Published in final edited form as:
Comput Stat Data Anal. 2010 April 1; 54(4): 1109–1116. doi:10.1016/j.csda.2009.10.022.

A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model

Ling Chen[a,b,*] and **Jianguo Sun**[b]

[a] Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Ave, St. Louis, MO 63110, USA

[b] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, USA

Abstract

This paper discusses regression analysis of interval-censored failure time data, which occur in many fields including demographical, epidemiological, financial, medical, and sociological studies. For the problem, we focus on the situation where the survival time of interest can be described by the additive hazards model and a multiple imputation approach is presented for inference. A major advantage of the approach is its simplicity and it can be easily implemented by using the existing software packages for right-censored failure time data. Extensive simulation studies are conducted which indicate that the approach performs well for practical situations and is comparable to the existing methods. The methodology is applied to a set of interval-censored failure time data arising from an AIDS clinical trial.

1. Introduction

Interval-censored failure time data occur in many areas including demographical, epidemiological, financial, medical, and sociological studies (Sun, 2006). In these situations,the exact time of failure cannot be observed but is known to fall between two observation times. Interval-censored failure time data may arise in several ways. For instance, an individual may miss one or more observation times that have been scheduled to clinically observe possible changes in a disease status and then return with a changed status. Alternatively, individuals may visit clinical centers at times convenient to them rather than at predetermined observation times. In both situations, the data on change in status are interval-censored.

A special case of interval-censored failure time data, case I interval-censored data arise when there is only one monitoring time and each subject is known only to experience the onset of the event either before or after this observation time. In other words, the observation on the failure time of interest is either left- or right- censored (Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003). Left (right) censoring can be considered as an

* Corresponding author at: Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Ave, St. Louis, MO 63110, USA. ling@wubios.wustl.edu (L. Chen)..

extreme case of interval censoring in which the left (right) endpoint is 0 (∞). Case I interval-censored data are often referred to as current status data. In this paper, we focus on general or sometimes referred to as case II interval-censored failure time data.

Many authors have considered regression analysis of interval-censored failure time data since mid-1980s. Finkelstein (1986) studied the use of the proportional hazards (PH) model for interval-censored data followed by many others. In particular, Huang and Wellner (1997) discussed the asymptotic properties of the PH model along with other models. Several other semiparametric models have also been considered for this type of data. These include the proportional odds model studied by Huang and Rossini (1997) and the accelerated failure time model proposed by Rabinowitz et al. (1995). Also for this problem, Kooperberg and Clarkson (1997) discussed linear spline models and Sun (1997) investigated the logistic model. Furthermore, Younes and Lachin (1997) and Zhang et al. (2005) proposed to use the linear transformation model. Bacchetti and Quale (2002) and Zeng et al. (2006) considered the additive hazards model.

For inference, three approaches are commonly used and they are the maximum likelihood approach (Huang, 1996), the estimating equation approach (Lin et al., 1998) and the imputation approach (Pan, 2000). In theory, the maximum likelihood approach applies to any model and is the most efficient method. However, it may be complicated both in terms of investigation of its properties and its implementation. Alternatively, one may consider profile likelihood approach (Huang and Wellner, 1997). But the asymptotic validity of the profile likelihood approach needs to be verified. For semiparametric model, the efficient score function approach (Huang and Wellner, 1997; Bickel et al., 1993; van der Vaart and Wellner, 1996) is commonly used, but it requires estimating an infinite-dimensional nuisance parameter such as the cumulative hazard function in the additive hazards model. For case II interval-censored data problem, a general approach to develop estimating equations or inference procedures is to transfer them to current status data assuming a simple inference procedure is available in the case of current status data. However, the asymptotic validity and properties of the resulting methods and their efficiency are difficult to investigate.

In this paper, we develop a multiple imputation approach for fitting the additive hazards model to interval-censored failure time data. A major advantage of the imputation approach over the existing methods is its simplicity and it can be easily implemented. Among others, Wei and Tanner (1991) developed multiple imputation algorithms for right-censored failure time data under the location-scale regression model. This algorithm imputes censoring times by sampling from the current estimate of the conditional distribution of the error. Once the censoring times have been imputed, least squares is applied to estimate the regression parameters. Pan (2000) investigated multiple imputation approach for interval-censored failure time data under the PH model. The basic idea of his approach is to impute exact survival times from interval-censored data and reduce the analysis of interval-censored failure time data to that of right-censored failure time data. However, under the PH model the effect of covariates is restricted to be multiplicative on hazards of failure and the estimates of regression parameters are not easy to interpret. We adopted the idea of Pan (2000) but allow the covariate effect on hazards of failure to be additive. The additive

hazards model was argued to be more reasonable for certain instances such as dose on risk or hazard (Breslow and Day, 1987). In particular, we considered time-dependent covariates as they often arise in practice but most of the inference approaches developed for this type of data only apply to time-independent covariates.

Before presenting the multiple imputation approach, we will first briefly introduce in Section 2 the additive hazards model and the inference procedure proposed by Lin and Ying (1994) for right-censored failure time data. The imputation approach is then presented in Section 3 and the key idea behind the approach is to generate right-censored data conditional on the observed data. Then the estimation procedure given by Lin and Ying (1994) can be applied. Results from an extensive simulation study are reported in Section 4 for assessing the developed approach and they indicate that the presented multiple imputation procedure performs well for the situations considered compared to the existing methods. Section 5 applies the method to a set of well-known interval-censored failure time data arising from an AIDS clinical trial and Section 6 contains some concluding remarks.

2. The additive hazards model and right-censored failure time data

Consider a survival study and let T denote the failure time of interest and Z a vector of covariates that may depend on time t. We assume that given Z, the hazard function of T has the form

$$\lambda(t;Z) = \lambda_0(t) + \beta' Z(t), \quad (1)$$

where $\lambda_0(t)$ denotes the unknown baseline function and β is the vector of regression coefficients. That is, T follows the additive hazards model (Cox and Oakes, 1984). The goal is to make inference about β.

In this section, we assume that right-censored failure time data are available and given by $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ from n independent subjects. Here X_i denotes the observed failure time defined as the minimum of the true failure time T_i and the censoring time for subject i and $\delta_i = 1$ if the true failure time is observed and 0 otherwise. It is assumed that the failure time and the censoring time are independent given covariates. Define $Y_i(t) = I(X_i \geq t)$, the risk indicator process, and $N_i(t) = I(X_i \leq t, \delta_i = 1)$, a counting process, $i = 1, \ldots, n$.

To estimate β, Lin and Ying (1994) proposed to use the following estimating equation

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i(t) - \bar{Z}(t) \right\} \left\{ \mathrm{d}N_i(t) - Y_i(t) \beta' Z_i(t) \, \mathrm{d}t \right\} = 0,$$

where

$$\bar{Z}(t) = \frac{\sum_{i=1}^{n} Y_i(t) Z_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

It can be easily shown that the solution to the equation above has the explicit form

$$\hat{\beta}=\left[\sum_{i=1}^{n}\int_0^\infty Y_i(t)\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}\mathrm{d}t\right]^{-1}\left[\sum_{i=1}^{n}\int_0^\infty\left\{Z_i(t)-\bar{Z}(t)\right\}\mathrm{d}N_i(t)\right], \quad (2)$$

where $a^{\otimes 2}=aa'$ for a vector a. Furthermore, they showed that $n^{1/2}\left(\hat{\beta}-\beta_0\right)$ converges weakly to a normal vector with mean zero and a covariance matrix that can be consistently estimated by $\Sigma = A^{-1}BA^{-1}$, where β_0 denotes the true value of β,

$$\begin{aligned} A &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty Y_i(t)\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}\mathrm{d}t, \\ B &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}\mathrm{d}N_i(t). \end{aligned}$$

Given $\hat{\beta}$, a natural estimate of the baseline cumulative hazard function $\Lambda_0(t)=\int_0^t\lambda_0(s)\,\mathrm{d}s$ is given by

$$\hat{\Lambda}_0\left(t;\hat{\beta}\right)=\int_0^t\frac{\sum_{i=1}^{n}\left\{\mathrm{d}N_i(s)-Y_i(s)\hat{\beta}'Z_i(s)\,\mathrm{d}s\right\}}{\sum_{i=1}^{n}Y_i(s)}. \quad (3)$$

Note that the estimate (3) may not always be monotone in t. For this, Lin and Ying (1994) suggested using $\hat{\Lambda}_0^*\left(t;\hat{\beta}\right)=max_{s\le t}\quad\hat{\Lambda}_0\left(s;\hat{\beta}\right)$. In the next section, we will discuss the use of the estimates (2) and (3) in estimation of β when only interval-censored failure time data are available.

3. A multiple imputation approach for estimation of β

Now we consider situations where instead of right-censored data, one observes only interval-censored data given by $\{(L_i, R_i], Z_i, i = 1, \ldots, n\}$, where $L_i \le R_i$. Here $L_i = 0$ represents a left-censored observation, $R_i = \infty$ corresponds to right-censored one, and Z_i is the same as before. Note that for right-censored data, one may know T_i exactly, while for interval-censored data, all available information about T_i is the interval $(L_i, R_i]$ with $T_i \in (L_i, R_i]$. Our proposal is to impute the exact failure times from finite interval-censored (left-censored or interval-censored) observations but not right-censored observations. Then the problem is reduced to analyzing the imputed (right-censored) data, which can be handled with the inference procedure in Section 2.

To estimate β, we propose the following estimation procedure. Let K be a prespecified integer.

Step 1. Choose initial estimates $\hat{\beta}^{(0)}$ and $\hat{S}_0^{(0)}$ of β and the baseline survival function $S_0(t) = \exp\{-\Lambda_0(t)\}$, respectively.

Step 2. At the *l*th iteration, let $\hat{\beta}^{(l-1)}$ and $\hat{S}_0^{(l-1)}$ denote the estimates of β and $S_0(t)$ obtained at the $(l-1)$th iteration and generate K sets of right-censored data $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n, k = 1, \ldots, K\}$ as follows. If $R_i = \infty$ (right-censored), define $X_{ik} = L_i$ and $\delta_{ik} = 0$; if $R_i < \infty$, define $\delta_{ik} = 1$ and X_{ik} to be a random number drawn from the survival function

$$\hat{S}_0^{(l-1)}\left(t;\hat{\beta}^{(l-1)}\right)\exp\left\{-\hat{\beta}^{(l-1)'}Z_i^*(t)\right\}$$

given $L_i \le X_{ik} < R_i$, where $Z_i^*(t) = \int_0^t Z_i(s)\,\mathrm{d}s$. The imputed value was drawn this way: suppose that, in interval $(L_i, R_i]$, $\hat{S}_0^{(l-1)}\left(t;\hat{\beta}^{(l-1)}\right) exp\left\{-\hat{\beta}^{(l-1)'}Z_i^*(t)\right\}$ has probability mass $\{p_1, \ldots, p_{S_i}\}$ at time points $\{t_1, \ldots, t_{S_i}\}$, then X_{ik} was randomly drawn from $\{t_1, \ldots, t_{S_i}\}$ with probability proportional to $\{p_1, \ldots, p_{S_i}\}$. Also for all i and k, define $Z_{ik} = Z_i$.

Step 3. First define the estimate $\hat{\beta}_k^{(l)}$ as $\hat{\beta}$ given in (2) with $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ replaced by $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$. Then determine the estimate $\hat{S}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right) = exp\left\{-\hat{\Lambda}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right)\right\}$, where $\hat{\Lambda}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right)$ is given by (3) with $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ replaced by $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$. Also calculate the covariance matrix $\Sigma_k^{(l)}$ as Σ given in Section 2 based on $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$.

Step 4. Define the updated regression estimate $\hat{\beta}^{(l)}$ and the estimate of baseline survival $\hat{S}_0^{(l)}$ as

$$\hat{\beta}^{(l)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k^{(l)}, \qquad \hat{S}_0^{(l)}\left(t;\hat{\beta}^{(l)}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{S}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right),$$

and the covariance matrix can be estimated by

$$\hat{\Sigma}^{(l)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}_k^{(l)} + \left(1+\frac{1}{K}\right)\sum_{k=1}^{K}\frac{\left(\hat{\beta}_k^{(l)} - \hat{\beta}^{(l)}\right)\left(\hat{\beta}_k^{(l)} - \hat{\beta}^{(l)}\right)'}{K-1}.$$

Step 5. Return to step 2 until the convergence is achieved.

To implement the procedure above, one needs to choose the initial estimates of β and S_0. For β, a simple choice is to let $\hat{\beta}^{(0)} = 0$ or use the estimate (2) based on the right-censored data $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$. To generate the X_i's, one can simply impute $X_i = (L_i + R_i)/2$ for the left- and interval-censored observations. Instead of this simple imputation, an alternative is to generate a uniform random variable from the interval (L_i, R_i). For $\hat{S}_0^{(0)}$, as with the second approach for β, one could apply the estimate (3) based on the same set of right-censored data

and the relationship $S_0(t) = \exp\{-\Lambda_0(t)\}$. For the convergence of the procedure, since the main purpose is to estimate β, one can apply the criterion

$$\left|\frac{\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)}}{\hat{\beta}^{(l-1)}}\right| \leq \epsilon \quad \text{or} \quad \left|\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)}\right| \leq \epsilon$$

for a given positive constant ε. Alternatively, one could judge the convergence based on both $\hat{\beta}^{(l)}$ and $\hat{S}_0^{(l)}\left(t;\hat{\beta}^{(l)}\right)$.

Note that as the estimate (3), the estimate $\hat{S}_0^{(l)}\left(t;\hat{\beta}^{(l)}\right)$ may not be monotone. In this case, we can replace it by $min_{s\leq t} \quad \hat{S}_0^{(l)}\left(s;\hat{\beta}^{(l)}\right)$ in step 4. Let $\hat{\beta}=\hat{\beta}^{(L)}$ denote the final estimate of β given by the multiple imputation procedure above. Then one can expect that when n is large, the distribution of $\hat{\beta}$ can be approximated by the normal distribution with mean β_0 and the covariance matrix that can be estimated by

$$\hat{\Sigma}^{(L)}=\frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}_k^{(L)}+\left(1+\frac{1}{K}\right)\sum_{k=1}^{K}\frac{\left(\hat{\beta}_k^{(L)}-\hat{\beta}^{(L)}\right)\left(\hat{\beta}_k^{(L)}-\hat{\beta}^{(L)}\right)'}{K-1}.$$

4. Simulation studies

Two simulation studies were conducted to assess the performance of the multiple imputation approach presented in Section 3. In the first study, the hazard function for the underlying failure time T was taken to be $0.1 + \beta Z$, which gives the density and survival functions

$$S(t;Z)=\exp\left[-(0.1+\beta Z)\,t\right], \qquad f(t;Z)=(0.1+\beta Z)\exp\left[-(0.1+\beta Z)\,t\right],$$

respectively, for a subject with covariate Z. In the second study, we allowed a time-dependent covariate with the hazard function $0.1 + \beta Zt$, which gives the density and survival functions

$$S(t;Z)=\exp\left[-\left(0.1t+0.5\beta Zt^2\right)\right], \qquad f(t;Z)=(0.1+\beta Zt)\exp\left[-\left(0.1t+0.5\beta Zt^2\right)\right],$$

respectively, for a subject with covariate Zt. For the observation times, to mimic the pattern of many medical follow-up studies, the time interval between two examinations was assumed to be constant as $len = 0.25$. It was assumed that there are potentially eight examinations in total for each subject. Denote $\tau_0 = 0$ and $\tau_9 = \infty$. Suppose τ_1 is the random baseline examination time, let $\tau_1 = C_i$, $i = 1, \ldots, n$, and C_i was generated from the uniform distribution $(0, a)$, where a was arbitrarily chosen to give a prespecified percentage of observations right-censored. Then the seven follow-up times are calculated as $\tau_k = \tau_1 + (k -$

1) × *len*, $k = 2, 3, \ldots, 8$; but a subject may miss the scheduled examination times with probabilities 0.3 for the first four and 0.5 for the latter four. We can obtain an interval-censored observation $(L_i, R_i]$, where $L_i = \tau_j$ and $R_i = \tau_k$ for some $0 \le j < k \le 9$ and (τ_j, τ_k) is the shortest interval covering T_i such that the subject did not miss the examinations at τ_j and τ_k. The covariate Z was generated from the Bernoulli distribution with success probability $p = 0.5$.

For comparison, in the simulation study, we also studied the parametric maximum likelihood estimation procedure in addition to the imputation approach. This procedure assumes that one knows that the distribution of the T_i's with the hazard function $\lambda(t; Z) = \lambda_0 + \beta Z$ or $\lambda(t; Z) = \lambda_0 + \beta Z t$ with λ_0 and β as unknown parameters. In this case, β can be easily estimated by its maximum likelihood estimate with the variance estimated based on the Fisher information matrix. The results given below are based on 2000 replications.

Table 1 presents the simulation results for estimation of β and for the situation where the covariates Z_i's are time-independent with the true value, β_0, of β taken to be 0.01, 0.1, 0.3, 0.5 and 0.7, the sample size $n = 100$ and 200, respectively. For each simulation setup, the average percentage of right censoring (%RC) is recorded. The=results include the average bias of the estimates of β (BIAS), the means of the estimated standard deviations (ESD), the sample standard errors of the estimated β (SSE), and the 95% empirical converge probabilities (CP). It can be seen from the table that the multiple imputation approach performs as well as the parametric estimation approach as the two methods gave similar biases and variance estimates. Note that the parametric estimation approach serves as a benchmark here as it is the most efficient method for the situations. It can also be seen from the table that as expected, both bias and variance decrease when the sample size increases.

Table 2 gives the simulation results obtained under the same setups as in Table 1 except that the covariates were assumed to change with time. Overall it gave similar conclusions to those from Table 1. For all situations considered in the simulation study, similar performances were given by the parametric approach and the imputation approach. It can also be seen from both tables that even with about 50% purely interval-censored observations, the imputation approach still performs quite well.

We also investigated the performance of our proposed method with different percentages of exact failure times and various choices of initial estimates. The simulation results based on 2000 replications indicate our estimating procedure is not sensitive to how many exact failure times are observed or choice of initial estimates (shown in Table 3).

For the simulation results given above, we used $K = 10$ for the imputation procedure. To investigate the effect of K on the performance of the procedure, we also performed the simulation with $K = 5$, 50 and 100 and obtained similar results. To assess the normality of $\hat{\beta}$, we obtained the quantile plots of the standardized $\hat{\beta}$ against the standard normal variable. Figs. 1 and 2 presented two such plots and suggest that the asymptotic normality seems reasonable. The plots for other setups were similar.

5. An application

We now apply the presented imputation procedure to a set of interval-censored data arising from an AIDS clinical trial, AIDS Clinical Trial Group (ACTG) 181 on 204 HIV-infected individuals (Goggins and Finkelstein, 2000). In this study, patients were scheduled to provide urine samples at clinic visits every 4 weeks. The Urine samples were tested for the presence of the cytomegalovirus (CMV), which is often referred to as shedding of the virus. Since CMV shedding is not accompanied by any symptoms and only detectable in the laboratory test, it cannot be observed immediately when it exists and only possibly observed at scheduled clinic visits. For some patients the shedding times are left-censored because the shedding had already occurred when they entered the study. For some patients the shedding times are right-censored since they had not started shedding by the end of the study. For the other patients, the shedding times are given by observed intervals with the last negative and first positive urine tests, respectively. Thus the observations on virus shedding times in urine are interval-censored. In addition to the observed information about CMV shedding times, the data also include information about the patients' baseline CD4 cell counts. In particular, the patients are classified into two groups with $Z = 1$ if the baseline CD4 cell count was less than 75 cells/μl and $Z = 0$ otherwise. We are interested in determining whether the baseline CD4 cell count is predictive of CMV shedding in urine.

To analyze the data above, we applied the multiple imputation procedure with $K = 10$ to the observations of CMV shedding times in urine. For this event, there are 49, 67 and 88 left-censored, interval-censored, and right-censored observations, respectively. The procedure yielded $\hat{\beta}=0.072$ with the estimated standard error of 0.0175, The p-value is close to zero for testing $\beta = 0$, no group difference. With $K = 5$ and $K = 50$, the procedure yielded p-values zero. These results suggest that the patients with baseline CD4 cell count below 75 cells/μl have significantly higher risk of CMV shedding in urine than those with baseline CD4 cell count above 75 cells/μl. This result is similar to those given in Sun (2006) under the PH model.

The presented method and analysis above assumed that the CMV shedding time in urine can be reasonably described by the additive hazards model (1) and one may question the appropriateness of the model for the data. For comparison and model checking, we include in Fig. 3 the model-based estimators of the survival functions for the patients in the two groups and the nonparametric maximum likelihood estimators (NPMLEs) of the same survival functions obtained by the self-consistency algorithm of Turnbull (1976). Fig. 3 suggests that the additive hazards model seems to provide a reasonable fit to the data.

6. Concluding remarks

This paper discussed regression analysis of interval-censored failure time data generated under the additive hazards model and for the analysis, a multiple imputation approach was presented and investigated. Compared to the existing methods for the problem, a major advantage of the imputation approach is its simplicity as it can be easily implemented by using the existing software packages for right-censored failure time data. Although there exist other approaches for regression analysis of interval-censored failure time data, they are

either less efficient or complicated. The numerical studies showed that the imputation approach presented here is efficient and performs well for practical situations.

Time-dependent covariates often arise in practice. However, most of the inference approaches developed for interval-censored failure time data only apply to time-independent covariates (Sun, 2006). This paper generalized the imputation approach to situations where covariates are time-dependent.

One potential limitation of the study is that the imputation procedure is likely to be improper. According to Rubin (1987), if the statistics of $\hat{\beta}$ and $\hat{\Sigma}$ from infinite-K imputation procedures provide valid random-response randomization based inferences for the posterior distributions of β and Σ, then the imputation is proper. We adopted Wei and Tanner's Poorman's Data Augmentation (PMDA) and this imputation procedure is likely to be improper because it omits the step of drawing parameters from their posterior distributions. Instead, we update the regression estimates by taking the average over the K augmented data sets, thus multiple imputations are conditional on the current guess. In this case the true between-imputation variability may be underestimated when the missingness is severe. However, in our procedure, repeated imputations were generated under the true model for the response mechanism and the true model for the data, and the complete-data inference equals the complete-data inference derived under the same models. Under these situations, the conditions for repeated-imputation inferences to yield valid random-response randomization based inference will be satisfied asymptotically under mild regularity conditions (Rubin, 1987). This means the average of the estimators is a consistent, asymptotically normal estimator, and an estimator of its asymptotic variance is given by a simple combination of the average of the complete-data variance estimators (within-imputation variance) and the empirical variance of the K estimators (between-imputation variance). Our simulation studies suggest PMDA works reasonably well even with 50% of missingness (finite interval censoring). However, future research is warranted to investigate algorithms that specify random drawings of parameter values to facilitate random generation of multiple imputations.

There exists several other directions for future research. One is that although the simulation suggests that the normal approximation seems reasonable to the distribution of $\hat{\beta}$ and $\hat{\beta}$ is efficient, it would be helpful to provide rigorous justification to these. Also one may want to consider the generalization of the imputation approach to situations where failure times and observation times are dependent and bivariate failure time data as in the example. It would also be interesting to develop a formal procedure for model comparison between the proportional hazards model and the additive hazards model as such a procedure has not been established so far.

References

Bacchetti P, Quale C. Generalized additive models with interval-censored data and time-varying covariates: Application to human immunodeficiency virus infection in Hemophiliacs. Biometrics. 2002; 58:443–447. [PubMed: 12071419]

Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press; Baltimore, MD: 1993. Efficient and Adaptive Estimation for Semiparametric Models..

Breslow, N.; Day, NE. The Design and Analysis of Cohort Studies. Vol. 2. World Health Organization; Lyon: 1987. Statistical Methods in Cancer Research..

Cox, DR.; Oakes, D. Analysis of Survival Data. Chapman & Hall; London: 1984.

Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics. 1986; 42:845–854. [PubMed: 3814726]

Goggins WB, Finkelstein DM. A proportional hazards model for multivariate interval-censored failure time data. Biometrics. 2000; 56:940–943. [PubMed: 10985240]

Huang J. Efficient estimation for the proportional hazards model with interval censoring. The Annals of Statistics. 1996; 24:540–568.

Huang J, Rossini JA. Sieve estimation for the proportional odds model with interval-censoring. Journal of the American Statistical Association. 1997; 92:960–967.

Huang, J.; Wellner, JA. Interval censored survival data: A review of recent progress.. In: Lin, D.; Fleming, T., editors. Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Springer-Verlag; New York: 1997. p. 123-169.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd ed.. John Wiley; New York: 2002.

Klein, JP.; Moeschberger, ML. Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag; New York: 2003.

Kooperberg C, Clarkson DB. Hazard regression with interval-censored data. Biometrics. 1997; 53:1485–1494. [PubMed: 9423263]

Lin DY, Oakes D, Ying Z. Additive hazards regression with current status data. Biometrika. 1998; 85:289–298.

Lin DY, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994; 81:61–71.

Pan W. A multiple imputation approach to Cox regression with interval-censored data. Biometrics. 2000; 56:199–203. [PubMed: 10783796]

Rabinowitz D, Tsiatis AA, Aragon J. Regression with interval-censored data. Biometrika. 1995; 82:501–513.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; New York: 1987.

Sun J. Regression analysis of interval-censored failure time data. Statistics in Medicine. 1997; 16:497–504. [PubMed: 9089958]

Sun, J. The Statistical Analysis of Interval-censored Failure Time Data. Springer; New York: 2006.

Turnbull BW. The empirical distribution with arbitrarily grouped censored and truncated data. Journal of the Royal Statistical Society, Series B. 1976; 38:290–295.

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. Springer; New York: 1996.

Wei GCG, Tanner MA. Applications of multiple imputation to the analysis of censored regression data. Biometrics. 1991; 47:1297–1309. [PubMed: 1786320]

Younes N, Lachin J. Link-based models for survival data with interval and continuous time censoring. Biometrics. 1997; 53:1199–1211.

Zeng D, Cai J, Shen Y. Semiparametric additive risks model for interval-censored data. Statistica Sinica. 2006; 16:287–302.

Zhang Z, Sun L, Zhao X, Sun J. Regression analysis of interval-censored failure time data with linear transformation models. The Canadian Journal of Statistics. 2005; 33(1):61–70.

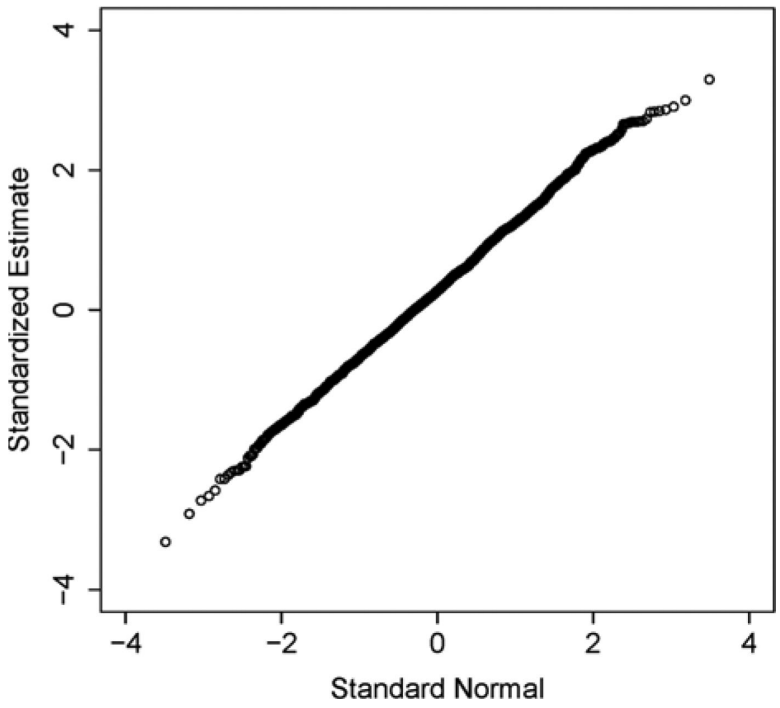

Fig. 1.
Quantile plot of the estimates with time-independent covariates ($\beta_0 = 0.01$, $n = 200$).

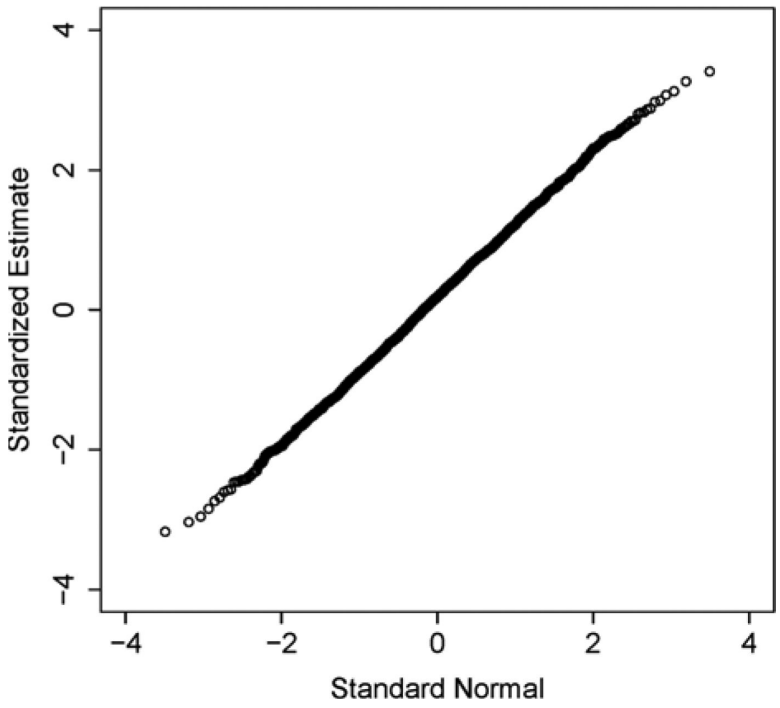

Fig. 2.
Quantile plot of the estimates with time-dependent covariates ($\beta_0 = 0.01$, $n = 200$).

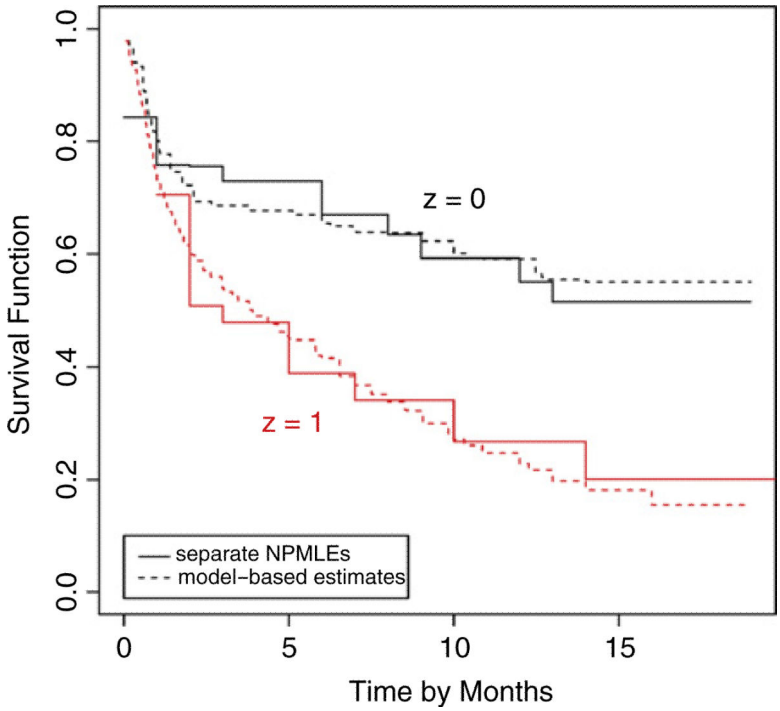

Fig. 3.
Estimates of survival functions of time to CMV shedding in urine.

Table 1

Estimation of β when the covariates are time-independent.

β_0	n	%RC	Parametric MLE				Multiple imputation			
			BIAS	ESD	SSE	CP	BIAS	ESD	SSE	CP
0.7	100	51.3	0.0064	0.1376	0.1332	0.961	–0.0152	0.1361	0.1299	0.950
0.5	100	52.5	0.0098	0.1083	0.1108	0.945	–0.0073	0.1074	0.1066	0.945
0.3	100	52.6	0.0164	0.0784	0.0792	0.958	0.0039	0.0777	0.0756	0.958
0.1	100	59.0	0.0142	0.0490	0.0518	0.947	0.0108	0.0491	0.0505	0.953
0.01	100	64.7	0.0057	0.0464	0.0456	0.953	0.0056	0.0465	0.0496	0.957
0.7	200	51.2	0.0038	0.0967	0.0986	0.944	–0.0100	0.0961	0.0967	0.943
0.5	200	52.6	0.0025	0.0757	0.0769	0.944	–0.0088	0.0752	0.0750	0.943
0.3	200	52.4	0.0116	0.0547	0.0562	0.947	0.0036	0.0541	0.0548	0.953
0.1	200	59.0	0.0124	0.0342	0.0350	0.943	0.0102	0.0341	0.0343	0.948
0.01	200	65.0	0.0043	0.0326	0.0324	0.946	0.0044	0.0327	0.0320	0.949

Table 2

Estimation of β when the covariates are time-dependent.

β_0	n	%RC	Parametric MLE				Multiple imputation			
			BIAS	ESD	SSE	CP	BIAS	ESD	SSE	CP
0.7	100	52.2	–0.0086	0.1449	0.1369	0.951	–0.0193	0.1540	0.1504	0.933
0.5	100	51.4	0.0042	0.1079	0.1065	0.951	–0.0057	0.1155	0.1167	0.939
0.3	100	48.7	0.0141	0.0692	0.0711	0.953	0.0122	0.0751	0.0801	0.943
0.1	100	53.2	0.0104	0.0325	0.0347	0.944	0.0091	0.0342	0.0364	0.949
0.01	100	61.6	0.0047	0.0142	0.0138	0.962	0.0047	0.0146	0.0142	0.967
0.7	200	52.2	–0.0057	0.1022	0.1029	0.943	–0.0168	0.1090	0.1121	0.931
0.5	200	51.3	0.0018	0.0755	0.0778	0.946	–0.0053	0.0812	0.0855	0.930
0.3	200	48.5	0.0125	0.0487	0.0508	0.945	0.0099	0.0529	0.0559	0.943
0.1	200	53.2	0.0083	0.0228	0.0244	0.940	0.0075	0.0239	0.0257	0.938
0.01	200	61.6	0.0044	0.0100	0.0101	0.941	0.0049	0.0102	0.0102	0.939

Table 3

Estimation of β using different initial estimates.

β_0	n	Methods	Multiple imputation			
			BIAS	ESD	SSE	CP
0.5	100	Impute uniform variable	–0.0073	0.1074	0.1066	0.945
		$\hat{\beta}^{(0)} = 0$	–0.0093	0.1072	0.1026	0.955
		Impute midpoint	–0.0066	0.1071	0.1059	0.951
0.01	100	Impute uniform variable	0.0056	0.0465	0.0496	0.957
		$\hat{\beta}^{(0)} = 0$	0.0044	0.0469	0.0447	0.961
		Impute midpoint	0.0081	0.0465	0.0453	0.961