Product-limit estimators and Cox regression with missing cause-of-failure information

Ian W. McKeague
Department of Statistics
Florida State University

Sundar Subramanian
Department of Mathematics & Statistics
University of Massachusetts

Abstract

The Kaplan–Meier estimator of a survival function is used when cause of failure (censored or non-censored) is always observed. A method of survival function estimation is developed under the assumption that the failure indicators are missing completely at random (MCR). The resulting estimator is a smooth functional of the Nelson–Aalen estimators of certain cumulative transition intensities. The asymptotic properties of this estimator are derived. A simulation study shows that the proposed estimator has greater efficiency than competing MCR-based estimators. The approach is extended to the Cox model setting for the estimation of a conditional survival function given a covariate.

Key words: Nelson–Aalen estimators, right censorship, incomplete data, product integral, counting processes

1 Introduction

In this article we study inference from right-censored survival data in which the cause-of-failure can be missing for some individuals. Let T denote the failure time of interest, Z a covariate vector, and let C be a censoring time that is conditionally independent of T given Z. We are given data on $X = T \wedge C$ and also on any covariates Z that are to be included in the analysis, but the failure indicator $\delta = I(T \leq C)$ is allowed to be missing at random. The problem is to estimate the survival function of T and to account for the covariate effect.

Cause-of-failure information δ may be missing for a variety of reasons, e.g., in a bioassay experiment some subjects might not be autopsied to save expense, or the results of an autopsy may be inconclusive. The simplest approach is to assume that the failure indicators are *missing completely at random* (MCR) in the sense that the mechanism for missingness is

[1] *Running head:* Missing cause-of-failure

independent of everything else. The MCR mechanism and more general patterns of missingness have been studied by Rubin (1976), Little and Rubin (1987), Heitjan and Rubin (1991), Jacobsen and Keiding (1995), and Gill, Laan and Robins (1995), among others.

We first consider the problem in the absence of covariates. The survival function of T, denoted S, can be consistently estimated under MCR by simply ignoring the missing data and applying the usual Kaplan–Meier estimator to the complete data. However, such a procedure (called the *complete case estimator*) would be highly inefficient if there is a significant degree of missingness.

The first attempt to improve upon the complete case estimator was made by Dinse (1982) who used the EM algorithm to obtain a nonparametric maximum likelihood estimator (NPMLE). Lo (1991) showed that there are infinitely many NPMLEs, and each of them is self-consistent in the sense of Efron (1967). However, Lo noted that some of the NPMLEs may be inconsistent. He constructed two alternative estimators, one of which is consistent and asymptotically normal. Gijbels, Lin and Ying (1993), henceforth GLY, further improved upon this estimator by taking a product-integral of a convex combination of two estimates of the cumulative hazard function of T.

We shall construct a new estimator for S. The observations are considered to be i.i.d. replicates of (X, ξ, σ) where ξ is the indicator that δ is not missing and $\sigma = \xi\delta$. It is assumed that ξ is independent of T and C, and $\rho = P(\xi = 1) > 0$. We will find it convenient to view this set-up as a four state non–time–homogeneous Markov process having state space $\{0, 11, 10, 00\}$, where "0" represents the initial state in which no failure has occurred, "11" and "10" represent $(\xi, \delta) = (1, 1)$ and $(1, 0)$, respectively, and "00" represents a missing failure indicator. The process jumps at time X from state 0 into one of the absorbing states 11, 10, or 00, as in Figure 1.

The cumulative transition intensities Λ_{11}, Λ_{10}, Λ_{00} for the transitions $0 \to 11, 0 \to 10, 0 \to 00$ can be expressed in terms of the cumulative hazard function of T (denoted by Λ_T), ρ and the survival function of C. The cumulative hazard function Λ_T can in turn be expressed as the functional (see Section 2):

$$\begin{aligned}\Lambda_T(t) &= \phi(\Lambda_{10}, \Lambda_{11}, \Lambda_{00})(t) \\ &= \Lambda_{11}(t) + \pi(t)\Lambda_{00}(t), \qquad (1.1)\end{aligned}$$

where $\pi(\cdot) = \Lambda_{11}/(\Lambda_{10} + \Lambda_{11})$. We shall plug into (1.1) the Nelson–Aalen estimators $\hat{\Lambda}_{10}, \hat{\Lambda}_{11}$, and $\hat{\Lambda}_{00}$ of Λ_{10}, Λ_{11}, and Λ_{00}. Finally, by taking a product integral we will obtain our estimator $\hat{S}$. This estimator is easier to compute than the GLY estimator, which involves a further step of estimating an optimal convex combination of two 'sub-optimal' estimators, and has superior small sample performance, see Section 2. Moreover, its asymptotic properties are relatively straightforward to derive, being a consequence of (1.1) and standard results for Nelson–Aalen estimators, and a specially tailored analysis is not required.

Note that $\hat{\Lambda}_{11}$ underestimates Λ_T and we need to compensate by adding a suitably scaled estimate of Λ_{00}. The factor $\pi(t)$ represents the proportion of transitions to 00 that are uncensored failures. In the 'full data' case ($\rho = 1$) we have $\hat{\Lambda}_{00} = 0$, and $\rho = 1$, so $\hat{\Lambda}_T$ reduces to the Nelson–Aalen estimator, and $\hat{S}$ reduces to the Kaplan–Meier estimator.

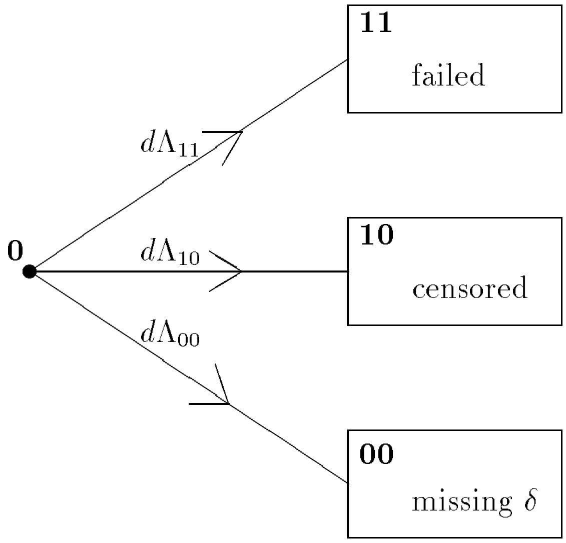

Figure 1: Transitions in the missing failure indicator model

We next consider an extension of our approach to the setting of Cox 's (1972) proportional hazards model in order to account for the effect of the covariate Z. Here the conditional hazard function of T given $Z = z$ takes the form $\lambda(t|z) = \lambda_0(t)\exp(\beta' z)$, where β is a vector of regression parameters and λ_0 is a baseline hazard function. We assume that ξ is independent of (T, C, Z). If only the complete case data are used, then Cox's maximum partial likelihood estimator of β is consistent and asymptotically normal, see Andersen and Gill (1982, henceforth AG). GLY introduced a more efficient estimator of β based on an optimal linear combination of estimating equations corresponding to the $0 \to 11$ and $0 \to 00$ transitions, and they showed that considerable improvements over the complete case estimator are possible. We propose an alternative estimating equation that treats the $0 \to 00$ transitions in a way that is similar to our derivation of $\hat{S}$. Our estimator improves upon the overall performance of GLY's estimator and it is computationally simpler. We also show how our basic estimator $\hat{S}$ can be extended to the Cox model setting for estimation of the conditional survival function of T given Z.

Another approach to the problem of missing cause-of-failure in the context of the Cox model has been taken by Goetghebeur and Ryan (1995) and Andersen, Goetghebeur and Ryan (1996). These authors considered a competing risks framework with two cause-specific hazard rates specified by Cox models having proportional baseline hazard functions. They obtained consistent estimators of the regression parameters in each Cox model. However, their approach assumes that the censoring indicators are always observed, which is not the case in the present setting.

Recently Laan and McKeague (1996) obtained an asymptotically efficient estimator of S in the absence of covariates, and essentially showed that any estimator that relies on the MCR assumption is inefficient. By relating the problem to the estimation of a bivariate survival function from bivariate right-censored data, they constructed an efficient estimator for S in the wider model in which ρ is a nonparametric function of time. Their estimator is asymptotically better than $\hat{S}$, but it does not extend to the Cox model setting, and it requires an artificial binning of the data that can be unappealing in practice, especially for small samples.

The paper is organized as follows. Section 2 deals with estimation of the survival function in the absence of covariates. We give a brief review of existing estimators in Section 2.1. The proposed estimator and its asymptotic properties are presented in Section 2.2, and some numerical results are discussed in Section 2.3. We treat estimation for the Cox model in Section 3. Existing estimators of β are reviewed in Section 3.1. We propose a new estimator of β and discuss its asymptotic properties in Section 3.2. An estimator of the conditional survival function of T given the covariate is introduced in Section 3.3. Some numerical results are given in Section 3.4. All proofs are placed in the Appendix.

2 Estimation of a survival function

2.1 Review of existing estimators

The observations consist of n i.i.d. replicates (X_i, ξ_i, σ_i) of the generic triple (X, ξ, σ) that was defined in the Introduction. Lo's (1991) estimator of the survival function S is given by

$$\hat{S}_L(t) \quad = \quad \prod_{X_i \le t} \left(1 - \frac{\sigma_i}{r(X_i)}\right)^{1/\hat{\rho}},$$

where $\hat{\rho} = \sum_{i=1}^n \xi_i/n$ is the proportion of observed failure indicators, and $r(t) = \#\{j\colon X_j \ge t\}$ is the 'size of the risk set' at time t. Like the complete case estimator, $\hat{S}_L$ jumps only at the uncensored failure times with known failure status. Note, however, that $\hat{S}_L$ uses the full size of the risk set that would be used by the Kaplan–Meier estimator if all the δ_i's were observed, so it makes more efficient use of the available data than the complete case estimator.

GLY observe that $\hat{S}_L$ does not use all of the information from individuals with $\xi_i = 0$, and they find a way to make better use of such information. They first estimate Λ_T and then estimate S by taking a product integral (as in (2.3)). Their estimator of Λ_T can be expressed as

$$\hat{\Lambda}_{GLY}(t) \quad = \quad \alpha(t)\hat{\Lambda}_1(t) + (1 - \alpha(t))\hat{\Lambda}_2(t), \tag{2.1}$$

where $0 \le \alpha(t) \le 1$ is specified,

$$\begin{aligned} \hat{\Lambda}_1(t) &= \hat{\Lambda}_{11}(t)/\hat{\rho}, \\ \hat{\Lambda}_2(t) &= \hat{\Lambda}_{00}(t)/(1-\hat{\rho}) - \hat{\Lambda}_{10}(t)/\hat{\rho}. \end{aligned}$$

They estimate the function $\alpha(t)$ that minimizes the asymptotic variance of estimators of the form (2.1), and plug that back into (2.1). They point out that the choice $\alpha(t) = 1$ is equivalent to Lo's estimator.

2.2 Proposed estimator

The distinction between our approach and those of Lo (1991) and GLY is that their estimators involve $\hat{\rho}$, whereas the estimator $\hat{S}$ to be proposed in this section does not. We believe that our approach is preferable because $\hat{\rho}$ is an ancillary statistic under the MCR model (its distribution is independent of the parameter of interest), so it does not contribute any information about S.

The distribution functions of T and C are denoted F and G, respectively, and the corresponding survival functions by $S = 1 - F$ and $R = 1 - G$. The subdistribution functions $H_{kl}(t) = P(X \leq t, \xi = k, \sigma = l)$ for $(k, l) \in \Gamma = \{(1,0), (1,1), (0,0)\}$, can be written as

$$\begin{aligned} H_{10}(t) &= \rho \int_0^t S_- dG, \\ H_{11}(t) &= \rho \int_0^t R_- dF, \\ H_{00}(t) &= (1-\rho) \int_0^t (S_- dG + R_- dF), \end{aligned} \tag{2.2}$$

where $S_-(t) = S(t-)$ is the left-continuous version of S. The range of integration in (2.2) is $[0, t]$. The survival function $\bar{H}$ of X may be expressed in terms of these subdistribution functions as $\bar{H} = 1 - \sum_{(k,l)\in\Gamma} H_{kl}$.

We shall consider estimation of S over an interval $[0, \tau]$, where $\bar{H}(\tau) > 0$. The three cumulative transition intensities for the non-time-homogeneous Markov process discussed in the Introduction are defined on $[0, \tau]$ by $d\Lambda_{kl} = dH_{kl}/\bar{H}_-$, for $(k, l) \in \Gamma$. Let Λ_X denote the cumulative hazard function of X. Some basic equations relating these functions on $[0, \tau]$ are

$$\begin{aligned} \Lambda_X &= \Lambda_{10} + \Lambda_{11} + \Lambda_{00}, \\ \rho\Lambda_X &= \Lambda_{10} + \Lambda_{11}, \\ \rho\Lambda_T &= \Lambda_{11}. \end{aligned}$$

These identities can be checked directly from (2.2), or more readily using the the interpretation of the Λ_{kl} as cumulative transition intensities (see Figure 1). The fundamental identity (1.1) is obtained by eliminating ρ from the last equation using the first two.

The Nelson–Aalen estimators of the cumulative transition intensities Λ_{kl} are given by

$$\hat{\Lambda}_{kl}(t) = \int_0^t \frac{d\hat{H}_{kl}(s)}{\hat{\bar{H}}(s-)},$$

where $\hat{\bar{H}}$ and $\hat{H}_{kl}$ are the empirical analogues of $\bar{H}$ and H_{kl}. The estimator proposed by Lo and the $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ estimators proposed by GLY can be obtained from our basic equations by plugging-in the relevant Nelson–Aalen estimators and $\hat{\rho}$. Specifically, Lo's estimator and $\hat{\Lambda}_1$ arise from our third basic equation $\rho\Lambda_T = \Lambda_{11}$, and $\hat{\Lambda}_2$ arises from $\Lambda_T = \Lambda_{00}/(1-\rho) - \Lambda_{10}/\rho$.

Our approach is to estimate Λ_T by plugging the Nelson–Aalen estimators $\hat{\Lambda}_{kl}$ into (1.1) to obtain

$$\hat{\Lambda}_T(t) = \hat{\Lambda}_{11}(t) + \hat{\pi}(t)\hat{\Lambda}_{00}(t),$$

where

$$\hat{\pi}(t) \quad = \quad \frac{\hat{\Lambda}_{11}(t)}{\hat{\Lambda}_{10}(t) + \hat{\Lambda}_{11}(t)}.$$

Here $\hat{\pi}(t)$ is defined to be 0 when the denominator vanishes. Finally, our estimator of the survival function S is the product-limit estimator

$$\hat{S}(t) \quad = \quad \prod_{(0,t]}(1 - d\hat{\Lambda}_T(s)), \tag{2.3}$$

where $\prod$ denotes the product integral as defined by Gill and Johansen (1990).

The following result gives the asymptotic distribution of $\hat{S}$.

Theorem 1 *Suppose that F and G are absolutely continuous. Then $\sqrt{n}(\hat{S} - S)$ converges weakly in $D[0,\tau]$ to $S{\cdot}L$, where L is a continuous zero-mean Gaussian process with covariance function*

$$\begin{aligned}
\mathrm{Cov}\,(L(u), L(v)) \quad = \quad & D_{11}(u)D_{11}(v)C_{00}(u) \\
& + D_{00}(u)D_{00}(v)D_{11}(u)D_{11}(v)C_{10}(u) \\
& + (1 + D_{00}(u)D_{10}(u))(1 + D_{00}(v)D_{10}(v))C_{11}(u)
\end{aligned} \tag{2.4}$$

for $u \leq v$, and where

$$C_{kl}(u) = \int_0^u \frac{d\Lambda_{kl}(s)}{\bar{H}(s-)}, \qquad D_{kl}(u) = \frac{\Lambda_{kl}(u)}{\Lambda_{10}(u) + \Lambda_{11}(u)}.$$

An asymptotic $100(1-\alpha)\%$ pointwise confidence interval for $S(t)$ is given by

$$\hat{S}(t) \pm z_{\alpha/2} n^{-1/2} \hat{V}(t),$$

where z_α is the upper α-quantile of the standard normal distribution and $\hat{V}(t)$ is a consistent estimate of the asymptotic variance of $\hat{S}(t)$ obtained by replacing the unknown functions in $V(t) = S(t)^2 \mathrm{Var}(L(t))$ by their estimates. Simultaneous confidence bands for S cannot readily be obtained from the above result, however, because of the complexity of the limiting covariance function.

The proof of Theorem 1, given in the Appendix, uses the functional delta method of Gill (1989) in a similar way to Gill and Johansen's (1990) derivation of the asymptotic distribution of the Kaplan–Meier estimator. This approach works because $\hat{S}$ is a smooth (compactly differentiable) functional of the Nelson–Aalen estimators $\hat{\Lambda}_{kl}$. Moreover, due to the interpretation of the missing failure indicator model as a non-time-homogeneous Markov process (as in Figure 1), we can appeal to the standard weak convergence result for Nelson–Aalen estimators in that context. This makes the proof relatively straightforward, and avoids the lengthy covariance calculations made by GLY.

2.3 Numerical results

In this section we report the results of a simulation study comparing the performance of our estimator with that of the GLY estimator. We based the comparison on mean integrated square error (MISE) over the follow-up period. The failure time and censoring distributions are taken to be exponential with parameters 1 and λ, respectively, with λ adjusted to give prescribed censoring rates.

Figure 2 gives plots of the MISE as a function of the probability of non-missingness ρ for censoring rates 90%, 75%, 50%, sample size 100, and follow-up period $[0,1]$. The MISE was estimated over a fine grid of values of ρ using $10,000$ samples at each point on the grid.

In all cases the proposed estimator is found to be more efficient than the GLY estimator. Moreover, the relative efficiency of $\hat{S}$ with respect to $\hat{S}_{GLY}$ increases as ρ decreases and as the censoring rate increases. The proposed estimator makes significant improvements over $\hat{S}_{GLY}$ when the censoring rate is at least 70% and at least 70% of the failure indicators are missing.

3 Adjustment for covariates using the Cox model

In many applications it is necessary to adjust estimates of the survival distribution for the presence of risk factors, and to assess the influence of those risk factors. The standard way of doing this is through a Cox proportional hazards model analysis, and in this section we consider an extension of our approach to deal with missing failure indicators in this model.

The observations now consist of n i.i.d. replicates $(X_i, \xi_i, \sigma_i, Z_i)$ of the generic (X, ξ, σ, Z), where Z is a $p \times 1$ vector of bounded covariates and the conditional hazard function of T given Z is specified by the Cox model, as defined in the Introduction. We begin with a review of existing estimators of the regression parameter β, the true value of which is denoted β_0.

3.1 Review of existing estimators

When none of the failure indicators are missing, the standard estimator of β is the maximum partial likelihood estimator $\hat{\beta}_P$ which solves the estimating equation $U(\beta) = 0$, where

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left(Z_i - \bar{Z}(\beta, t) \right) dN_i^u(t).$$

Here $N_i^u(t) = \delta_i N_i(t)$ with $N_i(t) = 1_{\{X_i \leq t\}}$,

$$\bar{Z}(\beta, t) \quad = \quad \frac{\sum_{i=1}^n Y_i(t) \exp(\beta' Z_i) Z_i}{\sum_{i=1}^n Y_i(t) \exp(\beta' Z_i)},$$

and $Y_i(t) = 1_{\{X_i \geq t\}}$ denotes the at-risk indicator for the i-th individual. AG showed that $\sqrt{n}\left(\hat{\beta}_P - \beta_0\right) \xrightarrow{D} N(0, \Sigma^{-1})$, where $\Sigma = -\lim_{n\to\infty} n^{-1} \partial U(\beta_0)/\partial\beta$ is assumed to be positive definite.

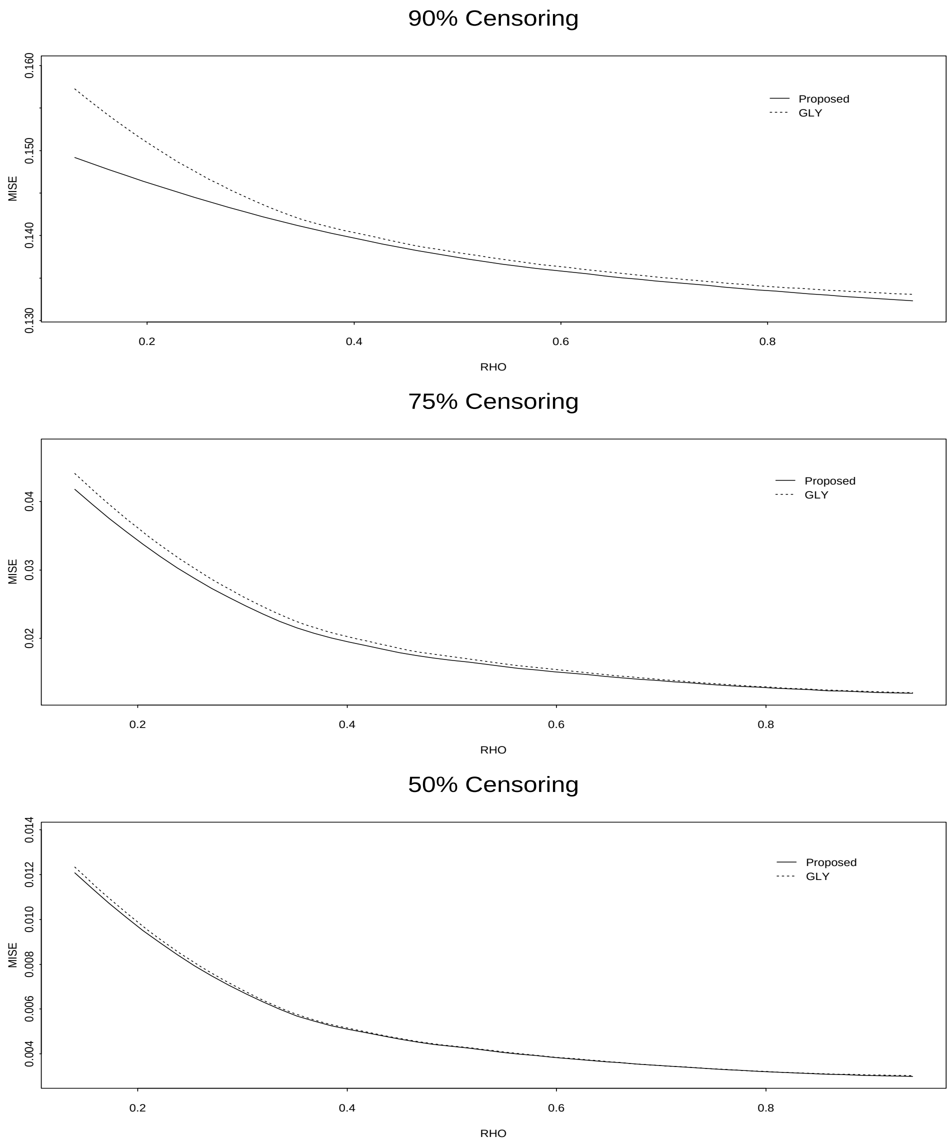

Figure 2: The MISE of the proposed estimator $\hat{S}$ (solid line) and of $\hat{S}_{GLY}$ (dotted line) are plotted against the probability of non-missingness.

In the case of missing failure indicators, GLY proposed the estimating equation

$$S_{11}(\beta,\infty)+DS_*(\beta,\infty)=0,$$

where D is a certain $p\times p$ matrix,

$$\begin{aligned} S_{11}(\beta,t) &= \sum_{i=1}^n \int_0^t \left(Z_i-\bar{Z}(\beta,s)\right)\xi_i dN_i^u(s),\\ S_*(\beta,t) &= \sum_{i=1}^n \int_0^t \left(Z_i-\bar{Z}(\beta,s)\right)\left((1-\xi_i)dN_i(s)-\frac{1-\hat\rho}{\hat\rho}\xi_i dN_i^c(s)\right), \end{aligned}$$

and $N_i^c(t)=(1-\delta_i)N_i(t)$. They show that solutions of this estimating equation are asymptotically normal, and they find the asymptotically optimal choice of D.

3.2 Proposed estimating function

The idea is to exploit the information in the $0\to 00$ transitions in a way that is similar to our approach in the absence of covariates. Define the estimating function corresponding to the $0\to 00$ transitions by

$$S_{00}(\beta,t) = \sum_{i=1}^n \int_0^t \left(Z_i-\bar{Z}(\beta,s)\right)(1-\xi_i)dN_i(s), \tag{3.5}$$

and the estimating function S_{10} corresponding to the $0\to 10$ transitions by integrating with respect to $\xi_i dN_i^c$ instead of $(1-\xi_i)dN_i$. We propose the estimating function

$$S(\beta,t) = S_{11}(\beta,t)+\hat\pi_P(\beta,t)S_{00}(\beta,t),$$

where

$$\hat\pi_P(\beta,t)=P(\beta,t)\left(P(\beta,t)+Q(\beta,t)\right)^{-1},$$

$P(\beta,t)=\mathrm{diag}(S_{11}(\beta,t))$, $Q(\beta,t)=\mathrm{diag}(S_{10}(\beta,t))$, $R(\beta,t)=\mathrm{diag}(S_{00}(\beta,t))$ and $\mathrm{diag}(v)$ is the diagonal matrix with diagonal v.

Our estimator $\hat\beta$ is a solution to the estimating equation $S(\beta,\infty)=0$. It can be shown that within any compact set containing β_0, a unique solution to this estimating equation exists with probability tending to 1 as the sample size increases, see the proof of consistency part of Theorem 2. To show that $\hat\beta$ is asymptotically normal we need the following notation:

$$\begin{aligned} W_i(t) &= Z_i-\bar{z}(\beta_0,t),\\ N_i^{CZ}(t) &= \int_0^t W_i(s)dN_i^c(s),\\ \bar{z}(\beta,t) &= \frac{E\left(Y_1(t)\exp\left(\beta' Z_1\right)Z_1\right)}{E\left(Y_1(t)\exp\left(\beta' Z_1\right)\right)},\\ N_i^Z(t) &= \int_0^t W_i(s)dN_i(s),\\ B(t) &= \mathrm{diag}\left\{E\left(N_{1j}^{CZ}(t)\right)/E\left(N_{1j}^Z(t)\right),\ j=1,\dots,p\right\}. \end{aligned}$$

Also, let $v^{\otimes 2}$ denote the cross product vv' for any vector v.

Theorem 2 *The estimator $\hat{\beta}$ is consistent and $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{D} N(0,V)$, where*

$$V = \Sigma^{-1} + (\rho^{-1}-1)\Sigma^{-1} E\left(N_1^{CZ}(\infty) - B(\infty) N_1^Z(\infty)\right)^{\otimes 2} \Sigma^{-1}.$$

The first term in V is the asymptotic variance of Cox's maximum partial likelihood estimator based on the full data, and the second term represents the effect of the missing failure indicators.

A consistent estimator of V can be obtained as follows. Note that if we suppose that the processes N_i^u and N_i^c are observed, each term in V can be consistently estimated by plugging $\hat{\beta}$ and $\hat{\rho}$ into the corresponding empirical estimator in place of the unknown β_0 and ρ. Replacing those (unobserved) processes by $\hat{\rho}^{-1}\xi_i N_i^u$ and $\hat{\rho}^{-1}\xi_i N_i^c$, respectively, results in a consistent estimator of V.

3.3 Estimation of the conditional survival function

In this section we briefly discuss estimation of the conditional survival function $S(t|z) = P(T > t|Z = z)$ under the Cox model with missing failure indicators. This will be done by plugging $\hat{\beta}$ and an estimate of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ into the Cox model based expression for $S(t|z)$.

The baseline hazard Λ_0 can be expressed in the same form as the basic equation (1.1) for Λ_T, except that the cumulative transition intensities Λ_{kl} are now replaced by the functions

$$\Lambda_{kl}^0(t) \quad = \quad \int_0^t \frac{dH_{kl}(s)}{E(Y_1(s)\exp(\beta_0' Z_1))}.$$

This leads to the following consistent estimator of the baseline hazard function:

$$\hat{\Lambda}_0(t) \quad = \quad \hat{\Lambda}_{11}^0(t) + \hat{\pi}^0(t)\hat{\Lambda}_{00}^0(t),$$

where

$$\hat{\pi}^0(t) \quad \equiv \quad \frac{\hat{\Lambda}_{11}^0(t)}{\hat{\Lambda}_{11}^0(t) + \hat{\Lambda}_{10}^0(t)}$$

is defined to be zero when the denominator vanishes, and $\hat{\Lambda}_{kl}^0$ is the empirical version of Λ_{kl}^0. The estimator $\hat{\Lambda}_0$ reduces to Breslow's estimator when there are no missing failure indicators.

The resulting estimator of the conditional survival function is the product integral

$$\hat{S}(t|z) \quad = \quad \mathcal{\pi}_{(0,t]}(1 - \exp(\hat{\beta}' z) d\hat{\Lambda}_0(s)),$$

cf. Andersen et al. (1993, p.509). It is straightforward to show that $\hat{S}(t|z)$ consistently estimates $S(t|z)$. It is also possible to prove asymptotic normality of $\hat{S}(t|z)$ using the techniques of Theorems 1 and 2, but the asymptotic variance is complicated.

3.4 Numerical results

We carried out a simulation study to compare the performance of the proposed estimator with that of the GLY and complete case estimators of β.

The underlying Cox model model was taken to be $\lambda(t|z) = \exp(\beta_0 z)$, for $\beta_0 = 0., 0.5, 1.0$, and the covariate Z was either standard normal or uniform on [0,1]. The censoring was exponential with the parameter adjusted to give prescribed censoring rates of 30% and 80%. In each case the mean square errors (MSE) of the various estimators were computed from 10,000 simulated samples of size $n = 100$. The 'full data' estimator is included for comparison. The results given in Tables 1–4 are classified according to the distribution of the covariate and the value of ρ (0.8 or 0.5).

The proposed estimator improves upon the GLY estimator in terms of MSE when the censoring is heavy (80%) and there is a low proportion of missing failure indicators ($\rho = 0.8$). The degree of the improvement increases as the covariate effect (β_0) diminishes. Although $\hat{\beta}$ can do slightly worse than the GLY estimator (under light censoring and $\rho = 0.5$; see the last three rows of Tables 2 and 4), it has the best overall performance. In terms of bias, the two estimators have very similar performance.

Table 1. $Z \sim U[0,1], \quad \rho = 0.8$.

Censoring rate	β_0	*Full data*		*Proposed*		*Gijbels-Lin-Ying*		*Complete Case*	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
	0.0	0.009	0.709	0.012	0.810	0.015	0.880	0.014	0.928
80%	0.5	0.504	0.737	0.504	0.881	0.499	0.905	0.499	0.960
	1.0	1.022	0.783	1.022	0.974	1.023	0.975	1.029	1.036
	0.0	-0.008	0.188	-0.005	0.208	-0.006	0.203	-0.003	0.241
30%	0.5	0.512	0.197	0.513	0.216	0.514	0.213	0.517	0.254
	1.0	1.016	0.202	1.015	0.220	1.013	0.217	1.013	0.260

Table 2. $Z \sim U[0,1], \quad \rho = 0.5$.

Censoring rate	β_0	*Full data*		*Proposed*		*Gijbels-Lin-Ying*		*Complete Case*	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
	0.0	0.004	0.724	0.006	1.330	-0.004	1.581	-0.008	1.878
80%	0.5	0.499	0.748	0.519	1.521	0.502	1.624	0.519	1.954
	1.0	1.002	0.763	1.033	1.734	1.025	1.693	1.053	2.093
	0.0	-0.008	0.188	-0.008	0.270	-0.010	0.247	-0.007	0.415
30%	0.5	0.512	0.194	0.513	0.274	0.516	0.257	0.523	0.444
	1.0	1.015	0.199	1.013	0.309	1.023	0.269	1.037	0.462

Table 3. $Z \sim N(0,1)$, $\rho = 0.8$.

Censoring rate	β_0	*Full data*		*Proposed*		*Gijbels-Lin-Ying*		*Complete Case*	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
80%	0.0	0.003	0.060	0.003	0.067	0.005	0.073	0.005	0.078
	0.5	0.514	0.071	0.514	0.083	0.514	0.087	0.516	0.093
	1.0	1.014	0.101	1.017	0.118	1.016	0.125	1.018	0.134
30%	0.0	-0.003	0.016	-0.002	0.023	-0.003	0.023	-0.002	0.037
	0.5	0.506	0.019	0.507	0.021	0.506	0.021	0.507	0.025
	1.0	1.011	0.027	1.102	0.031	1.102	0.030	1.014	0.035

Table 4. $Z \sim N(0,1)$, $\rho = 0.5$.

Censoring rate	β_0	*Full data*		*Proposed*		*Gijbels-Lin-Ying*		*Complete Case*	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
80%	0.0	0.003	0.060	0.004	0.106	0.002	0.124	0.004	0.161
	0.5	0.510	0.073	0.513	0.153	0.512	0.150	0.521	0.190
	1.0	1.025	0.101	1.046	0.202	1.042	0.203	1.057	0.254
30%	0.0	-0.003	0.016	-0.002	0.023	-0.003	0.021	-0.002	0.037
	0.5	0.508	0.020	0.512	0.029	0.512	0.026	0.519	0.044
	1.0	1.012	0.0277	1.014	0.041	1.016	0.036	1.024	0.060

Appendix

Proof of Theorem 1

The product-limit functional $\phi(\Lambda_T) = \prod(1-d\Lambda_T)$ is compactly differentiable with derivative $[d\phi(\Lambda_T)(l)](t) = -S(t)l(t)$, for $l \in D[0,\tau]$, see Gill and Johansen (1990). Here we have used the continuity of S. We will show below that $\sqrt{n}(\hat{\Lambda}_T - \Lambda_T)$ converges in distribution to the Gaussian process L in Theorem 1, so the functional delta-method implies that $\sqrt{n}(\hat{S} - S)$ converges in distribution to $-S \cdot L \stackrel{\mathcal{D}}{=} S \cdot L$. This will complete the proof.

Using (1.1) and the definition of $\hat{\Lambda}_T$ we may write

$$\begin{aligned}\sqrt{n}(\hat{\Lambda}_T(t) - \Lambda_T(t)) &= \left(1 + \frac{\Lambda_{00}(t)D_{10}(t)}{\hat{\Lambda}_{11}(t) + \hat{\Lambda}_{10}(t)}\right)\sqrt{n}(\hat{\Lambda}_{11}(t) - \Lambda_{11}(t)) \\ &+ \hat{D}_{11}(t)\sqrt{n}(\hat{\Lambda}_{00}(t) - \Lambda_{00}(t)) - \frac{\Lambda_{00}(t)D_{11}(t)}{\hat{\Lambda}_{11}(t) + \hat{\Lambda}_{10}(t)}\sqrt{n}(\hat{\Lambda}_{10}(t) - \Lambda_{10}(t)), \qquad (3.6)\end{aligned}$$

where $\hat{D}_{11}$ is the empirical version of D_{11}. Now we apply the standard weak convergence result for Nelson–Aalen estimators of cumulative transition functions: $(\sqrt{n}(\hat{\Lambda}_{kl} - \Lambda_{kl}(t)); (k,l) \in \Gamma) \xrightarrow{\mathcal{D}} (W_{kl}; (k,l) \in \Gamma)$ on $[0,\tau]$, where the W_{kl} are independent zero-mean continuous Gaussian martingales with covariance function $\mathrm{Cov}(W_{kl}(u), W_{kl}(v)) = C_{kl}(u \wedge v)$, see Andersen et al. (1993, p.198). The assumed absolute continuity of F and G allows us to use the latter result, which requires the existence of transition intensities. The filtration used in this connection is the the natural filtration of the four-state Markov process that underlies the missing failure indicator model. It follows immediately from (3.6) that $\sqrt{n}(\hat{\Lambda}_T - \Lambda_T)$ converges weakly on $[0, \tau]$ to the Gaussian process

$$L(t) = D_{11}(t)W_{00}(t) - D_{00}(t)D_{11}(t)W_{10}(t) + (1 + D_{00}(t)D_{10}(t))W_{11}(t)$$

which has the covariance function given in Theorem 1. □

Proof of Theorem 2

This proof follows the basic approach of AG in deriving the asymptotic distribution of the maximum partial likelihood estimator $\hat{\beta}_P$, and it also uses some of the steps taken by GLY.

Taylor expanding $S(\beta, \infty)$ about β_0 and using $S(\hat{\beta}, \infty) = 0$, it is seen that $n^{-1/2}S(\beta_0, \infty) = n^{-1}\mathcal{J}(\beta^*, \infty)n^{1/2}(\hat{\beta} - \beta)$, where β^* is on the line segment between $\hat{\beta}$ and β_0, and $\mathcal{J}(\beta, \infty)$ is the partial derivative of $S(\beta, \infty)$ with respect to β. Thus, once we have shown consistency of $\hat{\beta}$, asymptotic normality will follow if we show that $n^{-1/2}S(\beta_0, \cdot)$ converges weakly to a Gaussian process and $n^{-1}\mathcal{J}(\beta^*, \infty)$ converges in probability to the non-singular matrix Σ for any $\beta^* \xrightarrow{P} \beta_0$. We only consider the case of a one-dimensional covariate, as the general case is similar.

The following expression for $S(\beta, t)$ will be useful:

$$S(\beta, t) \quad = \quad S_{11}(\beta, t) + S_{11}^D(\beta, t),$$

where

$$S_{11}^D(\beta, t) \quad = \quad S_{00}(\beta, t) - \tilde{P}(\beta, t)S_{10}(\beta, t)$$

and $\tilde{P}(\beta, t) = R(\beta, t)\,(P(\beta, t) + Q(\beta, t))^{-1}$.

We begin by establishing the weak convergence of $n^{-1/2}S(\beta_0, \cdot)$. To simplify the notation, when an expression is evaluated at $\beta = \beta_0$ we shall suppress its dependence on β_0. First recall that the counting process $N_i^u(t)$ has intensity $\lambda_i(t) = \lambda(t|Z_i)Y_i(t)$ with respect to the filtration generated by the processes N_j^u, Y_j and Z_j, $j = 1, \ldots, n$. The corresponding martingale is $M_i(t) = N_i^u(t) - \int_0^t \lambda_i(s)\, ds$.

Note that $n^{-1/2}S_{11}(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{S}_{11}(t)$, where

$$\tilde{S}_{11}(t) \quad = \quad \sum_{i=1}^n \int_0^t W_i(s)\xi_i dM_i(s) + \sum_{i=1}^n \int_0^t (\xi_i - \rho)W_i(s)\lambda_i(s)\, ds,$$

see GLY (the decomposition of $S_1(\beta_0, t)$ in the proof of Theorem 1). This can be re-expressed as

$$\tilde{S}_{11}(t) \quad = \quad \rho\sum_{i=1}^n \int_0^t W_i(s)dM_i(s) + \sum_{i=1}^n (\xi_i - \rho)\int_0^t W_i(s)dN_i^u(s).$$

Next we obtain a suitable representation for $S^D_{11}(t)$. Note that

$$\sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\, \lambda_i(s)\, ds \;\;=\;\; 0. \tag{3.7}$$

Using $N_i = N_i^u + N_i^c$ we may write

$$\begin{aligned} S_{00}(t) &= \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\,(1-\xi_i) dM_i(s) + \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\,(1-\xi_i)\lambda_i(s)\, ds \\ &\quad + \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\,(1-\xi_i) dN_i^c(s). \end{aligned}$$

Then, since $\tilde{P}(t) = 1/\hat{\rho}_1(t) - 1$ where

$$\hat{\rho}_1(t) \;\;=\;\; \frac{\sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\, \xi_i dN_i(s)}{\sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\, dN_i(s)},$$

we have

$$\tilde{P}(t) S_{10}(t) \;\;=\;\; \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big) \left(\frac{1}{\hat{\rho}_1(t)} - 1\right) \xi_i dN_i^c(s).$$

Using (3.7), this gives $S^D_{11}(t) = A_1(t) + A_2(t)$ where

$$\begin{aligned} A_1(t) &= \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\,(1-\xi_i) dM_i(s) \\ &\quad + \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big)\,(1-\xi_i - (1-\rho))\, \lambda_i(s)\, ds, \end{aligned}$$

and

$$A_2(t) \;\;=\;\; \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big) \left(1 - \frac{\xi_i}{\hat{\rho}_1(t)}\right) dN_i^c(s).$$

GLY (proof of Theorem 2) show that $n^{-1/2} A_1(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{A}_1(t)$, where

$$\tilde{A}_1(t) \;\;=\;\; (1-\rho) \sum_{i=1}^n \int_0^t W_i(s) dM_i(s) - \sum_{i=1}^n (\xi_i - \rho) \int_0^t W_i(s) dN_i^u(s).$$

Next consider $A_2(t)$. Using conditions B and D of AG it can be shown that $\hat{\rho}_1(t) - \rho = O_p(n^{-1/2})$. In addition, applying the delta method gives

$$A_2(t) \;\;=\;\; \sum_{i=1}^n \int_0^t \Big(Z_i - \bar{Z}(s)\Big) \left(1 - \frac{\xi_i}{\rho}\right) dN_i^c(s)$$

$$
\begin{aligned}
& + \left(\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \int_0^t \left(Z_i - \bar{Z}(s)\right) dN_i(s)\right) \frac{\sum_{i=1}^n \int_0^t W_i(s) dN_i^c(s)}{\sum_{i=1}^n \int_0^t W_i(s) dN_i(s)} + r_n(t) \\
= & \sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \int_0^t \left(Z_i - \bar{Z}(s)\right) dN_i^c(s) \\
& + \left(\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \int_0^t \left(Z_i - \bar{Z}(s)\right) dN_i(s)\right) \frac{E\left(N_i^{CZ}(t)\right)}{E\left(N_i^{Z}(t)\right)} + r_n(t),
\end{aligned}
$$

where $\sup_t |r_n(t)| = o_p(n^{1/2})$. Using Lemma 1 of GLY it can be shown that $n^{-1/2}A_2(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{A}_2(t)$, where

$$
\begin{aligned}
\tilde{A}_2(t) &= -\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) N_i^{CZ}(t) + \sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) N_i^{Z}(t) \frac{E\left(N_i^{CZ}(t)\right)}{E\left(N_i^{Z}(t)\right)} \\
&= -\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \left(N_i^{CZ}(t) - B(t) N_i^{Z}(t)\right).
\end{aligned}
$$

We have shown that $n^{-1/2}S(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{S}(t)$, where

$$
\tilde{S}(t) = \tilde{S}_{11}(t) + \tilde{A}_1(t) + \tilde{A}_2(t),
$$

which is a sum of n i.i.d. processes. Convergence of the finite dimensional distributions of $n^{-1/2}S(\cdot)$ follows. Tightness can be established as in GLY using their Lemma 1. We now calculate the limiting covariance at (s, t). Note that

$$
E\left(\int_0^s W_1(v)\, dN_1^u(v) \int_0^t W_1(w)\, dN_1^u(w)'\right) = E\left(\int_0^{s \wedge t} W_1^{\otimes 2}(v)\, dN_1^u(v)\right). \tag{3.8}
$$

GLY (see the last step of the proof of Theorem 2, but using different notation from ours) have shown that $E\left(\tilde{S}_{11}(s)\tilde{A}_1(t)'\right) = 0$. Making use of (3.8) it can be shown that

$$
\begin{aligned}
E\left(\tilde{S}_{11}(s)\tilde{S}_{11}(t)'\right) &= n\rho E\left(\int_0^{s\wedge t} W_1^{\otimes 2}(v) dN_1^u(v)\right), \\
E\left(\tilde{S}_{11}(s)\tilde{A}_2(t)'\right) &= n(1-\rho) E\left(\int_0^{s\wedge t} W_1^{\otimes 2}(v) dN_1^u(v)\right) B(t), \\
E\left(\tilde{A}_1(s)\tilde{A}_1(t)'\right) &= n(1-\rho) E \int_0^{s\wedge t} W_1^{\otimes 2}(v) dN_1^u(v), \\
E\left(\tilde{A}_1(s)\tilde{A}_2(t)'\right) &= -n(1-\rho) E\left(\int_0^{s\wedge t} W_1^{\otimes 2}(v) dN_1^u(v)\right) B(t), \\
E\left(\tilde{A}_2(s)\tilde{A}_2(t)'\right) &= n\rho^{-1}(1-\rho) E\left(\left(N_1^{CZ}(s) - B(s)N_1^{Z}(s)\right)\left(N_1^{CZ}(t) - B(t)N_1^{Z}(t)\right)'\right),
\end{aligned}
$$

leading to the covariance

$$
\Sigma + (\rho^{-1} - 1) E\left\{\left(N_1^{CZ}(s) - B(s)N_1^{Z}(s)\right)\left(N_1^{CZ}(t) - B(t)N_1^{Z}(t)\right)'\right\}.
$$

We next prove consistency. Using the independence assumption of the MCR model, it is easily shown (cf. AG) that $n^{-1}S_{11}(\beta,\infty)$ converges uniformly in any compact set to $\rho m(\beta)$ almost surely, where

$$m(\beta) = E\int_0^\infty (Z_1 - \bar{z}(\beta,t))\,dN_1^u(t).$$

Similarly, using the decomposition

$$\begin{aligned} S_{11}^D(\beta,\infty) &= \sum_{i=1}^n \int_0^\infty \left(Z_i - \bar{Z}(\beta,s)\right)(1-\xi_i)\,dN_i^u(s) \\ &\quad + \sum_{i=1}^n \int_0^\infty \left(Z_i - \bar{Z}(\beta,s)\right)(1-\xi_i)\,dN_i^c(s) \\ &\quad - \frac{\sum_{i=1}^n \int_0^\infty \left(Z_i - \bar{Z}(\beta,s)\right)(1-\xi_i)\,dN_i(s)}{\sum_{i=1}^n \int_0^\infty \left(Z_i - \bar{Z}(\beta,s)\right)\xi_i\,dN_i(s)} \sum_{i=1}^n \int_0^\infty \left(Z_i - \bar{Z}(\beta,s)\right)\xi_i\,dN_i^c(s), \end{aligned} \tag{3.9}$$

it can be shown that $n^{-1}S_{11}^D(\beta,\infty)$ converges uniformly in any compact set to $(1-\rho)m(\beta)$, since the contribution from the sum of the last two terms in the decomposition is asymptotically negligible. Thus $n^{-1}S(\beta,\infty)$ is uniformly approximated by the continuous function $m(\beta)$, which has unique root β_0, and we conclude that $\hat{\beta}$ is consistent.

It remains to show that $n^{-1}\mathcal{J}(\beta^*,\infty)$ converges in probability to Σ. Along the lines of AG (p.1108), but also using the independence assumption of the MCR model, we have that $n^{-1}\partial S_{11}(\beta^*,\infty)/\partial\beta$ converges in probability to $\rho\Sigma$. Similarly, using the decomposition (3.9), in which only the first term makes a contribution asymptotically, it can be shown that $n^{-1}\partial S_{11}^D(\beta^*,\infty)/\partial\beta$ converges in probability to $(1-\rho)\Sigma$. This completes the proof. $\square$

References

Andersen, P. K., Borgan, Ø., Keiding, N. and Gill, R. D. (1993). *Statistical Models Based on Counting Processes.* Springer, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120.

Andersen, J. W., Goetghebeur, E. J. and Ryan, L. (1996). Analysis of survival data under competing risks with missing cause of death information: application and implications for study design. In: *Lifetime Data: Models in Reliability for Survival Analysis* (N. P. Jewell et al., eds.), 13–19, Kluwer, The Netherlands.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics* **38** 417–431.

Efron, B. (1967). The two sample problem with censored data. In: *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853.

Gijbels, I., Lin, D. Y. and Ying, Z. (1993). Non- and semi-parametric analysis of failure time data with missing failure indicators. *Tech. Report* 039-93, Mathematical Sciences Research Institute, Berkeley.

Gill, R. D. (1989). Non- and semi- parametric maximum likelihood estimators and the von Mises method (part 1). *Scand. J. Statist.* **16** 97–124.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555.

Gill, R. D., van der Laan, M. J. and Robins J. M. (1995). Coarsening at random: characterizations, conjectures and counter-examples. *Preprint.*

Goetghebeur, E. J. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82** 821–833.

Heitjan D. F. and Rubin D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253.

Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23** 774–786.

van der Laan, M. J. and McKeague, I. W. (1996). Efficient estimation from right-censored data when failure indicators are missing at random. *Submitted Ann. Statist.*

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *J. Multiv. Anal.* **39** 217–235.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–590.