BIOMETRICS 57, 74–80
March 2001

Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation

Michael G. Hudgens,[1,2,*] Glen A. Satten,[3] and Ira M. Longini, Jr.[1]

[1]Department of Biostatistics, Emory University, Atlanta, Georgia 30322, U.S.A.
[2]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MW-500,
PO Box 19024, Seattle, Washington 98109, U.S.A.
[3]Centers for Disease Control and Prevention, Atlanta, Georgia 30333, U.S.A.
* *email:* mhudgens@scharp.org

SUMMARY. We derive the nonparametric maximum likelihood estimate (NPMLE) of the cumulative incidence functions for competing risks survival data subject to interval censoring and truncation. Since the cumulative incidence function NPMLEs give rise to an estimate of the survival distribution which can be undefined over a potentially larger set of regions than the NPMLE of the survival function obtained ignoring failure type, we consider an alternative pseudolikelihood estimator. The methods are then applied to data from a cohort of injecting drug users in Thailand susceptible to infection from HIV-1 subtypes B and E.

KEY WORDS: Competing risks; Cumulative incidence function; HIV subtypes; Interval censoring; Nonparametric maximum likelihood; Pseudolikelihood; Truncation.

1. Introduction

In this article, we develop nonparametric estimation methods for competing risks survival data subject to interval censoring and truncation. These methods are motivated by the Bangkok Metropolitan Administration (BMA) injecting drug users (IDUs) cohort, which was established in 1995 to assess the feasibility of conducting a phase III HIV vaccine efficacy trial in the IDU population in Bangkok, Thailand (Vanichseni et al., 2001, in press). The study was designed to measure rates of successful follow-up and HIV incidence as well as to determine related risk factors to target more effective HIV prevention strategies. The cohort consisted of 1209 HIV seronegative IDUs enrolled in two time periods in May–November 1995 and May–December 1996 at 15 BMA drug treatment clinics providing methadone treatment as part of a comprehensive care program. At enrollment and approximately every 4 months thereafter, participants received HIV prevention counseling and were assessed for HIV seroconversion. As of December 1998, there were 1124 people with at least one follow-up visit, 133 HIV seroconversions, and approximately 2400 person years of follow-up. Of the 133 people observed to seroconvert, 27 were of subtype B and 99 of subtype E. The remaining seven seroconversions were of unknown subtype, which we assume would be subtype B or E if known. Further, we assume that infection with one subtype precludes infection with the other. Thus, for these data, times of seroconversion are interval censored, i.e., only known to lie in a certain time interval; individuals are subject to competing risks; and for analyses in calendar time rather than time in study, data are also subject to left truncation due to different enrollment times.

In the competing risks setting, it is of interest to estimate the cumulative incidence function, i.e., the cumulative probability of a specific failure type. The nonparametric maximum likelihood estimate (NPMLE) of the cumulative incidence function for right-censored, competing risks survival data is given in Aalen (1976) and Kalbfleisch and Prentice (1980). In the absence of competing risks, Peto (1973) first characterized the survival function NPMLE for interval-censored failure time data and used a constrained Newton–Raphson algorithm for estimation. Turnbull (1976) extended the work of Peto to allow for truncation while using a self-consistent algorithm. Later, Frydman (1994) modified Turnbull's method to account for truncation properly. Gentleman and Geyer (1994) gave the conditions under which Turnbull's estimator is indeed the NPMLE and is unique, although they do not address truncation. Frydman (1995) extended Turnbull's estimator to the special case of a Markov illness–death process, which has a competing risks element in the structure of the process. We extend Turnbull's estimator to the general setting of competing risks, allowing for any number of failure types and for each failure time to be subject to interval censoring and truncation.

We begin by introducing some notation and stating the problem formally. Let T be the random variable for survival time with corresponding survival function $S(t) = \Pr[t < T]$. Suppose that there are J possible types or causes of failure and, for $j = 1, \ldots, J$, let the cumulative incidence function for cause j be $I_j(t) = \Pr[T \leq t$, failure of type $j]$. This is the probability of failure of type j by time t when competing risks are present. Note that $\Sigma_j I_j(t) = 1 - S(t)$ for any time t.

Suppose we have n observations and, for the ith observation $(1 \leq i \leq n)$, let t_i be the failure time and j_i the cause of failure. Following Frydman (1994), let $B_i = (v_i, u_i)$ be the truncation set such that we observe t_i only if $t_i \in B_i$ and assume that $\Pr_S(\cup_i B_i) = 1$. Suppose that t_i is only known to be in the interval $A_i = [l_i, r_i]$, where $A_i \subseteq B_i$. For right-censored observations, we use the convention that l_i is the censoring time and $r_i = \infty$. Let δ_i be one if the failure type is known and zero otherwise. The failure type is unknown for right-censored observations and may also be unknown for interval-censored observations. We will also assume that the censoring and truncation mechanism for T and the missing mechanism for failure type are independent of the failure time.

Our goal is to find estimators for the functions $I_j(t)$, $j = 1, \ldots, J$. We first consider estimators, $\hat{I}_j(t)$, that maximize the likelihood

$$\prod_{i=1}^{n} \frac{[I_{j_i}(r_i+) - I_{j_i}(l_i-)]^{\delta_i} \left[\sum_j \{I_j(r_i+) - I_j(l_i-)\}\right]^{1-\delta_i}}{\sum_j \{I_j(u_i-) - I_j(v_i+)\}}, \tag{1}$$

subject to the constraints

$$I_j(t) \text{ nonnegative, nondecreasing for each } j, \text{ and}$$
$$\sum_j I_j(t) \le 1 \qquad \text{for all } t. \tag{2}$$

In Section 2, we characterize $\hat{I}_j(t)$ and give an EM algorithm for its estimation. In the absence of competing risks, our EM algorithm reduces to Turnbull's algorithm. We show that these estimators have an unexpected and undesirable property, namely the resulting estimator of the survival distribution, $\hat{S}(t) = 1 - \Sigma_j \hat{I}_j(t)$, is undefined over a potentially larger set of regions than the NPMLE of $S(t)$ ignoring failure type. Therefore, we explore a second nonparametric estimator in Section 3 that conditions on the NPMLE of the survival distribution ignoring failure type. In Section 4, the ideas of Sections 2 and 3 are illustrated through a simple example with three data points that highlights differences in the two estimators. Finally, in Section 5, we apply these methods to the IDU data from Thailand.

2. Nonparametric Maximum Likelihood Estimation

When failure type is the same for all individuals or differences in failure type are ignored, the NPMLE of the survival distribution, $\tilde{S}(t)$, is known for failure times subject to interval censoring and truncation. We will call $\tilde{S}(t)$ the marginal survival NPMLE. Frydman (1994) showed that characterization of the marginal survival NPMLE is determined by the set $C = \cup_{k=1}^{m}[q_k, p_k]$, the union of m closed intervals, where the q_k's and p_k's are chosen as follows. First, let $L = \{l_i, u_i : 1 \le i \le n\}$ and $R = \{r_i, v_i : 1 \le i \le n\}$. Then order all points in the sets L and R and let $[q_k, p_k]$ be assigned to the kth occurrence of an element of L followed immediately by an element of R such that $q_k \le p_k$. Frydman (1994) showed that the NPMLE of $S(t)$ will be constant outside C and will decrease on some or all of the intervals within C. Further, the likelihood is a function only of the amount $S(t)$ decreases on these intervals and not of how $S(t)$ decreases. Thus, the NPMLE is a collection of functions determined only by the amount of decrease on the intervals of C.

In a similar fashion, we now construct sets C_j, which characterize the NPMLE of I_j. For each event type $j = 1, \ldots, J$, let $C_j = \cup_{k=1}^{m_j}[q_{jk}, p_{jk}]$ be the union of m_j closed intervals, where the q_{jk}'s and p_{jk}'s are obtained as follows. For each j, let $N_j = \{i : \delta_i = 0 \text{ or } j_i = j\}$, i.e., N_j is an index set for those observations with an event of type j or who have a missing failure type. Now let $L_j = \{l_i : i \in N_j\} \cup \{u_i : 1 \le i \le n\}$ and $R_j = \{r_i : i \in N_j\} \cup \{v_i : 1 \le i \le n\}$. Order all points in the sets L_j and R_j and let $[q_{jk}, p_{jk}]$ be assigned to the kth occurrence of an element of L_j followed immediately by an element of R_j such that $q_{jk} \le p_{jk}$. We now state two lemmas that characterize the NPMLE of the $I_j(t)$'s.

LEMMA 1: *For $j = 1, \ldots, J$, any cumulative incidence function that increases outside the set C_j cannot be an MLE of $I_j(t)$.*

LEMMA 2: *For fixed values $I_j(q_{jk}-)$ and $I_j(p_{jk}+)$ ($j = 1, \ldots, J$; $k = 1, \ldots, m_j$), the likelihood is independent of the behavior of $I_j(t)$ within each interval $[q_{jk}, p_{jk}]$.*

The proof of Lemma 1 is given in the Appendix and the proof of Lemma 2 follows directly from inspection of the likelihood (1). These lemmas tell us that the NPMLE of $I_j(t)$ will be constant outside C_j and will increase on some or all of the intervals of C_j. Like the marginal NPMLE of $S(t)$, the NPMLE of $I_j(t)$ will be a collection of functions determined solely by the amount of increase on the intervals of C_j.

We can now rewrite likelihood (1) based on the lemmas and develop an EM algorithm for maximization. Let $\phi_{jk} = I_j(p_{jk}+) - I_j(q_{jk}-)$, $\boldsymbol{\phi} = \{\phi_{jk} : j = 1, \ldots, J; k = 1, \ldots, m_j\}$, and α_{ijk} be an indicator variable that is one if $[q_{jk}, p_{jk}] \subseteq A_i$ and $i \in N_j$ and is zero otherwise, i.e., α_{ijk} indicates whether it is possible that the ith observation had a failure of type j in the kth interval of C_j. Let $\beta_{ijk} = 1$ if $[q_{jk}, p_{jk}] \subseteq B_i$ and be zero otherwise. Then, from the lemmas, it follows that maximizing likelihood (1) subject to (2) is equivalent to maximizing

$$\prod_{i=1}^{n} \frac{\sum_j \sum_{k=1}^{m_j} \alpha_{ijk}\phi_{jk}}{\sum_j \sum_{k=1}^{m_j} \beta_{ijk}\phi_{jk}},$$

subject to the constraints $\phi_{jk} \ge 0$ for all j, k and

$$\sum_j \sum_{k=1}^{m_j} \phi_{jk} = 1.$$

To attain this goal, we use an EM algorithm as follows.

To compute the expectation step of the EM algorithm, first we define an indicator variable $\mathcal{I}_{ijk} = 1$ if $t_i \in [q_{jk}, p_{jk}]$ and $j_i = j$ and $\mathcal{I}_{ijk} = 0$ otherwise, i.e., $\mathcal{I}_{ijk}$ indicates whether the ith observation has an event of type j in the interval $[q_{jk}, p_{jk}]$. Note that we may not know $\mathcal{I}_{ijk}$, e.g., for $\delta_i = 0$, we do not even know j_i. Let the expected value of $\mathcal{I}_{ijk}$ be denoted by $\mu_{ijk}(\boldsymbol{\phi})$. Then, under $\boldsymbol{\phi}$,

$$\mu_{ijk}(\boldsymbol{\phi}) = \frac{\alpha_{ijk}\phi_{jk}}{\sum_{j'} \sum_{k'=1}^{m_{j'}} \alpha_{ij'k'}\phi_{j'k'}}. \tag{3}$$

Because of truncation, the ith observation can be thought of as representing a group of unknown size where all observations in that group are unobserved because their failure time lies outside of B_i. Turnbull refers to these observations as the ith observation's ghosts. Let $\mathcal{G}_{ijk}$ be the number in the group corresponding to the ith observation having failure type j in

$[q_{jk}, p_{jk}]$, the kth interval of C_j. Let the expected value of $\mathcal{G}_{ijk}$ be denoted by $\nu_{ijk}(\boldsymbol{\phi})$. Then

$$\nu_{ijk}(\boldsymbol{\phi}) = \frac{(1-\beta_{ijk})\phi_{jk}}{\sum_{j'}\sum_{k'=1}^{m_{j'}} \beta_{ij'k'}\phi_{j'k'}}. \tag{4}$$

This follows because the expected number of ghosts for the ith observation is $\Pr(B_i^c)/\Pr(B_i)$ and the probability of one of these ghosts failing in the kth interval of C_j is $(1-\beta_{ijk})\phi_{jk}/\Pr(B_i^c)$, where $\Pr(B_i) = \Sigma_j \Sigma_{k=1}^{m_j} \beta_{ijk}\phi_{jk}$ and $\Pr(B_i^c) = 1 - \Pr(B_i)$.

In the maximization step, we treat expected values as observed. The overall proportion of failures of type j in the kth interval of C_j is

$$\pi_{jk}(\boldsymbol{\phi}) = \frac{\sum_i \left\{\mu_{ijk}(\boldsymbol{\phi}) + \nu_{ijk}(\boldsymbol{\phi})\right\}}{\sum_i \sum_{j'} \sum_{k'=1}^{m_{j'}} \left\{\mu_{ij'k'}(\boldsymbol{\phi}) + \nu_{ij'k'}(\boldsymbol{\phi})\right\}}. \tag{5}$$

The EM algorithm iterates between equations (3), (4), and (5) after selecting initial estimates $\phi_{jk}^{(0)} > 0$ such that

$$\sum_j \sum_{k=1}^{m_j} \phi_{jk}^{(0)} = 1,$$

i.e., computes $\mu_{ijk}(\boldsymbol{\phi}^{(0)})$ and $\nu_{ijk}(\boldsymbol{\phi}^{(0)})$, updates $\boldsymbol{\phi}$ by $\phi_{jk}^{(1)} = \pi_{jk}(\boldsymbol{\phi}^{(0)})$, and repeats until convergence. The resulting self-consistent estimate of $\boldsymbol{\phi}$, which is a solution of the simultaneous equations $\phi_{jk} = \pi_{jk}(\boldsymbol{\phi})$ $(j = 1, \ldots, J; k = 1, \ldots, m_j)$, reduces exactly to Turnbull's (1976) estimate when $J = 1$. In other words, this a generalization of Turnbull's self-consistency algorithm to the competing risks setting.

We note that the EM algorithm is not guaranteed to converge to the MLE. However, we determine that we have attained a maximum by checking the Kuhn–Tucker conditions as described in Gentleman and Geyer (1994). (See Fletcher (1987) for general methods to assess uniqueness of a maximum in constrained optimization problems and Hudgens (2000) for the specific details related to this problem.) Further, when there is no truncation, we verify the uniqueness of the estimated masses following their same methods since the log likelihood is concave. When there is truncation, the log likelihood is not concave so that the Kuhn–Tucker conditions only ensure that we are at a local maximum. If the estimate is not unique, it may be that some of the masses are not identifiable. For example, it is possible that, for some $j \neq j'$ and some $k \in \{1, \ldots, m_j\}$ and $k' \in \{1, \ldots, m_{j'}\}$, the terms ϕ_{jk} and $\phi_{j'k'}$ appear in the likelihood only in sum. This occurs if $\alpha_{ijk} = \alpha_{ij'k'}$ and $\beta_{ijk} = \beta_{ij'k'}$ for all i, in which case, only $\phi_{jk} + \phi_{j'k'}$ is identifiable. Inspection of $\Sigma_{i=1}^n \alpha_{ijk}$ and $\Sigma_{i=1}^n \beta_{ijk}$ for different j and k can be helpful in determining possible unidentifiable parameters. Another method for finding unidentifiable parameters is to look for estimates that vary for different initial starting values of the EM.

Once we have $\hat{\boldsymbol{\phi}}$, the NPMLE of I_j is

$$\hat{I}_j(t) = \begin{cases} 0 & \text{if } t < q_{j1} \\ \hat{\phi}_{j1} + \hat{\phi}_{j2} + \cdots + \hat{\phi}_{jk} & \text{if } p_{jk} < t < q_{jk+1} \\ & \quad (1 \leq k \leq m_j - 1) \\ \hat{\phi}_{j1} + \hat{\phi}_{j2} + \cdots + \hat{\phi}_{jm_j} & \text{if } t > p_{mj}. \end{cases} \tag{6}$$

For $t \in [q_{jk}, p_{jk}]$, $\hat{I}_j(t)$ is undefined if $\hat{\phi}_{jk} > 0$ and $[q_{jk}, p_{jk}] \subseteq C_j$ and equals $\hat{\phi}_{j1} + \hat{\phi}_{j2} + \cdots + \hat{\phi}_{jk-1}$ otherwise. Note that, if $p_{jm_j} = \infty$ and $\phi_{jm_j} > 0$, then $I_j(t)$ is undefined for $t > q_{jm_j}$.

No formal discussion of asymptotic properties of (6) will be undertaken here. Based on previous work of nonparametric estimation in the presence of interval censoring (e.g., Groeneboom and Wellner, 1992; Yu, Li, and Wong, 1998), if the support of the censoring mechanism is discrete and finite (as in the case of the BMA data we consider here), the estimation of the cumulative incidence functions becomes a finite dimensional estimation problem and we expect the NPMLE to have the usual $n^{1/2}$ convergence rate. If the random variables dictating the censoring are treated as continuous, the rate of convergence of the NPMLE will likely not be $n^{1/2}$, and derivation of the limiting distribution will not be trivial. Likewise, we expect the bootstrap confidence intervals based on resampling data sets of size n to be valid in the discrete setting and possibly misleading in the continuous case (Wellner and Zhan, 1997). To account for undefined regions that may arise from the bootstrap samples in cases where bootstrapping is valid, we employ the following method. In computing the upper (lower) boundary of the 95% confidence intervals, for each bootstrap sample, we consider the estimate within the class of NPMLEs that places all of an interval's mass at the left-hand (right-hand) endpoint of the undefined region. This method may result in conservative (i.e., too wide) confidence intervals; however, we expect the bias to be negligible asymptotically.

3. Pseudolikelihood Estimate

When employing the NPMLE method, once we have estimated the $I_j(t)$'s, we can estimate the survival distribution by $\hat{S}(t) = 1 - \Sigma_j \hat{I}_j(t)$. Note, however, that $\hat{S}(t)$ is defined only for t where each $\hat{I}_j(t)$ is defined, i.e., $\hat{S}(t)$ may be undefined within all or part of $\cup_j C_j$. Recall the marginal NPMLE, $\tilde{S}(t)$, is undefined only within C, where $C \subseteq \cup_j C_j$. Thus, it is possible that $\hat{S}(t)$ is undefined on a larger region than ,$\tilde{S}(t)$. (For an example, see the next section which explores a data set having three observations.) In this section, we give estimators $\tilde{I}_j(t)$ such that

$$1 - \sum_j \tilde{I}_j(t) = \tilde{S}(t). \tag{7}$$

We do this by first calculating the NPMLE of the survival distribution, $\tilde{S}(t)$, ignoring failure type, and then impose the condition in equation (7). We will refer to this method and resulting estimator as the pseudolikelihood method and pseudolikelihood estimate (PLE), respectively.

Note that an equivalent form of (1) is given by

$$\prod_{i=1}^n \frac{[I_{j_i}(r_i+) - I_{j_i}(l_i-)]^{\delta_i}[S(l_i-) - S(r_i+)]^{1-\delta_i}}{S(v_i-) - S(u_i+)}. \tag{8}$$

Thus, maximizing (1) subject to the constraint (7) is equivalent to maximizing

$$\prod_{i=1}^n [I_{j_i}(r_i+) - I_{j_i}(l_i-)]^{\delta_i}. \tag{9}$$

By constraining the PLEs of the cumulative incidence functions by (7), $\tilde{I}_j(t)$ must be constant outside C for each j. This follows since the cumulative incidence functions are monotonically increasing and must sum to $1 - \tilde{S}(t)$, which is constant outside C. Whereas each cumulative incidence function estimate $\hat{I}_j(t)$ using the NPMLE approach is undefined on different regions, the PLEs will be undefined only on C for all types of failure. Thus, we let $\psi_{jk} = I_j(p_k+) - I_j(q_k-)$, $\boldsymbol{\psi} = \{\psi_{jk} : j = 1, \ldots, J; k = 1, \ldots, m\}$, and $\tilde{\psi}_k = \tilde{S}(q_k-) - \tilde{S}(p_k+)$. Also let α^c_{ijk} be an indicator variable that is one if $[q_k, p_k] \subseteq A_i$, $\delta_i = 1$, and $j_i = j$ and is zero otherwise. Following reasoning similar to the previous section, (9) is equivalent to

$$\prod_{i \in R} \sum_j \sum_{k=1}^{m} \alpha^c_{ijk} \psi_{jk}, \qquad (10)$$

where $R = \{i \mid \delta_i = 1\}$. The goal is to maximize (10) subject to the constraints that $\psi_{jk} \geq 0$ for all j, k and $\Sigma_j \, \psi_{jk} = \tilde{\psi}_k$. Again, we use an EM algorithm to attain this goal.

Define an indicator variable $\mathcal{J}_{ijk} = 1$ if $t_i \in [q_k, p_k]$ and $j_i = j$ and $\mathcal{J}_{ijk} = 0$ otherwise, i.e., $\mathcal{J}_{ijk}$ indicates whether the ith observation has an event of type j in the interval $[q_k, p_k]$. Let the expected value of $\mathcal{J}_{ijk}$ under $\boldsymbol{\psi}$ be denoted by $\mu_{ijk}(\boldsymbol{\psi})$ such that

$$\mu_{ijk}(\boldsymbol{\psi}) = \frac{\alpha^c_{ijk} \psi_{jk}}{\sum_{j'} \sum_{k'=1}^{m} \alpha^c_{ij'k'} \psi_{j'k'}}. \qquad (11)$$

Treating the expected values as observed, the proportion of failures of type j in the kth interval is

$$\pi_{jk}(\boldsymbol{\psi}) = \tilde{\psi}_k \frac{\sum_{i \in R} \mu_{ijk}(\boldsymbol{\psi})}{\sum_{i \in R} \sum_{j'} \mu_{ij'k}(\boldsymbol{\psi})} \qquad (12)$$

such that $\Sigma_j \, \pi_{jk}(\boldsymbol{\psi}) = \tilde{\psi}_k$.

The EM algorithm iterates between equations (11) and (12) after selecting initial estimates $\psi^{(0)}_{jk}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, m$ such that $\Sigma_j \, \psi^{(0)}_{jk} = \tilde{\psi}_k$, where $\psi^{(0)}_{jk} > 0$ if $\tilde{\psi}_k > 0$ and $\psi^{(0)}_{jk} = 0$ if $\tilde{\psi}_k = 0$. Once we have $\tilde{\boldsymbol{\psi}}$, we can compute the PLE $\tilde{I}_j(t)$ in a fashion similar to (6). As in the NPMLE case, we verify that we have reached a maximum via standard techniques for constrained optimization similar to the methods outlined in Gentleman and Geyer (1994).

4. Example

In this section, we consider a simple example to illustrate the potential different behaviors of the NPMLE and PLE. Consider the following data set of $n = 3$ observations without truncation:

i	A_i	δ_i	j_i
1	[1, 3]	1	1
2	[2, 5]	1	2
3	[4, 5]	1	2

(13)

From the data, we see that $C_1 = [1, 3]$ and $C_2 = [4, 5]$. Thus, to get the NPMLE, we maximize $\phi_{11}\phi_{21}\phi_{21}$, subject to constraints $\phi_{11} + \phi_{21} = 1$, $\phi_{11} \geq 0$, and $\phi_{21} \geq 0$, where $\phi_{11} = I_1(3+) - I_1(1-)$ and $\phi_{21} = I_2(5+) - I_2(4-)$. Following Gentleman and Geyer (1994), we use the Kuhn–Tucker conditions to find that the unique constrained maximum occurs at $\hat{\phi}_{11} = 1/3$ and $\hat{\phi}_{21} = 2/3$. Starting at $\phi^{(0)}_{11} = \phi^{(0)}_{21} = 1/2$, the EM converges to these values.

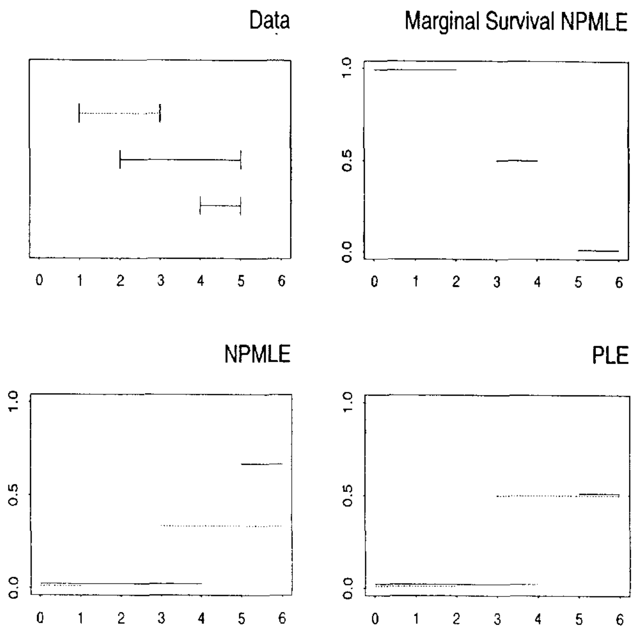

Figure 1. Example data given in Section 4. In all four panels, the horizontal axis is time in arbitrary units. Top left panel: graphical representation of data where the dotted and solid lines denote censoring intervals for failure types 1 and 2, respectively. Top right panel: marginal NPMLE of the survival distribution, $\tilde{S}(t)$. Bottom left panel: NPMLE of the cumulative incidence functions. The dotted and solid lines give $\hat{I}_1(t)$ and $\hat{I}_2(t)$, respectively. Bottom right panel: Cumulative incidence function estimates using the pseudolikelihood method. The dotted and solid lines give $\tilde{I}_1(t)$ and $\tilde{I}_2(t)$, respectively.

The set C consists of two regions, $[2, 3]$ and $[4, 5]$. The marginal NPMLE for the survival distribution is $\tilde{\psi}_1 = 1/2$ and $\tilde{\psi}_2 = 1/2$. For the pseudolikelihood method, we let $\psi_{j1} = I_j(3+) - I_j(2-)$ and $\psi_{j2} = I_j(5+) - I_j(4-)$ for $j = 1, 2$. The goal is to maximize $\psi_{11}(\psi_{21} + \psi_{22})\psi_{22}$ subject to the constraints $\psi_{1k} + \psi_{2k} = \tilde{\psi}_k = 1/2$ for $k = 1, 2$ and $\psi_{jk} \geq 0$ for $j, k = 1, 2$. The unique constrained maximum occurs at $\tilde{\psi}_{11} = \tilde{\psi}_{22} = 1/2$ and $\tilde{\psi}_{12} = \tilde{\psi}_{21} = 0$. Starting at $\psi^{(0)}_{jk} = 1/4$ for all j, k, the EM converges to these values. See Figure 1 for a graphical representation of this example.

There are two differences in the estimators illustrated by this simple example. First, the length of regions on which each estimator is undefined can be different. For example, here $\hat{S}(t)$ is undefined on $C_1 \cup C_2 = [1, 3] \cup [4, 5]$ whereas $\tilde{S}(t)$ is only undefined on the smaller region $C = [2, 3] \cup [4, 5]$. Second, even in the regions where both estimators are defined, they will not necessarily be equal. For example, see the regions $[3, 4]$ and $[5, \infty)$. Estimators $\hat{S}$ and $\tilde{S}$ may also be defined over a different number of regions, although that is not the case in this example.

The motivation behind the pseudolikelihood method is to find an estimator of $I_j(t)$ such that the resulting estimate of the survival distribution is defined over the same regions as

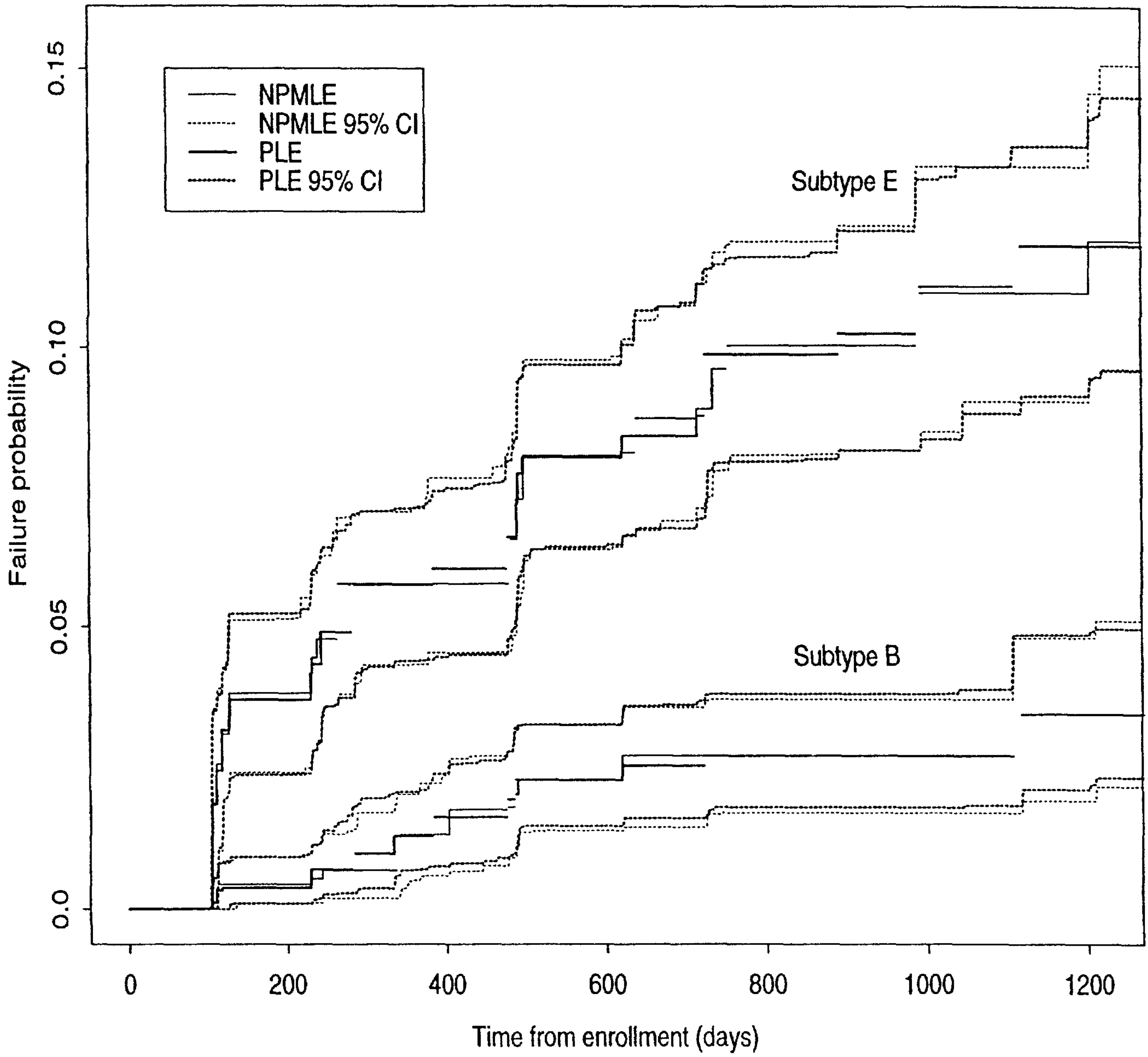

Figure 2. NPMLE and PLE of the subtype-specific cumulative incidence functions for time from enrollment to seroconversion in the BMA cohort. Ninety-five percent bootstrap confidence intervals (CI) for NPMLE and PLE are given by dotted lines.

$\tilde{S}(t)$, namely outside C. The pseudolikelihood method ensures not only that the estimated survival function is defined over these regions but also that it is equal to the marginal NPMLE of the survival function, i.e., $\tilde{S}(t) = 1 - \Sigma_j \tilde{I}_j(t)$ for all t outside C. A different estimator arises if we only assume that the estimate of $I_j(t)$ is defined and constant outside C but is not necessarily subject to (7). For example, suppose we observe the same data as (13) except that $j_3 = 1$. If we let $\psi_{j1} = I_j(3+) - I_j(2-)$ and $\psi_{j2} = I_j(5+) - I_j(4-)$ for $j = 1, 2$ and maximize $\psi_{11}(\psi_{21} + \psi_{22})\psi_{12}$ with the only constraint being $\Sigma_{j,k}\, \psi_{jk} = 1$, then the constrained maximum occurs at $\hat{\psi}_{11} = \hat{\psi}_{2\cdot} = \hat{\psi}_{12} = 1/3$, where $\psi_{2\cdot} = \psi_{21} + \psi_{22}$, i.e., ψ_{21} and ψ_{22} are not identifiable. In general, this estimator introduces additional parameters that are not identifiable without the constraints of the pseudolikelihood method and thus will not be considered further. Note that the parameters in both the NPMLE and PLE are identifiable for the data set given in (13) with the change $j_3 = 1$.

5. Application

In this section, we apply the above methods to the BMA cohort study, which is described in the Introduction and in more detail elsewhere (Vanichseni et al., 2001, in press). It

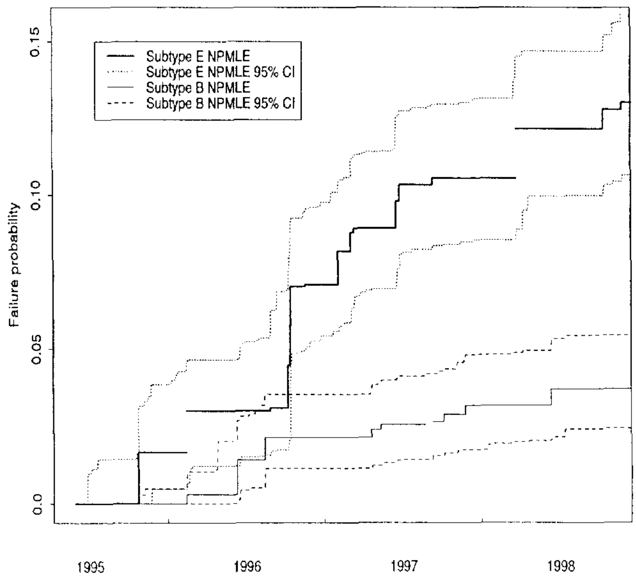

Figure 3. NPMLE and 95% bootstrap confidence intervals (CI) of the subtype-specific cumulative incidence functions for calendar time of seroconversion conditional on survival beyond day 0.

is unknown if subtypes B and E differ with regard to infectiousness or transmissibility in the Bangkok IDU population. A first step in addressing this issue is to compare the cumulative incidence of subtypes E and B in the cohort.

We first calculate the cumulative incidence of subtypes B and E infection as a function of time since study enrollment. This would be the appropriate analysis if all participants had been enrolled simultaneously; it is also valid if the hazard of infection by each HIV subtype is roughly constant. It is also a natural first analysis for these data, even if the more appropriate time scale for analysis is calendar time. The NPMLE and PLE of the cumulative incidence functions for subtypes B and E are given in Figure 2. Since the enrollment date, last seronegative, and first seropositive tests are measured in days, the support of the censoring mechanism is discrete and finite. Therefore, confidence intervals for the NPMLE and PLE were obtained using the bootstrap percentile method from 500 replacement samples of size $n = 1124$ from $\{(A_i, \delta_i, j_i) : i = 1, \ldots, n\}$.

The largest right-censored observation occurs at 1297 days after enrollment, which is greater than the right-hand endpoint of all interval-censored observations. Thus, for either method, estimates of $S(t)$, $I_B(t)$, and $I_E(t)$ are undefined beyond 1297. Furthermore, $I_B(\infty) - I_B(1297)$ and $I_E(\infty) - I_E(1297)$ are not identifiable for either method (i.e., we cannot estimate the ultimate proportion infected with B or E). Otherwise, the NPMLE masses $\hat{\phi}$ and the PLE masses $\tilde{\psi}$ are unique for this data set. Excluding the region $[1297, \infty)$, we compare the number and length of undefined regions for the two estimators in this example. These regions are denoted by the absence of vertical lines in the estimators in Figure 2. The NPMLEs of $\hat{I}_B(t)$ and $I_E(t)$ are both undefined on five distinct regions, so $\hat{S}(t)$ is undefined on 10 regions. The PLEs $\tilde{I}_B(t)$ and $\tilde{I}_E(t)$ are undefined in six and seven regions, respectively, and $\tilde{S}(t)$ is undefined in eight regions. Looking at the overall length of the undefined regions, the NPMLEs of $I_B(t)$, $I_E(t)$, and $S(t)$ are undefined for 21, 13, and 34 days, respectively, while the corresponding PLEs are undefined for 22, 26, and 28 days.

Because HIV incidence may vary over time, it is necessary to calculate the survival and cumulative incidence functions over calendar time. Since the enrollment time varies within the cohort, this results in left-truncated data. It is known that the NPMLE of the survival distribution is inconsistent and can severely underestimate the true survival probability in the presence of left truncation and interval censoring (Pan and Chappell, 1999). This is due to a small number of people being at risk at the early times of a study. For example, the marginal survival NPMLE for the BMA cohort data with truncation drops 0.17 on day 0 (i.e., May 30, 1995), resulting in an underestimate of the survival curve. Similarly, the NPMLE of the cumulative incidence function for subtype B increases 0.17 on day 0. To compensate, we conditioned on survival beyond day 0. Figure 3 gives the NPMLE of the cumulative incidence functions for subtypes B and E over calendar time, both conditional on not seroconverting on day 0. The plot (not shown) of the PLEs of $I_B(t)$ and $I_E(t)$ over calendar time is similar to that in Figure 3 when conditioning on survival beyond day 0. Confidence intervals for the NPMLE were obtained using 500 replacement bootstrap samples of size $n = 1124$. The PLEs of the cumulative incidence functions and marginal survival NPMLE are defined everywhere. The NPMLE of $I_B(t)$ and $I_E(t)$ are undefined on one and two regions, having total length 15 and 2 days, respectively, while $\hat{S}(t)$ is undefined on three intervals of length 17 days.

Figure 2 suggests that the cumulative incidence for subtype E is considerably higher than that for subtype B. However, looking at Figure 3, we see that, although the incidence of subtype E is higher throughout, the difference is not very pronounced until late 1996. The cause of this difference is under investigation using subtype-specific transmission probability models (Hudgens et al., unpublished manuscript).

6. Discussion

In this work, we characterized the NPMLE of cumulative incidence functions for competing risks survival data subject to interval censoring and truncation. Further, we developed an EM algorithm that, coupled with the Kuhn–Tucker conditions, provides a numerical algorithm for estimation. We considered a pseudolikelihood estimate as an alternative to the NPMLE to avoid introducing additional undefined regions to the survival function estimate. For the BMA example without truncation, the desired effect was achieved, namely, $\tilde{S}$ has fewer undefined regions of shorter total length than $\hat{S}$. However, there is a trade-off in that the PLEs of the cumulative incidence functions have more undefined regions of greater total length than the corresponding NPMLEs. When truncation is included, the results differ in that the nonparametric maximum likelihood methods result in estimators of the survival and cumulative incidence functions that are undefined over more regions of greater total length than the corresponding estimates from the pseudolikelihood method. Further research is needed to investigate the theoretical properties, such as consistency, rates of convergence, and asymptotic distributions, of the two proposed estimators.

Acknowledgements

We would like to thank the following investigators for providing information about the BMA cohort and for their participation and suggestions: Dr Kachit Choopanya, Dr Suphak Vanichseni, and Dr Suwanee Raktham of the BMA, Bangkok; Dr Dwip Kitayaporn of the HIV/AIDS Collaboration, Bangkok and the Faculty of Tropical Medicine, Mahidol University, Bangkok; Dr Timothy D. Mastro and Mr Philip A. Mock of the HIV/AIDS Collaboration, Bangkok; and Dr Dale J. Hu of the Division of HIV/AIDS, Centers for Disease Control and Prevention. We also would like to thank Dr M. Elizabeth Halloran of Emory University for her comments. This research was partially supported by NIH grants R01-AI32042, R01-AI40846, and T32-AI07442.

Résumé

Nous fournissons une estimation non-paramétrique du maximum de vraisemblance (NPMLE) de fonctions d'incidence cumulées pour des données de survie sujettes à des risques compétitifs et des censures par intervalle et des troncations. Puisque les estimations NPMLE de la fonction d'incidence cumulée produisent une estimation de la fonction de survie, qui peut être indéfinie sur des régions potentiellement plus larges que l'estimation NPMLE de la fonction de survie obtenue en ignorant le type de défaillance, nous considérons un estimateur alternatif de pseudo-vraisemblance. Ces méthodes sont alors appliquées à des données provenant d'une cohorte d' utilisateurs de drogues par injection en Thaïlande, susceptibles d'être infectés par les sous-types B et E du VIH.

References

Aalen, O. O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandanavian Journal of Statistics* **3,** 15–27.

Fletcher, R. (1987). *Practical Methods of Optimization.* New York: Wiley.

Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval censored and truncated observations. *Journal of the Royal Statistical Society, Series B* **56,** 71–74.

Frydman, H. (1995). Nonparametric estimation of a Markov 'illness–death' process from interval censored observations, with application to diabetes survival data. *Biometrika* **82,** 773–789.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81,** 618–623.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

Hudgens, M. G. (2000). HIV, interval censoring, and competing risks. Ph.D. thesis, Emory University, Atlanta.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Pan, W. and Chappell, R. (1999). A note on inconsistency of NPMLE of the distribution function for left truncated and case I interval censored data. *Lifetime Data Analysis* **5,** 281–291.

Peto, R. (1973). Empirical survival curves for interval censored data. *Applied Statistics* **22,** 86–91.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **52,** 290–295.

Vanichseni, S., Kitayaporn, D., Mastro, T. D., Mock, P., Raktham, S., Des Jarlai, D. C., Sujarita, S., Srisuwanvilai, L., Young, N. L., Wasi, C., Subbarao, S., Heyward, W. L., Esparza, J., and Choopanya, K. (2001). Continued high HIV-1 incidence in a vaccine trial preparatory cohort of injecting drug users in Bangkok, Thailand. *AIDS,* in press.

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* **92,** 945–959.

Yu, Q., Li, L., and Wong, G. Y. C. (1998). Asymptotic variance of the GMLE of a survival function with interval censored data. *Sankhyā* **60,** 184–187.

Received September 1999. *Revised May* 2000.
Accepted June 2000.

Appendix

Lemma 1. *For $j = 1, \ldots, J$, any cumulative incidence function that increases outside the set C_j cannot be an MLE of $I_j(t)$.*

Proof. Suppose we have an estimator $\hat{I}_j(t)$ satisfying the constraints given in (2). Further suppose there exists at least one j such that $\hat{I}_j(t)$ is not constant outside C_j. We will now show that $\hat{I}_j(t)$ cannot be the MLE, thus proving the lemma.

Construct another estimator $\bar{I}_j(t)$ as follows. For $k = 1, \ldots, m_j - 1$, choose $r_{jk} \in (p_{jk}, q_{jk+1})$ such that r_{jk} is greater than all elements of R_j and less than all elements of L_j that are in the interval. Then let $\bar{I}_j(t)$ be constant outside C_j such that, for $k = 1, \ldots, m_j - 1$,

$$\bar{I}_j(p_{jk}+) = \bar{I}_j(q_{jk+1}-) = \bar{I}_j(r_{jk}) = \hat{I}(r_{jk}).$$

Letting $\bar{I}_j(t) = \hat{I}_j(t)$ for all $t \in C_j$, it is easy to see that $\bar{I}_j(t)$ also satisfies the constraints given in (2). Further, likelihood (1) is greater at $\bar{I}_j(t)$ than $\hat{I}_j(t)$ since, for any $t \in \{R_j\}$,

$$\hat{I}_j(t+) \leq \bar{I}_j(t+), \tag{14}$$

and for any $t \in \{L_j\}$,

$$\hat{I}_j(t-) \geq \bar{I}_j(t-), \tag{15}$$

with at least one strict inequality holding in (14) or (15) since $\hat{I}_j(t)$ is not constant outside C_j. Therefore, $\hat{I}_j(t)$ is not an MLE.