BIOMETRICS 59, 1063–1070
December 2003

Estimation of Competing Risks with General Missing Pattern in Failure Types

Anup Dewanji* **and Debasis Sengupta****

Applied Statistics Unit, Indian Statistical Institute, 203, B. T. Road, Calcutta 700 035, India
**email:* dewanjia@isical.ac.in
***email:* sdebasis@isical.ac.in

SUMMARY. In competing risks data, missing failure types (causes) is a very common phenomenon. In this work, we consider a general missing pattern in which, if a failure type is not observed, one observes a set of possible types containing the true type, along with the failure time. We first consider maximum likelihood estimation with missing-at-random assumption via the expectation maximization (EM) algorithm. We then propose a Nelson-Aalen type estimator for situations when certain information on the conditional probability of the true type given a set of possible failure types is available from the experimentalists. This is based on a least-squares type method using the relationships between hazards for different types and hazards for different combinations of missing types. We conduct a simulation study to investigate the performance of this method, which indicates that bias may be small, even for high proportion of missing data, for sufficiently large number of observations. The estimates are somewhat sensitive to misspecification of the conditional probabilities of the true types when the missing proportion is high. We also consider an example from an animal experiment to illustrate our methodology.

KEY WORDS: Competing risks; Cause-specific hazard; Missing failure type; Missing at random; EM algorithm; Nelson-Aalen estimator.

1. Introduction

There is now considerable interest in the analysis of competing risks data with missing failure types. Ideally, in the absence of any such missingness, we have observation on possibly censored survival time T and the failure type J (exactly one of, say, m possible types) on each individual. However, due to inadequacy in the diagnostic mechanism, the experimentalists quite often are uncertain about the true failure type, or are reluctant to report any specific value of J for some individuals. The problem of missing failure type arises in carcinogenicity studies in which, besides deaths (failures) without tumor, we have deaths with tumor present due to either the tumor itself or some other causes. Often there is uncertainty in assigning this cause of death even if the presence of tumor can be ascertained (Dinse, 1986; Lagakos and Louis, 1988). In extreme situations, one cannot even ascertain presence or absence of tumor, because it is totally cannibalized or autolysed (see Section 5 for details).

Dinse (1982) was among the first to consider this uncertainty in the information on failure types while estimating survival due to different failure types. Information on failure type was either completely available (that is, observed as exactly one of m possible types) or not available at all (that is, unobserved failure type is any one of the m possible types). This problem with two failure types was considered by many authors (see Miyakawa, 1984; Racine-Poon and Hoel, 1984; Lo, 1991; Mukerjee and Wang, 1993). Goetghebeur and Ryan (1990, 1995) considered the regression problem using partial likelihood for two types of failures under the assumption that the cause-specific hazards for the two failure types are proportional. See also Dewanji (1992) and Lu and Tsiatis (2001).

Missingness in all this work meant that no information on failure type was available at all. However, while dealing with more than two failure types, the experimentalists quite often may be able to narrow down to fewer than m types to be responsible for failure (as in the motivating example mentioned before). In this work, we consider a general missing pattern so that for each individual failure we observe, in addition to survival time, a subset $g \subseteq \{1, \ldots, m\}$ to be the possible types of failure (exactly one of which is the true, but unobserved, cause of failure). When g is a singleton set, then the failure type is exactly observed, and when $g = \{1, \ldots, m\}$, then the missingness is total. Flehinger, Reiser, and Yashchin (1998) considered such general pattern of missing failure types for the purpose of estimating survival due to different types, with the strong assumption of proportional hazards due to different types. They also assumed that, for some of the observations with missing failure type, a second-stage diagnosis can be performed to pinpoint the type.

In this work, we consider nonparametric estimation of the different cause-specific hazards, based on data with the above-mentioned general missing pattern in failure types, and without the assumption of proportional hazards due to different types. In Section 2, we develop an EM algorithm for nonparametric maximum likelihood estimation. Quite often, given a set of possible causes of failure, experimentalists can provide some information on the probability that one of them is responsible. By incorporating this information, we suggest a counting process approach leading to a Nelson-Aalen type estimator, presented in Section 3. Section 4 presents a simulation study to investigate the finite-sample performance of this estimator. We illustrate the methods in Section 5 with an example of a carcinogenicity study conducted by the British Industrial Biological Research Association (Peto et al., 1984). Section 6 ends with some discussion.

2. Maximum Likelihood Estimation Using the EM Algorithm

Suppose, for the m competing causes of failure, the corresponding cause-specific hazard rates are given by

$$\lambda_j(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr[T \in [t, t+\Delta t), J = j \mid T \geq t], \qquad (1)$$

for $j = 1, \ldots, m$, where T denotes the failure time and J the failure type. Suppose the data consists of failure time, but only partial information about the failure type is available. For the lth individual, we observe the failure or censoring time, the censoring indicator δ_l (1 for failure and 0 for censoring) and, in case of failure, a set $g_l \subseteq \{1, \ldots, m\}$ representing the possible causes of failure. The survival function $S(t)$ can be written in terms of the cause-specific hazard rates as $S(t) = \exp[-\int_0^t \sum_{j=1}^m \lambda_j(u)\, du]$.

Let $t_1, \ldots, t_N$ denote the distinct observed failure or censoring times. Also, let A_i denote the set of labels for the individuals failed or censored at time t_i, and D_i the same for failed individuals only. Assume that the mechanism of observing the g's follows the missing-at-random assumption (Little and Rubin, 1987, p. 90), in the sense that the probability of observing g (given failure time, failure type, and censoring indicator $\delta = 1$) is the same for all the types contained in g. Formally, for j, $j' \in g$ with $j \neq j'$,

$$\Pr[G = g \mid T = t, J = j, \delta = 1] = \Pr[G = g \mid T = t, J = j', \delta = 1], \qquad (2)$$

where G denotes the observed set of possible causes of failure. Then, assuming independent censoring, the likelihood function for the observed data is proportional to

$$\prod_{i=1}^{N} \left[\prod_{l \in A_i} \left\{ \left(\sum_{j \in g_l} \lambda_j(t_i) \right)^{\delta_l} S(t_i) \right\} \right]. \qquad (3)$$

It is clear from (3) above that the nonparametric maximum likelihood estimates of cause-specific hazard rates have masses at most at the observed failure times, $s_1 < \cdots < s_K$, say. Then, writing λ_{ji} as the discrete cause-specific hazard of type j

$$\prod_{l \in D_i} \left(\sum_{j \in g_l} \lambda_{ji} \right) \Bigg\} \left(1 - \sum_{j=1}^{m} \lambda_{ji} \right)^{n_i - d_i} \Bigg], \qquad (4)$$

$|D_i|$, number of failures at s_i, and n_i= number iduals at risk at time s_i-. We use the EM algorithm (Dempster, Laird, and Rubin, 1977) to estimate λ_{ji}'s as follows.

For the lth individual, let us write

$$x_{jl} = \begin{cases} 1, & \text{if the } l\text{th individual fails due to cause } j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the x_{jl}'s are not always observed and, in the complete data version for the EM algorithm, we assume these x_{jl}'s to be observed (that is, the failure type for each individual is available). Then, the complete data likelihood is

$$L_C = \prod_{i=1}^{K} \left[\left\{ \prod_{l \in D_i} \left(\prod_{j=1}^{m} \lambda_{ji}^{x_{jl}} \right) \right\} \left(1 - \sum_{j=1}^{m} \lambda_{ji} \right)^{n_i - d_i} \right]$$
$$= \prod_{i=1}^{K} \left(\prod_{j=1}^{m} \lambda_{ji}^{d_{ji}} \right) \left(1 - \sum_{j=1}^{m} \lambda_{ji} \right)^{n_i - d_i}, \qquad (5)$$

where $d_{ji} = \sum_{l \in D_i} x_{jl}$, the number of individuals in D_i failing due to cause j (not observed). Note that $d_i = \sum_{j=1}^m d_{ji}$.

The E-step of the EM algorithm involves taking conditional expectation of d_{ji}'s, or x_{jl}'s, given the incomplete observed data and initial estimates of the λ_{ji}'s, denoted by $\lambda_{ji}^{(0)}$'s. This conditional expectation of x_{jl}, denoted by $x_{jl}^{(0)}$, is given by $\lambda_{ji}^{(0)}/[\sum_{k \in g_l} \lambda_{ki}^{(0)}]$, for $j \in g_l$, and 0 otherwise. The conditional expectation of d_{ji} is, therefore, calculated as $d_{ji}^{(0)} = \sum_{l \in D_i} x_{jl}^{(0)}$. The M-step maximizes conditional expectation of $\log L_C$ (see (5)), involving these $d_{ji}^{(0)}$'s, with respect to λ_{ji}'s, to obtain the improved estimates as $\lambda_{ji}^{(1)} = d_{ji}^{(0)}/n_i$ for all j and i.

By Theorem 1 of Dempster et al. (1977), the incomplete likelihood L_I in (4) increases at each EM iteration. Then, by the continuity of the mappings $\lambda_{ji}^{(k)} \to \lambda_{ji}^{(k+1)}$, the above algorithm converges to a local maximum. Now we prove that L_I has unique maximum under certain conditions. From (4), we see that the observed information matrix is a block diagonal matrix, with the blocks (each being a $m \times m$ matrix) corresponding to the observed failure times. The ith such block corresponding to the failures at time s_i is given by

$$-\frac{\partial^2 \log L_I}{\partial \lambda_{ji}^2} = \sum_{l \in D_i} \frac{I\{j \in g_l\}}{\left(\sum_{k \in g_l} \lambda_{ki} \right)^2} + \frac{n_i - d_i}{\left(1 - \sum_{k=1}^{m} \lambda_{ki} \right)^2},$$

$$-\frac{\partial^2 \log L_I}{\partial \lambda_{ji} \partial \lambda_{j'i}} = \sum_{l \in D_i} \frac{I\{j, j' \in g_l\}}{\left(\sum_{k \in g_l} \lambda_{ki} \right)^2} + \frac{n_i - d_i}{\left(1 - \sum_{k=1}^{m} \lambda_{ki} \right)^2},$$

for $j \neq j'$. For the lth individual, let us define g_l to be the $m \times 1$ vector with jth element being 1 if $j \in g_l$ and 0 otherwise. Then, the ith block, as given above, can be written as

$$I_i = \sum_{l \in D_i} \frac{\underset{\sim}{g_l} \underset{\sim}{g_l}^T}{\left(\sum_{k \in g_l} \lambda_{ki} \right)^2} + \frac{n_i - d_i}{\left(1 - \sum_{k=1}^{m} \lambda_{ki} \right)^2} \underset{\sim}{1} \underset{\sim}{1}^T, \qquad (6)$$

where $\underset{\sim}{1}$ denotes the $m \times 1$ vector of 1's. If this matrix I_i in (6) can be shown to be positive definite, then the likelihood function L_I of (4) is concave and hence has a unique maximum. Then, the EM algorithm described above converges to this unique maximum.

Clearly, d_i has to be at the least m so that it is possible to estimate all the cause-specific hazards. Let G_i be the $m \times d_i$ matrix formed by stacking side-by-side the vectors $\underset{\sim}{g_l}$'s, for $l \in D_i$. Then, we have the following theorem.

THEOREM 1: *When $d_i = n_i$, the matrix I_i in (6) is positive definite if and only if G_i is of full rank (that is, m). When $d_i < n_i$, the matrix is positive definite if and only if $(G_i : \underset{\sim}{1})$ is of full rank (that is, m).*

Proof. Let us first consider the case $d_i = n_i$. Let $\underset{\sim}{b}$ be any non-null $m \times 1$ vector. Since the denominators in all the terms of the first sum of (6) are positive, $\underset{\sim}{b}^T I_i \underset{\sim}{b} = 0$ if and only if $\underset{\sim}{b}^T \underset{\sim}{g_l} = 0$ for all $l \in D_i$. This is equivalent to $\underset{\sim}{b}^T G_i = 0$. Such a vector b exists if and only if the m rows of G_i are linearly dependent. Hence, the result is proved. When $d_i < n_i$, the above argument goes through with G_i replaced by $(G_i : \underset{\sim}{1})$.

Note that G_i is of rank m if and only if the d_i column vectors of G_i (that is, the $\underset{\sim}{g_l}$ vectors for $l \in D_i$) span $\mathcal{R}^m$. Therefore, if at the last observed failure time s_K we have $d_K = n_K$, then we need the g_l vectors to span $\mathcal{R}^m$. When $d_i < n_i$, then the $\underset{\sim}{g_l}$ vectors, along with the vector $\underset{\sim}{1}$, have to span $\mathcal{R}^m$. This is so because there is a contribution from survival beyond time s_i, allowing identifiability of $\sum_{j=1}^m \lambda_j$ corresponding to the vector $\underset{\sim}{1}$ in $\mathcal{R}^m$.

The EM algorithm described earlier allows one to estimate the different cause-specific hazard rates, or the cumulative cause-specific hazards, nonparametrically under the conditions of Theorem 1 above. However, the condition of Theorem 1 usually does not hold at each of the observed failure times and one needs to pool a number of such consecutive time points for the condition to hold. In practice, one chooses a partition of the study period so that the condition holds for the failures within each interval of the partition (see the example in Section 5). One then estimates for each interval the conditional probability of failure in that interval from a particular cause given survival up to the beginning of the interval; it is interpreted as a"cause-specific interval hazard." Prefixing this partition using prior knowledge has some advantage, since this results in a broad parametric model with finite number of discrete hazards. This also facilitates variance estimation using the formula of Louis (1982). See also the Appendix B of Dewanji and Kalbfleisch (1986) for a working formula based on Louis (1982).

3. Nelson-Aalen Type Estimator

Recall that an observation on failure type may be missing in the sense that when the true cause is j, it may be hidden with a number of other possible causes, so that a set $g \subseteq \{1, \dots, m\}$ containing j is observed as the set of possible causes. Consequently, let us consider the observation probabilities in (2), denoting its l.h.s. by $p_{gj}(t)$, for all $g \ni j$ and $j = 1 \dots, m$. If g does not contain j, this probability is zero, so for a fixed $j, \sum_g p_{gj}(t) = 1$. Assuming that the missing mechanism is independent of the censoring mechanism, the probability $p_{gj}(t)$ equals $\Pr[G = g \,|\, T = t,\ J = j]$, which is the conditional probability of observing $g \ni j$ as the set of possible causes, given failure at time t due to cause j. Noting that $\lambda_j(t)\,dt$ is the conditional probability of instantaneous failure due to cause j at time t, given survival up to time $t-$, it follows that the hazard rate for failure due to cause j at time t and with $g \ni j$ observed as the set of possible causes is $p_{gj}(t)\lambda_j(t)$. Hence, the hazard rate for failure at time t with g observed as the set of possible causes is

$$\lambda_g^*(t) = \sum_{j \in g} p_{gj}(t)\lambda_j(t). \tag{7}$$

As expected, the sum of the hazards given by (7) over all nonempty subsets g with $\{1, \dots, m\}$ is $\lambda.(t) = \sum_{j=1}^m \lambda_j(t)$, since in that sum, the coefficient of $\lambda_j(t)$, for a fixed j, is $\sum_{g \ni j} p_{gj}(t) = 1$.

Note that the observation probabilities $p_{gj}(t)$'s are usually not known and need to be estimated. The missing-at-random condition (2), though required in Section 2 for maximum likelihood estimation, is not needed here. Instead, in order to be able to estimate these probabilities in practice, we make a simplifying assumption that $p_{gj}(t)$ is independent of time t, though it may depend on g and j. Thus, the missing pattern here is allowed to be non-ignorable. We will write $p_{gj}(t)$ as p_{gj} in the subsequent discussion.

Let us consider the $(2^m - 1)$-dimensional counting process $\{N_g(t)\}_{g \in \mathcal{G}}$, where $\mathcal{G}$ consists of all non-empty subsets of $\{1, \dots, m\}$. $N_g(t)$ represents the number of failures up to time t, with g as the observed set of possible causes. Consider the multiplicative intensity model of Aalen (1978), in which the corresponding intensity process is given by $\alpha_g(t) = \lambda_g^*(t)Y(t)$, where $Y(t)$ is the number of individuals at risk at time $t-$ and $\lambda_g^*(t)$ is as in (7). We have, for each nonempty subset g of $\{1, \dots, m\}$,

$$dN_g(t) = \alpha_g(t)\,dt + dM_g(t),$$

where $M_g(t)$'s are local square integrable martingales. Hence, we have the Nelson-Aalen estimator of $\Lambda_g^*(t) = \int_0^t \lambda_g^*(s)ds$ given by (Andersen and Borgan, 1985)

$$\widehat{\Lambda}_g^*(t) = \int_0^t \frac{I\{Y(s) > 0\}}{Y(s)} dN_g(s), \tag{8}$$

which is also known to converge in distribution to a Gaussian process with mean $\Lambda_g^*(t)$ and a variance function which can be consistently estimated by

$$\widehat{\tau}_g(t) = \int_0^t \frac{I\{Y(s) > 0\}}{Y^2(s)} dN_g(s). \tag{9}$$

Writing $\underset{\sim}{\Lambda}^*$ as the $(2^m - 1) \times 1$ vector of $\Lambda_g^*(t)$'s and $\underset{\sim}{\Lambda}(t)$ as the $m \times 1$ vector of the cumulative cause-specific hazards $\Lambda_j(t)$'s, we have, from (7),

$$\underset{\sim}{\Lambda}^* = P\underset{\sim}{\Lambda}(t), \tag{10}$$

where P is the $(2^m - 1) \times m$ matrix of the p_{gj}'s.

Using (8) and (10), we have

$$\widehat{\underset{\sim}{\Lambda}}^*(t) = P\underset{\sim}{\Lambda}(t) + \underset{\sim}{\epsilon}(t), \tag{11}$$

where $\underset{\sim}{\widehat{\Lambda}}{}^*(t)$ is the vector of $\widehat{\Lambda}_g^*(t)$'s and $\underset{\sim}{\epsilon}(t)$ is a vector process converging to a vector of Gaussian martingales whose variance function is consistently estimated by the matrix $\mathrm{diag}(\widehat{\tau}_g(t))$ (Andersen and Borgan, 1985). Note that (11) can be seen as a linear model with the "design matrix" P to be estimated. Let $\widehat{P}$ denote a consistent estimate of P. Then, using the principle of weighted least squares and the consistent estimate (9), we have a consistent estimate of $\underset{\sim}{\Lambda}(t)$ as

$$\underset{\sim}{\widehat{\Lambda}}(t) = (\widehat{P}^T W(t) \widehat{P})^{-1} \widehat{P}^T W(t) \underset{\sim}{\widehat{\Lambda}}{}^*(t), \tag{12}$$

where $W(t)$ is the inverse of the estimated $(2^m - 1) \times (2^m - 1)$ diagonal covariance matrix of $\underset{\sim}{\widehat{\Lambda}}{}^*(t)$, as given by $W(t) = \mathrm{diag}(1/\widehat{\tau}_g(t))$.

Note that, for $g \ni j$,

$$p_{gj} = \Pr[G = g \mid J = j] = \frac{\Pr[J = j \mid G = g]\Pr[G = g]}{\sum\limits_{g' \ni j} \Pr[J = j \mid G = g']\Pr[G = g']}.$$

Therefore, we can estimate p_{gj} by $\widehat{p}_{gj} = f_g q_{jg} / \sum_{g' \ni j} f_{g'} q_{jg'}$, where f_g denotes the number of failures with G observed as g and $q_{jg} = \Pr[J = j | G = g]$. When the experimentalists can provide information on the q_{jg}'s, that can be used to estimate the p_{gj}'s as above, and then estimate $\underset{\sim}{\Lambda}(t)$ using (12). Note, however, that this estimate cannot be guaranteed to be nondecreasing, although it is expected to be so for large sample, because of its consistency proved later. In practice, one can use "pooling-the-adjacent-violators" algorithm to achieve monotonicity (See Figure 1).

Note that, if some of the $\{N_g(t)\}$'s are not observed to have any jump during the study, the corresponding $\widehat{\Lambda}_g^*(t)$'s and $\widehat{\tau}_g(t)$'s turn out to be zero; the corresponding rows of P are also estimated to be zero. The same estimation procedure goes through with the observed $\{N_g(t)\}$'s as long as the resulting $\widehat{P}$ is of full column-rank, except that the corresponding diagonal entries of $W(t)$ cannot be evaluated. We may assign arbitrary values to them, as these would be multiplied by the zero elements of $\widehat{P}$ anyway. Note that the assumption that the rank of $\widehat{P}$ is m entails much less restriction on the availability of data than that of Section 2. Even if the rank of $\widehat{P}$ is less than m, some components of $\underset{\sim}{\Lambda}(t)$ may be estimable.

When t is small, most of the elements of $\widehat{\Lambda}_g^*(t)$ are zero. The corresponding variance estimates are also zero. This may create a problem in the computation of $\underset{\sim}{\widehat{\Lambda}}(t)$. It is a good idea to assign a large value to all those components of $W(t)$, for small t. This large value may be replaced by $1/\widehat{\tau}_g(t)$ when $\tau_g(t)$ assumes a nonzero value. We used the value 50,000 for the example in the next section; values ranging from 10,000 to 500,000 gave similar results.

For a fixed t, since $\underset{\sim}{\widehat{\Lambda}}{}^*(t)$ converges in distribution to a $(2^m - 1)$-variate normal random vector with mean $\underset{\sim}{\Lambda}{}^*(t)$ and variance matrix estimated by $W^{-1}(t)$, we have, from (12) and the fact that $\widehat{P}$ is a consistent estimate of P, weak convergence of $\underset{\sim}{\widehat{\Lambda}}(t)$ (by Slutsky's theorem) to an m-variate normal random vector with mean $\underset{\sim}{\Lambda}(t)$ and variance matrix estimated by $(\widehat{P}^T W(t) \widehat{P})^{-1}$. This pointwise weak convergence has been used to find confidence limits in Section 5.

Estimators of the cause-specific hazards $\lambda_g(t)$ may be obtained via kernel smoothing as suggested by Ramlau-Hansen (1983). We first obtain smooth estimators of $\lambda_g^*(t)$, for all g, as

$$\widehat{\lambda}_g^*(t) = \frac{1}{h} \int_0^1 K\left(\frac{t-s}{h}\right) d\widehat{\Lambda}_g^*(s), \tag{13}$$

where $\widehat{\Lambda}_g^*(t)$ is as defined in (8). In the above expression, the whole observation range is transformed into [0,1], the kernel function $K(\cdot)$ is a bounded function with support [−1,1] and having integral 1, and h is a positive constant denoting the window length. The variance of this smooth estimate (13) may be estimated by

$$\widehat{\sigma}_g(t) = \frac{1}{h^2} \int_0^1 K^2\left(\frac{t-s}{h}\right) \frac{I\{Y(s) > 0\}}{Y^2(s)} dN_g(s),$$

so that the variance matrix of $\underset{\sim}{\widehat{\lambda}}{}^*(t)$, the vector of $\widehat{\lambda}_g^*(t)$'s, is estimated by the diagonal matrix $\widehat{\Sigma}(t) = diag(\widehat{\sigma}_g(t))$.

Using the relation $\underset{\sim}{\lambda}{}^*(t) = P\underset{\sim}{\lambda}(t)$ (see [7]), an estimate of the vector of cause-specific hazards $\underset{\sim}{\lambda}(t)$, as in (12), is given by

$$\underset{\sim}{\widehat{\lambda}}(t) = (\widehat{P}^T \widehat{\Sigma}^{-1}(t) \widehat{P})^{-1} \widehat{P}^T \widehat{\Sigma}^{-1}(t) \underset{\sim}{\widehat{\lambda}}{}^*(t), \tag{14}$$

where $\underset{\sim}{\widehat{\lambda}}{}^*(t)$ has components defined by (13). The variance matrix of $\underset{\sim}{\widehat{\lambda}}(t)$ is consistently estimated by $(\widehat{P}^T \widehat{\Sigma}^{-1}(t) \widehat{P})^{-1}$. Asymptotic normality of $\underset{\sim}{\widehat{\lambda}}{}^*(t)$, subject to some regularity conditions, follows from Ramlau-Hansen (1983); the same for $\underset{\sim}{\widehat{\lambda}}(t)$ follows from (14).

4. A Simulation Study

We conduct a simulation study to investigate the performance of the estimator (12) with $m = 3$ and constant cause-specific hazards 0.1, 0.2, and 0.3, respectively. Given the true cause j, the probability of missing failure type is $1 - p_{\{j\}j}$, which is taken as constant α, for all j, with values 0.1, 0.3, and 0.5, representing different degrees of missingness. Also, for $g \ni j$, p_{gj} is taken as $\alpha/3$, since there are three such g's. Once the data is simulated with sample size n (taken as 100, 200, and 500), the cumulative cause-specific hazards $\underset{\sim}{\Lambda}(t)$ is estimated using (12). However, we need specification of the q_{jg}'s, for which we consider three choices. First, we consider the true ones, which can be worked out, given the p_{gj}'s and the true cause-specific hazards. The true q matrix, in this case, turns out to be

$$q = \begin{bmatrix} 1 & 0 & 0 & 1/3 & 0 & 1/4 & 1/6 \\ 0 & 1 & 0 & 2/3 & 2/5 & 0 & 2/6 \\ 0 & 0 & 1 & 0 & 3/5 & 3/4 & 3/6 \end{bmatrix}.$$

Next we consider a slight deviation from the true ones by suitably adding (to type 3) and subtracting probability 0.05. Lastly, we consider a balanced case in which the q_{jg}'s are equal for all j in g. We carry out 1000 simulations based on which the following performance characteristics are investigated.

Since we believe, due to scarcity of data, there may be violation of monotonicity in the tail area, we consider the proportion of times (out of the 1000 simulations) this violation takes place before the 90th percentile of the observed failure times. It is desirable to have small values for this proportion. Our observation is that, when α is small (0.1), this proportion

Table 1

Simulation-based estimates of the cumulative cause-specific hazards at 60*th percentile of the true life distribution. The true values are* 0.153, 0.305, *and* 0.458, *respectively.*

q_{jg}'s	α	$n = 100$	$n = 200$	$n = 500$
True	.1	.139, .276, .418	.148, .296, .447	.151, .301, .456
		(.053, .078, .102)	(.036, .052, .062)	(.023, .031, .040)
	.3	.140, .288, .433	.150, .297, .449	.150, .303, .454
		(.054, .078, .095)	(.037, .052, .061)	(.023, .032, .038)
	.5	.145, .294, .440	.149, .300, .454	.153, .305, .455
		(.055, .078, .090)	(.038, .052, .064)	(.024, .034, .040)
Slight deviation	.1	.139, .279, .415	.147, .294, .449	.146, .302, .458
		(.051, .081, .109)	(.035, .051, .065)	(.022, .034,.040)
	.3	.135, .283, .450	.141, .293, .466	.142, .298, .472
		(.051, .079, .098)	(.035, .052, .067)	(.022, .032, .038)
	.5	.130, .281, .465	.135, .292, .479	.136, .294, .482
		(.049, .078, .093)	(.036, .051, .065)	(.021, .034, .040)
Balanced	.1	.141, .278, .422	.158, .295, .439	.163, .302, .444
		(.064, .079, .094)	(.038, .050, .062)	(.023, .032, .038)
	.3	.170, .284, .406	.182, .298, .419	.188, .303, .420
		(.065, .079, .087)	(.041, .053, .059)	(.025, .032, .035)
	.5	.197, .295, .386	.211, .300, .390	.214, .302, .392
		(.066, .077, .085)	(.047, .054, .060)	(.029, .032, .037)

is almost zero and, for higher α's, this approaches zero with increasing sample size, regardless of the choice of q_{jg}'s.

We also consider the mean of the estimates (over the 1000 simulations) at different quantiles of the true life distribution (*exponential* with hazard 0.6), presenting only those at the 60th percentile in Table 1. The true values of these cumulative cause-specific hazards are 0.153, 0.305, and 0.458, respectively. The corresponding standard errors are found by taking the mean of individual standard errors based on the formula of Section 3 and also based on the estimates from 1000 simulations. Since they give similar results, we report (in parentheses) only those obtained by the latter method. As expected, the standard error decreases with increasing sample size. However, this does not seem to depend on α, the proportion of missing data. For the true choice of q_{jg}'s, we notice the estimates to be closer to the true value with increasing sample size, as expected, regardless of the value of α. Even for the second choice with slight deviation, the estimates are quite satisfactory for $\alpha = 0.1$ and larger sample size. With higher α values, however, the performance worsens.

To check asymptotic normality, we compute the mean, median, variance, and the 95th and 97.5th percentiles of the 1000 simulation-based standardized estimates, to compare with those of the standard normal distribution. We find satisfactory results for the true choice of q_{jg}'s with increasing sample size, regardless of the value of α. For the other two choices, the results seem to be encouraging, except the bias factor, since the estimates converge to values that depend on the wrongly assumed q_{jg} values, resulting in some bias.

5. An Example

A large animal experiment with a total of 5000 rodents was conducted by the British Industrial Biological Research Association (Peto et al., 1984) to investigate the carcinogenicity of different nitrosamines administered in drinking water. Gart et al. (1986, pp. 58–66) reported details of the data set for the occurrence of pituitary tumors in male rats given N-nitrosodimethylamine (NDMA) in different concentrations. For our illustration, we consider only the control group having 192 animals. The data consists of the time to death (in days) for each animal and some information on the cause of each death, described as the "context" of death. There are essentially three causes of death: (1) death without tumor, (2) death due to tumor (fatal), and (3) death due to some other causes but with tumor present (incidental). Because of various pathological problems, observation on the actual cause of death is sometimes missing. The "context" of an observation gives this occasionally incomplete information. Out of seven such "contexts," three give the exact causes of death, mentioned above. Two other "contexts," *probably fatal* and *probably incidental*, are interpreted (for our analysis) as missing, with {2, 3} being the set of possible causes, since presence of tumor is observed. In yet another "context," the presence or absence of tumor is not ascertained, but death is known *not* to be caused by tumor; hence {1, 3} is taken as the set of possible causes. In the last "context," the cause of death is not at all ascertainable, and so {1, 2, 3} is taken as the set of possible causes. Exact cause of death is available for 183 out of the 192 animals, with 135, 25, and 23 deaths due to cause 1, 2, and 3, respectively. For the nine missing causes, 3, 2, and 4, respectively, are observed with {1, 3}, {2, 3} and {1, 2, 3} as the set of possible causes. There are 163 distinct times of death.

We first use the EM algorithm of Section 2. For this, we partition the range of observed failure times into six intervals. The identifiability condition of Theorem 1 is satisfied for the events in these six time intervals. The intervals were: (0,700], (700,800], (800,900], (900,1000], (1000,1100], (1100,∞). The corresponding estimates (of λ_{ji}'s) are interpreted as "cause-specific interval hazards," as discussed at the end of Section 2. The estimates with corresponding standard errors in parentheses are given in Table 2.

Table 2
Maximum likelihood estimates of cause-specific interval hazards

Time interval	Death type 1	2	3
(0,700]	.1342 (.0159)	.0117 (.0123)	0.0 –
(700,800]	.1265 (.0261)	.0316 (.0139)	.0126 (.0089)
(800,900]	.1537 (.0311)	.0368 (.0164)	.0154 (.0108)
(900,1000]	.3333 (.0454)	.0556 (.0220)	.0926 (.0279)
(1000,1100]	.4821 (.0676)	.1179 (.0450)	.0786 (.0376)
(1100,∞)	.6111 (.1843)	.1111 (.0786)	.2778 (.1242)

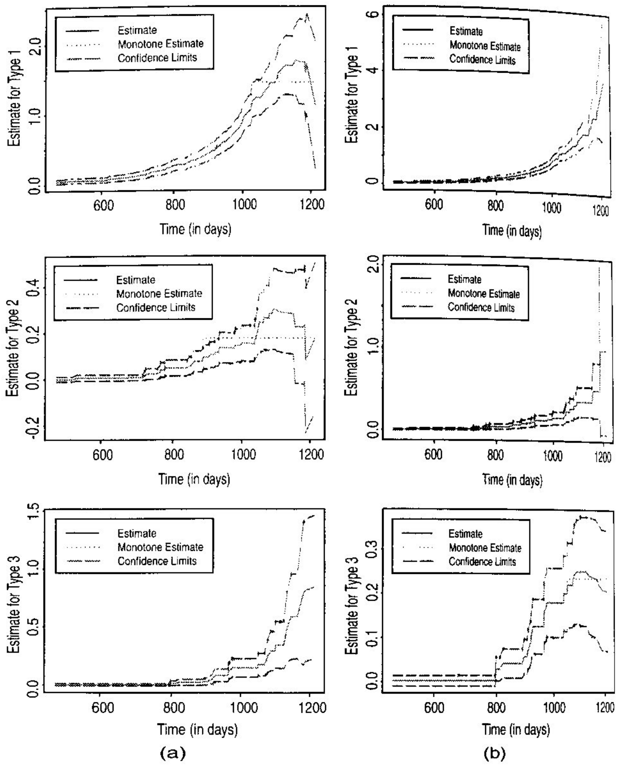

Figure 1. Estimated cumulative cause-specific hazards for cause (type) 1, 2, and 3 when (**a**) type 3 has low probability ($q_{3,\{1,2,3\}} = .02$, $q_{1,\{1,2,3\}} = q_{2,\{1,2,3\}} = .49$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = .02$ and $q_{1,\{1,3\}} = q_{2,\{2,3\}} = .98$) and (b) type 3 has high probability ($q_{3,\{1,2,3\}} = .98$, $q_{1,\{1,2,3\}} = q_{2,\{1,2,3\}} = .01$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = .98$ and $q_{1,\{1,3\}} = q_{2,\{2,3\}} = .01$). The solid line represents the estimate while the two dashed lines represent the 95% pointwise confidence limits. The dotted line gives the estimate obtained by using pooling-adjacent-violators algorithm.

In order to obtain the Nelson-Aalen type estimates for the cumulative cause-specific hazards of Section 3, we consider two extreme choices of the probabilities q_{jg}'s as follows.

(a) We choose $q_{3,\{1,2,3\}} = 0.02$, $q_{1,\{1,2,3\}} = q_{2,\{1,2,3\}} = 0.49$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = 0.02$ and $q_{1,\{1,3\}} = q_{2,\{2,3\}} = 0.98$. This case represents a situation where cause 3 has very low probability whenever it is observed with some other cause.

(b) We choose $q_{3,\{1,2,3\}} = 0.98$, $q_{1,\{1,2,3\}} = q_{2,\{1,2,3\}} = 0.01$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = 0.98$ and $q_{1,\{1,3\}} = q_{2,\{2,3\}} = 0.01$. This case represents a situation where cause 3 has very high probability whenever it is observed with some other cause.

We also consider the balanced choice (as in Section 4), which does not favor any particular cause. The results for this choice are generally found to lie in between those for the two extreme choices and are not reported in order to save space.

The estimates of the cumulative cause-specific hazards for the choices (a) and (b) are presented (solid lines) in Figures 1(a) and (b), respectively, along with 95% pointwise confidence limits (dashed lines). The three curves in each figure correspond to the three types of death, respectively. Note that, in some of these curves, monotonicity breaks down, but only in the tail area, where the estimates $\widehat{\Lambda}_g^*(t)$'s themselves are not precise (as seen by the widths of the confidence intervals in the tail area). In choice (a), the type 3 failure has low probability, and so the third curve mostly uses complete observations with type 3. Thus, it is not surprising that the third plot in Figure 1(a) is nondecreasing. Likewise, the other two plots of Figure 1(b) are nondecreasing. The estimates obtained by using pooling-adjacent-violators algorithm are also given by dotted lines. These estimates, as expected, coincide with the original ones (solid lines) in the three plots mentioned above.

Figures 2(a) and (b) show the plots of the three estimated cause-specific hazard rates for cases (a) and (b), along with the corresponding 95% pointwise confidence limits. For these estimates, we use the Epanechnikov's kernel function given by $K(x) = 0.75(1 - x^2)$, $|x| \leq 1$, and a window length h of approximately 100 days. This choice of window length corresponds to the one proportional to $n^{-1/2}$, where $n = 163$ is the number of observed time points, satisfying a sufficient condition for the asymptotic results (Ramlau-Hansen, 1983). We also work with another choice in which the window length (of approximately 220 days) is proportional to $n^{-1/3}$ and also satisfies the sufficient condition; the results are similar except for the greater smoothing effect.

6. Discussion

In this work, we consider a general pattern of missingness in failure types while dealing with competing risks data in which, instead of the true failure type, a set of possible types, including the true one, is observed. Under the missing-at-random assumption, we discuss maximum likelihood estimation via the EM algorithm and find that, in practice, only some interval cause-specific hazards are estimable, depending on the availability of data. In contrast, when information on the conditional probability of the true type, given a set of possible types, is available, a Nelson-Aalen type estimator is developed by assuming that the observation probabilities $p_{gj}(t)$'s

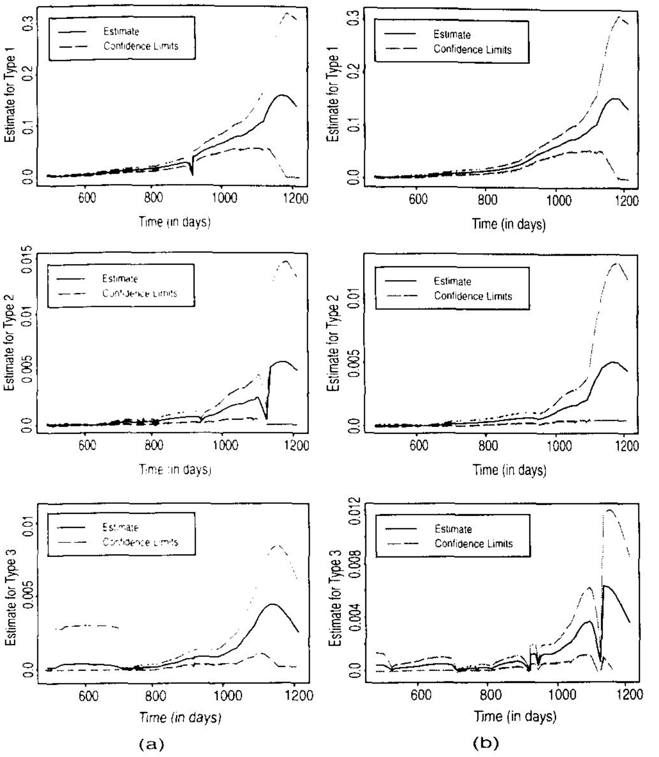

Figure 2. Estimated cause-specific hazard rates for cause (type) 1. 2. and 3 when (a) type 3 has low probability ($q_{3.\{1.2.3\}} = .02$. $q_{1.\{1.2.3\}} = q_{2.\{1.2.3\}} = .49$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = .02$ and $q_{1.\{1.3\}} = q_{2.\{2.3\}} = .98$) and (b) type 3 has high probability ($q_{3.\{1.2.3\}} = .98$. $q_{1.\{1.2.3\}} = q_{2.\{1.2.3\}} = .01$, $q_{3,\{1,3\}} = q_{3,\{2,3\}} = .98$ and $q_{1.\{1.3\}} = q_{2.\{2.3\}} = .01$). The solid line represents the estimate while the other two lines represent the 95% pointwise confidence limits.

are time-independent under the much less restrictive assumption that the matrix $\widehat{P}$ has full column-rank. Some of the cause-specific hazards may be estimable by the method of Section 3. even if this condition is violated. The two methods in Sections 2 and 3 have two different requirements on the missing mechanism. While the first makes a missing-at-random assumption (2). the second requires information on the q_{jg} probabilities. allowing the missing mechanism to be nonignorable. These q_{jg}'s are found to have some effect on the Nelson-Aalen type estimates and also on the cause-specific hazard rate estimates. Regarding the choice between maximum likelihood estimation and Nelson-Aalen type estimates, there remains the question of making either the missing-at-random assumption or the time-independence assumption for the nonignorable observation probabilities (p_{gj}'s), none of which can be tested.

The estimation procedure of the P matrix, as described in Section 3, leads to the ignoring of those components of $\underset{\sim}{\Lambda}^*(t)$ for which the corresponding $\{N_g(t)\}$'s are not observed to have jumps during the study. If, in particular, there is no missingness in the cause of death information (that is, only those $\{N_g(t)\}$'s with singleton g's are observed to have jumps), the P matrix is estimated to be an $m \times m$ identity matrix augmented by rows of zeros. As expected, this leads to the usual Nelson-Aalen estimator of $\underset{\sim}{\Lambda}(t)$ based on the observed $\{N_g(t)\}$'s. This, however, is not the case if the P matrix is known *not* to be of the above estimated form. In such a case, this extra information on the P matrix makes a modification of the Nelson-Aalen estimator via the least squares approach. One can think of a more restrictive assumption on the P matrix than the one used in Section 3. For example, we may assume, for a fixed set $g \ni j$ with $|g| = l$, the observation probability $p_{gj} = \theta_l$, which depends only on $|g| = l$, for $l = 1, \dots, m$. It is to be noted that this assumption, although not very realistic, conforms with the mechanism of missing at random. It can be seen, after some probabilistic arguments, that θ_l can be consistently estimated by

$$\binom{m-1}{l-1}^{-1} \times \frac{\text{number of observations with } |g| = l}{d.},$$

where $d. = \sum_{i=1}^{K} d_i$, total number of observed failures, for $l = 1, \dots, m$.

A set g is assumed to capture all the available information on the cause of death. However, we sometimes may have more information regarding the cause of death. In the example of Section 4, for instance, we have two descriptions of cause of death, given by "probably fatal" and "probably incidental," which are interpreted as missing with the same set $g_0 = \{2, 3\}$. Consequently, the probabilities $p_{g_0 j}$'s and q_{jg_0}'s, for $j = 2, 3$, cannot distinguish between the above two descriptions, although "probably fatal" is more likely to be due to cause 2 and "probably incidental" due to cause 3. In order to make this distinction, one needs to split g_0 into two cases and deal with them accordingly.

The estimators of Section 3 can serve as a basis for testing the completely specified null hypothesis given by $\mathcal{H}_0 : \Lambda_j(t) = \Lambda_j^0(t)$, for all j and a fixed $t = t'$. Using the weak convergence result of $\widehat{\underset{\sim}{\Lambda}}(t)$ (see Section 3), we have, under $\mathcal{H}_0$,

$$\left[\widehat{\underset{\sim}{\Lambda}}(t') - \underset{\sim}{\Lambda}^0(t')\right]^T \left[\widehat{P}^T W(t') \widehat{P}\right] \left[\widehat{\underset{\sim}{\Lambda}}(t') - \underset{\sim}{\Lambda}^0(t')\right]$$

as an asymptotic χ^2 variate with m degrees of freedom, where $\underset{\sim}{\Lambda}^0(t)$ denotes the vector of $\Lambda_j^0(t) = \int_0^t \lambda_j^0(s)ds, j = 1, \dots, m$. Similar tests can be constructed for more specific hypotheses such as $\Lambda_j(t') = \Lambda_j^0(t')$, for some of the j's.

Although we have dealt exclusively with competing risks data with partially missing information on failure types, the inference procedure would work for a general Markov chain with incomplete information on the observed types of transitions. Some work in this context with specific application would be interesting. Another relevant issue is incorporation of covariates that, in this context of general missing pattern, does not seem to be simple.

Acknowledgements

We would like to thank the associate editor and two anonymous referees for their careful reading and helpful suggestions which have led to improvement of this article.

Résumé

Dans des données de risques compétitifs, l'absence de certains types de défaillance est une situation commune. Dans ce travail nous envisageons un cadre général d'absence dans lequel si un type de défaillance n'est pas observé, il est cependant observé un ensemble de cas possibles incluant le type réel ainsi que le temps de défaillance. Nous envisageons d'abord une estimation au maximum de vraisemblance sous hypothèse d'absence aléatoire par l'algorithme EM. Nous proposons alors un estimateur de type Nelson-Aalen pour les situations dans lesquelles une information sur la probabilité conditionnelle du type réel étant donné un ensemble de types de défaillances possibles est disponible par les expérimentateurs. Ceci est basé sur une méthode de type moindres carrés utilisant les relations entre les risques de différents types et les risques pour différentes combinaisons des types absents. Une étude de simulation que nous avons réalisée pour étudier les performances de cette méthode montre que le biais peut être petit, même avec une proportion élevée de données manquantes, pour un nombre d'observations assez grand. Les estimations sont quelque peu sensibles aux mauvaises spécifications des probabilités conditionnelles des types réels lorsque la proportion de données manquantes est élevée. Nous illustrons aussi notre méthodologie avec un exemple d'expérimentation animale.

References

Aalen, O. O. (1978). Non-parametric inference for a family of counting processes. *Annals of Statistics* **6,** 701–726.

Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: A review. *Scandinavian Journal of Statistics* **12,** 97–158.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–22.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika* **79,** 855–857.

Dewanji, A. and Kalbfleisch, J. D. (1986). Nonparametric methods for survival/sacrifice experiments. *Biometrics* **42,** 325–341.

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

Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause of death data. *Journal of the American Statistical Association* **81,** 328–336.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures *Biometrika* **85,** 151–164.

Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E., and Wahrendorf, J. (1986). *Statistical Methods in Cancer Research.* Volume III: *The Design and Analysis of Long-Term Animal Experiments.* IARC Scientific Publications 79. Lyon: International Agency for Research on Cancer.

Goetghebeur, E. and Ryan, L. (1990). A modified log rank test for competing risks with missing failure type. *Biometrika* **77,** 207–211.

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

Lagakos, S. W. and Louis, T. A. (1988). Use of tumor lethality to interpret tumorigenicity experiments lacking cause-of-death data. *Applied Statistics* **37,** 169–179.

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

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57,** 1191–1197.

Miyakawa, M. (1984). Analysis of incomplete data in competing risks model. *IEEE Transactions in Reliability* **33,** 293–296.

Mukerjee, H. and Wang, J. L. (1993). Nonparametric maximum likelihood estimation of increasing hazard rate for uncertain cause-of-death data. *Scandinavian Journal of Statistics* **20,** 17–33.

Peto, R., Gray, R., Brantom, P., and Grasso, P. (1984). Nitrosamine carcinogenesis in 5120 rodents: Chronic administration of sixteen different concentrations of NDEA, MPYR and NPIP in the water of 4440 inbred rats, with parallel studies on NDEA alone of the effect of age of starting (3, 6 or 20 weeks) and of species (rats, mice or hamsters). In *N-Nitroso Compounds: Occurrence, Biological Effects and Relevance to Human Cancer,* I. K. O'Neill, R. C. Von Borstel, T. C. Miller, J. Long, and H. Bartsch (eds), 627–665. IARC Scientific Publications 57. Lyon: International Agency for Research on Cancer.

Racine-Poon, A. H. and Hoel, D. G. (1984). Nonparametric estimation of the survival function when cause of death in uncertain. *Biometrics* **40,** 1151–1158.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics* **11,** 453–466.