Published in final edited form as:
Biometrics. 2014 March ; 70(1): 1–9. doi:10.1111/biom.12109.

Parametric likelihood inference for interval censored competing risks data

Michael G. Hudgens, **Chenxi Li**, and **Jason P. Fine**
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, U.S.A

Jason P. Fine: jfine@bios.unc.edu

Summary

Parametric estimation of the cumulative incidence function (CIF) is considered for competing risks data subject to interval censoring. Existing parametric models of the CIF for right censored competing risks data are adapted to the general case of interval censoring. Maximum likelihood estimators for the CIF are considered under the assumed models, extending earlier work on nonparametric estimation. A simple naive likelihood estimator is also considered that utilizes only part of the observed data. The naive estimator enables separate estimation of models for each cause, unlike full maximum likelihood in which all models are fit simultaneously. The naive likelihood is shown to be valid under mixed case interval censoring, but not under an independent inspection process model, in contrast with full maximum likelihood which is valid under both interval censoring models. In simulations, the naive estimator is shown to perform well and yield comparable efficiency to the full likelihood estimator in some settings. The methods are applied to data from a large, recent randomized clinical trial for the prevention of mother-to-child transmission of HIV.

Keywords

Competing risks; Cumulative incidence function; Gompertz; HIV/AIDS; Maximum likelihood

1. Introduction

Breastfeeding accounts for up to half of infant HIV infections worldwide. Unfortunately, in resource poor settings where the burden of HIV infection is highest, non-breastfed babies face significantly increased morbidity and mortality from early childhood diseases like malnutrition and diarrhea due to alternative feeding methods, poor sanitation or lack of clean fresh water. This presents a dilemma for HIV positive mothers which has motivated studies of the prevention of mother-to-child transmission (PMTCT) of HIV through breast milk. In PMTCT studies, an infant can experience one of three events during the breastfeeding period: (i) HIV infection, (ii) weaning prior to HIV infection, or (iii) death prior to HIV

Correspondence to: Jason P. Fine, jfine@bios.unc.edu.

Supplementary Materials
The Web Appendices and Web Figures referenced in Sections 3, 5, 7, and 8 are available with this paper at the Biometrics website on Wiley Online Library.

infection or weaning. Typically the event times, especially the time of HIV infection, are not directly observed but only known up to some interval. PMTCT studies therefore give rise to interval censored competing risks data.

In the competing risks setting it is often of interest to estimate the probability of a particular event occurring by some time t as given by the cumulative incidence function (CIF). The CIF and the cause specific hazard function (CSHF) are basic identifiable quantities in the competing risks framework. In many settings the CIF may be preferred to the CSHF because the CIF has a simple interpretation as the cumulative risk of a specific event in the presence of competing risks, as opposed to the instantaneous rate of the event.

Nonparametric statistical methods have been studied for estimating the CIFs under interval censoring, with rigorous theory having been established for current status data with a single monitoring time. Hudgens et al. (2001) derived the nonparametric maximum likelihood estimator (NPMLE) of the CIFs for competing risks data subject to interval censoring. Jewell et al. (2003) studied the NPMLE of the CIF for current status data; they also introduced a naive estimator for current status data which only uses a subset of the observed data. Groeneboom et al. (2008b) derived the limiting distributions for the NPMLE and naive estimator of the CIF for current status data. Unfortunately nonparametric estimation has the disadvantage of being computationally intense, is difficult to implement using standard software, and may perform poorly in small samples owing to slow rates of convergence (Groeneboom et al., 2008a). Consequently, parametric models are attractive in this setting. When the model is correct, parametric estimation is usually more efficient than nonparametric estimation and permits extrapolation of long-term event probabilities. However, estimation of parametric models for the CIF for general interval censored competing risks data has not been investigated to date.

Jeong and Fine (2006) proposed parametric modeling of the CIF for right censored competing risks data. In this paper we extend the Jeong-Fine models to the general case of interval censored competing risks data. Both maximum likelihood estimators (MLEs) and a naive estimator are considered. The naive estimator enables separate estimation of models for each cause, unlike the MLEs where all models are fit simultaneously. This eases the computational burden, with standard software available for inference, and does not require correct specification of models for the competing causes. However, unlike the full likelihood, the validity of the naive likelihood is shown to depend on the particular interval censoring model assumed. These results have important practical implications for the use of the naive likelihood.

2. Competing risks model specification

Let the random variable $K \in \{1, 2, \ldots, n_K\}$ denote the cause of failure for an individual who can only experience one of n_K mutually exclusive competing causes. Let the non-negative random variable T denote the time of failure, which may be only known up to some interval. The CIF for events of type k is $F_k(t) = \Pr[T \leq t, K = k]$, i.e., the probability of experiencing an event of type k by time t in the presence of competing causes of failure. It is well known

that $F_k(t)=\int_0^t S(u)\lambda_k(u)du$ where $S(t)=\exp\{-\int_0^t \sum_{k=1}^{n_K}\lambda_k(u)du\}$ is the all cause survival probability and $\lambda_k(t) = \lim_{dt\to 0}\{\Pr(t \le T < t + dt, K = k|t \le T)/dt\}$ is the type k CSHF.

There are different ways to parametrically model the CIF. With right censored data the standard approach is by indirect parameterization via the CSHF (Prentice et al., 1978). Because of the form of the likelihood with right censored data, indirect modeling of CIF greatly simplifies estimation. Such simplification does not occur with interval censoring, in which case direct modeling of CIFs may be preferable as the likelihood can be more easily expressed using the CIFs (Section 3.1 below). The direct modeling approach (Jeong and Fine 2006) is appealing when the CIF is of primary interest because the assumed model has a natural interpretation in terms of the probability of an event of interest. In this case a separate parametric model $F_k(t; \Theta_k)$ is specified for each CIF such that Θ_k is distinct from Θ_j for all $j \neq k$. Assuming $n_K > 1$ and each cause occurs with non-zero probability, the CIF is an improper distribution function, i.e., $\lim_{t\to\infty} F_k(t) < 1$. Thus it is natural to model the CIF using cure-type models whereby the cure probability equals the probability of never having the event of interest. For example, in PMTCT studies there is interest in the probability of an infant never becoming HIV infected through breast milk.

Different cure-type models may be used to model the CIFs. For right censored data, Jeong and Fine (2006) considered the Gompertz model

$$F_k(t;\Theta_k)=1-\exp\left[\beta_k\left\{1-\exp(\alpha_k t)\right\}/\alpha_k\right] \quad (1)$$

with $\Theta_k = (\alpha_k, \beta_k)$ where $\beta_k > 0$ and $\alpha_k < 0$ ensure (1) is an improper distribution function. In this case, the probability of never having an event of type k equals $\lim_{t\to\infty}\{1-F_k(t; \Theta_k)\} = \exp(\beta_k/\alpha_k)$. Note if (1) holds for all k, the marginal distribution of T does not follow a Gompertz distribution; moreover $\Pr[T \le t;\Theta_1,\ldots,\Theta_{n_K}]=n_k-\sum_{k=1}^{n_K}\exp\left[\beta_k\left\{1-\exp(\alpha_k t)\right\}/\alpha_k\right]$ which does not reduce to a simple parametric form. An alternative parametric modeling approach entails letting $F_k(t; \Theta_k, \pi_k) = \pi_k \Pr(T \le t|K = k; \Theta_k)$ where $\pi_k = \Pr[K = k]$ and assuming $\Pr[T \le t|K = k; \Theta_k]$ follows a particular parametric distribution such as Weibull. However, unless the conditional distribution of T given $K = k$ follows a one-parameter model (e.g., exponential), this approach will generally be less parsimonious than (1).

In the sequel, regardless of parameterization, the models are assumed to satisfy

$$0<F_k(t;\Theta_k)<1 \text{ for all } t>0 \text{ and } k=1,\ldots,n_K; \quad (2)$$

$$F_k(t;\Theta_k) \text{ is monotonically increasing function of } t \text{ for } k=1,\ldots,n_K. \quad (3)$$

Additional constraints, such as assuming all individuals must eventually experience one of the n_K competing risks, i.e., $\lim_{t\to\infty}\sum_{k=1}^{n_K}F_k(t;\Theta_k)=1$, will be considered in Section 3.3.

3. Full likelihood estimation

3.1 Full likelihood

Interval censored data arise when subjects are inspected intermittently and the actual failure time is only known to lie between successive observation times. Formulations of the likelihood function for interval censored data without competing risks often assume the observation process is determined independently of the failure time, e.g., as in the "mixed case" interval censoring (Schick and Yu, 2000) model where each participant may have a different number of follow-up visits. Alternatively, Gruger, Kay, and Schumacher (1991) and Lawless (2003, §2.3.1) consider an independent inspection process (IIP) model for interval censored data which allows future observation times to possibly depend on the history of the observed data. The development below considers both mixed case and IIP models.

Let $V = (V_1, \ldots, V_M)$ denote the vector of ordered observations times where M is the random number of observation times for an individual. Let $V_0 = 0$ and $V_{M+1} = \infty$ such that $V_{l-1} < V_l$ for $l = 1, \ldots, M + 1$. Define $\Delta_{kl} = 1(V_{l-1} < T \leq V_l, K = k)$ for $k = 1, \ldots, n_K$ and $l = 1, \ldots, M$ where $1(\cdot)$ is the usual indicator function. In other words, Δ_{kl} equals 1 if an individual has an event of type k during the interval $(V_{l-1}, V_l]$, and 0 otherwise. Let $\Delta_{M+1} = 1 - \sum_{k=1}^{n_K} \sum_{l=1}^{M} \Delta_{kl}$. The event type is assumed to be unknown for right censored observations, i.e., when $\Delta_{M+1} = 1$. Assume we are not able to observe (T, K) directly, but we do observe copies of $Y = (M, V, \Delta)$ where $\Delta = (\Delta_{11}, \ldots, \Delta_{1M}, \Delta_{21}, \ldots, \Delta_{n_K M}, \Delta M_{+1})$.

Under the mixed case interval censoring model, the observation process is assumed to be independent of the time and cause of failure, i.e.,

$$(M, V) \perp (T, K). \quad (4)$$

This model might hold in studies of relatively healthy individuals where the competing events do not include death. For example, in longitudinal studies of women at risk for infection with various types of human papillomavirus (HPV), investigators are often interested in time until HPV infection, which is typically interval censored. Because the rate of subsequent clearance of infection and/or progression towards early stages of cervical cancer is also of interest, follow-up might continue on a regular schedule regardless of whether a woman becomes HPV infected, such that the mixed case model may be appropriate.

On the other hand, the mixed case model may be unreasonable in other settings. For example, in the PMTCT setting if an infant tests HIV positive at a particular visit, then typically no HIV testing is conducted at subsequent planned study visits. Or if an infant dies, then necessarily there will be no further study visits. In this case, the IIP model may be more applicable. Following Lawless (2003), for $l = 1, 2, \ldots$ define the history of observation times and failure information by $H_l = (V_0, V_1, \ldots, V_{l-1}, \Delta_{11}, \ldots, \Delta_{n_K 1}, \ldots, \Delta_{1,l-1}, \ldots, \Delta_{n_K l-1})$ where $H_1 = V_0 = 0$. Under the IIP model it is assumed

$$V_l \perp (T, K) | H_l, \quad (5)$$

i.e., the next observation time is independent of the failure time and cause given the history of observation times and failure information. As in Lawless (2003, page 65) in (5) it is implied that H_l includes information that the individual is alive and uncensored at V_{l-1}, i.e., (5) holds for $\Delta_{11} = \cdots = \Delta_{n_K 1} = \cdots = \Delta_{1,l-1} = \cdots = \Delta_{n_K l-1} = 0$. Assume that the IIP stops if a failure is detected, such that $\Delta_{jl} = 0$ for all $l < M$ and $j \in \{1, \ldots, n_K\}$.

Following the discussion in Section 2, inference about the CIF may be based on assuming parametric models for the CSHFs or by directly specifying parametric models for the CIF. In either case, for $k = 1, \ldots, n_K$ let $F_k(t; \Theta_k)$ denote the CIF for type k failure under the assumed model. Let $Y_1, \ldots, Y_n$ be a random sample of n independent and identically distributed copies of Y. The log likelihood function under either the mixed case or IIP model is given by the following lemma.

Lemma 1—Under the mixed case interval censoring model (4) or the IIP model (5), the log likelihood for $Y_1, \ldots, Y_n$ equals, up to a constant,

$$\log L(\Theta) = \sum_{i=1}^{n} \log \ell(Y_i; \Theta) \quad (6)$$

where Θ is the vector consisting of elements of $\Theta_1 \cup \cdots \cup \Theta_{n_K}$ and $\ell(Y; \Theta)$ is the likelihood contribution for a single observation, which equals

$$\ell(Y;\Theta) = \prod_{k=1}^{n_K} \prod_{l=1}^{M} \{F_k(V_l;\Theta_k) - F_k(V_{l-1};\Theta_k)\}^{\Delta_{kl}} \left\{1 - \sum_{k=1}^{n_K} F_k(V_M;\Theta_k)\right\}^{\Delta_{M+1}}.$$

Proofs of all lemmas are given in Web Appendix A. In the following sections we consider maximizing (6) under different constraints on Θ.

3.2 Unconstrained estimation

Define the unconstrained full likelihood (UFL) estimator $\hat{\Theta}$ to be the value of Θ which maximizes (6) under assumptions (2) – (3). Under suitable regularity conditions, $\sqrt{n}\{F_k(t;\hat{\Theta}_k) - F_k(t;\Theta_k)\}$ is asymptotically Normal with mean 0 and variance which may be estimated by

$$\widehat{var}\left\{F_k(t;\hat{\Theta}_k)\right\} = \left(\frac{\partial F_k(t;\Theta_k)}{\partial \Theta_k}\right) \widehat{\Sigma}_{\hat{\Theta}} \left(\frac{\partial F_k(t;\Theta_k)}{\partial \Theta_k}\right)' \Big|_{\Theta=\hat{\Theta}}$$

where $\hat{\Sigma}_{\hat{\Theta}}$ equals the inverse of the observed Fisher information, i.e., the matrix of the negative second derivatives of (6) with respect to Θ. Web Appendix B provides details for

model (1). A pointwise $(1 - \alpha)$ confidence interval (CI) for $F_k(t; \Theta_k)$ is $F_k(t;\hat{\Theta}_k) \pm z_{1-\alpha/2}\sqrt{\widehat{var}\left\{F_k(t;\hat{\Theta}_k)\right\}}$ where z_q is the q quantile of the standard Normal distribution.

The UFL estimator is unconstrained in that there are no restrictions on the parameters aside from those imposed by (2) – (3). Consequently, the UFL estimator has the property that the resulting estimator of the distribution of T, i.e., $\Pr[T \leq t;\hat{\Theta}]=\sum_{k=1}^{n_K} F_k(t;\hat{\Theta}_k)$ may be greater than one. Although, because $\hat{\Theta}$ maximizes (6), it follows $\sum_{k=1}^{n_K} F_k(t;\hat{\Theta}_k)<1$ for t less than or equal to the largest observation time V_M among right censored individuals, as otherwise evaluating (6) at $\hat{\Theta}$ would entail taking the log of a non-positive number. Thus, as with right censored competing risks data, with interval censored competing risks data, unconstrained full likelihood estimation only requires that the parametric models hold on the support of the observation time distribution and does not require modelling assumptions beyond the upper bound of this support.

3.3 Constrained estimation

In certain settings it may be known that all individuals must eventually experience one of the n_K competing risks such that

$$\lim_{t\to\infty}\sum_{k=1}^{n_K} F_k(t)=1. \quad (7)$$

This will be the case, for example, in studies where death is one of the competing risks. One may impose constraint (7) on the assumed parametric models. Define the constrained full likelihood (CFL) estimator to be the value of Θ which maximizes (6) under assumptions (2) – (3) subject to the equality constraint (7). Inference about Θ then follows from standard constrained maximum likelihood theory. In some cases, (7) can be enforced by solving for one parameter explicitly in terms of the other parameters, thereby reducing the number of model parameters by one; e.g., for model (1) when $n_K = 2$, let $\beta_2 = \alpha_2 \log\{1 - \exp(\beta_1/\alpha_1)\}$. In these cases the CFL estimator can be found by unconstrained maximum likelihood based of the reduced model.

In practice, it may not be known a priori whether the equality constraint (7) holds, particularly because there may not be information from the observable data about the tail of the distribution of T, e.g., when the maximum observation time V_M is bounded. For instance, in a cohort study of sexually active women at risk for infection with different HPV types, many women might not acquire HPV during the study yet may go on to acquire HPV subsequently. Moreover, in certain populations (e.g., commerical sex workers), whether all women will eventually contract HPV may not be known. If constraint (7) is not known to hold a priori, then the MLE for the full model could potentially be computed with respect to the inequality constraint

$$\lim_{t\to\infty}\sum_{k=1}^{n_K}F_k(t)\leq 1. \quad (8)$$

When the true model is off the boundary (i.e., (7) does not hold), then constrained MLEs obtained assuming (8) are asymptotically equivalent to the UFL and standard large sample results (as in Section 3.2) apply. However, when the true model is on the boundary (i.e., (7) holds), standard asymptotic results may not apply; e.g., the constrained MLEs assuming (8) will not in general have a Normal distribution asymptotically. Therefore, in situations where it is not known whether the true model lies on the boundary, we propose using the UFL estimator, which is as efficient as the constrained MLEs assuming (8) when the true model is off the boundary, but avoids inferential complexities when the true model is on the boundary. Simulations in Section 5 demonstrate that the UFL can have similar efficiency to the CFL when the true model is on the boundary.

Note the CFL estimator relies on modelling assumptions beyond the support of the distribution of observation times. Such assumptions cannot be checked with the observed data and may not be satisfied if the assumed models are misspecified beyond the support of the observation time distribution. Similar issues arise with right censored data, where parametric models cannot be checked beyond the support of the right censoring time distribution.

3.4 Partly interval censored data

In some instances events of certain types may be observed exactly whereas events of other types may be interval censored. For example, in the PMTCT study discussed below, failure times associated with two event types (HIV infection and weaning) are subject to interval censoring whereas failure times associated with the other event type (death) are observed exactly. In general, suppose for a subset $\mathcal{S}\subset\{1,\ldots,n_K\}$ of event types the corresponding times are observed exactly if $T\leq V_M$ and right censored otherwise with right censoring time V_M; otherwise if $K\notin\mathcal{S}$ assume T is subject to interval censoring. In this case the observed data are copies of $(M, V, \Delta, T1(T\leq V_M, K\in\mathcal{S}))$. Under either the mixed case or IIP models, the full likelihood contribution for a single observation is

$$\prod_{l=1}^{M}\left[\prod_{k\notin\mathcal{S}}\{F_k(V_l;\Theta_k)-F_k(V_{l-1};\Theta_k)\}^{\Delta_{kl}}\prod_{k\in\mathcal{S}}f_k(T;\Theta_k)^{\Delta_{kl}}\right]\left\{1-\sum_{k=1}^{n_K}F_k(V_M;\Theta_k)\right\}^{\Delta_{M+1}}, \quad (9)$$

where $f_k(t;\Theta_k)=\partial F_k(t;\Theta_k)/\partial t$ is the sub-density function for a type k failure. Here we assume $F_k(t;\Theta_k)$ is continuous at t; otherwise $f_k(t;\Theta_k)=F_k(t;\Theta_k)-F_k(t^-;\Theta_k)$ assuming in general $F_k(\cdot;\Theta_k)$ is right-continuous with left limits.

4. Naive likelihood estimation

Jewell et al. (2003) proposed a simple naive estimator for the CIF given current status competing risks data. The non-parametric naive estimator of the CIF is estimated separately for each failure type based on a reduced version of the observable data and has been shown

empirically to perform well relative to the full likelihood NPMLE. Jewell et al. (2003) also gave a brief discussion of using simple parametric models to estimate the CIF with current status data. Here a naive parametric estimator of the CIF is considered in the general case of interval censored competing risks data with a random number of observation times.

For $k \in \{1, \ldots, n_K\}$ let $\Delta_{k,M+1} = 1 - \sum_{l=1}^{M} \Delta_{kl}$ and $\Delta^k = (\Delta_{k1}, \ldots, \Delta_{kM}, \Delta_{k,M+1})$ denote the vector of indicator variables corresponding to cause k only. Let $Z_k = (M, V, \Delta^k)$ denote the observable random variables related to failure from cause k, with information about other causes of failure being ignored. The naive estimator defined below utilizes only the reduced data Z_k, essentially treating failures from other causes as right censored observations. Let $Z_{1k}, \ldots, Z_{nk}$ be a random sample of n independent and identically distributed copies of Z_k.

Lemma 2—Under mixed case interval censoring model (4), the log likelihood function for $Z_{1k}, \ldots, Z_{nk}$ equals, up to a constant,

$$\log L_k(\Theta_k) = \sum_{i=1}^{n} \log \ell_k(Z_{ik}; \Theta_k) \quad (10)$$

where

$$\ell_k(Z_k; \Theta_k) = \prod_{l=1}^{M} \{F_k(V_l; \Theta_k) - F_k(V_{l-1}; \Theta_k)\}^{\Delta_{kl}} \{1 - F_k(V_M; \Theta_k)\}^{\Delta_{k,M+1}}. \quad (11)$$

According to Lemma 2, under the mixed case model the naive likelihood has the same form as the usual likelihood for interval censored data in the absence of competing risks. Lemma 3 indicates that the naive log likelihood (10) is not valid under the IIP model.

Lemma 3—Under the IIP model (5), the log likelihood function for $Z_{1k}, \ldots, Z_{nk}$ equals, up to a constant,

$$\log L_k^{IIP}(\Theta_1, \ldots, \Theta_{n_K}) = \sum_{i=1}^{n} \log \ell_k^{IIP}(Z_{ik}; \Theta_1, \ldots, \Theta_{n_K}) \quad (12)$$

where

$$\ell_k^{IIP}(Z_k; \Theta_1, \ldots, \Theta_{n_K}) = \prod_{l=1}^{M} \{F_k(V_l; \Theta_k) - F_k(V_{l-1}; \Theta_k)\}^{\Delta_{kl}} \times \left\{1 - \sum_{j=1}^{n_K} F_j(V_{M-1}; \Theta_j) - F_k(V_M; \Theta_k) + F_k(V_{M-1}; \Theta_k)\right\}^{\Delta_{k,M+1}}.$$

Unlike the mixed case model, under the IIP model the naive log likelihood (12) includes parameters Θ_j for $j \neq k$. Thus the naive approach does not afford a simpler likelihood than the full likelihood under the IIP model. Therefore in the sequel only the naive likelihood under the mixed case model is considered.

For mixed case interval censoring, define the naive estimator $\tilde{\Theta}_k$ to be the value of Θ_k which maximizes (10) assuming $0 < F_k(t; \Theta_k) < 1$ for all t and $F_k(t; \Theta_k)$ is monotonically increasing in t. Under suitable regularity conditions $\sqrt{n}\left\{F_k(t;\tilde{\Theta}_k)-F_k(t;\Theta_k)\right\}$ is asymptotically Normal and pointwise $(1 - \alpha)$ CIs for $F_k(t; \Theta_k)$ can be computed as in Section 3.2. Like the UFL estimator, the naive estimators are unconstrained such that $\Pr[T \leq t;\tilde{\Theta}_1, \ldots, \tilde{\Theta}_{n_K}]=\sum_{k=1}^{n_K} F_k(t;\tilde{\Theta}_k)$ may be greater than 1. As in Section 3.4, under the mixed case model (11) can be generalized to allow for certain events to be observed exactly. For cause $k \in \mathcal{S}$, the set of event types where the corresponding times are observed exactly, the likelihood contribution for a single observation equals

$$\prod_{l=1}^{M} f_k(T;\Theta_k)^{\Delta_{kl}}\{1-F_k(V_M;\Theta_k)\}^{\Delta_{k,M+1}};$$

for cause $k \notin \mathcal{S}$, the naive likelihood contribution equals (11) as before.

5. Simulation study

Simulation studies were conducted to compare the UFL, CFL, and naive estimators under several scenarios. For all scenarios there were $n_K = 2$ causes of failure. In the first scenario failure time and type were simulated according to (1) with the parameters $\Theta = (\alpha_1, \alpha_2, \beta_1, \beta_2)$ chosen to satisfy the equality constraint (7) and observation times were simulated according to a mixed case interval censoring model. Event type and time were simulated utilizing the factorization $F_k(t; \Theta_k) = \Pr(T \leq t|K = k; \Theta_k) \Pr(K = k; \Theta_k)$, where the cause of failure K was first randomly generated from a multinomial distribution with cell probabilities $1 - \exp(\beta_k/\alpha_k)$ for $k = 1, 2$, and the failure time T was then simulated based on the conditional distribution of T given $K = k$ using the inverse probability transformation. The observation times $V_1 < \cdots < V_M$ were independently generated to mimic the PMTCT study described in Section 7. Specifically, study visits (i.e., observation times) were randomly generated to occur approximately every 4 weeks up to week 28, for a maximum of 7 study visits, where the observation times were uniformly distributed from week 3 to week 5, week 7 to week 9, and so on. For each scheduled visit, an individual missed the visit with probability 0.1, so the number of actual study visits M was often less than 7. Data sets were simulated for various sample sizes. For each simulated data set, the UFL, CFL, and naive estimators were computed. For comparison, NPMLEs of the CIFs were also computed, based on a full likelihood analogous to (6) and also a naive likelihood analogous to (10).

Results based on 5000 simulated data sets per sample size of $n = 500$, 1000, and 2000 are given in Tables 1 and 2. In terms of the parameters α_k and β_k, the UFL, CFL, and naive estimators were approximately unbiased, the model based variance estimates using the observed information were similar to the empirical variances of the estimators, and the corresponding 95% CIs exhibited approximately the correct coverage probability (Table 1). Similar results were obtained for the CIF estimators (Table 2, results for $n = 1000$ not shown). In comparison to the non-parametric estimators, the UFL, CFL, and naive

estimators exhibited smaller bias and variance which is not surprising given the CIF and observation process models were both correctly specified. Web Appendix C includes additional simulations conducted under alternative scenarios investigating the effect of model misspecification. In general these results suggest that the UFL, CFL, and naive estimators are not particularly robust to severe violations of the parametric assumptions of the CIF model such that assessment of model fit should be considered when using these estimators in practice.

6. Goodness-of-fit

The fit of a particular model for $F_k(t)$ can be assessed by comparing parametric estimates (e.g., $F_k(t; \hat{\Theta}_k)$) with non-parametric estimates such as the NPMLE that do not rely on any modeling assumptions. Formally deriving the properties of a goodness-of-fit statistic that compares nonparametric and parametric estimates of the CIF is challenging because under a continuous time model nonparametric estimators converge at a rate slower than $\sqrt{n}$ to nonstandard distributions (Groeneboom et al. 2008a, 2008b). On the other hand, the parametric and non-parametric estimates can be compared graphically as an informal assessment of fit since both estimators are consistent. An alternative approach to assessing fit is to consider a more general parametric model which includes as a special case the parametric model under consideration. For instance, a simple three parameter generalization of (1) is

$$F_k(t;\Theta_k)=1-\exp\left[\beta_k\{1-\exp(\alpha_k t^{\eta_k})\}/\alpha_k\right], \quad (13)$$

where here $\Theta_k = (\alpha_k, \beta_k, \eta_k)$ with the constraints $\eta_k > 0$, $\beta_k > 0$, and $\alpha_k < 0$ ensuring an improper distribution function. Under (13), $\lim_{t\to\infty} F_k(t; \Theta_k) = 1 - \exp(\beta_k/\alpha_k)$ as in the two parameter model. Because (13) reduces to (1) for $\eta_k = 1$, when using UFL or naive likelihood, a Wald, score or likelihood ratio test of $H_0 : \eta_k = 1$ provides a one degree of freedom goodness-of-fit test of (1). Haile (2008) also proposed a three parameter Gompertz model, although the form of $F_k(t; \Theta_k)$ under Haile's model is more complicated than (13).

7. The breastfeeding, antiretrovirals, and nutrition (BAN) study

The BAN study was a large randomized intervention trial of 2369 HIV-infected women and their infants conducted in Malawi (Chasela et al., 2010). The specific aims of the study included evaluating (i) the benefit and safety of antiretroviral (ARV) prophylaxis given either to infants or to their mothers for PMTCT of HIV during breastfeeding, and (ii) the feasibility of exclusive breastfeeding followed by early, rapid breastfeeding cessation. Eligible mother-infant pairs were randomized into one of three ARV arms: maternal ARV, infant ARV, or control. Blood for HIV testing was scheduled to be drawn from infants at birth and weeks 1, 2, 4, 6, 8, 12, 18, 24, 28. The actual timing of study visits often deviated from the scheduled times with some visits missed completely and some infants dropping out of the study.

One primary endpoint of BAN was infant HIV infection by week 28. Here there are $n_K = 3$ competing risks: HIV transmission ($k = 1$), death of an HIV-free breastfeeding infant ($k = 2$),

weaning prior to HIV infection ($k = 3$). By 28 weeks there were 184 HIV-infected infants and 29 HIV-uninfected deaths. The analysis below excludes 119 infants that were HIV infected in the first two weeks as these infections were likely due to transmission in utero or during delivery and thus were not of primary interest to investigators. Seven infants that died in the first two weeks were also excluded as breast milk transmission of HIV was unlikely for these infants. Finally 180 infants with no data on breastfeeding were excluded, yielding $n = 2063$.

Figure 1 depicts the NMPLE and UFL estimates of the CIF for the two (1) and three (13) parameter Gompertz models. The UFL estimates were computed using (9) with $\mathcal{S} = \{2\}$ since death times were known exactly. On the other hand, HIV infection times were interval censored, known only to be between the last visit where the infant tested HIV negative and the first visit where the infant tested positive. Likewise, weaning times were only known to be between the last visit where the mother reported breastfeeding and the first visit where she reported the infant had been weaned. In comparison to the NPMLE, the two parameter Gompertz model clearly provides a poor fit; indeed, likelihood ratio tests comparing the two and three parameter models were significant for all three study arms ($p < 0.001$). Conversely, agreement between the NPMLE and three parameter model estimates suggests the latter provides an adequate fit. Because there were no subsequent study visits after an infant died, the naive estimators are not recommended as the implied IIP invalidates these estimators. While all infants in the BAN study will eventually wean or die (i.e., $\lim_{t\to\infty}\sum_{k=1}^{n_K} F_k(t)=1$), the data only provide information on the first 28 weeks of life. Therefore the CFL estimates are also not recommended as (7) implies the unverifiable assumption that the Gompertz models hold for $t > 28$ weeks. Although not recommended, for comparison's sake the CFL and naive estimates are included in Web Figures 1 and 2; estimates of the CIFs are very similar to the UFL estimates in this case.

The parametric estimates of the CIFs provide a straightforward method to test for differences in the probability of a particular failure type by time t between two study arms. For example, let $F_k^g(t;\Theta_k^g)$ denote the CIF for a failure of type k for study arm $g = 1, 2$. Let $Z=\{F_k^1(t;\widehat{\Theta}_k^1)-F_k^2(t;\widehat{\Theta}_k^2)\}/\sqrt{\widehat{var}\{F_k^1(t;\widehat{\Theta}_k^1)\}+\widehat{var}\{F_k^2(t;\widehat{\Theta}_k^2)\}}$ where $\widehat{\Theta}_k^g$ are the UFL estimators computed separately for each study arm $g = 1, 2$. From Section 3.2 it follows that the Wald statistic Z will have a standard Normal distribution under the null hypothesis $H_0: F_k^1(t;\Theta_k^1)=F_k^2(t;\Theta_k^2)$. Analogous test statistics can be defined using the naive estimators. Wald statistics comparing the different arms of the BAN study at $t = 28$ suggest a significant difference in the probability of HIV infection by 28 weeks between the infant ARV and control arms (two-sided p-value $p < 0.001$), and between the maternal ARV and control arms ($p = 0.02$). There is also some indication the risk of HIV infection by 28 weeks is lower in the infant ARV arm compared to the maternal ARV arm ($p = 0.11$).

8. Discussion

Numerical studies suggest the NL, UFL, CFL perform quite well when the parametric models are correctly specified. In theory the CFL estimator should have smaller asymptotic variance than the UFL and naive estimators when (7) holds, although simulation results

suggest this gain may be small in practice. One appealing feature of the naive estimator is that under mixed case interval censoring the likelihood has the same form as in the absence of competing risks, such that existing software for interval censored data could potentially be utilized. In contrast, the UFL and CFL estimators will in general require additional programming. The simulations described in Web Appendix C suggest none of the estimators are particularly robust to severe violations of the assumed parametric CIF model, such that goodness-of-fit diagnostics will be important in practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

MGH was partially supported by NIH grant R01 AI029168. The authors thank the BAN investigators for access to their study data; the Editor, AE and reviewer for helpful comments; Peter Nyangweso for contributing to an earlier version of the manuscript (Nyangweso, Hudgens, Fine 2011); and Sooyoung Lee for assisting with the simulations and data analysis.

References

Chasela C, Hudgens MG, Jamieson DJ, Kayira D, Hosseinipour M, Kourtis AP, et al. Maternal antiretrovirals or infant nevirapine to reduce HIV-1 transmission. New England Journal of Medicine. 2010; 362:2271–81. [PubMed: 20554982]

Groeneboom P, Maathuis MH, Wellner JA. Current status data with competing risks: Consistency and rates of convergence of the MLE. Annals of Statistics. 2008a; 36:1031–1063.

Groeneboom P, Maathuis MH, Wellner JA. Current status data with competing risks: Limiting distribution of the MLE. Annals of Statistics. 2008b; 36:1064–1089. [PubMed: 19888358]

Gruger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. Biometrics. 1991; 47:595–605. [PubMed: 1912263]

Haile, SR. PhD thesis. Vol. 2008. University of Pittsburg; 2008. Inference on Competing Risks in Breast Cancer Data.

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]

Jeong J, Fine JP. Direct parametric inference for the cumulative incidence function. Journal of the Royal Statistical Society, Series C. 2006; 55:187–200.

Jewell NP, Van der Laan M, Henneman T. Nonparametric estimation from current status data with competing risks. Biometrika. 2003; 7:183–197.

Lawless, J. Statistical Models and Methods for Lifetime Data. Wiley; New York: 2003.

Nyangweso, P.; Hudgens, M.; Fine, J. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series. 2011. Parametric estimation of cumulative incidence functions for interval censored competing risks data. Working Paper 19

Prentice RL, Kalbfleisch JD, Peterson AV JR, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978; 34:541–554. [PubMed: 373811]

Schick A, Yu Q. Consistency of the GMLE with mixed case interval–censored data. Journal of the Royal Statistical Society Series B. 2000; 27:45–55.

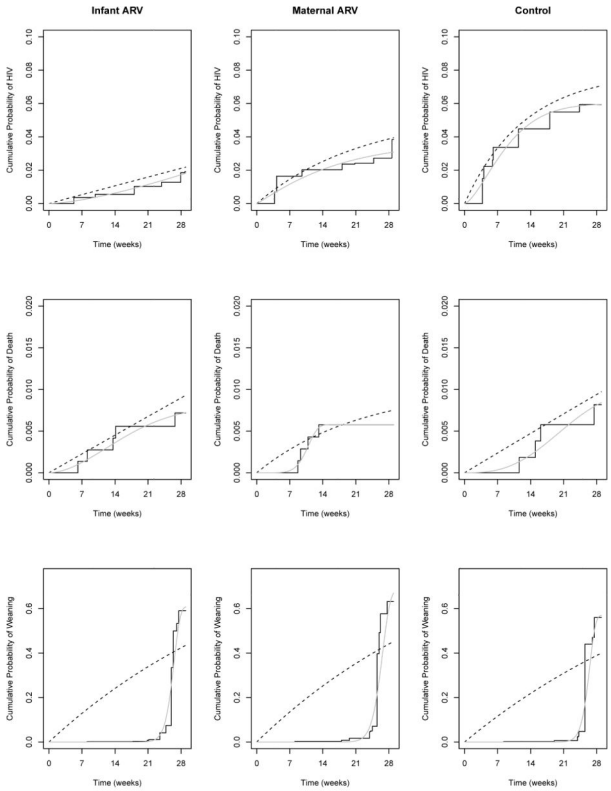

Figure 1.

Estimated CIFs for HIV, HIV-free death, and HIV-free weaning. The solid black line is the NPMLE; the dashed line is the UFL estimate for the two parameter Gompertz model; the solid gray line is the UFL estimate for the three parameter Gompertz model.

Table 1

Simulation study results for data generated by a mixed case interval censoring observation process and direct two-parameter Gompertz model of the CIFs with true parameter values α_k and β_k for $k = 1, 2$. Results are based on 5000 simulated data sets per sample size of n =500, 1000, and 2000. UFL denotes the unconstrained full likelihood estimator, CFL the constrained full likelihood estimator.

	Mean Bias ($\times 10^2$)			Average Variance ($\times 10^3$) Model Based			Average Variance ($\times 10^3$) Empirical			95% CI % Coverage		
Truth	**Naive**	**UFL**	**CFL**	**Naive**	**UFL**	**CFL**	**Naive**	**UFL**	**CFL**	**Naive**	**UFL**	**CFL**
$n = 500$												
$\alpha_1 = -0.058$	−0.075	−0.074	−0.158	0.324	0.324	0.267	0.320	0.319	0.263	95.3	95.3	91.9
$\alpha_2 = -0.035$	0.009	0.009	−0.075	0.058	0.058	0.020	0.057	0.057	0.019	95.0	95.1	94.7
$\beta_1 = 0.0093$	0.011	0.011	0.017	0.004	0.004	0.004	0.004	0.004	0.004	94.4	94.4	93.8
$\beta_2 = 0.067$	0.007	0.007		0.035	0.035		0.035	0.035		94.9	94.8	
$n = 1000$												
$\alpha_1 = -0.058$	−0.040	−0.040	−0.080	0.160	0.159	0.133	0.166	0.166	0.142	94.5	94.6	92.0
$\alpha_2 = -0.035$	0.001	0.001	−0.045	0.029	0.029	0.009	0.030	0.029	0.009	94.8	94.8	94.9
$\beta_1 = 0.0093$	0.006	0.006	0.009	0.002	0.002	0.002	0.002	0.002	0.002	94.1	94.0	93.4
$\beta_2 = 0.067$	0.003	0.004		0.017	0.017		0.018	0.018		94.5	94.4	
$n = 2000$												
$\alpha_1 = -0.058$	−0.055	−0.056	−0.070	0.079	0.079	0.068	0.082	0.082	0.070	94.8	94.8	93.7
$\alpha_2 = -0.035$	−0.002	−0.003	−0.019	0.014	0.014	0.004	0.014	0.014	0.004	95.1	95.1	94.4
$\beta_1 = 0.0093$	0.005	0.005	0.006	0.001	0.001	0.001	0.001	0.001	0.001	94.7	94.7	94.2
$\beta_2 = 0.067$	0.003	0.003		0.009	0.009		0.009	0.009		95.1	95.0	

Table 2

Simulation study results for data generated by a mixed case interval censoring observation process and direct two parameter Gompertz model of the CIFs with true values $F_k(t)$ for $k = 1, 2$. Results are based on 5000 simulated data sets per sample size of $n = 500$ and $n = 2000$. NL denotes the naive likelihood estimator, UFL unconstrained full likelihood estimator, CFL constrained full likelihood estimator, NNP naive nonparametric maximum likelihood estimator, NP full nonparametric maximum likelihood estimator.

	Mean Bias ($\times 10^2$)					Average Variance ($\times 10^3$) Model Based			Average Variance ($\times 10^3$) Empirical					95% CI % Coverage			Mean Squared Error ($\times 10^3$) Empirical				
Truth	**NL**	**UFL**	**CFL**	**NNP**	**NP**	**NL**	**UFL**	**CFL**	**NL**	**UFL**	**CFL**	**NNP**	**NP**	**NL**	**UFL**	**CFL**	**NL**	**UFL**	**CFL**	**NNP**	**NP**
$n = 500$																					
$F_1(5.07) = 0.04$	0.014	0.013	0.035	0.573	0.573	0.054	0.054	0.050	0.053	0.053	0.049	0.143	0.143	94.4	94.5	94.6	0.053	0.053	0.049	0.175	0.175
$F_1(12.7) = 0.08$	−0.026	−0.026	0.001	0.186	0.187	0.134	0.134	0.130	0.132	0.132	0.127	0.179	0.179	94.8	94.8	94.9	0.132	0.132	0.127	0.182	0.182
$F_1(27.5) = 0.12$	−0.039	−0.038	−0.051	−0.278	−0.266	0.210	0.210	0.209	0.204	0.204	0.204	0.210	0.211	94.3	94.4	94.7	0.204	0.204	0.204	0.217	0.218
$F_2(3.54) = 0.2$	−0.006	−0.006	0.092	−0.214	−0.214	0.194	0.194	0.146	0.196	0.196	0.145	0.772	0.772	94.7	94.8	95.5	0.196	0.196	0.146	0.776	0.777
$F_2(12.8) = 0.5$	−0.051	−0.050	0.029	1.037	1.037	0.420	0.420	0.397	0.424	0.424	0.397	0.820	0.820	94.8	94.8	95.0	0.424	0.424	0.397	0.927	0.927
$F_2(28.3) = 0.7$	−0.014	−0.013	−0.141	−0.004	−0.003	0.431	0.431	0.364	0.422	0.422	0.357	0.504	0.496	95.2	95.3	95.1	0.422	0.422	0.359	0.504	0.496
$n = 2000$																					
$F_1(5.07) = 0.04$	0.008	0.008	0.012	0.586	0.586	0.013	0.013	0.013	0.014	0.014	0.013	0.060	0.060	94.9	94.8	94.3	0.014	0.014	0.013	0.094	0.094
$F_1(12.7) = 0.08$	−0.007	−0.007	−0.002	0.103	0.103	0.034	0.034	0.033	0.034	0.034	0.033	0.048	0.048	95.0	94.9	94.7	0.034	0.034	0.033	0.049	0.049
$F_1(27.5) = 0.12$	−0.029	−0.029	−0.031	−0.163	−0.156	0.053	0.053	0.052	0.053	0.053	0.053	0.057	0.057	94.9	94.9	94.9	0.053	0.053	0.053	0.060	0.059
$F_2(3.54) = 0.2$	0.001	0.002	0.021	−0.037	−0.037	0.049	0.049	0.035	0.049	0.049	0.035	0.244	0.244	95.1	95.1	94.7	0.049	0.049	0.035	0.245	0.245
$F_2(12.8) = 0.5$	−0.014	−0.014	0.003	0.460	0.460	0.105	0.105	0.099	0.106	0.105	0.099	0.242	0.242	94.8	94.9	94.9	0.106	0.105	0.099	0.263	0.263
$F_2(28.3) = 0.7$	−0.014	−0.015	−0.041	−0.018	−0.017	0.108	0.108	0.090	0.107	0.107	0.089	0.131	0.129	95.0	95.1	95.1	0.107	0.107	0.089	0.131	0.129