infection among living CF patients are the event times of interest. The two infection times are naturally correlated, and their association is the focus of this study. The occurrence of one infection does not preclude the occurrence of the other. Hence, we treat the two infection times among living patients as observable bivariate event times that are subject to competing-risk censoring by death. Let T_1 (T_2) be the onset age of a *Pa* (*Sa*) infection in a patient, and let $\varepsilon_1 = 1$ ($\varepsilon_2 = 1$) if the subject obtained a *Pa* (*Sa*) infection. If a subject died without a *Pa* or *Sa* infection, then T_1 or T_2 would be the age at death with ε_1 or ε_2 being 2. (T_1, T_2) reported in the CFFR are subject to usual independent censoring by the end of the observational period. In our data analysis in Section 4, we will exclude those subjects who had lung infections at the study entry. Hence, (T_1, T_2) are also subject to left truncation by ages at entry to CFFR, because a subject has to qualify the following criteria to be included in the study: to be alive, diagnosed with CF, reported to CFFR and free of any lung infection by entry.

Though the two infection times are subject to the same independent censoring time and the same truncation time in our application, censoring or truncation times can be different in another type of bivariate competing risks data. For example, in a familial study of dementia, the onset age of dementia may be competing-risk censored by death, and the administrative censoring times of the two individuals in a mother–child pair may be different. Here, we adopt the notation that can incorporate both types of bivariate competing risks data. Let (C_1, C_2) be independent censoring times and (V_1, V_2) be left truncation times for a pair. One observes nothing if $T_1 \le V_1$ or $T_2 \le V_2$, and observes (X_1, δ_1, X_2, δ_2, V_1, V_2) if $T_1 > V_1$ and $T_2 > V_2$ (Andersen et al., 1993), where $X_j = \min(T_j, C_j)(j = 1, 2)$ and δ_j is equal to 1 if the individual developed a *Pa* or *Sa* infection, equal to 2 if the individual died without the infection, and 0 otherwise. Note that δ_j, $j = 1, 2$, are defined for the actual data with left-truncation and administrative right-censoring, and are closely related to the cause indicators ε_j, $j = 1, 2$, in that $\delta_j = \varepsilon_j$ if an event is observed, and 0 otherwise. The observed data contain n i.i.d. replicates of (X_1, δ_1, X_2, δ_2, V_1, V_2), denoted by $\{(X_{1i}, \delta_{1i}, X_{2i}, \delta_{2i}, V_{1i}, V_{2i}), i = 1, \ldots, n\}$.

2.2 Time varying association measures

Assume that (T_1, T_2) are independent of (C_1, C_2, V_1, V_2) and $P(V_j < C_j) = 1$, $j = 1, 2$. We first define an association measure based on the CSH functions. The CSH function in the bivariate setup is

$$\Lambda^{12}_{kl}(du, dv) = \frac{P(T_1 \in du, T_2 \in dv, \varepsilon_1 = k, \varepsilon_2 = l)}{P(T_1 \ge u, T_2 \ge v)}, k, l = 1, 2. \quad (1)$$

The notation Λ_{kl}^{12} looks rather complicated, where the superscripts correspond to individuals 1 and 2, which can be two event times from the same subject, and the subscripts correspond to the *k*-th failure from individual 1 and the *l*-th failure from individual 2. The similar notation will be used throughout the paper. We can write (1) as

$$\Lambda_{kl}^{12}(du,dv)=\frac{F_{kl}^{12}(du,dv)}{S(u-,v-)}, \qquad (2)$$

where $S(u, v) = P(T_1 > u, T_2 > v)$ is the joint survival function of (T_1, T_2) and $F_{kl}^{12}(u,v)=P(T_1 \le u, T_2 \le v, \varepsilon_1=k, \varepsilon_2=l)$ denotes the bivariate cause-specific CIF. Hence $\Lambda_{kl}^{12}(du,dv)$ is the instantaneous failure rate that individual 1 in the pair fails at time *u* and individual 2 fails at time *v*, given that the pair are free of any events by time (*u*, *v*). Similarly, the marginal CSH functions of T_j ($j = 1, 2$) can be written as

$$\Lambda_k^j(du)=\frac{F_k^j(du)}{S_j(u-)},$$

where $S_j(u) = P(T_j > u)$ is the marginal survival function of T_j and $F_k^j(u)=P(T_j \le u, \varepsilon_j=k)$ is the cause *k* marginal CIF for individual *j*, $k = 1, 2$.

The importance of CIFs is well recognized in analyzing competing risks data in the literature (Gray, 1988;Kalbfleisch and Prentice, 2002). Since $F_{kl}^{12}(u,v)$ quantifies the proportion of subjects failing from each of the cause-specific endpoints, $F_{kl}^{12}(u,v)$ is often preferred over CSH functions. In our study, we focus on association analysis of cause 1 events, which is of our primary interest. Note that in this application since the cause 2 events of death are the same for the paired data, $F_{11}^{12}(u,v)$ and $\Lambda_{11}^{12}(u,v)$ are well defined at any (*u*, *v*), but the cross-cause or cause 2 quantities may not be well defined at all time points. For example, $F_{12}^{12}(u,v)$ is only meaningful when $u \le v$ and $\Lambda_{22}^{12}(du,dv)$ is only defined when $u = v$. In a more general application where cause 1 events may be dependently censored at different cause 2 event times, our methods can be readily applied to cause 2 association as well as cross-cause associations.

Cheng et al. (2007) proposed two time-dependent association measures. One of them is based on CSH functions and given by

$$\phi(s,t)=\Lambda_{11}^{12}(s,t)\left\{\Lambda_1^1(s)\Lambda_1^2(t)\right\}^{-1},$$

where $\Lambda_1^j(s)=\int_0^s \Lambda_1^j(du)(j=1,2)$ and $\Lambda_{11}^{12}(s,t)=\int_0^s\int_0^l \Lambda_{11}^{12}(du,dv)$ are the univariate and bivariate cumulative CSH functions with respect to cause 1 events. As $\Lambda_{11}^{12}(s,t), \Lambda_1^1(s)$, and

$\Lambda_1^2(t)$ take on values from 0 to ∞, $\phi(s, t)$ ranges from 0 to ∞, where the value of 1 indicates independence on the cumulative hazards at (s, t), values > 1 suggest positive associations, and values between 0 and 1 correspond to negative associations.

The other association measure is based on CIFs. Notice that $F_{11}^{12}(u,v)$ captures the identifiable aspects of cause 1 association between T_1 and T_2. Hence the cause 1 association measure is given by

$$\psi(s,t)=F_{11}^{12}(s,t)\left\{F_1^1(s)F_1^2(t)\right\}^{-1}.$$

$\psi(s, t)$ takes on values from 0 to ∞, with $\psi = 1$ corresponding to independence on the CIFs at (s, t), and $\psi >1$ (or $0 < \psi < 1$) stands for a positive (or negative) association. As CIFs have direct probability interpretations, $\psi(s, t)$ can be thought of as

$$P(T_1 \le s, \varepsilon_1=1|T_2 \le t, \varepsilon_2=1)/P(T_1 \le s, \varepsilon_1=1),$$

which in our example measures the excessive (or prohibitive) risk for a CF patient to acquire a *Pa* infection before time s contributable to the fact that this patient has acquired a *Sa* infection before time t. In contrast, $\phi(s, t)$ is defined based on cumulative CSH functions and may not have straightforward interpretations. However, the cause k CIFs may be affected by noncause k events through their influence on the overall survival function of times to first events. Hence the strength of ψ may be affected by the association in failures from other causes. On the other hand, ϕ for cause k events is not affected by noncause k events. Though ϕ may not be as appealing as ψ in terms of interpretations, the comparison of ϕ and ψ gives us insight into how different causes interact with each other. In addition, the estimation of $\psi(s, t)$ naturally leads to the estimation of $\phi(s, t)$. Therefore, in this paper we will adapt both time-varying association measures $\psi(s, t)$ and $\phi(s, t)$ to bivariate left-truncated competing risks data.

Next, we briefly discuss the identifiability of $\psi(s, t)$ and $\phi(s, t)$. For any distribution function H, denote the left and right endpoints of its support by $a_H = \inf\{t : H(t) > 0\}$ and $b_H = \inf\{t : H(t) = 1\}$, respectively. For $j = 1, 2$, let F_j, Q_j and G_j denote the distribution functions of T_j, C_j and V_j, respectively. Assume that $a_{G_j} \le \min(a_{F_j}, a_{Q_j})$ and $b_{G_j} \le \min(b_{F_j}, b_{Q_j})$ for $j = 1, 2$. Woodroofe (1985) pointed out that F_j, Q_j, and G_j are all identifiable if the assumptions hold.

Furthermore, when the distributions of T and V have the same lower bound, that is $a_{F_j} = a_{G_j}$ $(j = 1, 2)$, we consider the conditional association measures for cause 1 event defined as

$$\phi(s,t|a)=\Lambda_{11}^{12}(s,t|a)\left\{\Lambda_1^1(s|a)\Lambda_1^2(t|a)\right\}^{-1},(s \wedge t=\min(s,t) \ge a),$$

and

$$\psi(s,t|a)=F_{11}^{12}(s,t|a)\{F_1^1(s|a)F_1^2(t|a)\}^{-1}. \quad (3)$$

Here, we define $\Lambda_{11}^{12}(du,dv|a)=\frac{F_{11}^{12}(du,dv|a)}{S(u=,v-|a|)}$ and $\Lambda_1^j(du|a)=\frac{F_1^j(du|a)}{S_j(u-|a)}$, where $F_{11}^{12}(u,v|a)=P(T_1 \leq u, T_2 \leq v, \varepsilon_1=1, \varepsilon_2=1|T_1 \wedge T_2>a)$ is the bivariate conditional CIF and $S(u, v|a) = P(T_1 > u, T_2 > v|T_1 \wedge T_2 > a)$ is the bivariate conditional survival function. Similarly, we define the marginal conditional CIFs and survival functions $F_1^j(u|a)=P(T_j \leq u, \varepsilon_j=1|T_1 \wedge T_2 \geq a)$ and $S_j(u|a) = P(T_j > u|T_1 \wedge T_2 \geq a)$, $j = 1, 2$.

2.3 Estimation of cause-specific association measures

Estimating ϕ(s, t)—First, we consider the estimation of $\phi(s, t)$ based on the observed left-truncated competing risks data $\{(X_{1i}, \delta_{1i}, X_{2i}, \delta_{2i}, V_{1i}, V_{2i}), i = 1, \ldots, n\}$. Nelson–Aalen type of estimators (Nelson, 1972; Aalen, 1978) will be constructed to estimate the involved bivariate and univariate cumulative CSH functions in $\phi(s, t)$. For this purpose, we define the cause-specific double-event process $H_{kl}^{12}(u,v)=I\{X_1 \leq u, \delta_1=k, X_2 \leq v, \delta_2=l\}$ for $k, l = 1, 2$, and the at-risk process $R(u, v) = I\{V_1 < u \leq X_1, V_2 < v \leq X_2\}$. Conditional on the observations being left truncated, the expectation of $R(u, v)$ is $E\{R(u, v)|T_1 > V_1, T_2 > V_2\}$ which equals

$$P(V_1<u \leq X_1, V_2<v \leq X_2|T_1>V_1, T_2>V_2)=p^{-1}K(u,v)S(u-,v-), \quad (4)$$

where $K(u, v) = P(V_1 < u \leq C_1, V_2 < v \leq C_2)$ and $p = P(T_1 > V_1, T_2 > V_2)$ is the un-truncated probability. Similarly, the conditional expectation of $H_{kl}^{12}(u,v)$ is

$$\begin{aligned} & E\left\{H_{kl}^{12}(u,v)|T_1>V_1, T_2>V_2\right\}=p^{-1}P(V_1<T_1 \leq u, C_1 \geq T_1, \varepsilon_1 \\ & \quad =k, V_2<T_2 \leq v, C_2 \geq T_2, \varepsilon_2=l). \end{aligned}$$

Thus,

$$\begin{aligned} & E\left\{H_{kl}^{12}(du,dv)|T_1>V_1, T_2>V_2\right\}=P(X_1 \in du, \delta_1=k, X_2 \in dv, \delta_2=l|T_1>V_1, T_2>V_2) \\ & =p^{-1}P(T_1 \in du, C_1 \geq u, \varepsilon_1=k, T_2 \in dv, C_2 \geq v, \varepsilon_2=l, V_1<u, V_2<v) \\ & =p^{-1}K(u,v)F_{kl}^{12}(du,dv). \end{aligned}$$

Henceforth, we will simply denote the conditional expectations as EH_{kl}^{12} and ER if there is no ambiguity. Since

$$\Lambda_{kl}^{12}(s,t)=\int_{a_{F_1}}^{s}\int_{a_{F_2}}^{t}\frac{F_{kl}^{12}(du,dv)}{S(u-,v-)}=\int_{a_{F_1}}^{s}\int_{a_{F_2}}^{t}\frac{EH_{kl}^{12}(du,dv)}{ER(u,v)}, \qquad (5)$$

it can be estimated by plugging in the corresponding empirical processes for the unknown population quantities. Let $\hat{H}_{kl}^{12}(u,v)=\frac{1}{n}\sum_{i=1}^{n}I\{X_{1i}\le u,\delta_{1i}=k,X_{2i}\le v,\delta_{2i}=l\}$ and $r(u,v)=\frac{1}{n}\sum_{i=1}^{n}I\{V_{1i}<u\le X_{1i},V_{2i}<v\le X_{2i}\}$. Then

$$\hat{\Lambda}_{kl}^{12}(s,t)=\int_{a_{F_2}}^{t}\int_{a_{F_1}}^{s}\frac{\hat{H}_{kl}^{12}(du,dv)}{r(u,v)}.$$

Similarly, we define the single-event processes $H_k^1(u,v)=I\{X_1\le u,\delta_1=k,V_2<v\le X_2\}$ and $H_l^2(u,v)=I\{V_1<u\le X_1,X_2\le v,\delta_2=l\}$. Their corresponding empirical processes are denoted by $\hat{H}_k^1(u,v)$ and $\hat{H}_l^2(u,v)$. The univariate cumulative hazard functions are estimated by

$$\hat{\Lambda}_k^1(s,a_{F_2})=\int_{a_{F_1}}^{s}\frac{\hat{H}_k^1(du,a_{F_2})}{r(u,a_{F_2})}\text{ and }\hat{\Lambda}_l^2(a_{F_1},t)=\int_{a_{F_2}}^{t}\frac{\hat{H}_l^2(a_{F_1},dv)}{r(a_{F_1},v)}.$$

Therefore, when $a_{Fj} > a_{Gj}$, a consistent estimator of $\phi(s, t)$ is given by

$$\hat{\phi}(s,t)=\hat{\Lambda}_{11}^{12}(s,t)\left\{\hat{\Lambda}_1^1(s,a_{F_2})\hat{\Lambda}_1^2(a_{F_1},t)\right\}^{-1}.$$

Note that we cannot estimate $\Lambda_1^1(s)$ using $\hat{\Lambda}_1^1(s)=\int_{a_{F_1}}^{s}\frac{\hat{H}_1^1(du)}{r^1(u)}$, where $\hat{H}_1^1(u)=\frac{1}{n}\sum_{i=1}^{n}I\{X_{1i}\le u,\delta_{1i}=1\}$ and $r^1(u)=\frac{1}{n}\sum_{i=1}^{n}I\{V_{1i}<u\le X_{1i}\}$, since the estimator $\hat{H}_1^1(du)/r^1(u)$ actually estimates $\frac{\int_{a_{F_2}}^{\infty}P(V_1<u\le C_1,V_2<v)F_{1.}^{12}(du,dv)}{\int_{a_{F_2}}^{\infty}P(V_1<u\le C_1,V_2<v)S(u-dv)}$, where $F_{1.}^{12}(u,v)=P(T_1\le u,\varepsilon_1=1,T_2\le v)$. Similarly, we cannot estimate $\Lambda_1^2(t)$ using $\hat{\Lambda}_1^2(t)=\int_{a_{F_2}}^{t}\frac{\hat{H}_1^2(dv)}{r^2(v)}$, where $\hat{H}_1^2(v)=n^{-1}\sum_{i=1}^{n}I\{X_{2i}\le v,\delta_{2i}=l\}$ and $r^2(v)=n^{-1}\sum_{i=1}^{n}I\{V_{2i}<v\le X_{2i}\}$.

Since the data are left-truncated, we are not able to test whether or not $a_{Fj} > a_{Gj}$. In applications, some $a > a_{Fj}$ is selected so that the size of the observed risk set at a is not too small (Gross and Lai, 1996), and both $\hat{\Lambda}_1^1(s,a)$ and $\hat{\Lambda}_1^2(a,t)$ are defined. Hence, a consistent estimator of $\phi(s, t|a)$ is given by

$$\hat{\phi}(s,t|a)=\hat{\Lambda}_{11}^{12}(s,t|a)\{\hat{\Lambda}_1^1(s,a)\hat{\Lambda}_1^2(a,t)\}^{-1},$$

where $\hat{\Lambda}_{11}^{12}(s,t|a)=\int_a^s\int_a^t\frac{\hat{H}_{11}^{12}(du,dv)}{r(u,v)}$, $\hat{\Lambda}_1^1(s,a)=\int_a^s\frac{\hat{H}_1^1(du,a)}{r(u,a)}$, and $\hat{\Lambda}_1^2(a,t)=\int_a^t\frac{\hat{H}_1^2(a,dv)}{r(a,v)}$. $\hat{\Lambda}_{11}^{12}(s,t|a)$ is similar to $\hat{\Lambda}_{11}^{12}(s,t)$ except that the integration is now from a instead of the lower end of support. Using the approach of Cheng et al. (2007), we consider the integrated weighted averages $\hat{\phi}^*$ as the test statistics for $\phi(s, t) = 1$ for all $s \in [a, \tau_1]$, $t \in [a, \tau_2]$:

$$\hat{\phi}^*=\int_a^{\tau_1}\int_a^{\tau_2}\overline{W}(s,t)\hat{\phi}(s,t|a)dsdt,$$

where $\tau_1 > a$ and $\tau_2 > a$, and $\tilde{W}(s, t)$ is a stochastic weighting function which is bounded between 0 and 1 and converges in probability to a deterministic weighting function.

Estimating ψ(s, t)—Next, we consider the other association measure that was defined in (3).

In order to estimate $F_{11}^{12}(u,v|a)$ in (3), we need to estimate both $S(u, v|a) = P(T_1 > u, T_2 > v|T_1 \wedge T_2 > a)$ and $\Lambda_{11}^{12}(u,v|a)$. In this paper, we propose a Dabrowska type estimator (Dabrowska, 1988) for the conditional survival function $S(u, v|a)$. The estimator is similar to the one that was considered in Shen and Yan (2008), except that the lower bound of the integration is from a instead of 0.

Define the Dabrowska estimator

$$\hat{S}_D(s,t|a)=\hat{S}_1(s|a)\hat{S}_2(t|a)\prod_{a\le y\le t a\le x\le s}\prod[1-\hat{L}(dx,dy)],$$

where the expressions for $\hat{S}_1(s|a)$, $\hat{S}_2(t|a)$, and $\hat{L}$ are given in the Appendix. Thus, we estimate $F_{kl}^{12}(s,t)$ as follows:

$$\hat{F}_{D,kl}^{12}(s,t)=\int_a^t\int_a^s\hat{S}_D(u-,v-|a)\hat{\Lambda}_{kl}^{12}(du,dv|a).$$

Notice that when there is no truncation $\hat{F}_{D,kl}^{12}(s,t|a)$ is reduced to a nonparametric estimator considered in Cheng et al. (2007). Similarly, for $j = 1, 2$, we obtain the marginal estimators $\hat{F}_{D,k}^1(s|a)=\int_a^s\hat{S}_D(u-,a)\hat{\Lambda}_k^1(du,a)$ and $\hat{F}_{D,l}^2(t|a)=\int_a^t\hat{S}_D(a,v-)\hat{\Lambda}_l^2(a,dv)$, for $k, l = 1, 2$. Plugging the univariate and bivariate estimators into (3), we have

$$\hat{\psi}_D(s,t|a)=\hat{F}^{12}_{D,11}(s,t|a)\{\hat{F}^{1}_{D,1}(s|a)\hat{F}^{2}_{D,1}(t|a)\}^{-1}.$$

Based on $\hat{\psi}_D(s, t|a)$, we obtain the following integrated weighted average as the test statistic for $\psi(s, t|a) = 1$ for all $s \in [a, \tau_1]$, $t \in [a, \tau_2]$:

$$\hat{\psi}^*_D=\int_a^{\tau_1}\int_a^{\tau_2}\tilde{W}(s,t)\hat{\psi}_D(s,t|a)dsdt,$$

where $\tilde{W}$ is some stochastic weighting function as the one defined for $\hat{\phi}^*$.

2.4 Asymptotic properties

The asymptotic properties of the above estimators and tests were established using the empirical process theories; see the technical appendix of the Supplementary Material for details. For any $(s, t) \in [a, \tau] = [a, \tau_1] \times [a, \tau_2]$, we have $\sqrt{n}\{\hat{\Lambda}^{12}_{11}(s,t|a)-\Lambda^{12}_{11}(s,t|a)\}$ and $\sqrt{n}\{\hat{F}_{11}(s,t|a)-F_{11}(s,t|a)\}$ converge weakly to mean zero Gaussian processes. Coupled with the asymptotic properties of the estimators for marginal quantities, we have weak convergence of $\sqrt{n}\{\hat{\psi}(s,t|a)-\psi(s,t|a)\}$ and $\sqrt{n}\{\hat{\phi}(s,t|a)-\phi(s,t|a)\}$ and asymptotic normality of $\sqrt{n}(\hat{\psi}^*-\psi^*)$ and $\sqrt{n}(\hat{\phi}^*-\phi^*)$. Bootstrap validity also holds for both the estimators and the tests.

3 Simulation studies

The modified Dabrowska estimator—We first conduct a simulation study to evaluate the finite sample performance of the modified Dabrowska estimator for bivariate left-truncated and right-censored data. We adopt a simulation setting similar with the one in Shen (2010). More specifically, we simulate bivariate event times (T_1, T_2) from a bivariate survival function $S(u_1, u_2) = \exp\{-(u_1 + u_2) - \max(u_1, u_2)\}$. Marshall and Olkin (1967) introduced the so-called Marshall–Olkin bivariate exponential model assuming that the joint survival function of T_1 and T_2 has the form $S(t_1, t_2) = \exp\{-\lambda_1 t_1 - \lambda_2 t_2 - \lambda_{12} \max(t_1, t_2)\}$. For simplicity, we let $\lambda_1 = \lambda_2 = \lambda_{12} = 1$. The left truncation times (V_1, V_2) are generated from an exponential distribution with mean 0.1. To make sure $C_j > V_j$, $j = 1, 2$, we let $C_j = V_j + W_j$, where W_j are simulated from an exponential distribution with mean 4. For each dataset, we generate 50 or 100 pairs of data where $T_j > V_j$, $j = 1, 2$. The proportion of censoring is about 10%.

In Table 1, we present averages of the estimates (EST) of the bivariate survival function at different time points based on 2000 datasets. Their empirical standard errors (ESE), averages of bootstrap standard errors (BSE), and coverage rates (Cov) of 95% confidence intervals based on the BSE and asymptotic normality are also given in Table 1. The averages of survival function estimates are close to the true values that range from 0.26 to 0.64. The empirical standard errors agree with the bootstrap standard errors, especially when $n = 100$. The standard errors decrease and coverage rates slightly improve when the sample sizes increase from 50 to 100. The proposed modified Dabrowska estimator is easy to implement and performs well for a sample of size 50 in this simulation setting. In contrast, the existing methods (Shen, 2006; Shen and Yan, 2008; Shen, 2010) require iterative algorithms and are computationally extensive. On the other hand, our estimator is developed for the conditional survival function, while the iterative methods (Shen, 2006; Shen and Yan, 2008; Shen, 2010) provide estimation of both unconditional and conditional survival functions. Hence, our simulation results are not directly comparable to those reported in the literature. However, the magnitude of bias appears comparable to those reported in Shen (2010).

Tests based on $\phi(s, t)$ and $\psi(s, t)$: Another simulation study is conducted to examine the performance of $\hat{\phi}(s, t)$ and $\hat{\psi}_D(s, t)$. We adopt a simulation setting similar with that used in Cheng et al. (2007). More specifically, let (T_1^1, T_1^2) denote the paired failure times of interest and (T_2^1, T_2^2) denote the paired failure times for the competing risks. The vector of random variables ($\log(T_1^1), \log(T_2^1), \log(T_1^2), \log(T_2^2)$) is drawn from a normal distribution with mean zero and covariance matrix $\sum = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_3 & \rho_2 \\ \rho_2 & \rho_3 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$. In this study, we explore five combinations of ρ_1, ρ_2 and ρ_3. The truncation variables V_1 and V_2 are independently drawn from an exponential distribution with mean θ. For $j = 1, 2$, the C_j's are defined as $C_j = D_j + V_j$, such that $P(V_j < C_j) = 1$, where D_j's are independent of V_j's and are uniformly distributed on the interval (u_1, u_2). For each dataset, we generate 200 or 300 replicates of $(X_1, \delta_1, X_2, \delta_2, V_1, V_2)$, where $X_j = \min(T_1^j, T_2^j, C_j)$ and $\delta_j = I\{T_j = T_1^j\} + 2I\{T_j = T_2^j\}$ for $j = 1, 2$.

For each scenario, we generate 500 datasets, and for each dataset, we compute $\hat{\phi}(s, t|a)$ and $\hat{\psi}_D(s, t|a)$. The value of a is chosen to be 0.1. The integrated association measures are estimated over time points [0.4, 1] × [0.4, 1]. The associations over the region are of equal importance, hence we have used a simple uniform weight function here. Other weight functions such as weighting proportionally by number of subjects at risk may be useful in other applications that focus on associations at certain time points. The test statistics log $\hat{\phi}^*$ and $\log \hat{\psi}_D^*$ are calculated and their bootstrap standard errors are computed based on 250 bootstrap samples. The tests are at level .05, rejecting H_0 if the absolute standardized statistic exceeds 1.96. The results are summarized in Table 2.

Under the first scenario of $\rho_1 = \rho_2 = \rho_3 = 0$, both $\phi^* = \psi^* = 1$, and the rejection rates are close to the nominal significant level 0.05. Under the other five scenarios, the powers of rejecting the null hypotheses $H_0 : \phi^* = 1$ and $H_0 : \psi^* = 1$ vary. Since the censoring proportions range from 0.15 to 0.19 for the five alternatives, the varying powers mainly depend on the strength of association in ϕ^* and ψ^* and sample sizes. The powers increase when the sample size increases from 200 to 300 across the five alternatives. When there is strong positive association between paired cause 1 event times ($\rho_1 = 0.5$) and weaker dependence between two competing events within a subject ($\rho_2 = 0.3$), the two association measures are about 2 ($\phi^* = 1.78$ and $\psi^* = 2$). That is, given one subject has developed the cause 1 event, the other subject in the same pair would be twice likely to develop the cause 1 event, as compared with the case that the two subjects act independently. With the presence of strong association in cause 1 events, the powers are 0.93 or above for the sample size of 200. When there is weaker association between paired cause 1 event times (ρ_2 = 0.3) and stronger positive association between two competing events within a subject (ρ_1 = 0.5), the rejection rates decrease from 90s percent to 60s percent for the sample size of 200. When ρ_2 takes on a negative value (ρ_2 = −0.3) and ρ_1 = 0.5 and ρ_3 = 0, we observe negative associations ϕ^* = 0.46 and ψ^* = 0.42. The powers further decrease to be in low 50s and high 40s percent for a sample of 200. However, for a sample of 300, the powers are still reasonable. When there is some weak cross-cause association, for example ρ_3 = 0.1 or ρ_3 = −0.1, the powers are lower than the cases when there is no cross-cause association (ALT 2 vs. ALT 4 and ALT 3 vs. ALT 5).

4 Cystic fibrosis study

We applied our time-varying association measures $\psi(s, t)$ and $\phi(s, t)$ to the CFFR data. Our study focused on quantifying the association between *Pa* and *Sa* infections among living CF patients using the CFFR data collected during 1986–2007. Specifically, we examined the association between the onset of first *Pa* and the onset of first *Sa* in living CF patients. Those patients who had *Pa* and/or *Sa* infection at study entry were excluded. The analyses were restricted to 15,176 patients who were reported to CFFR and infection free at the study entry. Hence the onset ages in these 15,176 patients were left truncated at the patient ages at entry. They were also subject to independent administrative censoring or dependent censoring by death. This falls into the paradigm of left-truncated bivariate competing risks data when the lung infections are of interest.

The two time-varying estimates discussed in the method section were computed for this CFFR dataset, where the two event times of *Pa* and *Sa* were competing-risk censored by the same event of death. When only the cause 1 association is concerned, our estimating procedures are valid for the CFFR application, even though the procedures have been developed for the general bivariate competing risks data. For the association measure $\phi(s, t)$, it is not directly affected by the cause 2 event since we are focusing on the cause 1 cumulative hazard functions. For $\psi(s, t)$, we are using the Dabrowska estimator for the bivariate overall survival function in estimating the bivariate CIF. The Dabrowska estimator was developed under the general bivariate censoring setting, but it performs well under univariate censoring based on our unreported simulation results. In addition, the Dabrowska estimator can handle discrete cases when (T_1, T_2) have positive probability along the diagonal. Therefore, the methods developed in Section 2 can be readily applied to the CFFR application to quantify the association between the two lung infection times.

We aimed to examine the association measures between s = 1.5 and t = 20 (years), during which most first infections were acquired. If we defined the at-risk set as $Y(t)=\sum_{i=1}^{n} I(T_i \geq t>V_i)$, then Y(1.5) = 4431, accounting for only 29% of the entire sample. The median truncation time is 2.3 years and 75% subjects had truncation times that were greater than 0.5. To avoid that the at-risk set at early ages is too small, we set a to be 0.5. Figure 1 provides the marginal CIFs and cumulative CSH functions of *Pa* and *Sa* infections between 1.5 and 20 years. We can see that most first infections occurred during this period of time and the cumulative risk of acquiring either infection by age 20 is above 0.9. In contrast, the incidence of death is low during this period of time, with a cumulative risk of death among infection-free children around 0.01 by age 20. Hence, ignoring the competing risks censoring may not change the results much. However, we adopt the competing risks framework to emphasize that infection times are only well defined among living population. In this application, death plays a more important role in left truncation, as those subjects who died earlier were not included in the CFFR. In addition, both hazard functions of first *Pa* and *Sa* infections appear constant over time. This, coupled with low incidence of deaths, implies that the associations quantified based on the CIFs and cumulative CSH may be similar.

Figure 2 presents the estimated bivariate CIF during the period of 1.5 to 20 years, conditional on both infections occurring after 0.5 years after birth. The conditional bivariate CIF starts with 0.06 at (1.5, 1.5) and increases gradually over time approaching 0.72 at (20, 20). In contrast, the CIF of *Pa* infection at year 20 is 0.93 and that of *Sa* infection is 0.96, which suggests negative association between the two infections at age 20. The association between the two infections is more apparent in Fig. 3 that presents the association estimates at the diagonal points. The top panel is the plot of $\hat{\phi}(t, t)$, $t = 1.5 – 20$, based on the cumulative hazard functions. The bottom panel is the plot of $\hat{\psi}_D(t, t)$, which is based on CIFs with the bivariate CIF estimated by using the Dabrowska estimator of the bivariate overall survival function. The estimates and their 95% pointwise confidence intervals are given for each panel. The associations in the cumulative hazards functions have a similar pattern as the associations based on the CIFs though the former is noticeably more variable than the latter.

It is as expected since the hazard functions for the infections are stable and the incidence of death is low. Both curves started with positive associations at early ages and switched to negative associations at age 3.5 years. After age 5, we observe significantly negative associations between the two infection onset ages. The negative association at late ages may suggest bacterial competition. *Pa* and *Sa* infections at late ages tend to be persistent, and chronic colonization of one bacteria organism in lower respiratory track or lungs competes with another for space. On the other hand, early *Pa* and *Sa* infections are more likely to be transient and therefore, competition of these two pathogens would not have been established. *Pa* and *Sa* infections are likely to be positively associated during this early, transient phase, because both *Pa* and *Sa* are common environmental pathogens and hence, patients infected by *Pa* are likely to be those who had poorer clinical status (such as malnutrition) that rendered them to be more susceptible to other pathogens such as *Sa*.

Finally, we wanted to examine the importance of considering left truncation in the association analysis based on left truncated data. We naively applied the original method in Cheng et al. (2007) to this dataset, which completely ignored left truncation and the integration bound was set to be zero. The resulting association estimates are given in Fig. 4, which are noticeably higher than what are shown in Fig. 3. Therefore, it is crucial to properly take into account the effect of left truncation in the analysis in order to remove the spurious positive dependence between the two event times that is introduced by the common left truncation time.

5 Discussion

In this paper, we have developed two nonparametric association estimators for bivariate survival data when both components are subject to left truncation and competing-risk censoring. Since left-truncation and competing event death are very common, especially in diseases associated with aging such as heart disease, cancer, stroke, and Alzheimer's disease, there has been a need to develop novel statistical methods for these complex medical data. The present study provided a sophisticated statistical approach to analyzing left-truncation competing risks data in CFFR. An alternative way of approaching the CFFR data is to formulate the association in terms of transition intensities and transition probabilities under the multistate framework. However, due to the coexisting nature of the two lung infections, transition intensities and transition probabilities may not be as appealing as our proposed association measures that are closely related to standard association measures based on cumulative distribution functions.

Our approach can also be extended to obtain an estimator as the generalization of Prentice-Cai (1992) estimator. This extension would further complicate the computation. In some applications, one might parameterize the distribution of truncated variables as $G(x; \theta)$ (Wang, 1989), where $\theta \in \Theta \subset R^q$, and θ is a q-dimensional vector. Further investigation is required for obtaining semiparametric association estimators.

The proposed method can also be extended tomultivariate competing risks data with left truncation. For an example, if we are interested in quantifying the familial association in ages of onset of first *Pa* infection among CF siblings, we may reasonably assume that the onset ages of infections among siblings have the same distribution. Under this assumption, our method can be extended to left-truncated multivariate competing risks data following the line of research in Tsai (1990); Cheng et al. (2009). It is also worth pointing out our current analysis is based on the independent assumption between (V_1, V_2) and (T_1, T_2), which may be relaxed to be quasi-independence, that is factorization of the joint density of failure and truncation times into a product proportional to the individual densities in the observable region. Quasi-independence has been considered by Tsai (1990) and Martin and Betensky (2005), among others, for left-truncated univariate or bivariate survival times. The quasi-dependence may be quantified through Kendall's concordance measure and estimated based on pairs where the event times are within the observation region and the concordance status is determinable. It will be interesting to develop our association analysis under thisweaker assumption and to formally test whether the quasi-independence holds in the CFFR application. These will be future research topics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Preston Campbell from the US Cystic Fibrosis Foundation for providing the registry data for this research. This work was partially supported by DMS-1207711 (to Cheng) and National Institutes of Health grant DK072126 (to Lai).

References

Aalen O. Nonparametric inference for a family of counting processes. The Annals of Statistics. 1978; 6:701–726.

Andersen, PK., Borgan, O., Gill, RD., Keiding, N. Statistical Models Based on Counting Processes. Springer-Verlag; New York, NY: 1993.

Bandeen-Roche K, Liang K. Modelling multivariate failure time associations in the presence of a competing risk. Biometrika. 2002; 89:299–314.

Bandeen-Roche K, Ning J. Nonparametric estimation of bivariate failure time associations in the presence of a competing risk. Biometrika. 2008; 95:221–232. [PubMed: 20305739]

Burke MD. Estimation of a bivariate survival function under random censorship. Biometrika. 1988; 75:379–382.

Campbell G. Nonparametric bivariate estimation with randomly censored data. Biometrika. 1981; 68:417–422.

Cheng Y, Fine JP. Nonparametric estimation of cause-specific cross hazard ratio with bivariate competing risks data. Biometrika. 2008; 95:233–240.

Cheng Y, Fine JP, Kosorok MR. Nonparametric analysis of bivariate competing risks data. Journal of the American Statistical Association. 2007; 102:1407–1416.

Cheng Y, Fine JP, Kosorok MR. Nonparametric association analysis of exchangeable clustered competing risks data. Biometrics. 2009; 65:385–393. [PubMed: 18549422]

Dabrowska DM. Kaplan–Meier estimate on the plane. The Annals of Statistics. 1988; 16:1475–1489.

Dabrowska DM. Kaplan–Meier estimate on the plane: weak convergence, LIL, and the bootstrap. Journal of Multivariate Analysis. 1989; 29:308–325.

Dai HS, Fu B. A polar coordinate transformation for estimating bivariate survival functions with randomly censored and truncated data. Journal of Statistical Planning and Inference. 2012; 142:248–262.

FitzSimmons SC. The changing epidemiology of cystic fibrosis. Journal of Pediatrics. 1993; 22:19.

Flume PA, Mogayzel PJJ, Robinson KA, Goss CH, Rosenblatt RL, Kuhn RJ, Marshall BC. Clinical practice guidelines for pulmonary therapies committee. Cystic fibrosis pulmonary guidelines: treatment of pulmonary exacerbations. American Journal of Respiratory and Critical Care Medicine. 2009; 180:802–808. [PubMed: 19729669]

Flume PA, O'Sullivan BP, Robinson KA, Goss CH, Mogayzel PJJ, Willey-Courand DB, Bujan J, Finder J, Lester M, Quittell L, Rosenblatt R, Vender RL, Hazle L, Sabadosa K, Marshall B. Cystic fibrosis foundation, pulmonary therapies committee. Cystic fibrosis pulmonary guidelines: chronic medications for maintenance of lung health. American Journal of Respiratory and Critical Care Medicine. 2007; 176:957–969. [PubMed: 17761616]

Gijbels I, Gürler U. Covariance function of a bivariate distribution function estimator for left truncated and right censored data. Statistica Sinica. 1998; 8:1219–1232.

Gray RJ. A class of *K*-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988; 16:1140–1154.

Gross ST, Lai TL. Nonparametric estimation and regression analysis with left truncated and right-censored data. Journal of the American Statistical Association. 1996; 91:1166–1180.

Gürler U. Bivariate estimation with right truncated data. Journal of the American Statistical Association. 1996; 91:1152–1165.

Gürler U. Bivariate distribution and hazard functions when a component is randomly truncated. Journal of Multivariate Analysis. 1997; 60:20–47.

Hougaard, P. Analysis of Multivariate Survival Data. Springer; New York, NY: 2000.

Huang J, Vieland VJ, Wang K. Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation. Statistica Sinica. 2001; 11:1047–1068.

Kalbfleisch, JD., Prentice, RL. The Statistical Analysis of Failure Time Data. 2. JohnWiley & Sons; New York, NY; Chichester, UK: 2002.

Kosorok MR, Zeng L, West SE, Rock MJ, Splaingard ML, Laxova A, Green CG, Collins J, Farrell PM. Acceleration of lung disease in children with cystic fibrosis after pseudomonas aeruginosa acquisition. Pediatric Pulmonology. 2001; 32:277–287. [PubMed: 11568988]

Lai HJ, Cheng Y, Cho H, Kosorok MR, Farrell PM. Association between initial disease presentation, lung disease outcomes and survival in patients with cystic fibrosis. American Journal of Epidemiology. 2004; 159:537–546. [PubMed: 15003957]

Marshall AW, Olkin I. A multivariate exponential distribution. Journal of the American Statistical Association. 1967; 62:30–44.

Martin EC, Betensky RA. Testing quasi-independence of failure and truncation times via conditional Kendall's Tau. Journal of the American Statistical Association. 2005; 100:484–492.

Nelson W. Theory and applications of hazard plotting for censored failure data. Technometrics. 1972; 14:945–966.

Prentice RL, Cai J. Covariance and survivor function estimation using censoredmultivariate failure time data. Biometrika. 1992; 79:495–512.

Pruitt, RC. Technical Report. Department of Statistics, University of Minnesota; 1991. Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis; p. 543

Rogers G, Hoffman L, Whiteley M, Daniels T, Carroll M, Bruce K. Revealing the dynamics of polymicrobial infections: implications for antibiotic therapy. Trends in Microbiology. 2010; 18:357–364. [PubMed: 20554204]

Shen PS. An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right-censoring. Journal of Statistical Planning and Inference. 2006; 136:4365–4384.

Shen PS. Nonparametric estimators of the bivariate survival function under left truncation and right censoring. Communications in Statistics – Theory and Methods. 2010; 39:2877–2889.

Shen PS, Yan YF. Nonparametric estimation of the bivariate survival function with left-truncated and right-censored data. Journal of Statistical Planning and Inference. 2008; 138:4041–4054.

Tsai WY. Testing the assumption of independence of truncation time and failure time. Biometrika. 1990; 77:169–177.

van der Laan MJ. Efficient estimation in the bivariate censoring model and repairing NPMLE. The Annals of Statistics. 1996a; 24:596–627.

van der Laan MJ. Nonparametric estimation of the bivariate survival function with truncated data. Journal of Multivariate Analysis. 1996b; 58:107–131.

Wang MC. A semiparametric model for randomly truncated data. Journal of the American Statistical Association. 1989; 84:742–748.

Woodroofe M. Estimating a distribution function with truncated data (Corr: V15 p883). The Annals of Statistics. 1985; 13:163–177.

Appendix

We now discuss how to construct a modified Dabrowska estimator based on the bivariate left truncated and right censored data. Consider the overall double-event process $W_{11}(u, v) = I\{X_1 \leq u, \delta_1 \neq 0, X_2 \leq v, \delta_2 \neq 0\}$ and its conditional expectation $E\{W_{11}(u,v)|T_1>V_1,T_2>V_2\}=p^{-1}\int_0^v\int_0^u K(x,y)S(dx,dy)$. Thus, we have

$$S(du,dv)=p\frac{E\{W_{11}(du,dv)|T_1>V_1,T_2>V_2\}}{K(u,v)}. \quad (1)$$

By (4) and (6), we have

$$\Lambda_{11}(du,dv)\equiv\frac{S(du,dv)}{S(u-,v-)}=\frac{E\{W_{11}(du,dv)|T_1>V_1,T_2>V_2\}}{E\{R(u,v)|T_1>V_1,T_2>V_2\}}.$$

Similarly,

$$\Lambda_{10}(du,v)\equiv\frac{-S(du,v-)}{S(u-,v-)}=\frac{E\{W_{10}(du,v)|T_1>V_1,T_2>V_2\}}{E\{R(u,v)|T_1>V_1,T_2>V_2\}},$$

where $W_{10}(u,v) = I\{X_1 \le u, \delta_1 \ne 0, V_2 < v \le X_2\}$, and

$$\Lambda_{01}(u,dv)\equiv\frac{-S(u,dv)}{S(u-,v-)}=\frac{E\{W_{01}(u,dv)|T_1>V_1,T_2>V_2\}}{E\{R(u,v)|T_1>V_1,T_2>V_2\}},$$

where $W_{01}(u,v) = I\{X_2 \le v, \delta_2 \ne 0, V_1 < u \le X_1\}$. Note that Λ_{11}, $\Lambda_{10,}$ and Λ_{01} are the conditional failure rates from both subjects or from a single subject, regardless of the causes. Hence, $\Lambda_{11}=\sum_k\sum_l\Lambda_{kl}^{12}$, where Λ_{kl}^{12} are cause-specific hazard functions defined in (2). Now we define

$$\hat{W}_{11}(u,v)=\frac{1}{n}\sum_{i=1}^{n}I\{X_{1i}\le u,\delta_{1i}\ne 0,X_{2i}\le v,\delta_{2i}\ne 0\},$$
$$\hat{W}_{10}(u,v)=\frac{1}{n}\sum_{i=1}^{n}I\{X_{1i}\le u,\delta_{1i}\ne 0,V_{2i}<v\le X_{2i}\},\text{ and}$$
$$\hat{W}_{01}(u,v)=\frac{1}{n}\sum_{i=1}^{n}I\{X_{2i}\le v,\delta_{2i}\ne 0,V_{1i}<u\le X_{1i}\}.$$

Hence, a consistent estimator of $L(dx, dy)$ in the Dabrowska representation is given by

$$\hat{L}(dx,dy)=\frac{\hat{\Lambda}_{10}(dx,y)\hat{\Lambda}_{01}(x,dy)-\hat{\Lambda}_{11}(dx,dy)}{[1-\hat{\Lambda}_{10}(dx,y)][1-\hat{\Lambda}_{01}(x,dy)]},$$

where $\hat{\Lambda}_{11}(dx,dy)=\frac{\hat{W}_{11}(dx,dy)}{r(x,y)}$, $\hat{\Lambda}_{10}(dx,y)=\frac{\hat{W}_{10}(dx,y)}{r(x,y)}$ and $\hat{\Lambda}_{01}(x,dy)=\frac{\hat{W}_{01}(x,dy)}{r(x,y)}$. The conditional marginal survival functions $S_1(s|a)$ and $S_2(t|a)$ can be estimated by

$$\hat{S}_1(s|a)=\prod_{a\le u\le s}(1-\hat{\Lambda}_{10}(du,a))\text{ and }\hat{S}_2(t|a)=\prod_{a\le v\le t}(1-\hat{\Lambda}_{01}(a,dv)).$$

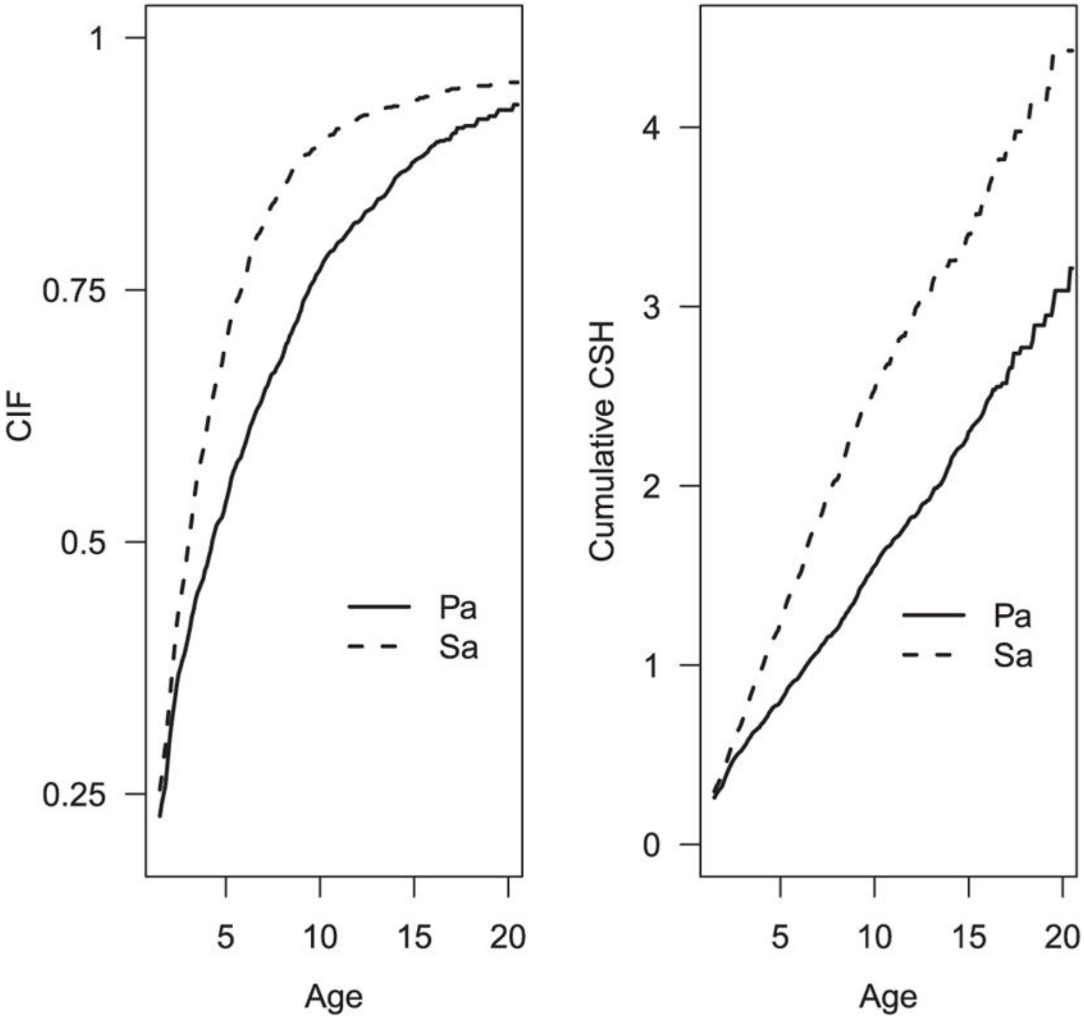

Figure 1.

Marginal CIFs and cumulative CSHs of *Pa* and *Sa* infections over time.

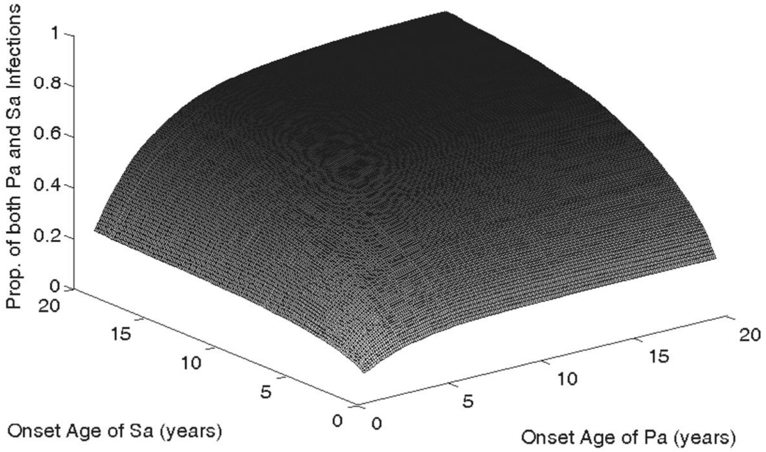

Figure 2.

Bivariate CIF of *Pa* and *Sa* infections over time.

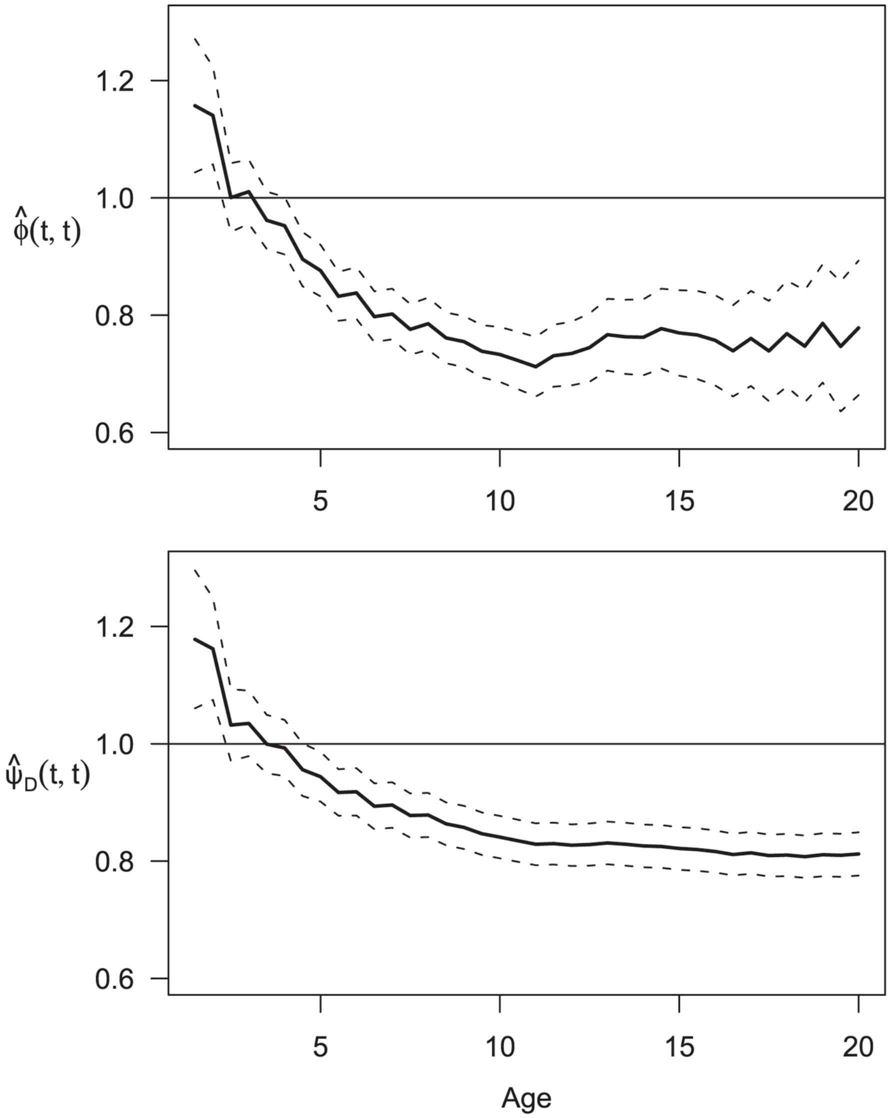

Figure 3.

Time varying association estimates between *Pa* and *Sa* infection. Solid line is point estimate and dash line is 95% pointwise confidence interval. Top: Association analysis based on cumulative CSH. Bottom: Association analysis based on CIFs using the Dabrowska method.

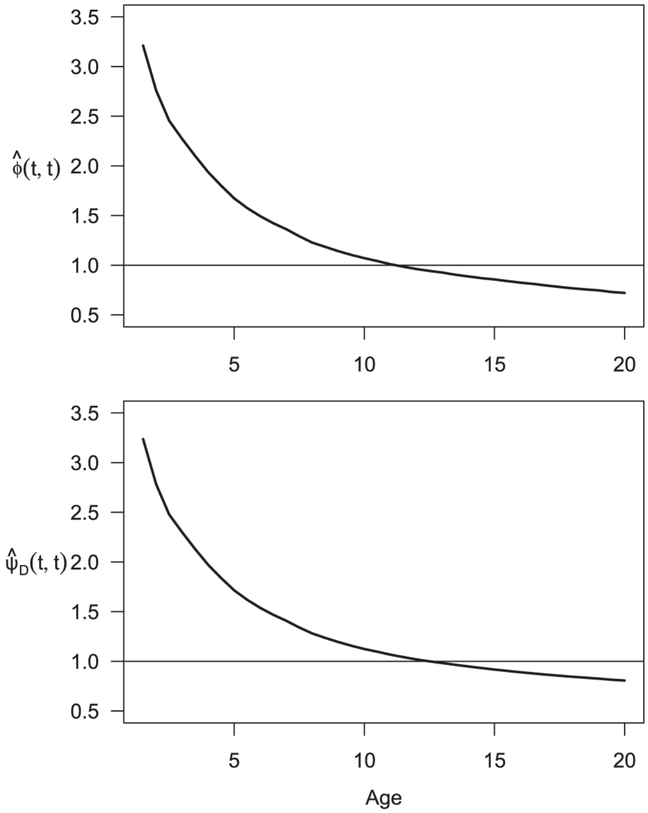

Figure 4.

Naive time varying association estimates between *Pa* and *Sa* infection while ignoring left truncation. Top: Association analysis based on cumulative CSH. Bottom: Association analysis based on CIFs using the Dabrowska method.

Table 1

Simulation results for $\hat{S}_D$ of the conditional bivariate survival function where a = (.05, .05), Est and BSE are averages of the estimate and bootstrap standard error of $\hat{S}_D$, ESE is the empirical standard error of the estimator and Cov is the empirical coverage of a .95 Wald confidence interval based on the BSE.

	n	***S*(.2, .2|*a*)**	***S*(.2, .3|*a*)**	***S*(.3, .3|*a*)**	***S*(.3, .4|*a*)**	***S*(.4, .4|*a*)**	***S*(.4, .5|*a*)**	***S*(.5, .5|*a*)**
True	–	.64	.52	.47	.39	.35	.29	.26
Est	50	.64	.53	.47	.39	.35	.28	.25
	100	.63	.52	.46	.38	.34	.28	.25
ESE	50	.13	.12	.12	.11	.11	.10	.10
	100	.09	.09	.09	.08	.08	.07	.07
BSE	50	.13	.13	.13	.12	.12	.11	.10
	100	.09	.09	.09	.08	.08	.07	.07
Cov	50	.93	.94	.94	.94	.94	.94	.93
	100	.94	.94	.95	.95	.94	.94	.93

Table 2

Rejection rates for nominal 0.05 level tests based on $\hat{\boldsymbol{\phi}}^{*}$ and $\hat{\psi}^{*}_{D}$.

		Rejection rates					
		NULL	**ALT 1**	**ALT 2**	**ALT 3**	**ALT 4**	**ALT 5**
		$\rho_1 = 0$	$\rho_1 = 0.3$	$\rho_1 = 0.5$	$\rho_1 = 0.5$	$\rho_1 = 0.5$	$\rho_1 = 0.5$
DIM	Proportion	$\rho_2 = 0$	$\rho_2 = 0.5$	$\rho_2 = 0.3$	$\rho_2 = -0.3$	$\rho_2 = 0.3$	$\rho_2 = -0.3$
		$\rho_3 = 0$	$\rho_3 = 0$	$\rho_3 = 0$	$\rho_3 = 0$	$\rho_3 = 0.1$	$\rho_3 = -0.1$
	censoring	0.10	0.15	0.19	0.19	0.19	0.19
	$\boldsymbol{\phi}^{*}$	1	1.78	1.61	0.46	1.44	0.61
	$\boldsymbol{\psi}^{*}$	1	2	1.71	0.42	1.55	0.55
200	rejection $\hat{\boldsymbol{\phi}}^{*}$	0.04	0.93	0.68	0.50	0.47	0.26
	rejection $\hat{\psi}^{*}_{D}$	0.04	0.96	0.65	0.46	0.46	0.28
300	rejection $\hat{\boldsymbol{\phi}}^{*}$	0.05	0.99	0.89	0.80	0.64	0.44
	rejection $\hat{\psi}^{*}_{D}$	0.04	0.996	0.85	0.73	0.63	0.49