STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:769–781
Published online 5 June 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2581

Multiple imputation for interval censored data with auxiliary variables

Chiu-Hsieh Hsu[1,*,†], Jeremy M. G. Taylor[2,‡], Susan Murray[2,§] and Daniel Commenges[3,¶]

[1] *Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, Tucson, AZ 85724, U.S.A.*
[2] *Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
[3] *INSERM E0338 Biostatistics, ISPED, Bordeaux 2 University, Bordeaux 33000, France*

SUMMARY

We propose a non-parametric multiple imputation scheme, NPMLE imputation, for the analysis of interval censored survival data. Features of the method are that it converts interval-censored data problems to complete data or right censored data problems to which many standard approaches can be used, and that measures of uncertainty are easily obtained. In addition to the event time of primary interest, there are frequently other auxiliary variables that are associated with the event time. For the goal of estimating the marginal survival distribution, these auxiliary variables may provide some additional information about the event time for the interval censored observations. We extend the imputation methods to incorporate information from auxiliary variables with potentially complex structures. To conduct the imputation, we use a working failure-time proportional hazards model to define an imputing risk set for each censored observation. The imputation schemes consist of using the data in the imputing risk sets to create an exact event time for each interval censored observation. In simulation studies we show that the use of multiple imputation methods can improve the efficiency of estimators and reduce the effect of missing visits when compared to simpler approaches. We apply the approach to cytomegalovirus shedding data from an AIDS clinical trial, in which CD4 count is the auxiliary variable. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: auxiliary variables; interval censored; multiple imputation

[*] Correspondence to: Chiu-Hsieh Hsu, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, Tucson, AZ 85724, U.S.A.
[†] E-mail: phsu@azcc.arizona.edu
[‡] E-mail: jmgt@umich.edu
[§] E-mail: skmurray@umich.edu
[¶] E-mail: Daniel.Commenges@isped.u-bordeaux2.fr

Contract/grant sponsor: NIH; contract/grant number: AI29196

Received 7 October 2005
Accepted 13 March 2006

1. INTRODUCTION

There is a large literature on statistical methods to estimate the survival function for interval-censored data. For example, Peto [1] and Turnbull [2] proposed the non-parametric maximum likelihood estimator (NPMLE) to estimate the survival function. Frydman [3] modified Turnbull's method. Finkelstein and Wolfe [4], Satten [5], and Goggins *et al.* [6] used a Cox proportional hazards model to analyse interval-censored data. Most of these methods used intensively iterative computation to obtain measures of uncertainty, i.e. the standard error of the estimator.

In survival analysis, the event times for interval censored observations can be regarded as missing event times [7]; hence multiple imputation, a tool for handling missing data, can be applied to handle interval-censored observations. After imputation, the interval-censored data will be simplified to complete or right-censored data. Then standard statistical methods can be performed on the imputed data sets. As a result, estimates and measures of uncertainty can be easily obtained by following well established rules described in Rubin and Schenker [8]. Examples of imputing event times for interval censored observations can be found in References [9–13]. Brookmeyer and Goedert [9] and Law and Brookmeyer [10] imputed the AIDS infection time by the midpoint of the censored interval. Pan [11] drew imputed values derived from a non-parametric distribution. Pan [12, 13] imputed failure times using the data augmentation technique [14] based on a Cox regression model iteratively fitted to the imputed data.

A common situation where interval censored data arises is in a screening study where participants are observed for the presence of a characteristic at scheduled visits. The censored interval for a subject is the time interval during which the characteristic changes from negative to positive. If the scheduled visits are widely spread or if participants miss visits then the width of the censored interval could be considerable. It is also typically the case that some subjects will be right censored in such a study, if they remain negative at all visits.

Besides the interval-censored data, in many studies there is other information obtained about subjects, and such data are often informative about the health condition of the subjects. Some examples of this are CD4 counts and viral load in studies of HIV and AIDS. These markers are often associated with the event times and, therefore, may be treated as auxiliary variables that can help recover some of the lost information, due to the uncertainty about the event times, for interval censored subjects. In this paper, our interest is in estimating the marginal survival distribution; thus the relationship between the auxiliary variable and the event time is not of primary interest, but it will be used to provide some additional information on endpoint occurrence times for interval censored observations. Therefore, while we try to simplify interval censored data problems to right censored data problems, at the same time, we are also interested in recovering information for interval-censored observations using the auxiliary variables.

The published work on interval censored data is either concerned with estimating the marginal survival distribution [1–3, 10, 11] or focused on discovering the association between the event times and the auxiliary variables [4–6, 9, 12], but does not consider incorporating auxiliary variables into the estimate of the marginal survival distribution. In addition, most of the methods have used either parametric or partially parametric models. We will focus on non-parametric techniques to handle and analyse interval censored data that incorporate the auxiliary variables.

Taylor *et al.* [15] and Hsu *et al.* [16, 17] have studied multiple imputation for right censored data in the one sample case [15] and with additional covariates [16, 17]. Taylor *et al.* [15] showed how imputation schemes can reproduce the standard Kaplan–Meier (KM) estimates, thus providing a theoretical foundation for non-parametric imputation of event times. Hsu *et al.* [16, 17] considered the situation of possibly multiple time-independent or time-dependent continuous covariates. In Hsu *et al.* [16, 17] two risk scores derived from two working proportional hazards (PH) models, one for the failure time and one for the censoring time, were used to define a neighbourhood for each censored case. Then the event time was drawn from a non-parametric distribution based on this neighbourhood. By incorporating predictive auxiliary variables into the multiple imputation method one can both increase efficiency and reduce bias due to dependent censoring of the marginal survival distribution. Hsu *et al.* [16, 17] showed conditions under which the non-parametric imputation enhanced estimate is consistent and reproduces the weighted Kaplan–Meier estimator [18], a method for incorporating categorical auxiliary variables.

In this paper we adapt and generalize the ideas in References [15–17] to handle the case of interval censored data. We propose fitting a working failure-time PH model to reduce the auxiliary variables into a single scalar index of risk that is a combination of the auxiliary variables. This index is then used to define the imputing risk set for each case of interval censoring. Based on the imputing risk set, non-parametric multiple imputation methods are then conducted. If the auxiliary variables used to define the imputing risk set are predictive of the event times, the analyses based on the multiply-imputed data should be more efficient than the analyses based on the data without imputation.

This paper is organized as follows. In Section 2, we review the NPMLE of the survival function for interval censored data. In Section 3, we describe the imputation procedures. In Section 4, we study properties of imputation procedures for survival analysis in finite sample sizes through a simulation study. In Section 5, we apply the techniques to cytomegalovirus (CMV) shedding data. A discussion follows in Section 6.

2. THE NPMLE FOR INTERVAL CENSORED DATA

A key component of multiple imputation is to draw a value for each missing observation from an appropriately chosen distribution. For right censored data, Taylor *et al.* [15] selected an event time using a Kaplan–Meier estimator of the distribution of event times among those still at risk for each censored subject. For interval censored data, we propose to select an event time using a NPMLE of the distribution of event times, analogous to the KM estimates derived from right censored data, among those with similar risk to the censored subject. This section thus provides a review of the NPMLE of the survival distribution for interval censored data.

Let T denote time to the outcome of interest, with c.d.f. $F(t)$. T is said to be censored into a non-zero interval, if we only know that T falls in some interval (L,R), where $L<T<R$. Right censoring is equivalent to $R=\infty$. Let $S(t)=1-F(t)$, where $S(t)$ is the survival function for T. Let (L_i,R_i) denote the observable random interval and (l_i,r_i) denote the observed time interval for each subject under study. The observed data are thus $\mathbf{Y}=\{(l_1,r_1),\ldots,(l_n,r_n)\}$, from a random sample. Under the survival function S, the likelihood for the ith observation is $\{S(L_i)-S(R_i^-)\}$ and the likelihood for all the data is $L(S)=\prod_{i=1}^{n}\{S(L_i)-S(R_i^-)\}$.

Peto [1] used a two-step procedure to obtain the NPMLE, i.e. $\hat{S}$, of S, which is the maximizer of $L(S)$. In the first step, the support of $\hat{S}$ is characterized as a finite number of disjoint intervals. The endpoints of these intervals are elements of the set $\{l_1, l_2, \ldots, l_n, r_1, \ldots, r_n\}$, thus there are at most $2n+1$ disjoint intervals. The set of probabilities associated with these disjoint intervals determines S. In the second step, a constrained Newton–Raphson (NR) method is used to compute $\hat{S}$. In contrast, Turnbull [2] proposed a self-consistency algorithm, a special case of the EM algorithm, to compute $\hat{S}$. It needs intensive computation to obtain measures of uncertainty of the survival estimator. The computational algorithms and large sample properties of the NPMLE can be found in Reference [19].

3. IMPUTATION PROCEDURES

In this section, we describe how to calculate risk scores, how to select the imputing risk set using the risk scores, and two strategies for non-parametric multiple imputation with censored survival data.

3.1. Calculating risk scores

Let $\mathbf{Z} = \{z_1, \ldots, z_n\}$ denote the values of auxiliary variables for the n subjects. For imputation methods, these auxiliary variables are only used to define the imputing risk set. We propose to combine the auxiliary variables into a scalar summary variable (risk score) that measures an individual's risk of disease or death. This is done by fitting a working proportional hazards (PH) model that gives risk scores summarizing the association between the auxiliary variables and the failure time. For the purpose of fitting the working PH model we modify the data to make it right censored. Right censored subjects remain right censored at l_i. For interval censored subjects, we use the midpoint (m_i) of the observed time interval as the hypothetical failure time, i.e. $m_i = (l_i + r_i)/2$. The modified data set is then used to fit the working PH model with the regression coefficients estimated by maximizing the partial likelihood. Because the PH model uses auxiliary variables as covariates, each risk score is then a linear combination of $\mathbf{Z}$.

We fit this working PH model to the available data to obtain a risk score defined as $\mathrm{RS}_f = \hat{\beta}_f Z$, where $\hat{\beta}$ denotes the estimates of the parameters of the PH model for failure times. Each risk score is centred and scaled by subtracting the mean and dividing by the standard deviation of the risk scores. The centred and scaled risk score is denoted as $\mathrm{RS}_f^* = \{\hat{\beta}_f Z - \mathrm{mean}(\hat{\beta}_f \mathbf{Z})\}/\mathrm{SD}(\hat{\beta}_f \mathbf{Z})$. This strategy summarizes the multi-dimensional structure of the auxiliary variables into one dimension. We note that in the case with one auxiliary variable the risk score is equivalent to the covariate itself. Therefore, there is no need to fit this working model.

3.2. Defining the imputing risk set

The scale-free risk score is used to measure the distance between subjects. The distance, based on the original data, between subject j and k is defined as

$$d(j,k) = \{\mathrm{RS}_f^*(j) - \mathrm{RS}_f^*(k)\}^2$$

For each censored subject j, this distance is then employed to define a set of nearest neighbours. The neighbourhood consists of all subjects who have a distance from the censored subject j smaller than d. Note that we did not include in the definition of nearest neighbour a condition that the neighbour k had to survive longer than censored subject j, e.g. $r_k > l_j$, because this would have created a selection bias problem since an individual with a wider interval is more likely to be selected. This nearest neighbourhood for the censored interval, (l_j, r_j), is defined as the imputing risk set $R(j, d)$. Instead of specifying d to be the same for each interval, we choose NN, the size of the nearest neighbourhood, to control the closeness between subjects. For example, $R(j, \mathrm{NN} = 10)$ consists of ten subjects who have the 10 nearest distances from the censored subject j. In the rare case where all subjects in the nearest neighbourhood are interval censored earlier than l_j, we recommend increasing the number in the neighbourhood to ensure some individuals are at risk in a way that overlaps subject j's risk interval.

3.3. *Imputation schemes*

We propose two multiple-imputation schemes to impute the event time for an interval-censored observation. Once the new data set is created, the procedure can be independently repeated M times to obtain multiple imputed data sets for use in estimation. In this paper, the survival estimates for each augmented data set are computed using the KM method and combined to give final estimates. The methods for analysing multiply imputed data sets follow well established rules as described in Reference [8].

3.3.1. Uniform imputation (UNII). For each of the censored intervals, (l_j, r_j), the UNII method simply imputes a event time drawn at random from Uniform(l_j, r_j). For the right censored observations, the UNII method does not impute event times, they remain as right censored. Hence for each censored interval, (l_j, r_j), the UNII method does not use an imputing risk set based on the available auxiliary variables.

3.3.2. NPMLE imputation (NPMLEI). An alternative method that does use the information in the auxiliary variables draws an event time utilizing the NPMLE of the distribution of event times among those in the imputing risk set. The NPMLE is defined on the whole line, but for interval censored subject j we are only interested in the portion between l_j and r_j. Thus we draw an event time from the NPMLE conditional on $t \in (l_j, r_j)$. As mentioned in Reference [11], the NPMLE based on interval-censored data tends to have a smaller number of jumps and hence larger jump sizes than the empirical distribution function based on complete data. Therefore, we propose to use a linear interpolation of the NPMLE to impute for interval-censored observations. Specifically, for each censored interval, (l_j, r_j), a NPMLE survival curve (right continuous), $\hat{S}(j, t)$, is estimated from among those individuals in $R(j, \mathrm{NN})$ with the linearly interpolated version denoted as $\hat{S}^*(j, t)$. Then the NPMLEI method imputes a value t_j^*, which satisfies $l_j < t_j^* < r_j$, from the corresponding linearly interpolated cumulative distribution function $1 - \hat{S}^*(j, t)$. We note that if there are no jumps in the time interval (l_j, r_j), i.e. $\hat{S}(l_j) = \hat{S}(r_j^-)$, for $\hat{S}(j, t)$, then the NPMLEI method just randomly draws an event time from Uniform(l_j, r_j). If there are no individuals at risk in the imputing risk set for the censored subject j, the NPMLEI method will randomly draw an event time from Uniform(l_j, r_j). For a right censored subject j, there is a probability $S^*(j, R_M)/S^*(j, l_j)$ that the NPMLEI method will treat the subject j as right censored at R_M, where $R_M = \max(r_1, r_2, \ldots, r_n)$. There is a probability $1 - (S^*(j, R_M)/S^*(j, l_j))$ that the NPMLEI method will impute a value t_j^*, which satisfies $l_j < t_j^* < R_M$, from the corresponding linearly interpolated cumulative distribution function $1 - \hat{S}^*(j, t)$. When there are no auxiliary variables, the NPMLE for imputation is estimated by using the whole data set with no need to define the nearest neighbourhood.

3.3.3. Bootstrap imputation procedure. Procedures for imputing event times, such as the NPMLEI, by themselves do not incorporate the full uncertainty in the imputes, because they regard the distribution from which the imputes are drawn as known, rather than as an estimate with uncertainty. Therefore, they would not be viewed as proper multiple imputation schemes. In a parametric imputation scheme this can be rectified by including a first stage corresponding to an initial parameter draw. The NPMLEI procedure can be enhanced by including a Bootstrap stage in the procedure, which is designed to make it proper [8]. Consider the bootstrap sample $\{(l_1, r_1)^{(B)}, \ldots, (l_n, r_n)^{(B)}\}$ selected with replacement from the original data set. A PH model for failure time is fitted to this bootstrap sample. Based on this model, a risk score, $\mathrm{RS}_f^{(B)} = \hat{\beta}_f^{(B)} Z^{(B)}$ can be obtained. After centring and scaling, it is denoted as $\mathrm{RS}_f^{(B)*}$. The distance between the censored subject j, we want to impute for, in the original data and the subject k in the bootstrap sample is defined as

$$d^{(B)}(j,k) = \{\mathrm{RS}_f^*(j) - \mathrm{RS}_f^{(B)*}(k)\}^2$$

A nearest neighbourhood $R^{(B)}(j, \mathrm{NN})$ consists of NN subjects who have the NN nearest distances from the censored subject j in the Bootstrap sample. Then the imputing risk set for the censored interval, (l_j, r_j), is this nearest neighbourhood. For the censored interval, (l_j, r_j), the NPMLEI method incorporating the bootstrap method, hereafter denoted as the NPMLEIB method of imputation, imputes a value $t_j^{(B)*}$ from the smooth estimated distribution function, $\{1 - \hat{S}^{(B)*}(j, t)\}$, from the risk set $R^{(B)}(j, \mathrm{NN})$ conditional on the interval (l_j, r_j). Multiple imputations are created by independently repeating the bootstrap stage for each of the M data sets. The inclusion of a Bootstrap stage has been shown to improve the properties of multiple imputation procedures [8, 15, 20].

4. SIMULATION STUDY

We perform several simulation studies to investigate the properties of the multiple imputation based procedures under a variety of parameter combinations. First, we consider situations without any auxiliary variables, which is aimed at comparing the KM estimates from the imputation based analyses and the NPMLE. Second, we consider the situation with several time-independent continuous auxiliary variables. In both situations, for the survival estimates we investigate bias, variance and coverage rates of confidence intervals, and how these are affected by the probabilities of missing four follow-up examinations, and by the inclusion of the bootstrap stage in the multiple imputation procedure. In addition, the effect of the size of the nearest neighbourhood on survival estimates is investigated in cases with continuous auxiliary variables.

4.1. Data generation

A subject is enrolled at the admission time $\tau_0(0)$. For each enrolled subject, the first post-baseline examination is conducted at time τ_1, treated as random. After the first post-baseline examination, there are four follow-up examinations, i.e. $\tau_k = \tau_1 + (k-1) * \text{len}$, $k = 2, 3, 4, 5$. To mimic the pattern of the CMV shedding data described in the next section, the time interval between two adjacent examinations is considered to be constant, e.g. len $= 0.25$. An enrolled subject may miss any of the four follow-ups with some probability, but will not miss the admission time at τ_0 and the first visit at τ_1. Specifically, a random interval-censored sample is generated as follows: Step 0: Specify the probabilities of missing each of the four follow-up visits, e.g. 0.1, 0.1, 0.2, 0.2. Step 1: For $i = 1$ to n repeat Steps 2–4. Step 2a: Generate auxiliary variables (Z_i) from some specified distributions, e.g. $\text{U}(0, 1)$, and then linearly combine them such that the hazard function of the event time is a function of auxiliary variables (Z), e.g. $\beta_1 Z_1 + \beta_2 Z_2$. Step 2b: Generate the event time T_i from some specified distribution, which could be a function of auxiliary variables (Z). Step 3: Generate the first post-baseline examination time τ_{i1} from some specified distribution. Step 4: Calculate other τ_{ik} $(k = 2, \ldots, 5)$ as described above and let $\tau_{i6} = \infty$. We then obtain an interval-censored observation (L_i, R_i), where $L_i = \tau_{ij}$ and $R_i = \tau_{ik}$ for some $0 \leqslant j < k \leqslant 6$ and (τ_{ij}, τ_{ik}) is the shortest interval covering T_i such that the subject did not miss the examinations at τ_{ij} and τ_{ik}. The distribution of τ_{i1} $(i = 1, \ldots, n)$ is $\text{Uniform}(0, \alpha)$, where α is chosen such that about 25 or 35 per cent of subjects are right censored at their last visits. For the probabilities of missing visits, we consider two settings. One is (0,0,0,0), i.e. each subject will not miss any of the four follow-ups. One is (0.1,0.1,0.2,0.2), i.e. a subject may miss any of the four follow-ups and is more likely to miss a latter visit.

4.2. Imputation and analysis

For the 'Fully-Observed' (FO) analysis (the gold standard), we apply KM estimation to each data set before any censoring is applied. For the 'Partially-Observed' (PO) analysis, we apply NPMLE to each data set with random interval censoring. For the multiple imputation methods, for each simulated data set, we multiply impute times for each censored subject as described in Section 2. We then compute Kaplan–Meier estimates for each augmented data set and combine the results to give final estimates. We focus on $S(t)$ at two fixed time points, chosen so that $S(t)$ is equal to, or close to, 0.5 or 0.35.

4.3. Results

4.3.1. Without covariates. Table I shows the survival estimates at the two time points and their associated operating characteristics. For the situation with no missing visits at the four follow-ups, the results indicate that the FO, NPMLEI, and NPMLEIB methods produce point estimates very close to the true values, sometimes closer than the PO analysis gets. The coverage rates for both the NPMLEIB and the PO method tend to be slightly lower than the nominal level. The NPMLEI method without the inclusion of the bootstrap stage produces a low coverage rate. There is no difference in efficiency, as measured by SD, between PO, NPMLEI and NPMLEIB. The UNII method produces biased point estimates and a substantially lower coverage rate than the other methods. These trends also manifest themselves as the

Table I. Monte Carlo results without covariates: survival estimates. The event times ~ exponential with mean 4.0 and a right censoring rate of 25 per cent. Results based on 500 replications and $M = 10$.

		Sample size = 50; NN = 50				Sample size = 200; NN = 200			
Method	True value	Average	SD*	SE†	CR‡	Average	SD	SE	CR
Missing visit probabilities = (0.0, 0.0, 0.0, 0.0)									
FO	0.50	0.50	0.070	0.070	94	0.50	0.034	0.035	95
PO§		0.52	0.151	0.145	91	0.52	0.074	0.074	94
UNII		0.64	0.046	0.086	70	0.51	0.072	0.045	77
NPMLEI		0.50	0.155	0.085	68	0.64	0.023	0.043	2
NPMLEIB		0.50	0.154	0.145	88	0.51	0.074	0.075	92
FO	0.35	0.35	0.067	0.067	95	0.35	0.033	0.034	96
PO		0.36	0.129	0.124	95	0.35	0.068	0.065	93
UNII		0.50	0.050	0.089	69	0.35	0.068	0.042	78
NPMLEI		0.34	0.133	0.078	73	0.49	0.025	0.044	2
NPMLEIB		0.33	0.121	0.117	86	0.35	0.067	0.065	91
Missing visit probabilities = (0.1, 0.1, 0.2, 0.2)									
FO	0.50	0.50	0.069	0.070	94	0.50	0.036	0.035	94
PO		0.53	0.147	0.147	94	0.51	0.080	0.076	91
UNII		0.64	0.046	0.086	67	0.50	0.079	0.045	75
NPMLEI		0.50	0.147	0.086	72	0.64	0.024	0.044	2
NPMLEIB		0.51	0.145	0.148	90	0.50	0.077	0.075	90
FO	0.35	0.35	0.067	0.067	95	0.35	0.035	0.034	93
PO		0.37	0.131	0.128	92	0.35	0.071	0.066	92
UNII		0.50	0.050	0.089	65	0.35	0.071	0.042	75
NPMLEI		0.34	0.136	0.079	74	0.50	0.025	0.045	2
NPMLEIB		0.33	0.126	0.121	86	0.35	0.068	0.066	91

*Empirical standard deviation.

†Estimated standard error based on Greenwood's formula for FO, UNII, NPMLEI, and NPMLEIB and standard error estimated from 500 bootstrap samples for PO.

‡Coverage rate of 95 per cent confidence interval calculated as estimate $\pm t_v^{(0.975)}$ standard error.

§Based on NPMLE.

probabilities for missing visits at the four following times increase, although the bias of the UNII method is more apparent than before.

Overall, the results in Table I for the case without covariates show that NPMLEIB estimates target the point estimate a bit better, but with a slightly lower coverage rate than the PO estimates. As the sample size increases, the similarity in results between the PO (NPMLE) approach and the NPMLEIB approach mimics that seen when comparing non-parametric imputation based methods in Taylor *et al.* [15] and Kaplan–Meier estimation. The results where the right censoring was increased to 35 per cent are similar (results not shown).

4.3.2. Continuous time-independent covariates. We primarily focus on the effects of the sizes of the nearest neighbourhood (NN), sample size and model misspecification. To have better understanding of these effects, we conduct more than one set of simulations. The general

Table II. Monte Carlo results with two covariates, Z_1 and Z_2 from U$(0,1)$: survival estimates. Results based on 500 replications, $M=10$, and missing visit probabilities at the four follow-ups $(0.1, 0.1, 0.2, 0.2)$.

		Exponential[*]				Lognorm[†]			
Method	True value	Average	SD[‡]	SE[§]	CR[¶]	Average	SD	SE	CR
Sample size = 100									
FO	0.50	0.50	0.050	0.050	93	0.50	0.051	0.050	94
PO[‖]		0.53	0.112	0.111	92	0.50	0.074	0.070	93
UNII		0.67	0.030	0.060	9	0.51	0.047	0.055	98
NPMLEI	(NN = 10)	0.53	0.084	0.057	78	0.48	0.062	0.052	89
NPMLEIB		0.57	0.056	0.075	88	0.48	0.053	0.056	94
NPMLEI	(NN = 20)	0.48	0.105	0.057	70	0.49	0.065	0.053	87
NPMLEIB		0.51	0.078	0.092	96	0.48	0.058	0.058	93
NPMLEI	(NN = 50)	0.50	0.115	0.061	68	0.49	0.069	0.053	85
NPMLEIB		0.50	0.106	0.105	93	0.49	0.062	0.062	93
Sample size = 200									
FO	0.50	0.50	0.036	0.035	94	0.50	0.033	0.035	94
PO		0.52	0.083	0.081	93	0.50	0.054	0.067	99
UNII		0.66	0.023	0.043	0	0.51	0.030	0.039	98
NPMLEI	(NN = 10)	0.52	0.057	0.040	81	0.48	0.038	0.037	93
NPMLEIB		0.57	0.040	0.053	78	0.48	0.033	0.040	96
NPMLEI	(NN = 20)	0.48	0.073	0.041	72	0.49	0.040	0.037	92
NPMLEIB		0.51	0.056	0.064	96	0.48	0.036	0.041	96
NPMLEI	(NN = 50)	0.50	0.080	0.043	72	0.49	0.043	0.038	92
NPMLEIB		0.49	0.074	0.071	91	0.49	0.039	0.044	97

[*] The event time $\sim F(t) = 1 - \exp[-t * (0.3Z_1 + 0.25Z_2)]$.

[†] The event time $\sim$ lognorm$(0.2Z_1 - 0.6Z_2)$.

[‡] Empirical standard deviation.

[§] Estimated standard error based on Greenwood's formula for FO, UNII, NPMLEI, and NPMLEIB and standard error estimated from 500 bootstrap samples for PO.

[¶] Coverage rate of 95 per cent confidence interval calculated as estimate $\pm t_v^{(0.975)}$ standard error.

[‖] Based on NPMLE.

results are similar across different scenarios. We, therefore, only report two of the simulation studies in Table II. The results where the event time is from an exponential distribution, as expected, indicate that the biases of the UNII method are consistently greater than that of other methods (FO, PO, NPMLEI, NPMLEIB) in all situations. The bias results in low coverage rates for the UNII method. In both situations, i.e. sample size 100 and 200, when the size of the NN is small, e.g. 10, the NPMLEI and NPMLIEIB methods both produce a small degree of bias that is corrected for large sizes of NN. This implies that a reasonable size of the NN is needed to provide a good NPMLE for the distribution of event times for imputation.

However, as the size of the NN increases from 20 to 50, the coverage rate for the NPMLEIB method decreases a little due to lost efficiency in estimation. For example, the coverage rate ($n = 200$) decreases from 95.8 to 91.4 per cent. This indicates that the nearest neighbours are not being identified well for very large sizes of NN, which are too inclusive. When these issues relating to the choice of NN are balanced appropriately, the NPMLEIB can improve efficiency in estimation compared to the PO method. For example, at the 50th percentile of the survival function, the NPMLEIB (NN = 20) gains about 50 per cent of efficiency compared to the PO method in terms of the standard deviation (SD). In addition, we also note that the big difference in bias between the NPMLEI method and the NPMLEIB method decreases as the size of the NN increases. For example, these two produce comparable point estimates as the size of the NN increases to 50. When the working model is misspecified (i.e. the event time is from a lognormal distribution), the NPMLEIB method still can produce reasonable survival estimates and improve efficiency in estimation compared to the PO method.

5. APPLICATION TO CMV SHEDDING DATA

ACTG-181 clinical trial [6, 21] was a substudy of ACTG-081 [22]. In this trial, each patient was tested at regular intervals to determine whether he/she was shedding CMV in their urine. Urine samples were taken every four weeks. Therefore, the time of onset of CMV shedding for each patient is only known to fall in some interval. In addition, for each patient, several baseline characteristics (e.g. gender and race) were measured and CD4 counts were measured at two different time points, i.e. the beginning and end of the trial. We apply the non-parametric multiple imputation schemes to the interval-censored urine samples. We are interested in obtaining the distribution of CMV shedding-free survival. Since CD4 count is a critical aspect of the immune system, with low values indicating more severe immune deficiency, we incorporate CD4 count at the beginning ($\mathrm{CD4_b}$) and end ($\mathrm{CD4_e}$) of the trial as auxiliary variables for estimating the distribution of CMV shedding-free survival. The two CD4 counts are used as time-independent covariates in the working PH model. For patients who had at least one positive test for CMV virus in the urine, their last CD4 counts were measured at the end of the trial after their events have occurred. In this situation, directly incorporating a patient's last CD4 count into survival analysis gives regression coefficients that are hard to interpret. However, in this paper, we only incorporate a patient's last CD4 count to help define a set of nearest neighbours for each interval censored observation, thus the lack of interpretation of the regression coefficients is less of a concern.

There were 210 patients (out of 232 randomized to the trial) who were tested for CMV shedding at least once before or during the trial. Of these, 127 were interval censored or right censored based on their urine samples. Since our approach is designed to handle interval-censored or right-censored data, we restrict our analysis to these 127 patients. We fit the working model, $\lambda(t) = \lambda_0(t)\mathrm{e}^{\beta_1 \mathrm{CD4_b} + \beta_2 \mathrm{CD4_e}}$, for the failure times to calculate risk scores to choose 20 subjects who have the 20 nearest distances from the censored subject. The event time is drawn from the NPMLE based on the 20 subjects.

The results at two fixed time points, six months and one year, are shown in Table III. This table provides the NPMLE from the partially observed (PO) analysis, that is the analysis of the observed interval-censored event time data using the NPMLE method, and also provides the KM estimates from the multiple imputation analyses, including UNII, NPMLEI,

Table III. Estimates of CMV shedding-free survival probabilities and estimated standard errors based on interval-censored data (NPMLE) and multiply-imputed data (UNII: Uniform imputation, NPMLEI: NPMLE imputation, NPMLEIB: NPMLE-Based imputation using Bootstrap, NN = 20, and $M = 10$).

Method	$\hat{S}(180)^*$	$(\hat{SE}_{180})^\dagger$	$\hat{S}(365)$	$(\hat{SE}_{365})$
PO‡	0.818	0.0360	0.674	0.0448
UNII	0.805	0.0382	0.580	0.0543
NPMLEI	0.792	0.0394	0.589	0.0677
NPMLEIB	0.813	0.0391	0.650	0.0531

*KM survival estimate of remaining CMV shedding-free at six months.

†Estimated standard error based on Greenwood's formula for UNII, NPMLEI, and NPMLEIB andstandard error estimated from 500 bootstrap samples for PO.

‡Based on NPMLE.

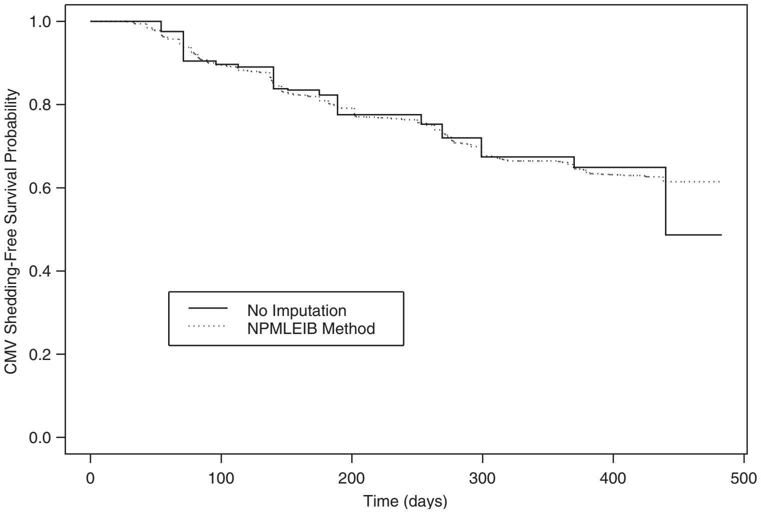

Figure 1. Comparison of CMV shedding-free curves based on the interval censored data (No Imputation) and based on NPMLEIB method using the baseline and last CD4 counts as the auxiliary variables.

and NPMLEIB using the earliest and latest observed CD4 counts as the auxiliary variables. As can be seen in this table and Figure 1, the PO and NPMLEIB methods produce comparable estimates of survival. The results indicates that about 81 per cent of patients will remain CMV shedding-free after six months and 67 per cent of patients will remain CMV shedding-free after one year. The UNII and NPMLEI methods produce a little lower survival estimates than other methods, especially in the tail.

6. DISCUSSION

The research in this paper provides a direct approach, non-parametric multiple imputation, to handle interval censored data. This approach converts interval-censored data problems to complete data or right censored data problems to which standard methods can be applied. This is an attractive feature of multiple imputation approaches. Another attractive feature is that the measures of uncertainty can be easily obtained using well established rules described in Reference [8].

The idea of imputing event times for interval censored data was discussed in Reference [11]. However, our method differs because we impute for right censored observations and also incorporate auxiliary variables into the imputation schemes to improve analysis. When there are no auxiliary variables, our approach behaves similarly to Pan's. When there are auxiliary variables, our approach does recover information for interval-censored observations by incorporating the auxiliary variables into the imputation. As can be seen in the simulation studies, the use of this non-parametric multiple imputation method can lead to improved performance of estimators when auxiliary variables exist. In general, the NPMLEI and NPMLEIB multiple imputation point estimates are closer to the truth than are the estimates produced by randomly imputing event times (UNII) from the censored intervals without using the auxiliary variables. The NPMLEIB has the most attractive operating characteristic of the imputation methods studied. To the extent that the risk scores correctly identify appropriate nearest neighbours, these methods also reduce the effects of dependent censoring on estimation. These methods can also be extended to allow the choice of nearest neighbours to depend on a second working PH model for the censoring distribution as Hsu *et al.* [16, 17] described for the right censored case.

In the situations with auxiliary variables, we use the midpoints of the censored intervals as the event times in order to fit a working model. The midpoint is only used as a convenience in calculating the risk score to choose the imputing risk set. More sophisticated and computationally intensive approaches for fitting the working model could be used, such as proportional hazard models for interval-censored data [4–6], but we suspect would only lead to marginal improvement in the endpoint imputation, but not the bias, which is the major concern for midpoint imputation. We believe that the bias in the regression coefficients may only be a problem if many of the censoring intervals are wide and the conditional event time distribution in the interval is highly skew. In addition, parametric assumptions connected with statistical models are only employed to define the imputing risk set. As a result, the reliance on the statistical model is weaker for our non-parametric multiple imputation schemes than that of parametric multiple imputation schemes. Due to this weak reliance on a model, the potential gains due to the multiple imputation will be largest when the auxiliary variable is strongly associated with the event time. The estimated event time distribution from which the imputes are drawn is derived from the NPMLE. Hence, the performance of imputation procedures will highly depend on the performance of the NPMLE. In small sample size, the NPMLE can be biased. This creates a small bias for the imputation methods in a case with a small nearest neighbourhood. Simulations also suggest the size of NN is very important. Future research could focus on this issue.

In addition to its robustness in this application, the general approach of multiple imputation methods has features that make it attractive. One such feature is that after imputation the data analyst is now free to choose and can easily perform any analysis appropriate for the goals of their study. Conditions for the appropriateness of this philosophy are discussed in Reference [23].

ACKNOWLEDGEMENTS

The authors thank Dianne Finkelstein for providing the CMV shedding data. This work was partially supported by NIH grant AI29196.

REFERENCES

1. Peto R. Experimental survival curves for interval-censored data. *Applied Statistics* 1973; **22**:86–91.
2. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, *Series B* 1976; **38**:290–295.
3. Frydman H. A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society*, *Series B* 1994; **56**:71–74.
4. Finkelstein DM, Wolfe RA. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 1985; **41**:933–945.
5. Satten GA. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* 1996; **83**:355–370.
6. Goggins WB, Finkelstein DM, Schoenfeld DA, Zaslavsky AM. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* 1998; **54**:1498–1507.
7. Heitjan DF. Ignorability in general incomplete-data models. *Biometrika* 1994; **81**:701–707.
8. Rubin DB, Schenker N. Multiple imputations in health-care database: an overview and some applications. *Statistics in Medicine* 1991; **10**:585–598.
9. Brookmeyer R, Goedert JJ. Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* 1989; **45**:325–335.
10. Law C, Brookmeyer R. Effects of midpoint imputation on the analysis of doubly censored data. *Statistics in Medicine* 1992; **11**:1569–1578.
11. Pan W. A comparison of some two-sample tests with interval censored data. *Journal of Nonparametric Statistics* 1999; **12**:133–146.
12. Pan W. A multiple imputation approach to Cox regression with interval censored data. *Biometrics* 2000; **56**:192–203.
13. Pan W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 2001; **57**:1245–1250.
14. Wei GCG, Tanner MA. Applications of multiple imputation to the analysis of censored regression data. *Biometrics* 1991; **47**:1297–1309.
15. Taylor JMG, Murray S, Hsu C-H. Survival estimation and testing via multiple imputation. *Statistics and Probability Letters* 2002; **58**:221–232.
16. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* 2006, in press.
17. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Technical Report*, University of Michigan Department of Biostatistics. http://www.bepress.com/umichbiostat/paper27
18. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; **52**:137–151.
19. Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, vol. 126. Birkhäuser: Basel, 1992.
20. Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991; **40**:13–29.
21. Finkelstein DM, Goggins W, Schoenfeld DA. Analysis of failure time data with dependent interval censoring. *Biometrics* 2002; **58**:298–304.
22. Bozzette SA, Finkelstein DM, Spector SA, Frame P, Powderly WG, He W, Phillips L, Craven D, van der Horst C, Feinberg J. A randomized trial of three anti-pneumocystis agents in patients with advanced HIV infection. *New England Journal of Medicine* 1995; **332**:693–699.
23. Meng XL. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* 1994; **9**:538–573.