

Nonparametric estimation of conditional cumulative hazards for missing population marks

Amalia Jácome Pumar ^{*}, Dipankar Bandyopadhyay [†]

Abstract

In this article, we study a novel Nelson-Aalen type estimator of the conditional cumulative hazard function, suitable for the competing risk setup where the population membership (mark) information can be possibly missing for some individuals due to random right-censoring. The standard Nelson-Aalen estimator is not appropriate for this setting. We propose to use imputed population marks for the censored individuals through fractional risk sets that estimate the underlying risk set for the process. Some asymptotic properties such as a strong iid representation and uniform strong consistency of the estimator are established. We study the practical performance of this estimator through simulation and apply it to a real data set for illustration.

Keywords: Competing risks; Fractional risk set; Nelson-Aalen; Right censoring.

^{*}Department of Mathematics, Facultad de Ciencias, Universidade da Coruña, 15071, Spain

[†]Corresponding author, Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC 29425, USA; Email: bandyopd@musc.edu, Tel.: +1 843 876 1603; Fax: +1 843 876 1126

1 Introduction

Consider a multistate model where a healthy (state 0) subject may end up in one of the J states corresponding to different types of failure. Occurrence of any one of the failure precludes the occurrence of the other ones. This setup is commonly referred to as a ‘competing risk’. This model is relevant in several disciplines, such as medicine, demography, actuarial science (as multiple decrement models), economics and manufacturing. Crowder (2001) and Lindqvist (2006) gives us a nice overview of the theory and methods of competing risks. More details on the application of the competing risk model to biostatistics appear in the monographs by Andersen et al. (1993) and Pintilie (2006). However, the data available for the analysis of competing risks are frequently right censored. This may be due to the termination of the study before all the individuals fail, some subjects may die from a cause not related to the study or are lost of follow-up. In this situation, the failure times, along with the population membership (the cause of failure an individual is assigned to), are unknown for the right censored individuals.

An important question, under a competing risk framework, might be to study the subpopulations corresponding to different failure types. For example, in cancer studies, common competing risks are relapse and death in remission (or treatment related mortality). Interest often lies in estimating the rate of occurrence of the competing risks, comparing these rates between treatment groups and modeling the effect of covariates on the rate of occurrence of the competing risks. A nonparametric maximum likelihood estimator (NPMLE) for the competing risk problem, along with martingale interpretations, was proposed by Aalen (1976) under the name ‘multiple decrement models’. These models can be thought of as a special case of the Aalen-Johansen theory of estimation of time-nonhomogeneous Markov processes (Aalen and Johansen, 1978). But we need to confront a potential problem, i.e. the membership/mark information can only be found after the individual actually fails (i.e., through autopsy, etc). Unfortunately, when some individuals are censored, we cannot classify their failure types. As a result, during an observational study these important subpopulation marks will be unavailable for individuals who were right censored and we are unable to assign them to their appropriate ‘at-risk’ sets.

For a better explanation of this problem, consider the following real-life example. In a

study popularly known as the Stanford Heart Transplantation Program (Crowley and Hu, 1977), patients were admitted to the Stanford program for heart replacement. The transplant recipients were subjected to mainly two sub-populations of failure, (here death), viz., transplant rejection, or other causes. However, this data set contains right-censored observations for whom the (eventual) cause of failure was not available. One might be interested in comparing the survival behavior of those patients who died of ‘transplant rejection’ to those who died of ‘other causes’ since the heart transplant. More details in this context and the associated problem of testing the equality of two (or more) survival curves can be found in Bandyopadhyay and Datta (2007) and Bandyopadhyay (2006). Another related study involves a randomized clinical trial of estrogen diethylstilbestrol (DES) (Cheng et al., 1998; Escarela and Carrière, 2003) where patients with Stage 3 and 4 prostate cancer were assigned to four treatment groups. Patients died either due to (a) prostate cancer or (b) other causes, along with right-censored individuals whose membership mark was unknown. One would be interested in studying the overall survival performance of the patients who died due to (a) prostate cancer and (b) other causes, which are the two competing risks in action. We revisit this data set in Section 6.

In competing risks, one of the primary quantities of interest is the cause specific hazard function (for the j th cause) defined as the instantaneous rate of death at time t from cause j among individuals who are still alive at time t in the presence of all causes of failure. Under random right censoring, the classical estimator of the associated cumulative cause specific hazard function is the so-called Nelson-Aalen (NA) estimator (Andersen et al., 1993). Aalen (1978a,b) derived some theoretical asymptotic properties of this estimator, including its consistency and asymptotic normality. However, they may not be appropriate if one desires to understand the summary probability of the different causes of failure given that failures have already occurred due to the competing risks. Despite being intuitively attractive, the cause specific hazard function can only be expressed in terms of observable functions of failure times (viz. marginal hazards) under the assumption of independent competing risks, which may not always be reasonable (Escarela and Carrière, 2003). For our study, we focus on the estimation of the j th conditional cumulative hazard function, defined in (2) as the net or marginal hazard function of the lifetimes of those individuals failing due to the j th risk (i.e. the cumulative hazard of the conditional distribution function F_j^* defined in the

next section). This function is appropriate when the target is the complete distribution and not the sub-distribution function of the lifetimes associated with each risk separately.

The rest of the paper is organized as follows. In Section 2, we introduce the Nelson-Aalen type estimator of the conditional cumulative hazard function based on fractional risk sets, and in Section 3 we give some asymptotic properties, including a strong representation and consistency. Section 4 addresses the case when the failure time and the cause of failure are independent. In Section 5, a small simulation study is carried out to assess the finite sample performance of the novel estimator, and it is applied to a real data in Section 6. The paper ends with a discussion section (Section 7) followed by an Appendix which contains the proofs of the main results.

2 Estimation based on fractional risk sets

We consider the competing risk network as a multistate continuous time stochastic process $\{Z(t), t \in \mathcal{T}\}$ with a finite state space $\mathcal{S} = \{1, \dots, J, 0\}$ having a tree topology and right-continuous sample paths: $Z(t+) = Z(t)$ where we assume that the states $1, \dots, J$ are absorbing whereas state 0 is transient (the root node). Here $\mathcal{T} = [0, \tau]$ where τ is a large possibly observed time point ($\leq \infty$). Typically, for applications, τ will be taken to be the largest time where some event (failure) took place. Let T_i^* be the time the i th individual leaves stage 0 for a failure (stage j , say), with distribution function F , and let X_i^* denote the stage occupied by the i th individual at time T_i^* (i.e., its failure type). A key analytical difficulty that often occurs with time-to-event data is the presence of right censored observations. So, besides the failure time T^* and the failure type X^* , we introduce also the censoring time C , with distribution function G . Hence, while the variables of interest are T^* and X^* , one cannot observe (T_i^*, X_i^*) , but (T_i, δ_i, X_i) , with $T_i = T_i^* \wedge C_i$ the right censored failure time with distribution function H , $\delta_i = \mathbf{1}(T_i^* \leq C_i)$ the failure/censoring indicator, and

$$X_i = X_i^* \delta_i = \begin{cases} j & \text{if } T_i^* \leq C_i \text{ and } X_i^* = j, \\ 0 & \text{if } T_i^* > C_i. \end{cases}$$

Note that δ_i and X_i are observable quantities for every individual, but X_i^* is observed only for the uncensored data. When X_i^* is observed, then it is equal to X_i . It is further assumed

that the censoring variable C is independent of (T^*, X^*) and all the random variables are independent and identically distributed (i.i.d) across the n individuals.

The joint distribution of the pair (T^*, X^*) is completely specified by the cumulative incidence function due to risk j :

$$F_j(t) = P(T^* \leq t, X^* = j),$$

i.e. the sub-distribution function of the individuals failing due to cause j , and the cause specific cumulative hazard function:

$$\Lambda_j(t) = \int_0^t \frac{dF_j(v)}{1 - F(v)}, \quad (1)$$

where $F(t) = \sum_{j=1}^J F_j(t)$ is the distribution function of T^* . The nonparametric maximum likelihood estimator (NPMLE) of Λ_j is given by the well-known Nelson-Aalen estimator (see Andersen et al., 1993):

$$\hat{\Lambda}_j^{NA}(t) = \sum_{T_i \leq t} \frac{\mathbf{1}(X_i = j)}{Y(T_i)} \text{ where } Y(T_i) = \sum_{k=1}^n \mathbf{1}\{T_k \geq T_i\}.$$

Since the main interest in competing risks is often the distribution of lifetime for cause j , we consider, apart from F_j and Λ_j , the following conditional cumulative hazard function

$$\Lambda_j^*(t) = \int_0^t \frac{dF_j^*(v)}{1 - F_j^*(v-)} \quad (2)$$

corresponding to the conditional distribution function

$$F_j^*(t) = P(T^* \leq t | X^* = j), \quad 1 \leq j \leq J,$$

that is, to the failure time distribution due to cause j . These functions are very useful to describe the distribution of the lifetimes due to the j th cause of failure.

The novelty of this paper lies in the fact that there is not any data-driven estimator of Λ_j^* proposed in literature. One may be tempted to estimate Λ_j^* using a Nelson-Aalen type estimator based on the failure times due to cause j . Let $N_j(t)$ be the counting process

counting the number of observed failures of type j (i.e., number of transitions into stage j) in the time interval $[0, t]$:

$$N_j(t) = \sum_{i=1}^n \mathbf{1}(T_i^* \leq t, \delta_i > 0, X_i^* = j) = \sum_{i=1}^n \mathbf{1}(T_i \leq t, X_i = j),$$

and let $Y_j(t)$ denote the number of individuals at risk of failing due to cause j or of getting censored:

$$Y_j(t) = \sum_{i=1}^n \mathbf{1}(T_i \geq t, X_i^* = j). \quad (3)$$

As a stochastic process, Y_j is predictable, i.e. $Y_j(t)$ is \mathcal{F}_{t-} measurable, where $\mathcal{F}_t = \sigma(\{N_j(s), \sum \mathbf{1}(T_i^* \leq s, \delta_i = 0, X_i^* = j) : 0 \leq s \leq t, j \geq 1\})$. Thus, the Nelson-Aalen estimator (Andersen et al., 1993) of cumulative hazards of failure amongst individuals of subpopulation (or failure type) j is the following:

$$\hat{\Lambda}_j^*(t) = \int_0^t \frac{\mathbf{1}(Y_j(v) > 0)}{Y_j(v)} dN_j(v) = \sum_{T_i \leq t} \frac{\mathbf{1}(X_i = j)}{Y_j(T_i)} \text{ with } Y_j(T_i) = \sum_{k=1}^n \mathbf{1}\{T_k \geq T_i, X_k^* = j\} \quad (4)$$

and, using the one-to-one relationship (product integral mapping) between the cumulative hazard function and the survival function, the estimator of F_j^* is given by

$$\hat{F}_j^*(t) = 1 - \prod_{v \leq t} \left(1 - \frac{dN_j(v)}{Y_j(v)}\right) = 1 - \prod_{T_i \leq t} \left(1 - \frac{\mathbf{1}(X_i = j)}{Y_j(T_i)}\right). \quad (5)$$

It is not difficult to see that N_j is computable from the observed data (T_i, X_i) $i = 1, \dots, n$. However, the size of the subpopulation at-risk set Y_j , on the other hand, is not computable, since we cannot classify the failure types of the censored individuals due to the unavailability of all the subpopulation marks X_i^* . In the absence of such an identifier, we may still assign a probability of each individual being in one of the J subpopulations (something like an imputed subpopulation identifier). Once these probabilities are known, we proceed with the supposition that the data be divided into J subpopulations, the risk set of each subpopulation now contains fractional observations with the fractional mass specified by an estimate of the probability that the observation belongs to a particular subpopulation. Thus, we estimate Y_j by Y_j^f , the ‘fractional risk set’ corresponding to the j th cause of failure (Satten and Datta,

1999; Datta et al., 2000) defined as follows

$$Y_j^f(t) = \sum_{i=1}^n \hat{\phi}_{ij} \mathbf{1}(T_i \geq t). \quad (6)$$

Here, $\hat{\phi}_{ij}$ is the estimated probability that the i th individual belongs to the j th subpopulation according to its failure type. A reasonable choice for $\hat{\phi}_{ij}$ is the following:

$$\hat{\phi}_{ij} = \begin{cases} 1, & \text{if } X_i = j \\ 0, & \text{if } X_i > 0, \quad X_i \neq j, \\ \hat{P}_j(T_i, \infty), & \text{if } X_i = 0 \end{cases}$$

where $\hat{P}_j(T_i, \infty)$ is the nonparametric maximum likelihood estimator (NPMLE) of the transition probability

$$P_j(s, t) = P\{T^* \leq t, X^* = j | T^* > s\} = P(\text{fail type } j \text{ by time } t | \text{alive at time } s) \quad (7)$$

evaluated at $s = T_i, t = \infty$. Then, $Y_j^f(t)$ gives the estimated fractional mass in the j th failure type group remaining at risk of failure at time t , counting ($\hat{\phi}_{ij} = 1$) the observations that are uncensored and failed due to cause j , discarding ($\hat{\phi}_{ij} = 0$) the observations that are uncensored and failed due to a cause other than j and, for the censored observations, estimating the probability of that observation being in the j th failure type group ($\hat{\phi}_{ij} = \hat{P}_j(T_i, \infty)$). The ‘fractional risk set’ concept has been recently used in the literature involving multistate models, for example Satten and Datta (1999), Datta and Satten (2000), Bandyopadhyay (2006) and Bandyopadhyay and Datta (2007).

Note that $\hat{P}_j(T_i, \infty)$ is the Aalen-Johansen estimator of the probability of eventually failing due to cause j given being alive at time T_i . This estimator is obtained specializing Andersen et al. (1993) results for a Markov chain to a competing risk setup,

$$\hat{P}_j(s, t) = \int_{(s,t]} \left\{ \prod_{(s,u)} \left(1 - \frac{dN_j(v)}{Y(v)} \right) \right\} \frac{dN_j(u)}{Y(u)}. \quad (8)$$

In the expression (8), $Y(t) = \sum_{i=1}^n \mathbf{1}(T_i \geq t)$ is the size of the ‘at-risk’ set irrespective of failure

types, $N(t) = \sum_{j=1}^J N_j(t)$ is the total number of observed failures of all types (total number of stages entered) by time t . Since N_j and N are discrete with jumps only at the failure times T_i 's, the above integral can be replaced by sums leading to the following simpler expression:

$$\hat{P}_j(s, t) = \sum_{s < T_i \leq t} \left\{ \frac{1 - F_n^{KM}(T_i-)}{1 - F_n^{KM}(s)} \right\} \left\{ \frac{\Delta N_j(T_i)}{Y(T_i)} \right\}, \quad (9)$$

where F_n^{KM} is the Kaplan-Meier (KM) estimator of the distribution function of the failure time due to all causes T^* (see Kaplan and Meier, 1958), and $\Delta N_j(T_i)$ is the number of failures of type j at time T_i . For a typical individual who had not failed up to and including time s , the first term $(1 - F_n^{KM}(T_i-))/(1 - F_n^{KM}(s))$ in (9) computes the probability that such an individual had not failed until time T_i , and the second term $\Delta N_j(T_i)/Y_j(T_i)$ is the probability of its failing at time T_i and the failure is of type j given survival until that time.

Considering expressions (4) and (5), and replacing Y_j with (6), the estimators of Λ_j^* and F_j^* based on fractional risk sets can be easily derived as:

$$\hat{\Lambda}_j^{*f}(t) = \int_0^t \frac{\mathbf{1}(Y_j(v) > 0)}{Y_j^f(v)} dN_j(v) \quad \text{and} \quad \hat{F}_j^{*f}(t) = 1 - \prod_{v \leq t} \left(1 - \frac{dN_j(v)}{Y_j^f(v)} \right).$$

Hence, the failure times among those failing from cause j can be analyzed using a Nelson-Aalen and Kaplan-Meier type estimators, just considering data as partitioned into J groups with fractional masses specified by an estimate of the probability that the observation belongs to a particular group. In fact, the proportion of the sample which fails due to cause j can be estimated as follows:

$$\hat{P}_j(0, \infty) = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_{ij}, \quad j = 1, \dots, J.$$

Finally, note that \hat{F}_j^{*f} coincides with the nonparametric maximum likelihood estimator (NPMLE) of F_j^* (see Satten and Datta, 1999):

$$\hat{F}_j^*(t) = \frac{\hat{P}_j(0, t)}{\hat{P}_j(0, \infty)},$$

with $P_j(s, t)$ given in (7) and estimated by (9).

3 Asymptotic properties

We introduce now some functions before we state the main results for the proposed estimator of Λ_j^* . Define

$$H^{nc}(t) = P(T \leq t, \delta > 0) \text{ and } H_j^{nc}(t) = P(T^* \leq t, \delta > 0, X^* = j) = P(T \leq t, X = j) \quad (10)$$

the (observable) sub-distribution functions of the failure times that are not censored, independently of the cause (H^{nc}) and due to the j th cause (H_j^{nc}),

$$H^c(t) = P(T \leq t, \delta = 0)$$

the sub-distribution function of the censored lifetimes, and consider

$$H_j(t) = P(T \leq t, X^* = j) \text{ and } \bar{H}_j(t) = P(X^* = j) - H_j(t-) = P(T \geq t, X^* = j) \quad (11)$$

the (not observable) sub-distribution functions of the failure times for cause j .

The importance of these functions is clear, since the empirical estimator $\hat{P}(X^* = j) = n_j/n$, where $n_j = \sum_{i=1}^n \mathbf{1}(X_i^* = j)$ is the (not observable) total number of failures due to cause j , together with the empirical estimators

$$\begin{aligned} H_{nj}^{nc}(t) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}(T_i \leq t, X_i = j) = \frac{1}{n} N_j(t), \\ H_{nj}(t-) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}(T_i < t, X_i^* = j) = \frac{1}{n} (n_j - Y_j(t)), \end{aligned} \quad (12)$$

lead to the estimator of the cumulative hazard function in (4):

$$\hat{\Lambda}_j^*(t) = \int_0^t \frac{dH_{nj}^{nc}(v)}{\frac{n_j}{n} - H_{nj}(v-)}.$$

This expression of the estimator $\hat{\Lambda}_j^*$ will be very useful in the derivation of the strong

representation of the ‘fractional risk set’ estimator $\hat{\Lambda}_j^{*f}$, since

$$\hat{\Lambda}_j^{*f}(t) = \hat{\Lambda}_j^*(t) + \int_0^t \left(\frac{1}{Y_j^f(v)} - \frac{1}{Y_j(v)} \right) dN_j(v). \quad (13)$$

Fix $b_H < \sup\{t : H(t) < 1\}$. As a first step, we shall now study some asymptotic properties of the estimator of the transition probabilities $\hat{P}_j(s, t)$ given in (8). We will give first an iid representation and then a consistency result. They are analogue to Theorem 1 in Aalen (1978b) for partial transition probabilities, and Theorem 5.1 in Fleming (1978) for nonhomogeneous Markov processes.

Theorem 3.1 *If the distribution functions F and G are continuous, then*

$$\hat{P}_j(s, t) - P_j(s, t) = \frac{1}{n} \sum_{i=1}^n \zeta_j(T_i, X_i^*, \delta_i, s, t) + r_n(s, t)$$

with

$$\begin{aligned} \zeta_j(T, X^*, \delta, s, t) &= \frac{1}{1 - F(s)} \times \left\{ \frac{1 - F(T)}{1 - H(T)} \mathbf{1}(s \leq T \leq t, \delta > 0, X^* = j) \right. \\ &\quad - \int_s^t \frac{1 - F(v)}{(1 - H(v))^2} \mathbf{1}(T \geq v, X^* = j) dH^{nc}(v) \\ &\quad - \frac{1}{1 - H(T)} \int_s^t (1 - F(v)) \mathbf{1}(s \leq T \leq v, \delta > 0) d\Lambda_j(v) \\ &\quad \left. + \int_s^t (1 - F(v)) \left[\int_0^v \frac{\mathbf{1}(T \geq u)}{(1 - H(u))^2} dH^{nc}(u) \right] d\Lambda_j(v) \right\} \\ &\quad - P_j(s, t) \int_0^s \frac{\mathbf{1}(T \geq v)}{(1 - H(v))^2} dH^{nc}(v) \end{aligned} \quad (14)$$

and $\sup_{0 \leq s \leq t \leq b_H} |r_n(s, t)| = O(n^{-1} \ln n)$ with probability one.

Remark 1 *This result generalizes the iid representation for the Aalen-Johansen estimator evaluated at (t, ∞) , which is a key in the study of many statistical properties in competing risks:*

$$\hat{P}_j(t, \infty) - P_j(t, \infty) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 - F(t)} \int_t^\infty \frac{1 - F(v)}{1 - H(v)} [dM_{ji}(v) - P_j(v, \infty) dM_{.,i}(v)] + r_n(t)$$

where

$$M_{ji}(t) = \mathbf{1}(T_i \leq t, \delta_i > 0, X_i^* = j) - \int_0^t \mathbf{1}(T_i \geq s, X_i^* = j) d\Lambda(s), \quad M_{.,i}(t) = \sum_{j=1}^J M_{ji}(t)$$

and $\sup_{0 \leq t \leq b_H} |r_n(t)| = O(n^{-1} \ln n)$ with probability one.

Corollary 3.1 *The estimator of the transition probabilities satisfies*

$$\sup_{0 \leq s \leq t \leq b_H} n^{1/2} (\ln n)^{-1/2} \left| \hat{P}_j(s, t) - P_j(s, t) \right| \rightarrow 0.$$

Theorem 3.2 *If the distribution functions F and G are continuous, then*

$$\frac{1}{n} \left[Y_j(t) - Y_j^f(t) \right] = \frac{1}{n} \sum_{i=1}^n \rho_j(T_i, X_i^*, \delta_i, t) + s_n(t)$$

with

$$\rho_j(T, X^*, \delta, t) = \mathbf{1}(T \geq t, \delta = 0) [\mathbf{1}(X^* = j) - P_j(T, \infty)] - \int_t^\infty \zeta_j(T, X^*, \delta, v, \infty) d\bar{H}^c(v) \quad (15)$$

where ζ_j given in (15) and $\sup_{0 \leq t \leq b_H} |s_n(t)| = O(n^{-1} \ln n)$ with probability one.

Corollary 3.2 *The fractional risk set estimator satisfies*

$$\sup_{0 \leq t \leq b_H} n^{-1/2} (\ln n)^{-1/2} \left| Y_j(t) - Y_j^f(t) \right| \rightarrow 0.$$

The following theorem gives a representation of the ‘fractional risk set’ estimator $\hat{\Lambda}_j^{*f}$ as a sum of iid variables plus a remainder term. It is based on the strong representation for the transition probabilities in Theorem 3.1.

Theorem 3.3 *If the distribution functions F and G are continuous, then*

$$\hat{\Lambda}_j^{*f}(t) - \Lambda_j^*(t) = \frac{1}{n} \sum_{i=1}^n \xi_j(T_i, X_i^*, \delta_i, t) + R_n(t)$$

where

$$\xi_j(T, X^*, \delta, t) = \frac{\mathbf{1}(T \leq t, X^* = j)}{\overline{H}_j(T)} - \int_0^t \frac{\mathbf{1}(T \leq v, X^* = j)}{\overline{H}_j^2(v)} dH_j^{nc}(v) + \int_0^t \frac{\rho_j(T, X^*, \delta, v)}{\overline{H}_j^2(v)} dH_j^{nc}(v)$$

with ρ_j given in (15) and $\sup_{0 \leq t \leq b_H} |R_j(t)| = O(n^{-1}(\ln n)^3)$.

Corollary 3.3 *The ‘fractional risk set’ estimator of Λ_j^* satisfies*

$$\sup_{0 \leq t \leq b_H} \left(\frac{n}{(\ln n)^3} \right)^{1/2} \left| \hat{\Lambda}_j^{*f}(t) - \Lambda_j^*(t) \right| \rightarrow 0 \quad \text{with probability 1.}$$

Remark 2 *The analogous results for the conditional distribution function estimator \hat{F}_j^{*f} can be easily derived considering the one-to-one mapping relation between the survival function and the cumulative hazard function:*

$$1 - F(t) = \exp(-\Lambda_c(t)) \prod_{u \leq t} (1 - \Delta\Lambda(u)),$$

where Λ_c is the continuous part of Λ , and $\Delta\Lambda(u) = \Lambda(u) - \Lambda(u-)$. The iid representation in Theorem 3.3 can also be applied to kernel-type density and hazard function estimation. The derivation is similar to that of Gijbels and Wang (1993) for LTRC data, although a sharper bound for the remainder term will be needed.

4 Independence of the variables T^* and X^*

The nature of the dependence between T^* and X^* is very useful. Under independence, the lifetimes T^* and the causes of failure X^* can be studied separately, which simplifies the analysis of competing risks to a great extent. As for the cause specific cumulative hazard function Λ_j in (1) and the conditional cumulative hazard function Λ_j^* in (2) studied in this paper, note that

$$\Lambda_j^*(t) = \frac{1}{P(X^* = j)} \Lambda_j(t).$$

Besides, the conditional Λ_j^* reduces to the cumulative hazard function Λ of the lifetimes T^* regardless of the cause of failure, since $F_j^*(t) = F(T^* \leq t)$. In this case, the classical

Nelson-Aalen estimator of Λ :

$$\hat{\Lambda}_n^{NA}(t) = \sum_{T_i \leq t} \frac{\mathbf{1}(\delta_i > 0)}{Y(T_i)} \quad \text{with} \quad Y(T_i) = \sum_{k=1}^n \mathbf{1}(T_k \geq T_i), \quad (16)$$

is the most efficient estimator of Λ (and hence of Λ_j^* for any $j = 1, \dots, J$). This efficiency can be easily be seen taking into account that the process $n^{1/2} \left[\hat{\Lambda}_n^{NA}(t) - \Lambda(t) \right]$ converges weakly to a mean zero Gaussian process with covariance structure (see Lo et al., 1989)

$$\int_0^{t_1 \wedge t_2} \frac{dH^{nc}(v)}{(1 - H(v))^2}.$$

The analogue result for the ‘fractional risks set’ estimator $\hat{\Lambda}_j^{*f}$ when the variables T^* and X^* are independent is given in the following proposition.

Proposition 4.1 *If the variables T^* and X^* are independent, then the process $n^{1/2} \left[\hat{\Lambda}_j^{*f}(t) - \Lambda_j^*(t) \right]$ converges weakly to a mean zero Gaussian process with covariance structure*

$$\frac{1}{P(X^* = j)} \int_0^{t_1 \wedge t_2} \frac{dH^{nc}(v)}{(1 - H(v))^2} + \left(\frac{1}{P(X^* = j)} - 1 \right) \int_0^{t_1} \int_0^{t_2} \frac{g(x_1, x_2) dH^{nc}(x_1) dH^{nc}(x_2)}{(1 - H(x_1))^2 (1 - H(x_2))^2}$$

where

$$g(x_1, x_2) = \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{m(v_1 \vee v_2) dH^c(v_1) dH^c(v_2)}{(1 - F(v_1))(1 - F(v_2))} \quad \text{and} \quad m(v) = \int_v^{\infty} \left(\frac{1 - F(u)}{1 - H(u)} \right)^2 dH^{nc}(u).$$

Dewan et al. (2004) proposed several tests for testing independence between time to failure T^* and the cause of failure X^* based on conditional probabilities involving T^* and X^* , when there is no censoring. According to Dewan et al. (2004), the variables T^* and X^* are independent if and only if the conditional probability

$$\phi_j(t) = P(X^* = j | T^* > t) = P_j(t, \infty) \text{ is a constant, that is, } \phi_j(t) = P(X^* = j).$$

Dykstra et al. (1998) and Kochar and Proschan (1991) provided some restricted tests for censored observations in a competing risks framework. In the competing risks setup with censored observations, the conditional cumulative hazard functions Λ_j^* can be used to test

independence without any restriction since, under independence, the difference between Λ_j^* and Λ_k^* will be close to zero for every $j, k = 1, \dots, J$. This suggests a hypothesis testing setup motivated by obtaining the FRS estimates of Λ_j^* for any $j = 1, \dots, J$, and analyzing if the estimates are close each other. The same is expected for the difference between the distribution functions F_j^* and F . One could think of a Kolmogorov-Smirnov (KS) type test:

$$T_{KS} = \sup_{t \in \mathbb{R}} \max_{j=1, \dots, J} |\hat{F}_j^{*f}(t) - F_n^{KM}(t)|, \quad (17)$$

or a Cramer Von-Mises (CM) test:

$$T_{CM} = \max_{j=1, \dots, J} \int \left(\hat{F}_j^{*f}(t) - F_n^{KM}(t) \right)^2 \omega(t) dt.$$

The limit distribution of these tests follows consequently from the limit distribution of the estimator \hat{F}_j^* , in Proposition 4.1, and that of F_n^{KM} (see Breslow and Crowley, 1974). The study of these tests, though interesting themselves, is out of the scope of this paper and will be considered in the future.

5 Simulation study

We have carried out a small simulation study, in order to assess the practical performance of the ‘fractional risk set’ estimator $\hat{\Lambda}_j^{*f}$ of the conditional cumulative hazard function Λ_j^* . For the sake of comparison, we have computed the Nelson-Aalen estimator $\hat{\Lambda}_n^{NA}$ in (16). Note that $\hat{\Lambda}_n^{NA}$ is an estimator of Λ_j^* with nice theoretical properties when the variables T^* and X^* are independent, since in such a case $\Lambda_j^* = \Lambda$. However, in many situations, this assumption is not always appropriate, and $\hat{\Lambda}_j^{*f}$ will be the only available estimator of Λ_j^* .

We have considered two models with $J = 2$ competing risks. For Model 1 (El-Nouty and Lancar, 2004), we assume the independence between T^* and X^* , whereas in Model 2 the variables T^* and X^* are dependent.

Model 1 The variables T^* and X^* are independent.

$$\begin{aligned} F(t) &= 1 - (1 - t/\tau)^{3/4} \exp(-t^2/2\tau^2) \text{ and } G(t) = 1 - (1 - t/\tau)^{1/4} \exp(t^2/2\tau^2) \\ F_1(t) &= F_2(t) = 0.5F(t) \\ \Lambda_j^*(t) &= t^2/2\tau^2 - 3/4 \ln(1 - t/\tau) \text{ for } j = 1, 2. \end{aligned}$$

Model 2 The variables T^* and X^* are dependent.

$$\begin{aligned} F(t) &= t/\tau \text{ and } G(t) = 1 - (1 - t/\tau)^2 \\ F_1(t) &= t/2\tau(1 - (1 - t/\tau)^2) \text{ and } F_2(t) = t/2\tau(1 + (1 - t/\tau)^2) \\ \Lambda_1^*(t) &= -\ln[1 - t/\tau(1 - (1 - t/\tau)^2)] \text{ and } \Lambda_2^*(t) = -\ln[1 - t/\tau(1 + (1 - t/\tau)^2)] \end{aligned}$$

Note that, for Model 1, $\phi_j(t) = P(X^* = j) = \frac{1}{2}$ for $j = 1, 2$ and therefore, the variables T^* and X^* are independent. However, for Model 2, we have

$$\phi_1(t) = \frac{1}{2} \left(1 + \frac{t}{\tau} - \frac{t^2}{\tau^2} \right) \text{ and } \phi_2(t) = \frac{1}{2} \left(1 - \frac{t}{\tau} + \frac{t^2}{\tau^2} \right),$$

and hence, T^* and X^* are dependent.

We have chosen $\tau = 1461$ (4 years), $\lambda = 1/730$ and $\alpha = 1$. In this case, the percentage of censoring is about 8.33% in Model 1 and 33% in Model 2.

We have obtained the conditional distribution functions F_j^* and F for both models, and computed the KS test in (17) using 1000 Monte Carlo samples assuming a sample size $n = 100$. Figure 1 shows the histogram plots of the values of the KS tests for both Models 1 and 2.

PUT FIGURE 1 ABOUT HERE

As expected, the KS test takes much lower values in Model 1 where T^* and X^* are independent, than in Model 2. This shows that the KS and CM tests are promising to test the independence between T^* and X^* .

6 Example: Prostate Cancer Data

We now illustrate the fractional risk set based estimator of the crude cumulative hazard function using the prostate cancer data described in the introduction.

We consider the same data set studied by Byar and Green (1980) and published in Andrews and Herzberg (1985). In those papers, the randomized trial was aimed to compare the different levels of diethylstilbestrol (DES), a drug to treat prostate cancer with respect to patient survival. A total of 506 Stage 3 and 4 prostate cancer patients were assigned to four treatment groups, viz. placebo, 0.2 mg DES/day, 1 mg DES/day and 5 mg DES/day. Because of the potentially fatal cardiovascular adverse effect from DES (Escarela and Carrière, 2003), the assessment of risk-benefit analysis of DES (Cheng et al., 1998) must take into account not only the death time from (a) prostate cancer, but also (b) other competing causes of death, which includes death due to cardiovascular related causes (treatment related mortality). Although we do not consider the set of covariates viz. age, weight index, performance rating, cardiovascular disease history, etc, in this data, a set of 23 patients having incomplete covariate information were removed from our analysis. Out of the 483 patients with complete covariate information, there were 125 (about 26%) deaths from prostate cancer, 219 (about 45%) deaths from ‘other causes’, along with 139 (about 29%) right censored observations whose subpopulation membership is unknown.

We are interested in the distribution of time to death T^* from (a) prostate cancer ($X^* = 1$) and (b) other different causes including cardiovascular related causes ($X^* = 2$), and also to verify the dependence/independence between the variables T^* and X^* . The log-rank test defined in Bandyopadhyay and Datta (2007) tests for the null hypothesis

$$H_0 : F_1^*(t) = \dots = F_J^*(t) (\equiv F^*(t) \text{ say}) \quad \forall 1 \leq j \leq J.$$

Note that H_0 is equivalent to the independence between T^* and X^* . The log-rank test statistic for this data set is 5.989, with a bootstrap estimated standard error of 14.275 leading to a χ^2 statistic of 0.176, which is not significant at 5% level. This implies that we cannot reject the independence between T^* and X^* . To verify the conclusion based on log-rank tests, we have computed the classical Nelson-Aalen estimator of Λ_j and the FRS estimator of Λ_j^* with $j = 1, 2$. Table 1 illustrates the estimated cumulative hazards at the

quartiles for the two competing risks using both the fractional risk set and the usual NA estimator. It can be seen that the estimated cumulative hazards for the two competing risks using the FRS estimator are much closer in contrary to the traditional Nelson-Aalen estimates. This is also supported by the overlapping plots (Figure 2) of the estimated FRS based cumulative hazards for the two competing risks. This implies that conditionally on the cause of failure X^* , the lifetimes of the prostate cancer patients seem to have the same distribution which agrees to the fact that the failure time and failure cause are independent. So, we can conclude that the application of the drug DES doesn't provide any differential effect on the overall survival behavior of the patients who died due to prostate cancer from those who died due to other causes.

PUT TABLE 1 ABOUT HERE

PUT FIGURE 2 ABOUT HERE

7 Discussion

A Nelson-Aalen type estimator of the j th conditional cumulative hazard function Λ_j^* appropriate under a competing risk framework has been proposed when the population marks of the right censored individuals in the study are unknown. Such incomplete data are common in practice and are tackled by the concept of fractional or imputed risk membership. The key is to split the total risk set at every time point t into J possible sub at-risk sets (corresponding to J sub-populations) also called fractional risk-set (FRS). This NA type estimator is unique and the authors are not aware of any other estimator being proposed earlier in literature that deals with this problem of estimating the j th conditional cumulative hazard. In a related testing of hypothesis scenario, Bandyopadhyay and Datta (2007) showed that throwing away the censored observations (equivalent to contributing zero mass to each 'at-risk' set) results in loss of power though maintaining proper size. Thus, the 'fractional risk set' provides a way to include the censored observations in the appropriate 'at-risk' set with nice probability interpretations. We have studied the theoretical properties of the FRS based estimator of Λ_j^* although they lead to complicated asymptotics. We have illustrated with a simulation study that, apart from the interest of studying the conditional survival distributions of the subjects subjected to different competing risks on the overall, the FRS

based estimators of F_j^* and Λ_j^* can also be used to test the independence between T^* and X^* .

Besides, although we have restricted our attention to the estimation of the conditional cumulative hazards under a competing risk framework, this can be extended much further to more complicated multistate networks like the three-stage irreversible illness-death model (Anderson et al., 1993). The concept of ‘fractional-risk-set’ allow us to compare the conditional cumulative hazards among different states in the model. This is also a subject of future research.

Acknowledgements

The first author would like to acknowledge the economic support of the Grant MTM2005-00429 (FEDER funding included) of the Spanish Ministerio de Educación y Ciencia and XUGA Grant PGIDT03PXIC10505PN. Part of this work was done while the second author was a doctoral student at the University of Georgia, USA. He acknowledges the research support from the University of Georgia for awarding him a Graduate School Dissertation Completion Fellowship. This research was supported in part by NIH/NCRR Grant P20 RR017696-04. The authors thank Prof. Somnath Datta for his valuable comments and for drawing attention to the FRS methodology.

Appendix

Proof of Theorem 3.1. The transition probabilities in (7) can also be written as follows:

$$P_j(s, t) = \frac{1}{1 - F(s)} \int_s^t (1 - F(v)) d\Lambda_j(v),$$

with F the distribution function of T^* , and Λ_j the cause specific cumulative hazard function given in (1). For the Aalen-Johansen estimator of $P_j(s, t)$ in (8), we have

$$\hat{P}_j(s, t) = \frac{1}{1 - F_n^{KM}(s)} \int_s^t (1 - F_n^{KM}(v)) d\hat{\Lambda}_j^{NA}(v)$$

where F_n^{KM} is the product-limit Kaplan-Meier (KM) estimator, and $\hat{\Lambda}_j^{NA}$ the Nelson-Aalen (NA) estimator of Λ_j . Then, the iid representation of $\hat{P}_j(s, t) - P_j(s, t)$ comes from the iid representation of the KM estimator F_n^{KM} (see Lo et al., 1989) and that of the NA cumulative hazard function estimator $\hat{\Lambda}_j^{NA}$ (see Lo et al., 1989). ■

Proof of Corollary 3.1. Applying the strong uniform consistency results for the KM and NA estimators (see Lo et al., 1989), the result is straightforward. ■

Proof of Theorem 3.2. Recall the definition of the fractional risk set Y_j^f in (6) and the risk set Y_j in (3). Therefore,

$$\begin{aligned} \frac{1}{n} \left[Y_j(t) - Y_j^f(t) \right] &= \frac{1}{n} \sum_{i=1}^n \left[\mathbf{1}(X_i^* = j) - P_j(T_i, \infty) \right] \mathbf{1}(T_i \geq t, \delta_i = 0) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left[P_j(T_i, \infty) - \hat{P}_j(T_i, \infty) \right] \mathbf{1}(T_i \geq t, \delta_i = 0). \end{aligned} \quad (18)$$

The first term is a normalized sum of zero mean summands (see Lemma 3.1 in Satten and Datta, 2000). The second term in (18) can be written as follows:

$$- \int_t^\infty \left[\hat{P}_j(v, \infty) - P_j(v, \infty) \right] d\bar{H}_n^c(v) = -\frac{1}{n} \sum_{i=1}^n \int_t^\infty \zeta_j(T_i, X_i, \delta_i, v, \infty) d\bar{H}^c(v) + s_n(t)$$

with $\sup_{0 \leq t \leq b_H} |s_n(t)| = O(n^{-1} \ln n)$. This concludes the proof. ■

Proof of Corollary 3.2. Recall the decomposition (18). The triangular inequality allows to bound the first term by the sum:

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}(T_i \geq t, \delta_i = 0) \mathbf{1}(X_i^* = j) - P(T \geq t, \delta = 0, X^* = j) \right| \\ &+ \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}(T_i \geq t, \delta_i = 0) P_j(T_i, \infty) - P(T \geq t, \delta = 0, X^* = j) \right|. \end{aligned}$$

These summands are the absolute error of estimation of certain empirical distribution functions. An immediate consequence of the Dvoretzky-Kiefer-Woldfowitz (DKW) bound for empirical measures yields that the supremum of the first term in (18) is $O\left(n^{-1/2} (\ln n)^{1/2}\right)$

almost surely. The second term in (18) can be written as follows:

$$- \int_0^\infty \left[\hat{P}_j(v, \infty) - P_j(v, \infty) \right] \mathbf{1}(v \geq t) dH_n^c(v).$$

Then, applying Corollary 3.1, the supremo of the second term in (18) is $o(n^{-1/2} \ln n)$ almost surely. This concludes the proof. ■

The outline of the proof of Theorem 3.3 is similar to that of Theorem 1 of Major and Rejto (1988). We start with a few preliminary results which will be useful for some methods in the main body of the proof of Theorem 3.3 and Corollary 3.3. Consider the function \bar{H}_j in (11) and its empirical estimator

$$\bar{H}_{nj}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(T_i \geq t, X_i^* = j). \quad (19)$$

The following lemmas give some consistency results for \bar{H}_{nj} .

Lema 1 $\sup_{0 \leq t \leq \infty} |\bar{H}_j(t) - \bar{H}_{nj}(t)| = O\left(n^{-1/2} (\ln n)^{1/2}\right)$ *a.s.*

Proof It is an immediate consequence of the DKW bound for empirical measures. ■

Lema 2 $\sup_{k: T_k \leq b_H} \bar{H}_j(T_k) / \bar{H}_{nj}(T_k) = O(\ln n)$ *with probability one.*

Proof The proof follows the same steps as in the proof of Lemma 1.1 in Stute (1993). ■

Proof of Theorem 3.3. Recall the definition of Λ_j^* from (2) and the representation (13) of the estimator based on fractional risk set. Therefore,

$$\begin{aligned} \hat{\Lambda}_j^{*f}(t) - \Lambda_j^*(t) &= \int_0^t \frac{d[H_{nj}^{nc}(v) - H_j^{nc}(v)]}{\bar{H}_j(v)} - \int_0^t \frac{\bar{H}_{nj}(v) - \bar{H}_j(v)}{\bar{H}_j^2(v)} dH_j^{nc}(v) \\ &\quad + R_{j1}(t) + R_{j2}(t) + R_{j3}(t), \end{aligned} \quad (20)$$

where

$$\begin{aligned}
R_{j1}(t) &= \int_0^t \frac{\bar{H}_j(v) - \bar{H}_{nj}(v)}{\bar{H}_j^2(v)} d[H_{nj}^{nc}(v) - H_j^{nc}(v)], \\
R_{j2}(t) &= \int_0^t \frac{(\bar{H}_j(v) - \bar{H}_{nj}(v))^2}{\bar{H}_{nj}(v)\bar{H}_j^2(v)} dH_{nj}^{nc}(v), \\
R_{j3}(t) &= \int_0^t \mathbf{1}(Y_j(v) > 0) \left(\frac{1}{Y_j^f(v)} - \frac{1}{Y_j(v)} \right) dN_j(v)
\end{aligned}$$

with H_j^{nc} , H_{nj}^{nc} , \bar{H}_j and \bar{H}_{nj} given in (10), (12), (11) and (19) respectively.

The term R_{j1} can be decomposed into four terms:

$$R_{j1}(t) = \int_0^t \frac{dH_{nj}^{nc}(v)}{\bar{H}_j(v)} - \int_0^t \frac{\bar{H}_{nj}(v)}{\bar{H}_j^2(v)} dH_{nj}^{nc}(v) - \int_0^t \frac{dH_j^{nc}(v)}{\bar{H}_j(v)} + \int_0^t \frac{\bar{H}_{nj}(v)}{\bar{H}_j^2(v)} dH_j^{nc}(v). \quad (21)$$

The second integral in (21) is a V -statistic of order two. We work with it as follows:

$$\begin{aligned}
\int_0^t \frac{\bar{H}_{nj}(v)}{\bar{H}_j^2(v)} dH_{nj}^{nc}(v) &= \frac{n-1}{n} \int_0^t \frac{dH_{nj}^{nc}(v)}{\bar{H}_j(v)} + \frac{n-1}{n} \int_0^t \frac{\bar{H}_{nj}(v)}{\bar{H}_j^2(v)} dH_j^{nc}(v) \\
&\quad - \frac{n-1}{n} \int_0^t \frac{dH_j^{nc}(v)}{\bar{H}_j(v)} + Q_n(t)
\end{aligned} \quad (22)$$

where we split up the integral into two terms, its diagonal and off-diagonal part and obtain the Hájek projection of the U-statistic. Note that $Q_n(t)$ is a degenerate U-statistic of order two and then (see Section 5.3.3 in Serfling, 1980), for each $\delta > 3/2$

$$\sup_{0 \leq t \leq b_H} |Q_n(t)| = o(n^{-1} \ln n) \text{ with probability 1.}$$

Application of the SLLN to each of the remaining processes in (22) allows us to replace $(n-1)/n$ by 1, so that, from (21) we have

$$\sup_{0 \leq t \leq b_H} |R_{j1}(t)| = o(n^{-1} \ln n) \text{ with probability 1.} \quad (23)$$

The term R_{j2} can be bounded as follows:

$$R_{j2}(t) \leq \sup_{0 \leq v \leq \infty} |\bar{H}_j(v) - \bar{H}_{nj}(v)|^2 \sup_{0 \leq T_k \leq b_H} \frac{\bar{H}_j(T_k)}{\bar{H}_{nj}(T_k)} \int_0^t \frac{dH_{nj}^{nc}(v)}{\bar{H}_j^3(v)}.$$

Immediate consequence of Lemmas 1 and 2 and the SLLN is

$$\sup_{0 \leq t \leq b_H} |R_{j2}(t)| = O(n^{-1}(\ln n)^2) \quad \text{with probability 1.} \quad (24)$$

For R_{j3} , consider the decomposition

$$\begin{aligned} R_{j3}(t) &= n^{-1} \int_0^t \frac{Y_j(v) - Y_j^f(v)}{\bar{H}_j^2(v)} dH_j^{nc}(v) \\ &\quad + n^{-1} \int_0^t [Y_j(v) - Y_j^f(v)] \left(\frac{1}{n^{-2}Y_j^f(v)Y_j(v)} - \frac{1}{\bar{H}_j^2(v)} \right) dH_j^{nc}(v) \\ &\quad + n \int_0^t \frac{Y_j(v) - Y_j^f(v)}{Y_j^f(v)Y_j(v)} d[H_{nj}^{nc} - H_j^{nc}](v). \end{aligned} \quad (25)$$

The last term in (25) is, in absolute value, lower than

$$\sup_{0 \leq t \leq b_H} \left| \frac{1}{n} (Y_j(t) - Y_j^f(t)) \right| \left(\sup_{0 \leq t \leq b_H} \left| \frac{\bar{H}_j(t)}{\bar{H}_{nj}(t)} \right| \right)^2 \frac{1}{\bar{H}_j^2(b_H)} \sup_{0 \leq t \leq b_H} |H_{nj}^{nc}(t) - H_j^{nc}(t)|$$

and, by Corollary 3.2, Lemma 2 and the DKW bound for empirical measures, the rate is $O(n^{-1}(\ln n)^3)$ with probability one.

With respect to the second integral in (25), we proceed as follows:

$$\begin{aligned} &\frac{1}{n^{-2}Y_j^f(v)Y_j(v)} - \frac{1}{\bar{H}_j^2(v)} \\ &= \frac{1}{n^{-1}Y_j(v)} \left(\frac{\bar{H}_j(v) - n^{-1}Y_j^f(v)}{n^{-1}Y_j^f(v)\bar{H}_j(v)} \right) + \frac{1}{\bar{H}_j(v)} \left(\frac{\bar{H}_j(v) - n^{-1}Y_j(v)}{n^{-1}Y_j(v)\bar{H}_j(v)} \right). \end{aligned}$$

Now, applying Corollary 3.2, Lemma 1 and the DKW bound, we have

$$\sup_{0 \leq t \leq b_H} \left| n^{-1} \int_0^t [Y_j(v) - Y_j^f(v)] \left(\frac{1}{n^{-2}Y_j^f(v)Y_j(v)} - \frac{1}{\bar{H}_j^2(v)} \right) dH_j^{nc}(v) \right| = O(n^{-1} \ln n)$$

with probability one.

For the first integral in (25), the dominant term comes from the result in Theorem 3.2:

$$\frac{1}{n} \sum_{i=1}^n \int_0^t \frac{\rho_j(T_i, X_i^*, \delta_i, v)}{\bar{H}_j^2(v)} dH_j^{nc}(v) + \int_0^t \frac{s_n(v)}{\bar{H}_j^2(v)} dH_j^{nc}(v)$$

with $\sup_{0 \leq t \leq b_H} |s_n(t)| = O(n^{-1} \ln n)$ with probability 1. Therefore,

$$R_{j3}(t) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{\rho_j(T_i, X_i^*, \delta_i, v)}{\bar{H}_j^2(v)} dH_j^{nc}(v) + R_{j4}(t) \quad (26)$$

with $\sup_{0 \leq t \leq b_H} |R_{j4}(t)| = O(n^{-1} \ln n)$ with probability 1.

The proof is finished using the decomposition (20) and the rates (23), (24) and (26). ■

Proof of Corollary 3.3. Consider the decomposition

$$\begin{aligned} \hat{\Lambda}_j^{*f}(t) - \Lambda_j^*(t) &= \int_0^t \left(\frac{1}{\bar{H}_{nj}(v)} - \frac{1}{\bar{H}_j(v)} \right) dH_{nj}^{nc}(v) + \int_0^t \frac{d[H_{nj}^{nc}(v) - H_j^{nc}(v)]}{\bar{H}_j(v)} \\ &+ \int_0^t \left(\frac{1}{Y_j^f(v)} - \frac{1}{Y_j(v)} \right) dN_j(v). \end{aligned} \quad (27)$$

It follows from Lemmas 1 and 2 and the SLLN, that the absolute value of the first term in (27) is $O\left(n^{-1/2} (\ln n)^{3/2}\right)$ with probability 1.

For the second term in (27) we apply integration by parts and the DKW bound for empirical measures. Then, it is $O\left(n^{-1/2} (\ln n)^{1/2}\right)$ with probability one.

To prove that the third term in (27) is negligible, note that it can be written as follows:

$$n \int_0^t \frac{Y_j(v) - Y_j^f(v)}{n^{-2} Y_j^2(v)} dH_{nj}^{nc}(v) - n \int_0^t \frac{Y_j(v) - Y_j^f(v)}{n^{-2} Y_j^f(v) Y_j(v)} \left(\frac{Y_j^f(v)}{Y_j(v)} - 1 \right) dH_{nj}^{nc}(v). \quad (28)$$

The second term in (28) is negligible with respect to the first one, which can be bounded by

$$\sup_{0 \leq t \leq b_H} \left| \frac{1}{n} (Y_j(t) - Y_j^f(t)) \right| \left(\sup_{k: T_k \leq b_H} \left| \frac{\bar{H}_j(T_k)}{\bar{H}_{nj}(T_k)} \right| \right)^2 \frac{1}{\bar{H}_j^2(b_H)} \int_0^{b_H} dH_{nj}^{nc}(v).$$

Therefore, from Corollary 3.2, Lemma 2 and the SLLN, the second term in (28) is $O(n^{-1/2} \ln n)$ with probability one. This concludes the proof. ■

Proof of Proposition 4.1. Recall the iid representation of the FRS estimator in Theorem 3.3. Long, although straightforward calculations give $E[\xi_j(T, X^*, \delta, t)] = 0$ and the covariance structure

$$\int_0^{t_1 \wedge t_2} \frac{dH_j^{nc}(v)}{\overline{H}_j^2(v)} + \int_0^{t_1} \int_0^{t_2} \frac{E[\rho_j(T, X^*, \delta, x_1) \rho_j(T, X^*, \delta, x_2)]}{\overline{H}_j^2(x_1) \overline{H}_j^2(x_2)} dH_j^{nc}(x_1) dH_j^{nc}(x_2),$$

where

$$\begin{aligned} & E[\rho_j(T, X^*, \delta, x_1) \rho_j(T, X^*, \delta, x_2)] \\ &= \int_{x_1}^{\infty} \int_{x_2}^{\infty} E[\zeta_j(T, X^*, \delta, v_1, \infty) \zeta_j(T, X^*, \delta, v_2, \infty)] dH^c(v_1) dH^c(v_2). \end{aligned}$$

This covariance reduces to the one in Proposition 4.1 taking into account that, if the variables T^* and X^* are independent, then

$$d\Lambda_j(t) = P_j(t, \infty) d\Lambda(t).$$

This completes the proof. ■

References

- [1] Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scand. J. Statist.* **3**, 15–27.
- [2] Aalen, O. (1978a). Nonparametric estimation of partial transition probabilities in multiple decrecement models. *Ann. Statist.* **6**, 534–545.
- [3] Aalen, O. (1978b). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.
- [4] Aalen, O. Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Statist.* **5**, 141–150.
- [5] Andersen, P.K. Borgan, Ø. Gill, R.D. Keiding, N. (1993). *Statistical Models based on Counting Processes*. Springer Series in Statistics. Springer-Verlag, New York.
- [6] Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from many fields for the Student and Research Worker*, Springer-Verlag, New York.

- [7] Bandyopadhyay, D. (2006). Novel nonparametric methods for event time data. *Ph.D. dissertation*, University of Georgia, Athens, GA, U.S.A.
- [8] Bandyopadhyay, D. and Datta, S. (2007). Testing equality of survival distributions when the population marks are missing. *J. Statist. Plann. Inference*, doi:10.1016/j.jspi.2007.06.028.
- [9] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics* **2**, 437–453.
- [10] Byar, D.P. and Green, S.B. (1980). The choice of treatment for cancer patients beased on covariate information: Applications to prostate cancer. *Bulletin Cancer, Paris* **67**, 477-488.
- [11] Cheng, S.C., Fine, J. and Wei, L.J. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **54**, 219-228.
- [12] Crowder, M.J. (2001). Classical Competing Risks. Chapman and Hall, London.
- [13] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Amer. Statist. Assoc.* **72**, 27–36.
- [14] Datta, S. and Satten, G.A. (2000). Estimating future stage entry and occupation probabilities in a multistage model based on randomly right-censored data. *Statist. Probab. Lett.* **50**, 89–95.
- [15] Datta, S. and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statist. Probab. Lett.* **55**, 403–411.
- [16] Dewan, I. Deshpande, J.V. and Kulathinal, S.B. (2004). On testing dependence between time to failure and cause of failure via conditional probabilities. *Scand. J. Statist.* **31**, 79–91.
- [17] Dykstra, R., Kochar, S.C. and Robertson, T. (1998). Restricted tests for testing independence of time to failure and cause of failure in a competing risks model. *Canad. J. Statist.* **26**, 57–68.

- [18] El-Nouty, C. and Lancar, R. (2004). The presmoothed Nelson-Aalen estimator in the competing risk model. *Comm. Statist. Theory Methods* **33**, 135–151.
- [19] Escarela, G. and Carrière, J.F. (2003). Fitting competing risks with an assumed copula. *Stat. Methods Med. Res.* **12**, 333–349.
- [20] Fleming, T.R. (1978). Nonparametric estimation for nonhomogenous Markov processes in the problem of competing risks. *Ann. Statist.* **6**, 1057-1070.
- [21] Gijbels, I. and Wang, J.L. (1993). Strong representations of the survival function estimator for truncated and censored data with applications. *J. Multivariate Anal.* **47**, 210–229.
- [22] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.
- [23] Kochar, S.C. and Proschan, F. (1991). Independence of time and cause of failure in the multiple dependent competing risks model. *Statist. Sinica* **1**, 295–299.
- [24] Lindqvist, B.H. (2006). Competing Risks. In *Encyclopedia of Statistics in Quality and Reliability* (eds Ruggeri, F, Kenett, R. and Faltin, F.), Wiley, NY. (to appear)
- [25] Lo, S.H. Mack, Y.P. and Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Related Fields* **80**, 461–473.
- [26] Major, P. and Rejtö, L.(1988). Strong embedding of the estimator of the distribution function under random censorship. *Ann. Statist.* **16**, 1113–1132.
- [27] Pintilie, M. (2006). *Competing Risks: a Practical Perspective*, Wiley, NY.
- [28] Satten, G.A. and Datta, S. (1999). Kaplan-Meier representation of competing risk estimates. *Statist. Probab. Lett.* **42**, 299–304.
- [29] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York.
- [30] Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *Ann. Statist.* **21** 146-156.

Estimated cumulative hazards functions				
Time	Prostate (FRS)	Prostate (NA)	Other causes (FRS)	Other causes (NA)
1st Quartile	0.4185	0.1159	0.4732	0.2143
Median	1.0705	0.2632	1.1582	0.4739
3rd Quartile	1.7237	0.3752	1.9843	0.6845

Table 1: Estimated cumulative hazard functions for the Prostate Cancer data computed using FRS (fractional risk set) and NA (Nelson-Aalen) estimators for the two competing risks, viz. (a) Prostate cancer and (b) Other causes.

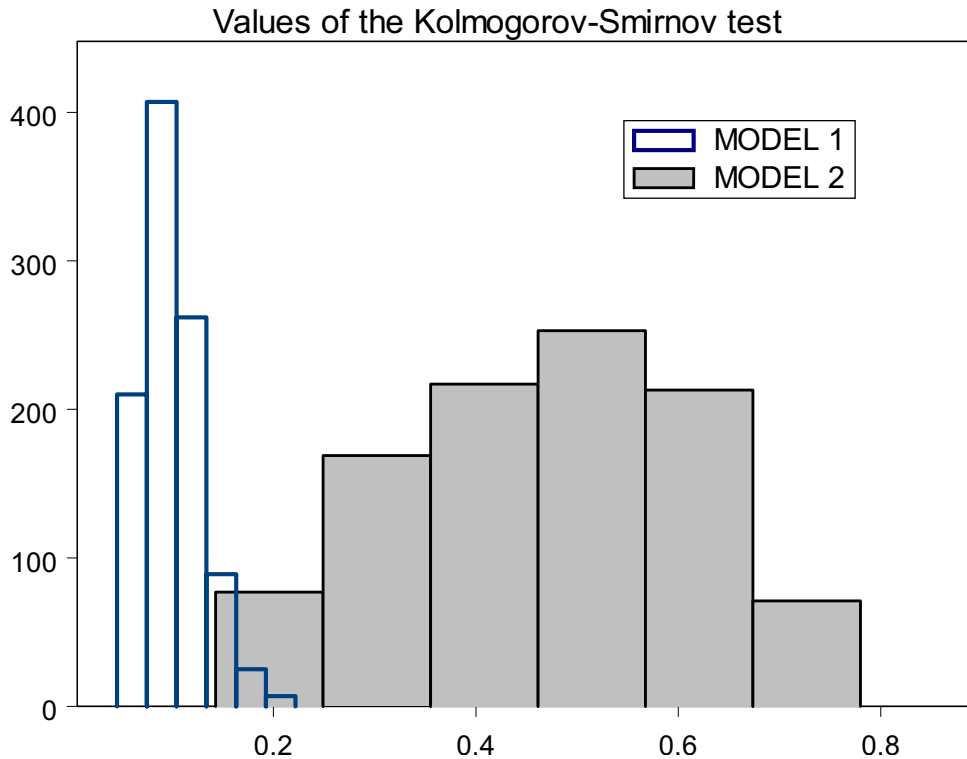


Figure 1: Values of the Kolmogorov-Smirnov test T_{KS} for Model 1 (under H_0 , T^* and X^* are independent) and Model 2 (under H_1).

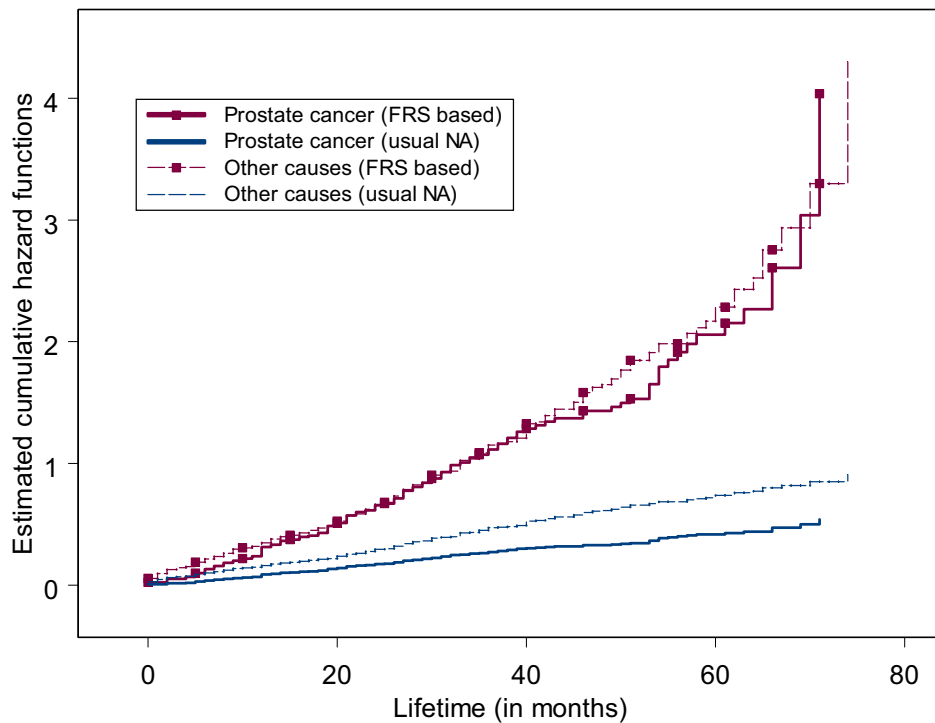


Figure 2: Estimated cumulative hazard functions for the Prostate Cancer data computed using FRS (fractional risk set) and NA (Usual Nelson-Aalen) estimators for the two competing risks, viz. (a) Prostate cancer and (b) Other causes.