University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 27

Survival Analysis Using Auxiliary Variables Via Nonparametric Multiple Imputation

Chiu-Hsieh Hsu* Jeremy Taylor†
Susan Murray‡ Daniel Commenges**

*University of Michigan Biostatistics
†University of Michigan, jmgt@umich.edu
‡University of Michigan Biostatistics, skmurray@umich.edu
**INSERM

http://biostats.bepress.com/umichbiostat/paper27

Survival Analysis Using Auxiliary Variables Via Nonparametric Multiple Imputation

Chiu-Hsieh Hsu, Jeremy Taylor, Susan Murray, and Daniel Commenges

Abstract

We develop an approach, based on multiple imputation, that estimates the marginal survival distribution in survival analysis using auxiliary variable to recover information for censored observations. To conduct the imputation, we use two working survival model to define the nearest neighbor imputing risk set. One model is for the event times and the other for the censoring times. Based on the imputing risk set, two nonparametric multiple imputation methods are considered: risk set imputation, and Kaplan-Meier estimator. For both methods a future event or censoring time is imputed for each censored observation. With a categorical auxiliary variable, we show that with a large number of imputes the estimates from the Kaplan-Meier imputation method correspond to the weighted Kaplan-Meier estimator. We also show that the Kaplan-Meier imputation method is robust to misspecification of either one of the two working models. In a simulation study with the time independent and time dependent auxiliary variables, we compare the multiple imputation approaches with an inverse probability of censoring weighted method. We show that all approaches can reduce bias due to dependent censoring and improve the efficiency. We apply the approaches to AIDS clinical trial data comparing ZDV and placebo, in which CD4 count is the time-dependent auxiliary variable.

Survival Analysis Using Auxiliary Variables Via Nonparametric Multiple Imputation

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

[1]Division of Epidemiology and Biostatistics, College of Public Health, University of Arizona, Tucson, AZ, 85721, USA

[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 develop an approach, based on multiple imputation, that estimates the marginal survival distribution in survival analysis using auxiliary variables to recover information for censored observations. To conduct the imputation, we use two working survival models to define a nearest neighbor imputing risk set. One model is for the event times and the other for the censoring times. Based on the imputing risk set, two nonparametric multiple imputation methods are considered: risk set imputation, and Kaplan-Meier imputation. For both methods a future event or censoring time is imputed for each censored observation. With a categorical auxiliary variable, we show that with a large number of imputes the estimates from the Kaplan-Meier imputation method correspond to the weighted Kaplan-Meier estimator. We also show that the Kaplan-Meier imputation method is robust to misspecification of either one of the two working models. In a simulation study with time independent and time dependent auxiliary variables, we compare the multiple imputation approaches with an inverse probability of censoring weighted method. We show that all approaches can reduce bias due to dependent censoring and improve the efficiency. We apply the approaches to AIDS

clinical trial data comparing ZDV and placebo, in which CD4 count is the time-dependent auxiliary variable.

KEY WORDS: Double robustness; Multiple Imputation; Nearest neighbor.

1. Introduction

In many clinical trials subjects take a long time to progress to the primary endpoint of interest and might drop out of a study during the trial or be event-free at the end of the study, creating censored observations. In survival analysis, the event times for censored observations can be regarded as missing data (Heitjan, 1994); in this sense there is a loss of information due to censoring. In many studies, there is other information obtained about subjects, and such data may be informative about their health condition. Some examples of this are CD4 count and viral load in studies of 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. 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 censored observations.

In a clinical study, not only does censoring result in a loss of efficiency of estimators, but there is the potential for bias too if the censoring mechanism is not independent of the event time mechanism. For example, in an AIDS study if people with low CD4 counts tended to drop out of the study before an event occurred, then a standard estimate of the marginal survival distribution could be biased. Thus incorporating auxiliary variables has potential to improve efficiency and reduce bias due to dependent censoring in estimating the marginal survival distribution. There are an increasing number of statistical methods (Fleming et al., 1994; Finkelstein and Schoenfeld, 1994; Robins and Rotnitzky, 1992; Malani, 1995; Gray, 1994; Murray and Tsiatis, 1996; Hubbard et al., 2000; Robins and Finkelstein, 2000) incorporating

auxiliary variables into survival analysis to improve survival estimates. Some of these methods have been limited to single or categorical covariates or have used parametric models. Our focus will be on using a less parametric method to incorporate the information in the possibly multiple, continuous or time-dependent auxiliary variables.

In the setting of time-dependent longitudinal auxiliary variable, where these methods are likely to have the most scientific relevance, a full understanding of the data would require consideration of both the stochastic process for the longitudinal variable and the event time process (Jewell and Kalbfleisch (1996)). In this paper our goal is to estimate the marginal survival distribution utilizing the longitudinal data, but without explicitly estimating the stochastic process for the longitudinal variable. Other authors have addressed the same problem, in particular the work of Robins and colleagues (Robins and Rotnitzky, 1992; Hubbard et al., 2000; Robins and Finkelstein, 2000). Their methods use inverse probability weighting to correct for possible bias and can involve extensions to improve efficiency. The methods in their simplest form consist of developing a model for the censoring mechanism and using the results of this model to reweight the observations in an estimating equation. With elegant mathematics, they show consistency of the estimators, under defined conditions, even if dependent censoring were to occur. In their work the initial emphasis is on bias correction if censoring is dependent. In our approach the primary emphasis will be on using the information in the data on the association between the auxiliary variables and the failure time to improve the estimate, while at the same time trying to minimize the impact of possible dependent censoring. Thus we will place more emphasis on modelling the failure time rather than the censoring time distribution.

One direct and transparent tool for handling missing data is multiple imputation (Rubin, 1987). It is a flexible method that has been used in survey research and many other areas of application. In multiple imputation the missing values are "filled-in" from appropriate

distributions, multiple times. The "filled-in" datasets are analysed and the results combined.

Taylor et al. (2002) considered the situation of one-sample survival estimation without auxiliary variables and developed nonparametric procedures to impute missing event times. They also showed that with a large number of imputes the estimates from these multiple imputation methods reproduce the Kaplan-Meier estimator. This provides a theoretical basis for investigating the use of nonparametric multiple imputation strategies in more complex situations, where the imputation strategy could depend on auxiliary variables. For example, Schenker and Taylor (1996) employed three auxiliary variables to determine the distribution of residual times from which to impute a time of AIDS. Faucett et al. (2002) used a parametric joint longitudinal-survival model to impute event times in an AIDS clinical trial. In this research, we propose using the auxiliary variables to define a nearest neighborhood of similar observations for each censored case, and then generate imputes from this set of neighbors.

The paper is organized as follows. In Section 2, we describe the multiple imputation procedures and in Section 3 we discuss their properties . In Section 4, we apply the techniques to data from an AIDS study. In Section 5, we give results from a simulation study. A discussion follows in Section 6.

2. Imputation Procedures

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

2.1 *Calculating risk scores*

Let $\{t_1, ..., t_n\}$ denote the observed times for the n independent subjects under study, with $\delta_i = 1$ if t_i is an event time and $\delta_i = 0$ if t_i is a censored time . Let $\mathbf{Z} = \{\mathbf{z_1}, ..., \mathbf{z_n}\}$ denote the values of the auxiliary variables and let $\mathbf{Y} = \{(t_1, \delta_1, \mathbf{z_1}), ..., (\mathbf{t_n}, \delta_\mathbf{n}, \mathbf{z_n})\}$.

For each censored observation we seek an imputing risk set consisting of subjects who are similar to the censored case. To define the imputing risk set, we first reduce the auxiliary

variables to a scalar index (risk score), which provides an indicator of an individual's risk of disease or death. We use a time-independent proportional hazards regression model to derive these risk scores, summarizing the association between the auxiliary variables and the failure time. When the model for survival is correctly specified, these risk scores define an imputing risk set that can be used to improve efficiency when censoring is independent and reduce bias when censoring is dependent upon the auxiliary variables. However, if the model for survival is misspecified, and censoring is dependent upon the auxiliary variables, bias may remain. Therefore, we will also investigate a second PH model that calculates risk scores by summarizing the association between the auxiliary variables and the censored time. Combinations of these risk scores based upon survival and censoring distributions will be studied to see to what extent a double robustness property for model misspecification can be established (Robins et al., 2000). Intuitively, if one of these two working models is correctly specified, conditional on these two risk scores, the event times are independent of the censoring times. Hence, within an imputing risk set that is defined using two risk scores, the event times are independent of the censoring times.

Since both models use auxiliary variables as covariates, each risk score is a linear combination of $\mathbf{Z}$. In the case with time-independent auxiliary variables, we can directly use all auxiliary variables as covariates to fit these two PH models. In the case with time-dependent auxiliary variables, we propose using the current observed value of the time-dependent auxiliary variable before the censored time we want to impute as the covariate, although earlier measurements could be included as additional covariates. The risk scores are defined as $\widehat{RS}_f = \hat{\beta}_f \mathbf{Z}$ and $\widehat{RS}_c = \hat{\beta}_c \mathbf{Z}$, where $\hat{\beta}_f$ denotes the estimates of the parameters of the PH model for failure times and $\hat{\beta}_c$ denotes the estimates of the parameters of the PH model for censored times. In cases with time-independent auxiliary variables, we fit these two PH models just once. However, in cases with time-dependent auxiliary variables, for every censored observation we

fit these two time-independent PH models to the data of those at risk at the censoring time using the currently available auxiliary variables as fixed covariates. Each risk score is centered and scaled by subtracting the mean and dividing by the standard deviation of the risk scores, denoted as $\widehat{RS}_f^* = \{\hat{\beta}_f \mathbf{Z} - mean(\hat{\beta}_f \mathbf{Z})\}/SD(\hat{\beta}_f \mathbf{Z})$ and $\widehat{RS}_c^* = \{\hat{\beta}_c \mathbf{Z} - mean(\hat{\beta}_c \mathbf{Z})\}/SD(\hat{\beta}_c \mathbf{Z})$, respectively. This strategy summarizes the multi-dimensional structure of the auxiliary variables into a two-dimensional structure. The hope is that this 2-dimensional summary contains most, if not all, the information about the future event and censoring times. We note that in the case of one auxiliary variable the two risk scores can be reduced to one risk score that is equivalent to the covariate itself. Therefore, there is no need to fit these two working models in that special case.

2.2 *Defining the imputing risk set*

The distance between subjects j and k is defined as

$$d(j,k) = \sqrt{w_f\{\widehat{RS}_f^*(j) - \widehat{RS}_f^*(k)\}^2 + w_c\{\widehat{RS}_c^*(j) - \widehat{RS}_c^*(k)\}^2}, \quad (1)$$

where w_f and w_c are non-negative weights that sum to one. Note that in the case of one covariate, the difference between covariate values defines distance between subjects. For each censored subject j, this distance is then employed to define a set of nearest neighbors. This neighborhood, $R(j^+, NN)$, consists of NN subjects who have longer survival time than the censoring time of subject j and a small distance from the censored subject j. For example, $R(j^+, NN = 10)$ consists of ten subjects with the 10 nearest distances from subject j amongst those who have longer survival time than the censoring time for subject j. When the number of individuals still at risk is less than NN, then they are all included in the imputing risk set. Non-zero weights for both w_f and w_c may be useful in reducing the bias resulting from model misspecification. Specifically, a small weight w_c (e.g. 0.2) will result in incorporating the risk scores from the censored time model into defining a set of nearest neighbors for censored subjects.

2.3 *Imputation schemes*

After the imputing risk set $R(j^+, NN)$ is defined, the four nonparametric imputation schemes developed in Taylor et al. (2002) and briefly described below can be easily used. The procedure can be independently repeated M times to obtain multiple imputed data sets for use in estimation. The methods for analyzing multiply imputed data sets follow well established rules as described in Rubin and Schenker (1991). In particular the final estimate is the average of the M estimates and the final variance is the sum of a between-imputation and a within-imputation component. In this case the estimate from each imputed dataset is a Kaplan-Meier estimate and the within-imputation variance is based on Greenwood's formula.

Risk set imputation (RSI) For each of the observed censored times t_j, the RSI method imputes a pair (t_j^*, δ_j^*) drawn at random from the observed pairs of those individuals in $R(j^+, NN)$. Hence for each censored time t_j the RSI method is equally likely to draw any of the observed failure or censored times from those individuals in the imputing risk set $R(j^+, NN)$.

Kaplan-Meier imputation (KMI) This method draws an event time from a KM estimator of the distribution of failure times based on the imputing risk set. Thus, the procedure imputes only observed failure times unless the longest time in the imputing risk set is censored, in which case some imputed times may include this censored time. Specifically, for each censored time t_j, a survival curve, $\hat{S}_{j^+}(t)$, is estimated from among those individuals in $R(j^+, NN)$. Then the KMI method imputes a value t_j^* from the corresponding estimated distribution function $1 - \hat{S}_{j^+}(t)$. Note that the KMI method will nearly always impute an event time, while RSI will frequently impute a censored time.

Bootstrap imputation procedure The RSI and KMI procedures by themselves do not incorporate the full uncertainty in the imputes, because they do not include a first stage of an initial parameter draw. Multiple imputation methods can be enhanced by including a Bootstrap stage, which has been shown to improve their properties (Rubin and Schenker, 1991; Heitjan

and Little, 1991). In Taylor et al. (2002), it was shown that when imputing event times, the inclusion of the Bootstrap stage improved the coverage rate of confidence intervals. A bootstrap sample is selected with replacement from the original data set. The two working models are fit to this Bootstrap sample. Based on these two models, two centered and scaled risk scores can be obtained. The distance between the censored subject j, we want to impute for, in the original data and the subject k (a potential impute) in the bootstrap sample is defined as in equation 1. The imputing risk set for the censored time t_j is the nearest neighborhood $R^{(B)}(j^+, NN)$ consisting of NN subjects with the NN nearest distances from the censored subject j amongst those in the Bootstrap sample who have longer survival time than t_j. For the censored time t_j, the KMI and RSI methods incorporating Bootstrap methods, denoted as KMIB and RSIB, impute a value $t_j^{(B^*)}$ from the estimated distribution function, or draw a pair $(t_j^{(B^*)}, \delta_j^{(B^*)})$ from $R^{(B)}(j^+, NN)$, respectively. Multiple imputations are created by repeating the bootstrap stage for each of the M data sets.

3. Properties of Kaplan-Meier Imputation Method

3.1 *Relationship between KMI and Weighted Kaplan-Meier (WKM) estimates*

For illustration, we assume the auxiliary variable Z is a baseline categorical covariate and takes on values $1, ..., K$. The survival function can be written as

$$S(t) = P(T > t) = \sum_{k=1}^{K} P(T > t|Z = k)P(Z = k) = \sum_{i=k}^{K} S_k(t)\theta_k,$$

where θ_k is the probability a subject has covariate value k, $(k = 1, ..., K)$ and $S_k(t)$ is the probability of survival conditional on having covariate value k. Based on the above expression, the WKM estimator (Murray and Tsiatis, 1996, Malani, 1995) is defined as $WKM(t) \stackrel{def}{=} \sum_{k=1}^{K} \hat{S}_k(t)\frac{n_k}{n}$, where $\hat{S}_k(t)$ is the KM estimator among those with covariate value k, n_k is the number of subjects with covariate value k, and $n = \sum_{k=1}^{K} n_k$. The WKM estimator is only defined up to a certain time. This time is defined as follows: let τ_k be the longest censored time

among subjects with $Z = k$ and for which the longest time is censored, let τ_k be infinite if the longest time is an event. Then WKM is only defined up to the minimum of the values of τ_k. The WKM estimator is consistent, if the event time and the censoring time are independent conditional on Z. Extensions of WKM to time-dependent covariates are also described by Murray and Tsiatis, 1996 and Malani, 1995.

For multiple imputation, the imputing risk set reduces to a risk set, $R(j^+, Z = z_j)$, consisting of those who have longer survival time than the censored time t_j and the same covariate value as z_j. For the observed censored time t_j, KMI imputes a pair (t_j^*, δ_j^*) drawn from the nonparametric survival curve for those individuals in $R(j^+, Z = z_j)$. Building on the theoretical foundation of KMI survival estimates obtained in Taylor et al. (2002), the property of KMI survival estimates $(\hat{S}_{KMI}(t))$ with a categorical covariate for a large number of imputes can be summarized in the following result. The proof is given in the Appendix.

Result 1: $E\{\hat{S}_{KMI}(t)|\mathbf{Y}\} = WKM(t)$.

In this expression the expectation is with respect to the distribution of possible imputes conditional on the observed data $\mathbf{Y}$. The above result shows that the KMI survival point estimates, with a large number of imputes, will on average reproduce the WKM survival estimate over the range of times where WKM is defined. The RSI imputation method, which tends to impute censored values more often than KMI, will not reproduce on average the WKM estimate. In more complex situations, such as the situations with multiple categorical covariates, multiple continuous covariates, or time-dependent covariates, the WKM may not be defined and when it is defined the KMI method will not necessarily reproduce the WKM estimate.

3.2 *Consistency of the KMI method*

Result 2: *If one of the two working models is correct, T and C are independent conditional on the two risk scores.* The conditions and proof are outlined in the Appendix.

This result is the key one, it enables us to use two risk scores to define an imputing risk set and within this imputing risk set the event times are asymptotically independent of the censoring times. Thus estimates of the residual time distribution, derived from observations in the imputing risk, are valid in large samples. Based on this property and appealing to the results in Dabrowska (1989), we have the following result, a sketch proof of which is outlined in the Appendix for the case of baseline covariates.

Result 3: *If one of two working PH models is correctly specified, the KMI method for estimating the distribution of T will have small bias in large samples for values of t prior to the first censored value in the imputed datasets.*

As a result, the KMI method has a large sample property of double robustness (Robins et al., 2000). Because we use two PH models to choose the imputing risk set, based on the above results, the survival estimate will be reasonable if only one of the two true models is from the PH model family and is correctly specified.

The method has a second robustness property for time-independent auxiliary variables, specifically if one of the two true models is from the accelerated failure time model family, then fitting two PH models still gives good estimates of the regression coefficients (Solomon (1984), Struthers and Kalbfleisch (1986)). Since it is only the regression coefficients, and not the link function that are used in defining the imputing risk set, the KMI procedure is robust to misspecification of the link function.

The above properties of the KMI method apply in large sample conditions. In small sample size situations, this nearest-neighborhood approach, which is analogous to a kernel-based method, could produce biased survival estimates even if one of the two working models is correctly specified, especially when the failure-time model is misspecified. This phenomenon is similar to that of the kernel-based method, as indicated in Pepe (1992). The bias is due to the lack of availability of suitable donor observations.

3.3 *Inverse Probability of Censoring Weighted method*

We will be including in the application and simulation study a comparison with the IPCW method as described in Robins and Finkelstein (2000). In particular we use the appropriate adaptation of equation 10 from their paper, which is a weighted Kaplan-Meier estimate, in which each event time is weighted by $1/K_i(t)$, where $K_i(t)$ is an estimate of the conditional probability that subject i is uncensored through time t, given the auxiliary data available. This estimate is obtained by fitting a time-dependent PH model to the censoring times, and requires estimation of both the regression coefficients and the baseline hazard in such a model. Standard errors are obtained using the expressions given in the appendix of Robins and Finkelstein (2000).

4. Application to AIDS Data

We apply the nonparametric multiple imputation schemes and the IPCW method to AIDS data from the ACTG-019 clinical trial (Volberding et al., 1990; Faucett et al., 2002). There are 1337 subjects, with 428 subjects in the placebo arm and 909 subjects in the treated arm, where this latter arm is a combination of two doses of ZDV. The censoring rates for the treatment and placebo groups are 97% and 92%, respectively. The percentage of the censored observations that were administratively censored due to study termination was 95% in the treatment group and 90% in the placebo group. The median follow-up time is 50 weeks. For each subject CD4 counts were measured at months 0, 3, 6, 9, 12, 18. We focus on the survival estimates at days 450 and 550 for both placebo and treated groups. Since CD4 count is a critical aspect of the immune system, with low values indicating more severe immune deficiency, we use it as an auxiliary variable in estimating the survival distribution of each group. For each observed censored time we use individuals who survived longer than the censored subject and who share the same treatment group to fit two working PH models for the censoring and failure time distributions. We consider different parameterizations of CD4

counts as covariates in these working models, e.g. using only the latest observed CD4 count before each censored time, and using both baseline CD4 count and the latest observed CD4 count before each censored time. Since most of the censoring is administrative, this is a study where we would expect to see little bias from dependent censoring and hope to see some gain in efficiency by using the auxiliary variables.

The results based on the latest observed CD4 as the only covariate are provided in Table 1 and Figure 1. Table 1 displays selected estimates from the partially-observed (PO) analysis, that is, the analysis of the observed censored event time data, from the multiple imputation analyses and from the IPCW method. Figure 1 displays the estimated survival curves based on the partially observed (PO) analysis and based on the KMIB multiple imputation method. In a situation with a single covariate, there is no need to fit the two working models, and therefore no weights are used for selecting the imputing risk set. For both the treatment and placebo groups, the imputation analyses produce very similar estimated survival to the PO method. This agrees with the conclusion in Faucett et al. (2002).

Besides the survival estimates, Table 1 also provides estimated SEs. RSI, KMI, RSIB, and KMIB tend to give a modest reduction in the estimated SEs compared to the PO analysis. The results for M=15 and 30 are similar to those for M=10, indicating that 10 imputations are reasonable for the analysis.

When both baseline CD4 count and the latest observed CD4 count before each censored time are in the working models, with $w_f = 0.8$ and $w_c = 0.2$, the multiple imputation methods again give similar results for both treated and placebo groups (results not shown).

The IPCW methods produced similar estimates of survival, but was slightly less efficient.

5. Simulation Study

We perform several simulation studies to investigate the properties of the multiple imputation based procedures and to compare with IPCW methods. We consider a binary auxiliary vari-

able, multiple time-independent auxiliary variables, and time-dependent auxiliary variables. We investigate bias, variance and coverage rates of confidence intervals, and how these are affected by the censoring mechanism, by the inclusion of the bootstrap stage, by the sample size, by model misspecification for calculating the risk scores, by the weights (w_f) and by the size (NN) of the nearest neighborhood.

5.1 *Data Generation*

For the situation with a binary covariate (Table 2), the event time and censoring time are both generated from an exponential distribution. For more complex situations (Tables 3-5), the event and censoring times are both generated from hypothetical PH models conditional on auxiliary variables.

For the situation with the failure time working model potentially misspecified (Tables 3-5), five hypothetical auxiliary variables $(Z_1, ..., Z_5)$ are independently generated from a $U(0,1)$ distribution. The event time is generated from the model $\lambda(t) = t^4 * exp(-2.0Z_1 + 0.5Z_2 - 2.0Z_3 + 2.0Z_4 + 2.0Z_5)$. The censoring time is generated from the model $\lambda_c(t) = t^3 * exp(-3.0Z_1 + 0.5Z_2 - 2.0Z_3 + 1.5Z_4 + 2.0Z_5)$ for dependent censoring or from $\lambda_c(t) = 0.6$ for independent censoring. When the working failure time or working censoring models are misspecified only the terms for Z_1, Z_2 and Z_3 are included.

For time-dependent covariates two models are considered, either a random effects (RE) model or a Brownian motion (BM) model. For the RE model, an auxiliary variable (Z_{ij}) is generated from the model, $Z_i(t) = b_{0i} + b_{1i} * t$, where t has units of days, $b_{0i} \sim N(0, 9)$, and b_{1i} distributed as $N(0, 0.005^2)$. For the BM model Z_i is generated from a model, $Z_i(t) = b_{0i} + b_1 * t + \sigma BM_i(t)$, where $b_{0i} \sim N(5.1696, 0.3)$, $b_1 = -0.2/365$, $BM_i(t)$ is a Brownian motion stochastic process and $\sigma^2 = 0.05/365$. The auxiliary variable is generated at time 0 and every 3 months for a total duration of 2 years and a maximum of 8 measurements. The event time T_i is generated from the model $\lambda_f(t) = \phi_0 exp\{\phi_1 l_i(t)\}$ and the random censoring time C_i is

generated from the model $\lambda_c(t) = \psi_0 exp\{\psi_1 l_i(t)\}$, where $l_i(t) = Z_i(t)$ for the RE model and $l_i(t) = Z_i(t)^2$ for the BM model. When simulating independent censoring, $\phi_0 = 0.0008 * e^{-0.15}$, $\phi_1 = -1.5$, $\psi_0 = e^{-6}$ and $\psi_1 = 0.0$ for the RE model and $\phi_0 = 5 * 10^8$, $\phi_1 = -1.0$, $\psi_0 = 0.0025$ and $\psi_1 = 0.0$ for the BM model. With dependent censoring, $\phi_0 = 0.0008 * e^{-0.15}$, $\phi_1 = -1.5$, $\psi_0 = e^{-6}$ and $\psi_1 = -0.5$ for the RE model and $\phi_0 = 8$, $\phi_1 = -0.3$, $\psi_0 = 1.0$ and $\psi_1 = -0.2$ for the BM model. There are 300 subjects generated. The latest observed covariate before each censored time is the only auxiliary variable used to define the nearest neighborhood.

We note that for the RE model the conditions necessary for the KMI method to eliminate bias are not satisfied, in particular there is not a single linear combination of current and past covariate values which determines the distribution of the future failure time. The reason for this is because the hazard depends on just the current value of the time-dependent variable, however both the current value and the slope are needed to determine the future values of the time-dependent auxiliary variable. For the BM model the conditions are satisfied because the current value contains all the information about both the future values of the time-dependent variable and the hazard of the failure time, and hence of the distribution of future event times.

5.2 *Imputation and Analysis*

For the "Fully-Observed" (FO) analysis (the gold standard), we apply KM estimation to each generated data set before any censoring is applied. For the "Partially-Observed" (PO) analysis, we apply KM estimation to each data set with random censoring. For the multiple imputation methods, for each simulated data set, we multiply impute future event or censored times for each observed censored time using auxiliary variables as described in Section 2. We compute KM estimates for each augmented data set and perform the multiple imputation analysis. We focus on $S(t)$ at a fixed time point t chosen so that the true $S(t)$ is equal to, or close to, 0.5.

5.3 *Results*

5.3.1 Binary time-independent covariate. In Table 2, we display the WKM, IPCW and multiple imputation estimates. In independent censoring cases, all the point estimates target the quantile correctly (almost identical to FO estimates). The SD and SE values for the PO method are higher than those of other methods. For example, the KMIB method gains about 5% efficiency compared to the PO method, and by comparing with the SD values for FO and PO we see that the methods recover about 1/3 of the information lost due to censoring. In addition, the KMI and RSI methods yield almost identical estimates and standard deviations as the WKM method and the IPCW method. The coverage rates are close to the nominal level. For dependent censoring, the KMI, KMIB, WKM and IPCW methods produce almost identical estimates as the FO estimates. However, the PO method yields biased survival estimates and the RSI and RSIB reduce but do not eliminate the bias.

5.3.2 Multiple time-independent covariates. Several different scenarios with continuous covariates were examined. In the cases of independent censoring (results not shown) all the imputation methods do not have bias, have improved efficiency compared to PO when the failure time model is correctly specified and have coverage rates close to the nominal level. The case of dependent censoring is more challenging, we show only the results for KMIB and IPCW because we find that RSI and RSIB are less good at reducing the bias. In Table 3 we show the impact of the size of the nearest neighbourhood (NN). The best choice of NN appears to be 10. The KMIB method substantially reduces the bias, but does not eliminate it. The coverage rates are close to the nominal level. Table 3 also shows the impact of changing the weights. As expected reducing the emphasis on the failure time model and increasing the emphasis on the censoring time model reduces the bias in cases where the working failure time model is misspecified and increases it when the censoring time model is misspecified. The

coverage rates are adequate and a reasonable choice for w_f would be 0.8 or 0.5. For KMIB, when the working failure time model is correctly specified, the bias is smaller compared to the situation with the working failure time model misspecified. Compared to the IPCW method, the KMIB estimator has greater bias but less variability. The squared difference ratio (SDR) column measures the squared difference of the estimate from the FO estimate for each dataset, relative to the PO method; thus higher numbers indicate a better estimator. It shows that both KMIB and IPCW give substantially closer estimates to the best (FO) estimate than the PO method, and that KMIB is closer to FO than IPCW is to FO when the failure time model is correctly specified and more weight is placed on this model.

Results for the effect of sample size on the bias of KMIB are provided in Table 4. As the sample size increases, the bias slowly decreases and the coverage rate improves.

5.3.3 Time-dependent covariates. Table 5 shows the results for KMIB and IPCW for both the random effects and the Brownian motion generation scheme. For the random effects model with independent censoring, KMIB and IPCW have similar properties. There is no bias and a slight gain in efficiency compared to PO and the coverage rates are good. For the Brownian motion model with independent censoring the KMIB has no bias, but the IPCW introduces bias for reasons we don't understand. With dependent censoring, for both the random effects and Brownian motion models the KMIB substantially reduces but does not eliminate the bias. The IPCW method is less successful at reducing the bias.

The bias from imputation methods is not eliminated completely because the time-dependent auxiliary variable is periodically measured, thus not necessarily at a time point close to the censored time point. We investigated this in a further simulation (results not shown) and found that the bias is reduced but not eliminated in finite samples if the auxiliary variables are measured much more frequently.

6. Discussion

The research in this paper provides a direct, simple and transparent approach, nonparametric multiple imputation, to using auxiliary variables to recover information for censored observations. The simulation study shows that the use of this multiple imputation method can lead to improved performance of estimators. In general, the multiple imputation point estimates are less variable and closer to the truth than the estimates produced by analyzing the observed data without using the auxiliary variables.

Of the imputation schemes, RSI removes less bias, if there is bias due to dependent censoring, compared to KMI. One feature of RSI is that both censored event times and uncensored values appearing in the imputing risk set are likely to be chosen for the newly imputed event times. KMI imputes censored values from the risk set with positive probability only if the longest observation in the imputing risk set is censored. Hence, after imputation, RSI will typically produce more censored observations in the augmented data set compared to KMI. The censoring that remains, although less prevalent, may still be informative and hence survival estimates based on the augmented data set may still be biased. Therefore, KMI is preferred to RSI. The major reason for the remaining bias in the KMI method in the case of dependent censoring is the sample size. In particular the nearest neighborhood contains some observations that are not close enough to the target value, so some remnants of dependent censoring remain within the neighborhood. This is likely to be more of a problem with high dimensional covariates compared to cases with less than say 5 auxiliary variables. An additional complication with high dimensional covariates is that it will be hard to obtain good estimates of the coefficients in the working models with many covariates, making it even harder to define a nearest neighborhood that is truely close to the target value.

Theoretical results for large samples indicate that the risk sets allow for good estimation of the imputation distribution of interest, even with dependent censoring. Numerical results

indicate that when the working model for the event time is misspecified the bias is greater than when it is correctly specified. The use of two working models does lead to a reduction in bias, from the double robustness property. One reason that a small bias remains is that the imputed dataset can contain a small number of censored observations. Another reason is that the nearest neighbors are not being chosen with enough precision. Both these sources of bias will diminish with increased sample size, however the rate of improvement is slow. We also observed more gains in efficiency when the failure time model is correctly specified. Thus, although double robustness is a very useful property, it should not be used as a replacement for trying to find reasonable fitting working models for both the failure time and the censoring time, rather it should be used in addition to seeking good models for the observed data.

The comparison between the KMIB method and the IPCW method showed that KMIB was not quite as effective as the IPCW method at reducing bias due to dependent censoring when there were multiple baseline covariates. KMIB appeared to be more efficient than IPCW in the application, but the difference in efficiency were smaller in the simulations. The results in Table 3 indicate that the KMIB method, with appropriae of weights, gives estimates closer to the best value, than the estimates from the IPCW method.

With time dependent covariates and dependent censoring the IPCW method was not as effective as reducing the bias as KMIB. This may be because the estimator depends crucially on the coefficients fit to the model of the censoring data. We found in the simulation that these coefficients were attenuated towards zero due to the fact that that the time dependent auxiliary variable was only measured at discrete times, rather than being measured continously.

An attractive aspect of the KMI procedure is its weak reliance on a statistical model, because the model is only used to identify a nearest neighborhood. Once this neighborhood is defined, the residual time distribution is obtained using nonparametric methods. Then the imputation is conducted on this estimated residual time distribution for censored observations.

After the imputation the analysis is based on the original data, augmented by the imputed data. As a result, the reliance on the statistical model is weak and any gains in efficiency or reduction in bias are derived mainly from the data, rather than from assumptions in the models.

In addition to its robustness in this application, the general approach of multiple imputation has features that make it attractive. One such feature is that after imputation the data analyst can perform other analyses appropriate for the goals of their study. A second attractive feature is that there is a standard way to obtain measures of uncertainty. Published work (Robins and Wang (2000)) indicates that the standard multiple imputation variance formula is only valid in simple situations, however the results in our simulations where the SE's were close to the SD's indicate that the formula is working well in our more complex situations.

The adequacy of imputation procedures will depend on the "nearness" of the imputing risk set and on the availability of possible donor observations, which diminishes in the tails of the survival distribution. The "nearness" of the imputing risk set will depend on the quality of the parameter estimates from the two working models. In situations where the working models are refit for every censored observation, the parameter estimates could be improved by assuming that they vary smoothly with time.

In this paper, we fix the size of the nearest neighborhood. Future research could employ a dynamic scheme to select the size of the nearest neighborhood dependent on the time of the censored observation. There other possible adaptations that might improve the KMIB method. For example, rather than equally weighting all the observations in the nearest neighborhood, one could give more weight to close observations. Instead of using the Kaplan-Meier estimate to summarize the residual time distribution in the imputing risk set, one could use a smoother estimate. By fitting two working models, one is essentially conditioning on two linear combinations of available covariates. In the case of a time-dependent auxiliary variable one

could condition on an additional linear combination designed to summarize the distribution of the possible future values of the longitudinal variable.

References

Andersen, P. K., Borgan, O. Gill, R. D. and Keiding, N. (1993). *Statistical Models based on Counting Processes*. Springer, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study (com: P1121-1124). *The Annals of Statistics* **10**, 1100–1120.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-meier estimate. *The Annals of Statistics* **17**, 1157–1167.

Faucett, C. L., Schenker, N. and Taylor, J. M. G. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58**, 37–47.

Finkelstein, D. M. and Schoenfeld, D. A. (1994). Analysing survival in the presence of an auxiliary variable. *Statistics in Medicine* **13**, 1747–1754.

Fleming, T. R., Prentice, R. L., Pepe, M. S. and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13**, 955–968.

Gray, R. J. (1994). A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* **81**, 527–539.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81**, 701–708.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics* **40**, 13–29.

Hubbard, A., van der Laan, M. J. and Robins, J. M. (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and

covariates in observational studies. In *Statistical Models in Epidemiology, the Environment and Clinical Trials, Ed. E. Halloran and D. Berry*, pages 135–177. Springer-Verlag: New York.

Jewell, N. P. and Kalbfleisch, J. (1996). Marker processes in survival analysis. *Lifetime Data Analysis* **2**, 15–29.

Malani, H. M. (1995). A modification of the redistribution to the right algorithm using disease markers. *Biometrika* **82**, 515–526.

Murray, S. and Tsiatis, A. A. (1996). Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* **52**, 137–151.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.

Robins, J. and Wang, N. (2000). Inference in imputation estimators. *Biometrika* **87**, 113–124.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests". *Biometrics* **56**, 779–788.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues, Ed. N. Jewell, K. Dietz and V. Farewell*, pages 297–331. Boston: Birkhauser.

Robins, J. M., Rotnitzky, A. and van der Laan, M. (2000). Comment on "on profile likelihood". *Journal of the American Statistical Association* **95**, 477–482.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* **10**, 585–598.

Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* **22**, 425–446.

Solomon, P. J. (1984). Effect of misspecification of regression models in the analysis of survival data (corr: V73 p245). *Biometrika* **71**, 291–298.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

Taylor, J. M. G., Murray, S. and Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics & Probability Letters* **58**, 221–232.

Volberding, P. A., Lagakos, S. W., Koch, M. A. and et al. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* **322**, 941–949.

Appendix

Relationship between KMI and WKM estimates

Proof of Result 1:

Consider time t for which there are no censored observations before t in each KMI imputed data set after imputation. For the m^{th} KMI imputed dataset, $m = 1, ..., M$, the KM survival estimator can be written as

$$\begin{aligned}\hat{S}^m_{KMI}(t) &= \sum_{i=1}^{n} \frac{\mathbf{I}(\text{patient } i \text{ is alive at } t \text{ in the } m^{th} \text{ imputed dataset})}{n} \\ &= \sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{\mathbf{I}(\text{patient } j \text{ with categorical variable } k \text{ is alive at } t \text{ in } m^{th} \text{ imputed dataset})}{n_k}\frac{n_k}{n} \\ &= \sum_{k=1}^{K} \hat{S}^m_k(t)\frac{n_k}{n},\end{aligned}$$

where $\hat{S}^m_k(.)$ is the KM survival estimate conditional on having covariate value k in the m^{th} KMI imputed dataset. Taylor et al. (2002) have shown that $E\{\hat{S}^m_k(t)|\mathbf{Y}\} = \hat{S}_k(t)$, $m = 1, ..., M$, where the expectation is with respect to the distribution of possible imputes conditional on $\mathbf{Y}$. The final survival estimator derived from M imputed datasets is

$\hat{S}_{KMI}(t) = \frac{1}{M}\sum_{m=1}^{M}\{\hat{S}^m_{KMI}(t)\} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\{\hat{S}^m_k(t)\frac{n_k}{n}\}$
and $E\{\hat{S}_{KMI}(t)|\mathbf{Y}\} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}[E\{\hat{S}^m_k(t)|\mathbf{Y}\}\frac{n_k}{n}] = \sum_{k=1}^{K}\{\hat{S}_k(t)\frac{n_k}{n}\} = WKM(t)$. Note that if t is after the the first censored time in the imputed dataset, $\hat{S}^m_{KMI}(t)$ can't be written as a sum and the first equality in the proof does not hold.

Properties of KMI Imputation Methods

Properties of probability models.

Consider first the case of q time-independent covariates, denoted by $\mathbf{Z}$. Let T and C denote the random variables for the event and censoring time respectively. Assume T and C are independent given $\mathbf{Z}$, i.e.

$$Pr(T > t_1, C > t_2|\mathbf{Z}) = Pr(T > t_1|\mathbf{Z})Pr(C > t_2|\mathbf{Z}) \tag{A.1}$$

for all t_1, t_2. Let $\lambda_T(t|\mathbf{Z}, C \geq t)$ denote the intensity of the counting process for the failure time T. Note that equation A.1 implies that $\lambda_T(t|\mathbf{Z}, C \geq t) = \lambda_T(t|\mathbf{Z})$, which is the usual assumption for independent censoring conditional on $\mathbf{Z}$ (Andersen et al. (1993)). Let $\mathbf{Z}_f$ and $\mathbf{Z}_c$ be vectors of covariates for the true failure-time and the true censoring-time models, respectively. These models are the true models in the sense that if they hold, then all the information in $\mathbf{Z}$ at time t about the distribution of T given $T > t$ is contained in linear combinations of $\mathbf{Z}_f$, with a similar statement for the distribution of C. Thus $Pr(T > t_1|\mathbf{Z}, T > t) = Pr(T > t_1|\beta_f\mathbf{Z}_f, T > t))$ and $Pr(C > t_2|\mathbf{Z}, C > t) = Pr(C > t_2|\beta_c\mathbf{Z}_c, C > t))$. Furthermore, we will specify these models to be proportional hazards of the form $\lambda_{0f}(t)exp(\beta_f\mathbf{Z}_f)$ and $\lambda_{0c}(t)exp(\beta_c\mathbf{Z}_c)$.

In practice, these two models may not hold, or may be misspecified. We define a working failure time model, and a working censoring time model based on covariates $\mathbf{Z}^*_f$ and $\mathbf{Z}^*_c$. These models have hazards of the form $\lambda_{0wf}(t)exp(\beta^*_f\mathbf{Z}^*_f)$ and $\lambda_{0wc}(t)exp(\beta^*_c\mathbf{Z}^*_c)$. Either of the two working models could be misspecified, examples of model misspecification include choosing

the wrong covariates, the wrong functional form of the covariates, e.g. using $\sqrt{Z_1}$ instead of Z_1, and the incorrect link function.

Here β_f (β_c) and β_f^* (β_c^*) denote the regression parameters for the true failure (censoring) time model and the working PH failure (censoring) time model, respectively. The parameter β_f^* is defined as the large sample limit of the estimate of β_f^* when the working model is fit in the absense of censoring, with a similar definition for β_c^*. Let $RS^* = (RS_f^*, RS_c^*) = (\beta_f^* \mathbf{Z}_f^*, \beta_c^* \mathbf{Z}_c^*)$, be the risk scores associated with the two working models.

Result 2. If either the working failure time model or the working censoring time model is correct then T and C are independent conditional on RS^*.

Proof: First, we consider the situation where the working censoring model is correct, i.e. $Pr(C > t_2|\mathbf{Z}) = Pr(C > t_2|\beta_c \mathbf{Z}_c)$, $\mathbf{Z_c^*} = \mathbf{Z}_c$ and $\beta_c^* = \beta_c$, and the working failure time model is incorrect. We further assume the parameter values are known. Hence, the censoring time C depends on $\mathbf{Z}$ only through $\beta_c \mathbf{Z}$ ($= \beta_c^* \mathbf{Z}_c^*$). Then we have

$$Pr(T > t_1, C > t_2|RS^*) = E_{Z|RS^*}\left[E\{I(T > t_1, C > t_2)|\mathbf{Z}, RS^*\}\right]$$

$$= E_{Z|RS^*}\left[Pr(T > t_1|\mathbf{Z}) Pr(C > t_2|\mathbf{Z})\right] = E_{Z|RS^*}\left[Pr(T > t_1|\mathbf{Z}) Pr(C > t_2|\beta_c^* \mathbf{Z}_c^*)\right]$$

$$= E_{Z|RS^*}\left[Pr(T > t_1|\mathbf{Z})\right] Pr(C > t_2|RS^*) = Pr(T > t_1|RS^*) Pr(C > t_2|RS^*)$$

The second equality holds because of the assumption of conditional independence given $\mathbf{Z}$, the third equality holds because of the assumption that the working censoring time model is correct. The proof for the other situation, where the failure time model is correctly specified, is similar. Thus we have shown if one of two working PH models is correct, T and C are independent conditional on two risk scores (RS^*). It further follows that conditioning on follow-up time, the residual time distribution of T and C are independent given RS^*, i.e. $Pr(T > t_1, C > t_2|RS^*, T \geq t, C \geq t) = Pr(T > t_1|RS^*, T \geq t) Pr(C > t_2|RS^*, C \geq t)$

A consequence of this result is that everyone who is censored at t has the same residual failure time as those people with the same risk score who are still at risk at t, ie censoring does

not add information about the future of the failure time process. This is a key property for the multiple imputation schemes in this paper, because it allows us to extend the follow-up for each censored observation with information derived from the imputing risk set. It implies that a Kaplan-Meier estimate of $Pr(T > t_1|RS^*, T > t)$, for $t_1 > t$, derived from people with the same risk score is a valid estimate, not biased by dependent censoring.

Note that if the failure time model is correctly specified then $Pr(T > t_1|\mathbf{Z}, T > t) = Pr(T > t_1|\beta_f \mathbf{Z}_f, T > t))$, i.e. all the information about the future distribution of T is contained in RS^*, and we would expect the estimator to be efficient. If the failure time model is incorrectly specified, then in general $Pr(T > t_1|\mathbf{Z}, T > t) \neq Pr(T > t_1|\beta_f \mathbf{Z}_f, T > t)$, thus some of the information about the distribution of T has been lost by conditioning on RS^*, this is likely to lead to less efficient estimators. We also note that if T and C are unconditionally independent, they will be independent given RS^* and hence the Kaplan-Meier estimator of $Pr(T > t_1|RS^*, T > t)$ is valid.

Properties of estimates.

Let $\hat{\beta}_f^*$ and $\hat{\beta}_c^*$ be the estimates from the two working models obtained by maximising the partial likelihood function of the two working models, and let $\widehat{RS}^* = (\hat{\beta}_f^* \mathbf{Z}_f^*, \hat{\beta}_c^* \mathbf{Z}_c^*)$. When the multiple imputation estimators are applied to finite sample size data, a number of other issues arise. One is that we know $\widehat{RS}^*$ rather than RS^*. A second is that RS_f^* and RS_c^* (or $\widehat{RS}_f^*$ and $\widehat{RS}_c^*$) can take a continuum of possible values, assuming that Z contains continuous rather than all categorical variables. Thus for each censored observation there is unlikely to be any other observations with identical values of RS^*, forcing us to use a neighborhood with similar, but not identical values of RS^*. The neighborhood will have to be small to minimize bias, but also large enough to keep variability low. We will provide a heuristic argument that suggests the bias of $\hat{S}_{KMI}(t)$ will be small in large samples.

A consequence of the result that T and C are independent conditional on RS^* if one of the

two working models is correct is that (Andersen and Gill (1982)) $\hat{\beta}_f^* \xrightarrow{pr} \beta_f^*$ and $\hat{\beta}_c^* \xrightarrow{pr} \beta_c^*$. Since $Pr(T > t_1, C > t_2|RS^*)$, $Pr(T > t_1|RS^*)$, and $Pr(C > t_2|RS^*)$ are continuous functions of RS^*, then by the continuous mapping theorem T and C are asymptotically independent conditional on $\hat{\beta}_f^* \mathbf{Z}_f^*$ and $\hat{\beta}_c^* \mathbf{Z}_c^*$.

For considering the large sample bias of $E\{\hat{S}_{KMI}(t)|\mathbf{Y}\}$ we start by assuming that both β_f^* and β_c^* known. Assume RS_f^* and RS_c^* are contained in the finite rectangle I in R^2. We then partition I into several (K) disjoint squares with q_n being the width of each square. This partition defines a categorical variable, I_{RS^*}, taking values 1 to K. The values of the risk scores within the same category are defined to be equal. In large samples, for each censored observation the imputation method uses an imputing risk set, all of which will have the same value of I_{RS^*}. As $n \to \infty$ we let $q_n \to 0$ and $K \to \infty$, in such a way that $nq_n \to \infty$, to ensure a large sample within each category. We noted earlier that $\hat{S}_{KMI}(t)$, with large M, was equivalent to WKM when there are no censored observations before time t in each KMI imputed data set. Using this result, the expression for the expectation of KMI conditional on the observed data $\mathbf{Y_n}$ can be rewritten as

$$E\{\hat{S}_{KMI}(t)|\mathbf{Y_n}\} = \sum_{k=1}^{K} \left[\hat{S}(t|I_{RS^*} = k) Pr(I_{RS^*} = k|\mathbf{Y_n})\right],$$

where $\hat{S}(t|I_{RS^*} = k)$ is the conditional KM estimator at time t among those in category k and $Pr(I_{RS^*} = k|\mathbf{Y_n})$ is the sample proportion, n_k/n, in category k. The above expression is analagous to a uniform kernel conditional KM estimate in Dabrowska (1989). Dabrowska shows that under certain conditions and assumptions $\hat{S}$, the kernel conditional KM estimate, is uniformly consistent. We appeal to this result and the previously shown result that T and C are independent conditional on RS^* to show that

$$|| \sum_{k=1}^{K} \left[\hat{S}(t|I_{RS^*} = k) Pr(I_{RS^*} = k|\mathbf{Y_n})\right] - \int S(t|RS^* = u) d\{F_{RS^*}(u)\} ||_I \longrightarrow 0,$$

where $||S(t|RS^*)||_I = sup\{|S(t, RS^*)| : RS^* \in I\}$, $S(t|RS^*)$ is the true survival rate given RS^* and $F_{RS^*}(u)$ is the distribution of RS^* on I. The marginal survival function $S(t)$ can be written as $\int S(t|RS^* = u)d\{F_{RS^*}(u)\}$, thus if the conditions are satisfied $\hat{S}_{KMI}(t)$ is a consistent estimate of $S(t)$, giving Result 3 when both β_f^* and β_c^* known. The same method can be applied if there is a single continuous covariate.

We note however that this argument only applies if there are no censored values in the imputed datasets less than t. In practice this is unlikely to be the case except for small t; depending on the value of t there are likely to be a small fraction of imputed censored values less than t. This suggests that in large samples the bias of $\hat{S}_{KMI}(t)$ will be small, but maybe not zero.

We now consider the case with both β_f^* and β_c^* unknown. The estimated risk scores are $\widehat{RS}^* = (\hat{\beta}_f^* \mathbf{Z}_f^*, \hat{\beta}_c^* \mathbf{Z}_c^*)$. Now the categorical indicator variable $I_{\widehat{RS}^*}$ based on the estimated risk scores differs from that based on RS^*. However, under the assumption that one of the two working models is correct, $\hat{\beta}_f^* \xrightarrow{pr} \beta_f^*$ and $\hat{\beta}_c^* \xrightarrow{pr} \beta_c^*$, and, conditional on $\widehat{RS}^*$, T and C are asymptotically independent. If $q_n \to 0$, as $n \to \infty$ in such a way that $\sqrt{n}q_n \to \infty$, then heuristically, it can be seen that as the sample size increases the number of observations in a category increases to infinity as the size of the neighborhood gets smaller and that the difference between $\widehat{RS}^*$ and RS^* is of smaller order than the the size of the neighborhood.

With a single continuous covariate $\widehat{RS}^* \equiv RS^*$, the working models and the regression coefficients $\hat{\beta}_f$ and $\hat{\beta}_c$ are not needed, so the fact that β_f^* and β_c^* are unknown is irrelevant.

Time-dependent covariates

Let Z_t denote all the information that is available up to time t, which includes current and previous values of Z and whether T and C are bigger than t. The key assumption is that T and C are conditionally independent given Z_t, ie $Pr(T > t_1, C > t_2|Z_t) = Pr(T > t_1|Z_t)Pr(C > t_2|Z_t)$, for all $t_1 > t$ and $t_2 > t$. In the multiple imputation method the working

failure time and censoring time models are refit at each time of a censored observation, thus the risk scores depend on time and are denoted by $RS_t^* = (RS_{ft}^*, RS_{ct}^*)$. The same proof as before leads to result 2 that T and C being conditionally independent given RS_t^*, if one of the two working models is correct, ie $Pr(T > t_1, C > t_2|RS_t^*) = Pr(T > t_1|RS_t^*)Pr(C > t_2|RS_t^*)$, $t_1 > t$, $t_2 > t$. Hence the Kaplan-Meier estimate derived from each imputing risk set is a valid estimate, not biased by dependent censoring, and can be used for imputing. Whether the method of Dabrowska can be extended to the case of time-dependent covariates, to prove the double robustness in result 3, is unclear. The statement that the working model is correctly specified is a stronger statement than in the case of time-independent covariates, because the distribution of the residual time depends on both the hazard of the event and the stochastic process for future values of Z. Thus the assumption is that information about both of these can be summarized in a single linear combination.

[Figure 1 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

Figure 1. Comparison of KM curves based on the partially-observed data (No Imputation) and based on KMIB method using the latest CD4 count as the auxiliary variable.

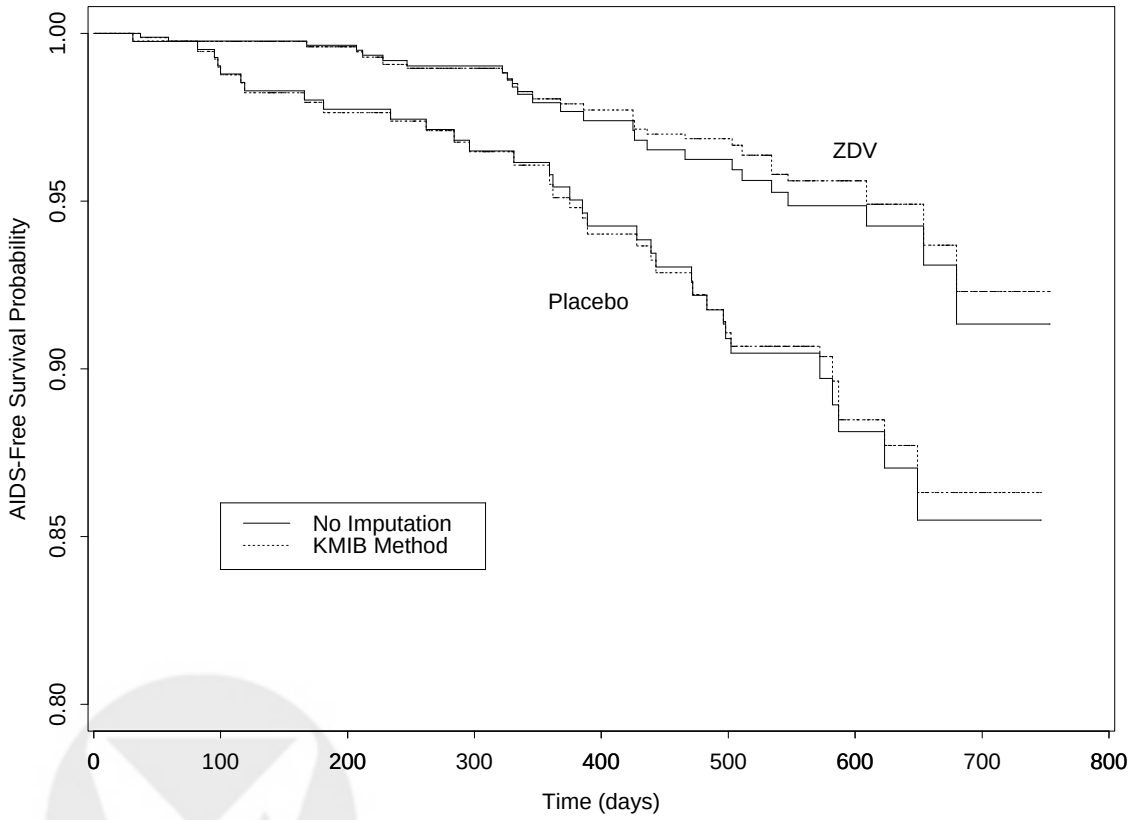

Table 1

Estimates of AIDS-free survival probabilities using the latest observed CD4 counts as covariates based on partially-observed data (PO), IPCW method and multiply-imputed data with M=10 and NN=10.

		Treated		
Method	$\hat{S}(450)^a$	$(\hat{SE}_{450})^b$	$\hat{S}(550)$	$(\hat{SE}_{550})$
PO	0.965	0.0087	0.949	0.0113
IPCW	0.964	0.0090	0.946	0.0118
KMI	0.970	0.0064	0.951	0.0081
RSI	0.965	0.0077	0.952	0.0098
KMIB	0.968	0.0068	0.951	0.0093
RSIB	0.968	0.0067	0.952	0.0108
		Placebo		
Method	$\hat{S}(450)$	$(\hat{SE}_{450})$	$\hat{S}(550)$	$(\hat{SE}_{550})$
PO	0.930	0.0147	0.905	0.0176
IPCW	0.929	0.0149	0.902	0.0180
KMI	0.932	0.0137	0.907	0.0171
RSI	0.929	0.0142	0.901	0.0166
KMIB	0.931	0.0140	0.901	0.0165
RSIB	0.933	0.0142	0.904	0.0173

[a] KM survival estimate of remaining AIDS-free at day 450.
[b] based on Greenwood's formula.

Table 2

Monte Carlo Results for a binary covariate: Survival estimates. The total sample size is 80, 40 for each group. Results based on 500 replications and M=50.

Independent Censoring; Censoring rate:0.47 Event times ~ Exp(1.0/0.1); Censoring times ~ Exp(0.28)					
Method	true value	average	SD^a	SE^b	CR^c
FO	0.50	0.497	0.0546	0.0556	95.8
PO		0.499	0.0633	0.0631	94.0
WKM		0.498	0.0600	0.0606	95.2
IPCW		0.498	0.0605	0.0595	95.0
RSI		0.498	0.0603	0.0599	94.8
KMI		0.497	0.0601	0.0590	94.8
RSIB		0.498	0.0600	0.0611	95.0
KMIB		0.497	0.0604	0.0606	95.0
Dependent Censoring; Censoring rate:0.50 Event times ~ Exp(1.0/0.1); Censoring times ~ Exp(0.5/0.2)					
Method	true value	average	SD	SE	CR
FO	0.50	0.497	0.0558	0.0555	95.2
PO		0.535	0.0645	0.0632	90.6
WKM		0.498	0.0651	0.0625	95.0
IPCW		0.497	0.0656	0.0611	93.2
RSI		0.508	0.0642	0.0607	93.6
KMI		0.498	0.0652	0.0594	93.4
RSIB		0.508	0.0639	0.0627	94.4
KMIB		0.498	0.0651	0.0626	95.0

[a] empirical standard deviation.

[b] estimated standard error based on Greenwood's formula.

[c] coverage rate of 95% confidence interval calculated as estimate +- $t_\nu^{(0.975)}$ standard error.

Table 3

Monte Carlo results for 5 time-independent covariates with dependent censoring:: the effects of the size of the nearest neighborhood and weights (w_f, w_c) on survival estimates (true value=0.5). Censoring Rate: 0.51. Sample size 200, 500 replications, M=10.

Failure[1]: correct; Censoring[2]: correct						
Method	NN	average	SD	SE	CR	SDR[3]
FO		0.501	0.0332	0.0353	96.4	
PO		0.568	0.0370	0.0395	58.6	1.00
IPCW		0.504	0.0421	0.0405	93.8	7.65
$(w_f, w_c) = (1.0, 0.0)$						
KMIB	5	0.511	0.0385	0.0406	95.4	10.00
KMIB	10	0.509	0.0385	0.0405	94.6	10.32
KMIB	20	0.511	0.0390	0.0406	95.2	9.80
KMIB	50	0.524	0.0392	0.0409	92.2	5.14
Failure: correct; Censoring: correct, NN=10						
Method	weights	average	SD	SE	CR	SDR
FO		0.494	0.0367	0.0353	94.4	
PO		0.562	0.0400	0.0396	62.0	1.00
IPCW		0.498	0.0439	0.0402	91.6	7.88
KMIB	(1.0,0.0)	0.502	0.0413	0.0403	95.2	9.94
KMIB	(0.8,0.2)	0.502	0.0410	0.0405	94.8	10.23
KMIB	(0.5,0.5)	0.503	0.0411	0.0405	94.8	9.71
KMIB	(0.0,1.0)	0.506	0.0413	0.0409	95.2	8.44
Failure: incorrect; Censoring: correct, NN=10						
FO		0.502	0.0361	0.0353	93.6	
PO		0.569	0.0397	0.0394	58.4	1.00
IPCW		0.505	0.0438	0.0401	93.4	6.88
KMIB	(1.0,0.0)	0.539	0.0411	0.0409	83.2	2.65
KMIB	(0.8,0.2)	0.521	0.0416	0.0407	91.0	6.21
KMIB	(0.5,0.5)	0.517	0.0405	0.0406	92.4	7.42
KMIB	(0.0,1.0)	0.514	0.0409	0.0409	93.6	8.45
Failure: correct; Censoring: incorrect, NN=10						
FO		0.497	0.0360	0.0353	94.0	
PO		0.565	0.0404	0.0395	61.2	1.00
IPCW		0.501	0.0443	0.0402	92.4	7.96
KMIB	(1.0,0.0)	0.505	0.0418	0.0403	93.4	10.51
KMIB	(0.8,0.2)	0.507	0.0421	0.0405	93.2	9.49
KMIB	(0.5,0.5)	0.510	0.0427	0.0405	92.4	8.24
KMIB	(0.0,1.0)	0.536	0.0428	0.0412	82.6	2.50

[1] working failure time model,[2] working censoring time model
[3] squared difference ratio = $\Sigma(PO - FO)^2/\Sigma(est - FO)^2$

Table 4

Monte Carlo results with 5 time-independent covariates with dependent censoring and the working failure time model incorrectly specified: the effects of sample size on survival estimates (true value=0.5). Censoring Rate: 0.51. Results based on 500 replications, $(w_f = 0.8, w_c = 0.2)$, NN=10, and M=10.

Method	sample size	average	SD	SE	CR
FO	200	0.499	0.0371	0.0353	93.2
PO	200	0.565	0.0407	0.0395	63.0
KMIB	200	0.516	0.0416	0.0407	90.8
FO	400	0.499	0.0238	0.0250	96.2
PO	400	0.566	0.0267	0.0279	33.8
KMIB	400	0.513	0.0283	0.0287	92.8
FO	800	0.499	0.0176	0.0177	95.0
PO	800	0.566	0.0194	0.0198	9.4
KMIB	800	0.509	0.0200	0.0204	93.2
FO	2000	0.499	0.0109	0.0112	95.4
PO	2000	0.566	0.0119	0.0125	0.0
KMIB	2000	0.506	0.0122	0.0129	95.3

Table 5

Monte Carlo results for time-dependent covariates. Results based on sample size 300, 500 replications, $NN = 10$, and M=10. (a) Random effects model, Survival estimates at one year (b) Brownian motion model, Survival estimates at six months.

(a) Random Effects Model				
Method	average	SD	SE	CR*
Independent censoring; Censoring rate:51%				
FO	0.445	0.0281	0.0286	96.8
PO	0.445	0.0350	0.0351	95.0
IPCW	0.445	0.0335	0.0336	94.6
KMIB	0.438	0.0339	0.0333	94.0
Dependent censoring; Censoring rate:57%				
FO	0.448	0.0302	0.0287	93.4
PO	0.543	0.0336	0.0325	17.8
IPCW	0.495	0.0369	0.0348	69.6
KMIB	0.466	0.0360	0.0340	89.6
(b) Brownian Motion Model				
Method	average	SD	SE	CR*
Independent censoring; Censoring rate:46%				
FO	0.514	0.0291	0.0288	94.4
PO	0.514	0.0307	0.0307	94.4
IPCW	0.537	0.0293	0.0288	86.6
KMIB	0.513	0.0301	0.0300	94.6
Dependent censoring; Censoring rate:62%				
FO	0.519	0.0291	0.0288	94.2
PO	0.615	0.0348	0.0331	19.2
IPCW	0.557	0.0368	0.0374	82.4
KMIB	0.535	0.0376	0.0362	91.6