Published in final edited form as:

Biometrics. 2017 September ; 73(3): 857–865. doi:10.1111/biom.12664.

Semiparametric Regression Analysis of Interval-Censored Competing Risks Data

Lu Mao^{1,*}, D. Y. Lin^{2,**}, and Donglin Zeng^{2,***}

¹Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53792, U.S.A

²Department of Biostatistics, CB 7420, University of North Carolina, Chapel Hill, North Carolina 27599-7420, U.S.A

Summary

Interval-censored competing risks data arise when each study subject may experience an event or failure from one of several causes and the failure time is not observed directly but rather is known to lie in an interval between two examinations. We formulate the effects of possibly time-varying (external) covariates on the cumulative incidence or sub-distribution function of competing risks (i.e., the marginal probability of failure from a specific cause) through a broad class of semiparametric regression models that captures both proportional and non-proportional hazards structures for the sub-distribution. We allow each subject to have an arbitrary number of examinations and accommodate missing information on the cause of failure. We consider nonparametric maximum likelihood estimation and devise a fast and stable EM-type algorithm for its computation. We then establish the consistency, asymptotic normality, and semiparametric efficiency of the resulting estimators for the regression parameters by appealing to modern empirical process theory. In addition, we show through extensive simulation studies that the proposed methods perform well in realistic situations. Finally, we provide an application to a study on HIV-1 infection with different viral subtypes.

Keywords

Cumulative incidence; Interval censoring; Nonparametric maximum likelihood estimation; Self-consistency algorithm; Time-varying covariates; Transformation models

1. Introduction

In clinical and epidemiological studies, the event of interest is often asymptomatic, such that the event time or failure time cannot be exactly observed but is rather known to lie in an interval between two examination times. An additional complication arises when there are

*lmao@biostat.wisc.edu

**lin@bios.unc.edu

***dzeng@email.unc.edu

Supplementary Materials

Web Appendix, referenced in Section 2.4, and the R code implementing the methodology are available with this paper at the Biometrics website on Wiley Online Library.

several distinct causes or types of failure. The resulting data are referred to as interval-censored competing risks data. Such data are commonly encountered in HIV/AIDS research, where seroconversion with different HIV-1 viral subtypes is determined through periodic blood tests (Hudgens et al., 2001). Such data are also encountered in cancer clinical trials, where different types of adverse events may occur during reporting periods and the follow-up for each patient is terminated upon occurrence of any adverse event.

The distribution function of the failure time with competing risks can be decomposed into the sub-distribution functions of the constituent risks. The sub-distribution function, also known as the cumulative incidence, represents the cumulative probability of failure from a specific cause over time (Kalbfleisch and Prentice, 2002, §8.2). This quantity is clinically relevant because it characterizes the ultimate experience of the subject.

Several methods are available to make inference about the cumulative incidence with right-censored competing-risks data (e.g., Gray, 1988; Fine and Gray, 1999; Kalbfleisch and Prentice, 2002, §8.2; Martinussen and Scheike, 2006, Ch. 10; Mao and Lin, 2016). Because none of the failure times are observed exactly under interval censoring, it is much more challenging, both theoretically and computationally, to deal with interval-censored than right-censored data. The literature on the cumulative incidence for interval-censored competing risks data has focused on one-sample estimation. Specifically, Hudgens et al. (2001) adapted the self-consistency algorithm of Turnbull (1976) to compute the nonparametric maximum likelihood estimator (NPMLE). Jewell et al. (2003) studied the NPMLE and other estimators with current status data, where each subject is examined only once. Groeneboom et al. (2008a; 2008b) established rigorous asymptotic theory for the NPMLE with current status data. Li and Fine (2013) studied kernel-smoothed nonparametric estimation in current status data.

In this article, we consider a general class of semiparametric regression models for competing risks data with potentially time-varying (external) covariates. This class of models encompasses both proportional and non-proportional sub-distribution hazards structures. We study the NPMLEs for these models when there is a random sequence of examination times for each subject and the cause of failure information may be partially missing. We develop a fast and stable EM-type algorithm by extending the self-consistency formula of Turnbull (1976). We establish that, under mild conditions, the proposed estimators for the regression parameters are consistent and asymptotically normal. In addition, the estimators attain the semiparametric efficiency bound with a covariance matrix that can be consistently estimated by the profile likelihood method (Murphy and van der Vaart, 2000). The proofs involve careful use of modern empirical processes theory (van der Vaart and Wellner, 1996) and semiparametric efficiency theory (Bickel et al., 1993) to address unique challenges posed by the combination of interval censoring and competing risks. We evaluate the operating characteristics of the proposed numerical and inferential procedures through extensive simulation studies. Finally, we describe an application to a clinical study of HIV/AIDS, in which a cohort of injecting drug users was followed for evidence of seroconversion with HIV-1 viral subtypes B and E.

2. Methods

2.1 Data and Models

Suppose that T is a failure time with K competing risks or causes of failure. Let $D \in \{1, \dots, K\}$ indicate the cause of failure, and let $\mathbf{Z}(\cdot)$ denote a p -vector of possibly time-varying external covariates. We formulate the effects of $\mathbf{Z}$ on (T, D) through the conditional cumulative incidence or sub-distribution functions

$$F_k(t; \mathbf{Z}) = \Pr(T \leq t, D=k | \mathbf{Z}), \quad k=1, \dots, K.$$

Equivalently, we consider the conditional sub-distribution hazard function

$$\lambda_k(t; \mathbf{Z}) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr\{t \leq T < t + \Delta t, D=k | T \geq t \cup (T \leq t \cap D \neq k), \mathbf{Z}\},$$

which is the conditional hazard function for the improper random variable

$T_k^* \equiv I(D=k)T + I(D \neq k)\infty$, where $I(\cdot)$ is the indicator function. It is easy to see that $F_k(t; \mathbf{Z}) = 1 - \exp\{-\Lambda_k(t; \mathbf{Z})\}$, where $\Lambda_k(t; \mathbf{Z}) = \int_0^t \lambda_k(s; \mathbf{Z}) ds$.

We adopt a class of semiparametric transformation models for each risk

$$\Lambda_k(t; \mathbf{Z}) = G_k \left\{ \int_0^t e^{\beta_k^T \mathbf{Z}(s)} d\Lambda_k(s) \right\}, \quad (1)$$

where $G_k(\cdot)$ is a known increasing function, β_k is a set of regression parameters, and $\Lambda_k(\cdot)$ is an arbitrary increasing function with $\Lambda_k(0) = 0$ (Mao and Lin, 2016). The choices of $G_k(x) = x$ and $G_k(x) = \log(1+x)$ correspond to the proportional hazards and proportional odds models, respectively, for the sub-distribution. When $\mathbf{Z}$ consists of only time-invariant covariates, equation (1) can be expressed in the form of a linear transformation model

$$g_k\{F_k(t; \mathbf{Z})\} = Q_k(t) + \beta_k^T \mathbf{Z},$$

where $g_k(\cdot)$ is a known increasing function, and $Q_k(\cdot)$ is an arbitrary increasing function (Cheng et al., 1995; Chen et al., 2002; Lu and Ying, 2004). The choices of $g_k(x) = \log\{-\log(1-x)\}$ and $\log\{x/(1-x)\}$ yield the proportional hazards and proportional odds models, respectively.

We consider a general interval-censoring scheme under which each subject has an arbitrary sequence of examination times and the information on the cause of failure is possibly missing. Specifically, let $U_1 < U_2 < \dots < U_J$ denote a random sequence of examination times, where J is a random integer. Define $\mathbf{U} = (U_1, \dots, U_J)^T$, where $U_j = I(U_{j-1} < T \leq U_j)$ ($j = 1, \dots, J$), and $U_0 = 0$. Also, write $\tilde{D} = D I(U_0 < T \leq U_1)$. In addition, let ξ indicate, by the values 1

versus 0, whether or not the cause of failure is observed. We set $\xi_i = 1$ if $\tilde{D}_i = 0$. Then, the observed data for a random sample of n subjects consist of $(J_i, U_{ij}, \xi_i, \xi_i \tilde{D}_i, \mathbf{Z}_i)$ ($i = 1, \dots, n$), where $U_i = (U_{i0}, U_{i1}, \dots, U_{iJ_i})^T$, and $i = (i_1, \dots, i_{J_i})^T$.

2.2 Nonparametric Maximum Likelihood Estimation

Write $\beta = (\beta_1^T, \dots, \beta_K^T)^T$ and $\Lambda = (\Lambda_1, \dots, \Lambda_K)$. We estimate β and Λ by the nonparametric maximum likelihood approach. Suppose that (T, D) is independent of (U, J) conditional on $\mathbf{Z}$ and that the cause of failure is missing at random (MAR). Then, the likelihood for (β, Λ) can be written as

$$\begin{aligned} L_n(\beta, \Lambda) = & \prod_{i=1}^n \left[\prod_{j=1}^{J_i} \prod_{k=1}^K \left(\exp \left[-G_k \left\{ \int_0^{U_{i,j-1}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right. \right. \\ & \left. \left. - \exp \left[-G_k \left\{ \int_0^{U_{ij}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right)^{I(\xi_i \tilde{D}_i = k, \Delta_{ij} = 1)} \right. \\ & \times \left\{ \sum_{k=1}^K \left(\exp \left[-G_k \left\{ \int_0^{U_{i,j-1}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right. \right. \\ & \left. \left. - \exp \left[-G_k \left\{ \int_0^{U_{ij}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right)^{I(\xi_i = 0, \Delta_{ij} = 1)} \right\} \\ & \times \left(\sum_{k=1}^K \exp \left[-G_k \left\{ \int_0^{U_{i,J_i}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] - K + 1 \right)^{I(\Delta_i = 0)} \Bigg], \end{aligned}$$

where the first and second factors correspond to $F_k(U_{ij}; \mathbf{Z}_i) - F_k(U_{i,j-1}; \mathbf{Z}_i)$ and

$\sum_{k=1}^K \{F_k(U_{ij}; \mathbf{Z}_i) - F_k(U_{i,j-1}; \mathbf{Z}_i)\}$, respectively, and the last factor corresponds to the

overall survival function $1 - \sum_{k=1}^K F_k(U_{i,J_i}; \mathbf{Z}_i)$. This likelihood is similar to that of the mixed-case censoring of Hudgens et al. (2014), but it pertains to semiparametric regression models instead of parametric distribution functions. Let $(L_j, R_j]$ denote the interval among $(U_{j0}, U_{j1}], \dots, (U_{j,J_j}, \infty]$ that contains T_j . Then, the above likelihood can be written as

$$\begin{aligned} & \prod_{k=1}^K \prod_{i: \xi_i \tilde{D}_i = k} \left(\exp \left[-G_k \left\{ \int_0^{L_i} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] - \exp \left[-G_k \left\{ \int_0^{R_i} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right) \\ & \times \prod_{i: \xi_i = 0} \left\{ \sum_{k=1}^K \left(\exp \left[-G_k \left\{ \int_0^{L_i} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] - \exp \left[-G_k \left\{ \int_0^{R_i} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right) \right\} \\ & \times \prod_{i: R_i = \infty} \left(\sum_{k=1}^K \exp \left[-G_k \left\{ \int_0^{L_i} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] - K + 1 \right). \end{aligned} \quad (2)$$

To maximize (2), we treat Λ_k as a right-continuous step function that jumps at the two ends of the intervals. Specifically, let $t_{k1} < \dots < t_{k,m_k}$ denote the distinct values of L_j and R_j with $\xi_j \tilde{D}_j = k$ or $\xi_j = 0$. In addition, let λ_{kj} denote the jump size of Λ_k at t_{kj} , and let $\mathbf{Z}_{ikj} = \mathbf{Z}_i(t_{kj})$. Then, the likelihood given in (2) becomes

$$\begin{aligned}
& \prod_{k=1}^K \prod_{i:\xi_i \tilde{D}_i=k} \left[\exp \left\{ -G_k \left(\sum_{t_{kj} \leq L_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} - \exp \left\{ -G_k \left(\sum_{t_{kj} \leq R_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} \right] \\
& \times \prod_{i:\xi_i=0} \left(\sum_{k=1}^K \left[\exp \left\{ -G_k \left(\sum_{t_{kj} \leq L_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} - \exp \left\{ -G_k \left(\sum_{t_{kj} \leq R_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} \right] \right) \\
& \times \prod_{i:R_i=\infty} \left[\sum_{k=1}^K \exp \left\{ -G_k \left(\sum_{t_{kj} \leq L_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} - K + 1 \right]. \quad (3)
\end{aligned}$$

2.3 Numerical Algorithm

Direct maximization of (3) is difficult due to the high dimensionality of λ_{kj} . This task is further complicated by the fact that, unlike the right-censoring case, the maximizers for some λ_{kj} are zero and thus lie at the boundary of the parameter space. To overcome such difficulties in the univariate case, Turnbull (1972) proposed a self-consistency formula for computation of NPMLE, which is essentially an EM algorithm based on completely observed data. Herein, we propose a novel EM algorithm that extends Turnbull's formula to regression analysis with competing risks.

Let $N_k(u, v]$ denote the number of events of the k th type that occur in the interval $(u, v]$. For the i th subject with $R_i < \infty$, let $s_{ik1} < s_{ik2} < \dots < s_{ik,j_{ik}}$ denote the distinct values of t_{kj} in the interval $(L_i, R_i]$, such that the interval $(L_i, R_i]$ is partitioned into a sequence of sub-intervals $(s_{ik0}, s_{ik1}]$, $\dots$, $(s_{ik,j_{ik}-1}, s_{ik,j_{ik}}]$, where $s_{ik0} = L_i$. In the "complete data", we know which sub-interval the failure time lies in, along with the cause of failure. That is, the complete data consist of the full paths of $N_k(0, \cdot]$ ($k = 1, \dots, K$), and the missing data are the unknown cause of failure and the precise location of each event within the observed interval. Then, the complete-data log-likelihood takes the form

$$\sum_{i=1}^n \left\{ \sum_{k=1}^K \sum_{j=1}^{j_{ik}} I(R_i < \infty) N_{ki}(s_{ik,j-1}, s_{ik,j}] \log \Delta F_k(s_{ik,j}; \mathbf{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) + I(R_i = \infty) \log S(L_i; \mathbf{Z}_i, \boldsymbol{\beta}, \Lambda) \right\},$$

where N_{ki} denotes the counting process N_k for the i th subject,

$S(t; \mathbf{Z}, \boldsymbol{\beta}, \Lambda) = 1 - \sum_{k=1}^K F_k(t; \mathbf{Z}, \boldsymbol{\beta}_k, \Lambda_k)$, $F_k(t; \mathbf{Z}, \boldsymbol{\beta}_k, \Lambda_k) = 1 - \exp\{-\Lambda_k(t; \mathbf{Z})\}$, and $F_k(t; \mathbf{Z}, \boldsymbol{\beta}_k, \Lambda_k)$ is the jump size of $F_k(\cdot; \mathbf{Z}, \boldsymbol{\beta}_k, \Lambda_k)$ at t .

In the M-step, we maximize

$$\sum_{i=1}^n \left\{ \sum_{k=1}^K \sum_{j=1}^{j_{ik}} \hat{w}_{ikj} \log \Delta F_k(s_{ik,j}; \mathbf{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) + I(R_i = \infty) \log S(L_i; \mathbf{Z}_i, \boldsymbol{\beta}, \Lambda) \right\}, \quad (4)$$

where $\hat{w}_{ikj}$ is the conditional probability that the i th subject experiences a failure of the k th cause in $(s_{ik,j-1}, s_{ik,j}]$ given the subject's failure information. If $\xi_i \tilde{D}_i = k'$, then

$$\hat{w}_{ikj} = E \{ N_{ki}(s_{ik,j-1}, s_{ikj}) | N_{k'i}(L_i | R_i) = 1 \} = I(k=k') \frac{\Delta F_k(s_{ikj}; \mathbf{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}{\sum_{l=1}^{j_{ik}} \Delta F_k(s_{ikl}; \mathbf{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}.$$

If $\xi_i = 0$, then

$$\hat{w}_{ikj} = E \left\{ N_{ki}(s_{ik,j-1}, s_{ikj}) \middle| \sum_{k'=1}^K N_{k'i}(L_i, R_i) = 1 \right\} = \frac{\Delta F_k(s_{ikj}; \mathbf{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}{\sum_{k'=1}^K \sum_{l=1}^{j_{ik'}} \Delta F_{k'}(s_{ik'l}; \mathbf{Z}_i, \boldsymbol{\beta}_{k'}, \Lambda_{k'})}.$$

If $R_i = \infty$, then $\hat{w}_{ikj} = 0$.

When maximizing (4), we update the parameters using a one-step self-consistency type formula. The first-order approximation to $F_k(t_{kj}; \mathbf{Z}, \boldsymbol{\beta}_k, \Lambda_k)$ is

$\tilde{G}_k(\sum_{j'=1}^j e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj'}} \lambda_{kj'}) e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj}} \lambda_{kj}$, where $\tilde{G}_k(x) = \exp \{-G_k(x)\} G'_k(x)$. Here and in the sequel, $f'(t) = df(t)/dt$ for any function f . Thus, the objective function in (4) can be approximated by

$$\sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^{m_k} \hat{w}_{ikj} \left\{ \log \lambda_{kj} + \boldsymbol{\beta}_k^T \mathbf{Z}_{ikj} + \log \tilde{G}_k \left(\sum_{j'=1}^j e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj'}} \lambda_{kj'} \right) \right\} + \sum_{i: R_i = \infty} \log S(L_i; \mathbf{Z}_i, \boldsymbol{\beta}, \Lambda),$$

(5)

where m_k is the total number of the t_{kj} . We set the derivative of (5) with respect to λ_{kj} to zero to obtain an updating formula for λ_{kj}

$$\lambda_{kj} = \left(\sum_{i=1}^n \hat{w}_{ikj} \right) \left[- \sum_{i=1}^n \sum_{j'=j}^{m_k} \hat{w}_{ikj'} e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj'}} \frac{\tilde{G}'_k}{\tilde{G}_k} \left(\sum_{j''=1}^{j'} e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj''}} \lambda_{kj''} \right) \right]^{-1} + \sum_{i: R_i = \infty, L_i \geq t_{kj}} S(L_i; \mathbf{Z}_i, \boldsymbol{\beta}, \Lambda)^{-1} \tilde{G}_k \left(\sum_{t_{kj'} \leq L_i} e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj'}} \lambda_{kj'} \right) e^{\boldsymbol{\beta}_k^T \mathbf{Z}_{ikj}}.$$

To update $\boldsymbol{\beta}$, we use a one-step Newton-Raphson algorithm based on (5).

We set the initial value of $\boldsymbol{\beta}$ to $\mathbf{0}$ and the initial value of λ_{kj} to n^{-1} . We iterate between the E- and M-steps until the sum of the absolute differences of the parameter estimates between two successive iterations is less than a small number. To increase the chance of reaching the global maximum, we suggest using a range of initial values for $\boldsymbol{\beta}$. The resulting estimators for $\boldsymbol{\beta}$ and Λ are denoted as $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}$, respectively.

The proposed algorithm has several desirable features. First, the conditional expectations in the E-step have simple analytic forms. Second, the M-step involves only a single analytic update for λ_{kj} and thus avoids large-scale optimization over high-dimensional parameters. Finally, the algorithm is applicable to any transformation model and time-varying covariates.

2.4 Asymptotic Properties

We show in the Web Appendix that, under mild regularity conditions, $\hat{\beta}$ and $\hat{\Lambda}$ are consistent, and $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically zero-mean normal with a covariance matrix that attains the semiparametric efficiency bound in the sense of Bickel et al. (1993). The estimator $\hat{\Lambda}$ is only $n^{1/3}$ -consistent, and its asymptotic distribution is unknown even in the univariate case (Huang and Wellner, 1997). Nonetheless, consistency alone allows one to predict the covariate-specific cumulative incidence.

Because $\hat{\Lambda}$ is not $\sqrt{n}$ -consistent, the variance of $\hat{\beta}$ cannot be estimated by the familiar Louis formula. Thus, we appeal to the profile likelihood method (Murphy and van der Vaart, 2000). The profile likelihood for β is defined as

$$pl_n(\beta) = \sup_{\Lambda \in \mathcal{C}} \log L_n(\beta, \Lambda),$$

where $\mathcal{C}$ is the space of Λ , in which Λ_k is a step function with non-negative jumps at t_{kj} . We propose to estimate the covariance matrix of $\hat{\beta}$ by the negative inverse of the matrix whose (j, k) th element is

$$\frac{pl_n(\hat{\beta}) - pl_n(\hat{\beta} + h_n \mathbf{e}_k) - pl_n(\hat{\beta} + h_n \mathbf{e}_j) + pl_n(\hat{\beta} + h_n \mathbf{e}_k + h_n \mathbf{e}_j)}{h_n^2}, \quad j, k = 1, \dots, p,$$

where $\mathbf{e}_j$ is the j th canonical vector in $\mathbb{R}^d$, h_n is a constant of the order $n^{-1/2}$, and p is the dimension of β . The consistency of this estimator follows from the profile likelihood theory of Murphy and van der Vaart (2000). To evaluate $pl_n(\beta)$, we adopt the EM algorithm described in Section 2.2 by using $\hat{\Lambda}$ as the initial value for Λ and holding β fixed in the iterations.

2.5 Reduced-Data Likelihood

We can estimate the parameters separately for each risk by using the reduced-data likelihood along the lines of Jewell et al. (2003) and Hudgens et al. (2014). Assume that there are no missing values on the causes of failure. In the reduced data, the subjects who experience the risk of interest and those who are right censored remain intact, whereas those who experience the other risks are treated as right censored at the last examination time U_J (Hudgens et al., 2014). For the k th risk, the reduced data consist of $\{J_i, U_i, \mathbb{I}(\tilde{D}_i = k), \mathbf{Z}_i\}$ ($i = 1, \dots, n$), and the corresponding likelihood for (β_k, Λ_k) is

$$\prod_{i=1}^n \left\{ \prod_{j=1}^{J_i} \left(\exp \left[-G_k \left\{ \int_0^{U_{i,j-1}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right. \right. \\ \left. \left. - \exp \left[-G_k \left\{ \int_0^{U_{ij}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right] \right)^{I(\tilde{D}_i=k)\Delta_{ij}} \right. \\ \left. \times \exp \left[-G_k \left\{ \int_0^{U_{i,J_i}} e^{\beta_k^T \mathbf{Z}_i(t)} d\Lambda_k(t) \right\} \right]^{I(\tilde{D}_i \neq k)} \right\},$$

which can be written as

$$\prod_{i:\tilde{D}_i=k} \left[\exp \left\{ -G_k \left(\sum_{t_{kj} \leq L_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} - \exp \left\{ -G_k \left(\sum_{t_{kj} \leq R_i} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\} \right] \\ \times \prod_{i:\tilde{D}_i \neq k} \exp \left\{ -G_k \left(\sum_{t_{kj} \leq U_{i,J_i}} \lambda_{kj} e^{\beta_k^T \mathbf{Z}_{ikj}} \right) \right\}. \quad (6)$$

We can maximize (6) by adapting the algorithm of Section 2.2. The resulting estimators are referred to as naive estimators.

With the exception of semiparametric efficiency, the asymptotic properties of the full-data NPMLEs carry over to the naive estimators by setting $K = 1$. The reason that efficiency is not attained by the naive estimators is because the reduced data do not contain all relevant information about the risk of interest. For example, the overall survival function, which combines the risk of interest with other risks in the full-data likelihood, is not included in the reduced-data likelihood.

In the reduced data for a particular risk, a subject who experiences a different risk is treated as right censored at the last potential examination time U_j . This examination time may be unknown if the available data consist only of (L_i, R_i) ($i = 1, \dots, n$). It might be tempting to replace U_{i,J_i} in (6) by R_i if the i th subject experiences a different risk and U_{i,J_i} is unavailable. However, the corresponding likelihood would involve parameters for other risks, such that the resulting naive estimators would be biased.

We have assumed that $\xi_i = 1$ for all $i = 1, \dots, n$. When there are missing values on the causes of failure, it is natural to consider only complete cases, i.e., subjects with non-missing values. (Note that right-censored observations are complete cases because their causes of failure are naturally unknown.) The complete-case analysis, however, is generally biased. Suppose, for instance, that missingness is completely random among subjects who are not right censored. Since shorter failure times are less likely to be right censored than longer failure times and thus are more likely to be associated with missing causes and discarded, the cumulative incidence will be underestimated. If the probability of right censoring depends on covariates, then the naive estimator for the regression parameter will also be biased.

3. Simulation Studies

We carried out simulation studies to assess the performance of the NPMLE and naive methods in realistic settings. We let $Z_1(t) = B_1 I(t \leq V) + B_2 I(t > V)$ and $Z_2 \sim \text{Unif}[0, 1]$, where B_1 and B_2 are independent Bernoulli(0.5), and V is $\text{Unif}[0, 3]$. In addition, we let U_1 and $U_2 - U_1$ be two independent random variables distributed as the minimum of 1.5 and an exponential random variable with hazard $e^{0.5Z_2 - 0.5}$. We considered $K = 2$ and $G_k(x) = r^{-1} \log(1 + rx)$ with $r = 0, 1$, and 0.5 . We set $\Lambda_1(t) = \Lambda_2(t) = 0.1(1 - e^{-t})$, $\beta_1 = (0.25, -0.25)^T$, and $\beta_2 = (-0.25, 0.25)^T$. We first assumed that the causes of failure are completely observed. Under these conditions, the event rate for each cause was roughly 15%. We generated 10,000 replicates for each scenario. Convergence criteria were met when the sum of the absolute differences of all the parameter estimates between two successive iterations fell below 10^{-4} . For both the NPMLE and the naive estimator, the algorithms converged in about 100 iterations. For variance estimation, we set $h_n = n^{-1/2}$. We also experimented with other values of h_n including $10n^{-1/2}$, and the results differed only at the third decimal place.

The results for $\beta_1 = (\beta_{11}, \beta_{12})^T$ are summarized in Table 1. For both the NPMLE and naive methods, the parameter estimators are virtually unbiased, and the standard error estimators reflect the true variability well. As a result, the empirical coverage probabilities of the confidence intervals are close to the nominal level. The NPMLE has smaller variance than the naive estimator. As shown in Figure 1, the NPMLE has little bias in estimating the cumulative hazard function, especially when $n = 500$.

Next, we simulated missing values on the causes of failure by assuming that the probability of $\xi = 0$ is 0.3 given that the subject is not right censored. The results for the estimation of β_1 by the NPMLE and the complete-case naive estimator are summarized in Table 2. The NPMLE continues to perform well. The naive estimator is substantially less efficient and is severely biased for β_{12} as a result of the dependence of the right censoring time on Z_2 .

4. An HIV Study

The Bangkok Metropolitan Administration (BMA) conducted a prospective study on a cohort of 1,209 initially HIV-seronegative injecting drug users (Hudgens et al., 2001). The study was designed to investigate risk factors for HIV incidence and develop better prevention strategies. The subjects were followed from 1995 to 1998 at 15 BMA drug treatment clinics. Blood tests were conducted on each participant approximately every 4 months post recruitment for evidence of HIV-1 seroconversion (i.e., detection of HIV-1 antibodies in the serum). By December 1998, there were 133 HIV-1 seroconversions and approximately 2,300 person-years of follow-up. Out of the 133 seroconversions, 27 were of viral subtype B, and 99 of subtype E. The subtypes for the remaining 7 were unknown but assumed to be either B or E.

We apply the proposed methods to the data derived from this study by treating the two HIV-1 subtypes as competing risks with partially missing information. We investigate the influence of potential risk factors on the cumulative incidence of HIV-1 seroconversion with time since recruitment as the principal time scale. The potential risk factors include age at

baseline (in years), gender (female vs male), whether the subject had a history of needle sharing (yes vs no), whether s/he had an imprisonment history before recruitment, and whether s/he had been imprisoned before recruitment and injected drug during imprisonment (yes vs no). We consider logarithmic transformation functions $G_1(x)=r_1^{-1}\log(1+r_1x)$ and $G_2(x)=r_2^{-1}\log(1+r_2x)$. The log-likelihood is maximized at $r_1 = 1.6$ and $r_2 = 0.2$, which is the combination of transformation functions that would be selected by the AIC criterion. Table 3 shows the results for this combination, as well as the results for $r_1 = r_2 = 0$ (proportional hazards) and $r_1 = r_2 = 1$ (proportional odds).

There are considerable differences in the parameter estimates between the NPMLE and naive methods; the latter tends to produce larger standard errors than the former. The effects of risk factors on the two competing risks are quite different. By the NPMLE, needle sharing significantly increases the incidence of seroconversion with HIV-1 subtype B, but its effect on subtype E is minimal. Younger age and imprisonment history significantly increase the incidence of seroconversion with subtype E; drug injection has a marginally significant effect on the incidence of subtype E. None of these risk factors, however, are significantly associated with the incidence of subtype B. To illustrate the joint inference by the NPMLE, we conduct a joint test for the effects of imprisonment history on the two viral subtypes. The χ^2_2 test statistic is 12.8, which is highly significant.

To illustrate prediction, we display in Figure 2 the NPMLE and naive estimates of the cumulative incidence function for a 32-year-old female with a history of needle sharing, drug injection, and imprisonment before recruitment. The cumulative incidence of seroconversion with HIV-1 subtype E is much higher than that of subtype B. The naive method yields noticeably different estimates than the NPMLE, especially for subtype E. The discrepancies are likely due to omission of failures with unknown causes, which are mostly assigned to subtype E by the NPMLE.

Table 4 shows the results from the approach of mid-point imputation, i.e., using the the mid-point of the interval to impute the failure time. The parameter estimates differ considerably from their NPMLE counterparts in Table 3, and the standard error estimates tend to be reduced. Thus, the significance levels for some of the risk factors are changed, indicating the importance of using statistically valid methods for analysis of interval-censored data.

5. Discussion

There is some literature on the NPMLE for interval-censored univariate failure time data; see Huang (1996), Huang and Wellner (1997), and Zeng et al. (2016). It is more difficult to establish the asymptotic properties of the NPMLE for interval-censored competing risks data with partially missing causes of failure. First, the multiplicity of failure types requires that the least favorable directions hold simultaneously for all nuisance tangent spaces. Second, as the probability of missing the cause of failure information may depend on the examination times and the covariates, the proof for the existence of a solution to the normal equations for the least favorable directions requires careful arguments for the smoothness of the score and information operators.

The computation of the NPMLE with interval-censored competing risks data also poses new challenges. The iterative convex minorant (ICM) algorithm (Huang and Wellner, 1997; Groeneboom and Jongbloed, 2014) is not applicable to time-varying covariates or incomplete cause-of-failure data because in such cases the nuisance parameters are entangled in the likelihood so that the diagonal approximation to the Hessian matrix becomes inaccurate. In addition, the ICM algorithm tends to be unstable for large datasets because it attempts to update a large number of parameters simultaneously using a quasi-Newton method (Wang et al., 2015). By sacrificing some computational speed, our EM algorithm updates each parameter using a self-consistency equation and is thus more reliable. We have not encountered any non-convergence in our extensive numerical studies. In the simulation studies, it took about 10 and 40 seconds to analyze a dataset with $n = 200$ and 500, respectively. We have included an R program in the Supplementary Materials.

In practice, it may be difficult to determine the causes of failure for all study subjects, especially when the ascertainment requires an extra (and possibly costly) step. In the BMA study, for instance, the HIV-1 subtype information required genotyping the viral DNA (Hudgens et al., 2001). The NPMLE approach enables one to make valid and efficient inference in the presence of missing information on the causes of failure.

We have studied both the NPMLE and naive estimators. The naive estimator may be preferable if the interest lies only in a subset of risks. However, the naive estimator is less satisfactory than the NPMLE for several reasons. First, the naive estimator is not statistically efficient. Second, it cannot properly handle unknown causes of failure. Finally, it does not provide simultaneous inference, which is often desirable because an increase in the incidence of one risk naturally reduces the incidence of other risks.

In independent work, Li (2016) proposed a spline-based method for the Fine and Gray (1999) model with interval-censored data. It is difficult to choose the number of knots and their locations. With right-censored data, one can use the quantiles of the observed failure times as the knots. This strategy is not applicable to interval-censored data due to the lack of exact observations. Specifically, because the examination times are not the actual event times, the cumulative baseline hazard functions may not jump at those locations. If there is no jump between two specified knots, the computation will be unstable. The NPMLE approach is advantageous in that it allows the jump points of the cumulative hazard functions to be determined in a data-adaptive manner and thus offers a fully automated solution.

In some applications, a subset of risks is interval censored while the rest is right censored. For example, in the Breastfeeding, Antiretrovirals, and Nutrition study, there were three competing risks: infant HIV infection, weaning, and infant death prior to infection or weaning (Hudgens et al. 2014). While the first two risks were interval censored, the time to death was known exactly or right censored. We plan to extend our work to this type of competing risks data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health awards R01GM047845, R01AI029168, R01CA082659, and P01CA142538. The authors thank the Editor, an Associate Editor, and a referee for their helpful comments.

References

- Bickel, PJ., Klaassen, CAJ., Ritov, Y., Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Baltimore: Johns Hopkins University Press; 1993.
- Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika*. 2002; 89:659–668.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82:835–845.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999; 94:496–509.
- Gray RJ. A class of K -sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*. 1988; 16:1141–1154.
- Groeneboom, P., Jongbloed, G. Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge: Cambridge University Press; 2014.
- Groeneboom P, Maathuis MH, Wellner JA. Current status data with competing risks: consistency and rates of convergence of the MLE. *The Annals of Statistics*. 2008a; 36:1031–1063.
- Groeneboom P, Maathuis MH, Wellner JA. Current status data with competing risks: limiting distribution of the MLE. *The Annals of Statistics*. 2008b; 36:1064–1089. [PubMed: 19888358]
- Huang J. Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*. 1996; 24:540–568.
- Huang, J., Wellner, JA. Interval censored survival data: a review of recent progress. In: Lin, DY., Fleming, TR., editors. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. New York: Springer-Verlag; 1997. p. 123–169.
- Hudgens MG, Satten GA, Longini IM. Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*. 2001; 57:74–80. [PubMed: 11252621]
- Hudgens MG, Li C, Fine JP. Parametric likelihood inference for interval censored competing risks data. *Biometrics*. 2014; 70:1–9. [PubMed: 24400873]
- Jewell NP, van der Laan M, Henneman T. Nonparametric estimation from current status data with competing risks. *Biometrika*. 2003; 90:183–197.
- Kalbfleisch, JD., Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2. Hoboken: John Wiley; 2002.
- Li C. The Fine-Gray model under interval censored competing risks data. *Journal of Multivariate Analysis*. 2016; 143:327–344. [PubMed: 26543275]
- Li C, Fine JP. Smoothed nonparametric estimation for current status competing risks data. *Biometrika*. 2013; 100:173–187.
- Lu W, Ying Z. On semiparametric transformation cure models. *Biometrika*. 2004; 91:331–343.
- Mao L, Lin DY. Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society, Series B*. 2016; doi: 10.1111/rssb.12177
- Martinussen, T., Scheike, TH. *Dynamic Regression Models for Survival Data*. New York: Springer-Verlag; 2006.

- Murphy SA, Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association*. 2000; 95:449–465.
- Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*. 1976:290–295.
- van der Vaart, AW., Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag; 1996.
- Wang L, McMahan CS, Hudgens MG, Qureshi ZP. A flexible, computationally efficient method for fitting the proportional hazards model to interval censored data. *Biometrics*. 2015; 72:22231.
- Zeng D, Mao L, Lin DY. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*. 2016; 103:253–271. [PubMed: 27279656]

Figure 1.

Estimation of the cumulative hazard function $\Lambda_1(\cdot)$ by the NPMLE. The true values and the mean estimates (based on 10,000 replicates) are shown by the solid and dashed curves, respectively.

Figure 2.

Estimated cumulative incidence of seroconversion in the BMA HIV-1 study for a 32-year-old woman with needle sharing, drug injection, and imprisonment before recruitment. The solid and dashed curves pertain to the NPMLE and naive methods, respectively.

Table 1Simulation results on the estimation of β_1 with complete data.

<i>n</i>	<i>r</i>	NPMLE				Naïve			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
200	0	β_{11}	0.004	0.298	0.303	0.961	0.003	0.312	0.318
		β_{12}	-0.003	0.663	0.666	0.955	0.001	0.691	0.693
	0.5	β_{11}	-0.010	0.302	0.307	0.955	-0.010	0.315	0.320
		β_{12}	0.012	0.670	0.675	0.958	0.010	0.699	0.700
500	1	β_{11}	-0.006	0.302	0.304	0.954	-0.009	0.314	0.320
		β_{12}	0.011	0.679	0.685	0.960	0.012	0.708	0.710
	0	β_{11}	0.002	0.187	0.188	0.950	0.011	0.201	0.205
		β_{12}	0.002	0.416	0.417	0.951	0.006	0.440	0.440
	0.5	β_{11}	-0.003	0.192	0.190	0.947	-0.010	0.197	0.197
		β_{12}	-0.003	0.426	0.430	0.953	0.005	0.438	0.439
	1	β_{11}	-0.002	0.194	0.196	0.952	0.000	0.199	0.200
		β_{12}	0.000	0.429	0.431	0.953	-0.002	0.449	0.452

Note: Bias and SE are the bias and standard error of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the variance of the naive estimator over that of the NPMLE. Each entry is based on 10,000 replicates.

Table 2

Simulation results on the estimation of β_1 with missing data.

<i>n</i>	<i>r</i>		NPMLE			Naive			CP	RE
			Bias	SE	SEE	Bias	SE	SEE		
200	0	β_{11}	0.010	0.319	0.322	-0.002	0.345	0.350	0.959	1.169
		β_{12}	0.002	0.703	0.705	0.122	0.767	0.771	0.930	1.190
	0.5	β_{11}	0.015	0.329	0.332	0.005	0.360	0.366	0.957	1.198
		β_{12}	-0.001	0.710	0.717	0.130	0.777	0.783	0.924	1.199
500	1	β_{11}	-0.002	0.329	0.332	-0.013	0.357	0.364	0.957	1.178
		β_{12}	0.007	0.721	0.724	0.122	0.781	0.787	0.921	1.174
	0	β_{11}	-0.003	0.198	0.202	0.002	0.216	0.217	0.953	1.187
		β_{12}	-0.005	0.444	0.443	0.131	0.489	0.490	0.821	1.213
	0.5	β_{11}	0.000	0.205	0.203	-0.004	0.224	0.226	0.952	1.198
		β_{12}	0.000	0.446	0.450	0.138	0.488	0.488	0.813	1.195
	1	β_{11}	0.005	0.206	0.206	-0.005	0.228	0.229	0.952	1.231
		β_{12}	0.000	0.453	0.451	0.119	0.498	0.499	0.834	1.206

Note: See the Note to Table 1.

Table 3

Analysis of the BMA HIV-1 study.

	NPMLE						Naive					
	Subtype B			Subtype E			Subtype B			Subtype E		
	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
Proportional hazards												
Age	-0.027	0.182	0.883	-0.268	0.099	0.007	0.009	0.199	0.963	-0.285	0.097	0.003
Gender	-0.116	0.444	0.793	0.618	0.277	0.026	0.020	0.551	0.970	0.628	0.377	0.096
Needle	1.038	0.240	<0.001	-0.031	0.196	0.875	1.033	0.301	0.001	-0.010	0.265	0.969
Drug	0.098	0.139	0.479	0.374	0.194	0.054	0.018	0.186	0.924	0.334	0.198	0.091
Prison	-0.557	0.424	0.189	0.727	0.216	0.001	-0.599	0.500	0.231	0.737	0.227	0.001
Proportional odds												
Age	0.085	0.184	0.646	-0.305	0.104	0.003	-0.010	0.194	0.959	-0.272	0.132	0.039
Gender	-0.143	0.454	0.752	0.677	0.355	0.057	0.022	0.531	0.967	0.729	0.431	0.091
Needle	0.975	0.238	<0.001	-0.064	0.222	0.775	1.071	0.282	<0.001	-0.055	0.277	0.841
Drug	0.118	0.113	0.300	0.425	0.194	0.028	-0.019	0.203	0.926	0.334	0.199	0.093
Prison	-0.588	0.450	0.191	0.757	0.193	<0.001	-0.513	0.543	0.344	0.717	0.244	0.003
Selected model												
Age	0.085	0.182	0.639	-0.264	0.090	0.003	-0.053	0.203	0.794	-0.256	0.119	0.032
Gender	-0.107	0.507	0.832	0.521	0.350	0.136	0.018	0.585	0.976	0.591	0.430	0.169
Needle	0.998	0.235	<0.001	-0.075	0.244	0.760	0.996	0.274	<0.001	-0.025	0.281	0.930
Drug	0.123	0.133	0.356	0.362	0.195	0.064	0.030	0.190	0.876	0.325	0.240	0.176
Prison	-0.552	0.421	0.190	0.769	0.188	<0.001	-0.508	0.520	0.329	0.716	0.241	0.003

Note: Est and SE denote the parameter estimate and (estimated) standard error.

Table 4

Analysis of the BMA HIV-1 study using mid-point imputation.

	Subtype B			Subtype E		
	Est	SE	p-value	Est	SE	p-value
Proportional hazards						
Age	0.020	0.155	0.899	-0.213	0.083	0.010
Gender	-0.216	0.453	0.633	0.414	0.320	0.196
Needle	1.140	0.214	<0.001	-0.002	0.216	0.994
Drug	-0.010	0.122	0.933	0.474	0.169	0.005
Prison	-0.524	0.354	0.138	0.613	0.171	<0.001
Proportional odds						
Age	0.052	0.165	0.752	-0.202	0.076	0.008
Gender	-0.304	0.429	0.479	0.341	0.314	0.278
Needle	1.008	0.202	<0.001	0.014	0.213	0.949
Drug	-0.056	0.120	0.643	0.490	0.173	0.005
Prison	-0.492	0.353	0.163	0.787	0.165	<0.001
Selected model						
Age	0.046	0.160	0.776	-0.166	0.079	0.036
Gender	-0.233	0.447	0.602	0.413	0.308	0.180
Needle	1.140	0.207	<0.001	0.053	0.215	0.803
Drug	-0.015	0.117	0.898	0.450	0.172	0.009
Prison	-0.474	0.371	0.201	0.696	0.166	<0.001

Note: See the Note to Table 3.