STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3503–3517
Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2452

Survival analysis using auxiliary variables via non-parametric multiple imputation

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

[1]*Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, Tucson, AZ 85724-5024, U.S.A.*
[2]*Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
[3]*INSERM E0338 Biostatistics, ISPED, Bordeaux 2 University, Bordeaux 33000, France*

SUMMARY

We 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 neighbour imputing risk set. One model is for the event times and the other for the censoring times. Based on the imputing risk set, two non-parametric 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 mis-specification 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. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: double robustness; multiple imputation; nearest neighbour

1. INTRODUCTION

In survival analysis, the event times for censored observations can be regarded as missing data [1]; in this sense there is a loss of information due to censoring. In many studies, there

*Correspondence to: C.-H. Hsu, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, 1515 N Campbell, PO Box 245024, Tucson, AZ 85724-5024, U.S.A.
†E-mail: phsu@azcc.arizona.edu

Received 13 June 2005
Accepted 3 October 2005

is other information obtained about subjects, and such data may be informative about their health condition, e.g. CD4 count in AIDS studies. These markers are often associated with the event times and 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. Incorporating information from auxiliary variables has the potential to reduce this bias, as well as improve efficiency. There are an increasing number of statistical methods [2–9] 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 [10]. 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 [2, 8, 9]. 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. 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 efficiency of 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 [11]. Taylor *et al.* [12] treated censored data as missing data and then used multiple imputation techniques to develop non-parametric procedures to impute missing event times in the situation of one-sample survival estimation without auxiliary variables. They showed that with a large number of imputes the estimates from these multiple imputation methods reproduce the Kaplan–Meier (KM) estimator. This provides a theoretical basis for investigating the use of non-parametric multiple imputation strategies in more complex situations, where the imputation strategy could depend on auxiliary variables. For example, Schenker and Taylor [13] employed three auxiliary variables to determine the distribution of residual times from which to impute a time of AIDS. Faucett *et al.* [14] 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 neighbourhood of similar observations for each censored case, and then generate imputes from this set of neighbours.

The paper is organized as follows. In Section 2, we describe two non-parametric 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

2.1. Calculating risk scores

Let $\{t_1,\ldots,t_n\}$ 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 $\mathbf{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ denote the values of the auxiliary variables and let $\mathbf{Y}=\{(t_1,\delta_1,\mathbf{z}_1),\ldots,(t_n,\delta_n,\mathbf{z}_n)\}$.

For each censored observation we seek an imputing risk set consisting of subjects who are similar to the censored case. To define each 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. This idea is analogous to predictive mean matching [15]. 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 mis-specified, 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. This idea is analogous to propensity score matching [16]. 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 mis-specification can be established [17]. 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 $\mathbf{Z}$. This model could include interactions between $\mathbf{Z}$'s or transformation of each $\mathbf{Z}$. In the case with time-independent auxiliary variables, we directly use all auxiliary variables as covariates to fit these two PH models. In the case with time-dependent auxiliary variables, we propose using a subset (e.g. baseline and latest observed values) of the measurements of the time-dependent auxiliary variable up to the censored time we want to impute as the covariates. Once the covariates, $\mathbf{Z}$, are chosen, the risk scores can be defined as $\widehat{RS}_{\mathrm{f}}=\hat{\boldsymbol{\beta}}_{\mathrm{f}}\mathbf{Z}$ and $\widehat{RS}_{\mathrm{c}}=\hat{\boldsymbol{\beta}}_{\mathrm{c}}\mathbf{Z}$, where $\hat{\boldsymbol{\beta}}_{\mathrm{f}}$ denotes the estimates of the parameters of the PH model for failure times and $\hat{\boldsymbol{\beta}}_{\mathrm{c}}$ 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 centred and scaled and denoted as $\widehat{RS}^*_{\mathrm{f}}=\{\hat{\boldsymbol{\beta}}_{\mathrm{f}}\mathbf{Z}-\mathrm{mean}(\hat{\boldsymbol{\beta}}_{\mathrm{f}}\mathbf{Z})\}/\mathrm{SD}(\hat{\boldsymbol{\beta}}_{\mathrm{f}}\mathbf{Z})$ and $\widehat{RS}^*_{\mathrm{c}}=\{\hat{\boldsymbol{\beta}}_{\mathrm{c}}\mathbf{Z}-\mathrm{mean}(\hat{\boldsymbol{\beta}}_{\mathrm{c}}\mathbf{Z})\}/\mathrm{SD}(\hat{\boldsymbol{\beta}}_{\mathrm{c}}\mathbf{Z})$, respectively. This strategy summarizes the multi-dimensional structure of the auxiliary

variables into a two-dimensional summary. The hope is that this two-dimensional summary contains most, if not all, the information about the future event and censoring times. 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

$$d(j,k)=\sqrt{w_{\mathrm{f}}\{\widehat{RS}^{*}_{\mathrm{f}}(j)-\widehat{RS}^{*}_{\mathrm{f}}(k)\}^2+w_{\mathrm{c}}\{\widehat{RS}^{*}_{\mathrm{c}}(j)-\widehat{RS}^{*}_{\mathrm{c}}(k)\}^2} \tag{1}$$

where w_{f} and w_{c} are non-negative weights that sum to one. For each censored subject j, this distance is then employed to define a set of nearest neighbours. This neighbourhood, $R(j^{+},\mathrm{NN})$, 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, $R(j^{+},\mathrm{NN}=10)$ consists of 10 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. The effect of the size of the nearest neighbourhood is explored in a simulation study. Non-zero weights for w_{c} may be useful in reducing the bias resulting from model mis-specification. 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 neighbours for censored subjects.

2.3. Imputation schemes

After the imputing risk set $R(j^{+},\mathrm{NN})$ is defined, the non-parametric imputation schemes developed in Reference [12] 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 analysing multiply imputed data sets follow well established rules as described in Reference [18]. 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 [11]. In this case the estimate from each imputed data set is a KM estimate and the within-imputation variance is based on Greenwood's formula. As pointed out in References [19–21], the variance estimator in Reference [11] under certain settings tends to overestimate the uncertainty of the parameter and may not be consistent [22]. For simplicity, the standard variance estimator [11] will be used and any significant bias monitored through comparing estimates of standard error (SE) and empirical standard deviation (SD).

Risk set imputation (*RSI*). For each of the observed censored times t_j, the RSI method imputes a pair $(t^{*}_{j},\delta^{*}_{j})$ drawn at random from the observed pairs in $R(j^{+},\mathrm{NN})$. 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 $R(j^{+},\mathrm{NN})$. The RSI method is analogous to hot deck imputation in survey research.

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 $R(j^+, \mathrm{NN})$. Then the KMI method imputes a value t_j^* by drawing at random 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. Multiple imputation methods can be enhanced by including a Bootstrap stage, which has been shown to improve their properties [18, 23]. In Reference [12], 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 centred 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 neighbourhood $R^{(B)}(j^+, \mathrm{NN})$ 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 $R^{(B)}(j^+, \mathrm{NN})$, 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 $1, \ldots, K$. The WKM estimator [6, 7] is defined as $\mathrm{WKM}(t) \stackrel{\mathrm{def}}{=} \sum_{k=1}^{K} \hat{S}_k(t)(n_k/n)$, 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 [6, 7].

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 non-parametric survival curve for those individuals in $R(j^+, Z = z_j)$. Using the result in Reference [12] that for a single sample the expectation of the KMI survival estimate with a large number of imputes equals the standard Kaplan–Meier estimate, we have the following result:

Result 1

$E\{\hat{S}_{\mathrm{KMI}}(t)|\mathbf{Y}\} = \mathrm{WKM}(t)$, where the expectation is with respect to the distribution of possible imputes conditional on the observed data $\mathbf{Y}$. The result follows by conditioning on covariate

values and summing (See the appendix of Reference [24] for details). The above result shows that the KMI survival 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 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 proof of this result involves first conditioning on Z, for which T and C are independent, the details are given in the appendix of Reference [24]. 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 Reference [25], we have the following result, a sketch proof of which is outlined in the appendix of Reference [24] for the case of time-independent 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 data sets.

As a result, the KMI method has a large sample property called double robustness [17]. 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 above properties of the KMI method apply in large sample conditions. In small sample size situations, this nearest-neighbourhood approach, which is analogous to a kernel-based method, could produce biased estimates even if one of the two working models is correctly specified, especially when the failure-time model is mis-specified. This phenomenon is similar to that of the kernel-based method [26], where the bias is due to the lack of availability of suitable donor observations.

3.3. Inverse probability of censoring weighted (IPCW) method

We will be including in the application and simulation study a comparison with the IPCW method as described in Reference [9]. 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 Reference [9].

4. APPLICATION TO AIDS DATA

We apply the non-parametric multiple imputation schemes and the IPCW method to AIDS data from the ACTG-019 clinical trial [14, 27]. 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 per cent, respectively. The percentage of the censored observations that were administratively censored due to study termination was 95 per cent in the treatment group and 90 per cent 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, or using both baseline CD4 count and the latest observed CD4 count. 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 I and Figure 1. Table I displays selected estimates from the partially observed (PO) analysis, that is, the standard 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. 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 Reference [14].

In addition, according to the estimates of SEs in Table I, RSI, KMI, RSIB, and KMIB tend to give a modest reduction in the estimated SEs compared to the PO analysis. Assuming there was no censoring before 2 years, the maximum possible gain of efficiency in a non-parametric sense, derived from a binomial distribution, is about 30 per cent for the treated group and 16 per cent for the placebo group at days 450. At days 450, KMIB gains about 22 per cent efficiency for the treated group and 5 per cent for the placebo group. This indicates that the KMIB method recovers about $\frac{7}{10}$ of the information lost due to censoring for the treated group and about $\frac{1}{3}$ for the placebo group. We also note that the SE's for KMIB and RSIB are in general only slightly larger than those for their counterparts without bootstrap, i.e. KMI and RSI.

When both baseline and the latest observed CD4 counts 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 method produces similar estimates of survival as the imputation methods, but was slightly less efficient.

Table I. 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 $\mathrm{NN} = 10$.

Method	$\hat{S}(450)^*$	$(\hat{\mathrm{SE}}_{450})^\dagger$	$\hat{S}(550)$	$(\hat{\mathrm{SE}}_{550})$
Treated				
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				
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

*KM survival estimate of remaining AIDS-free at day 450.
†Based on Greenwood's formula.

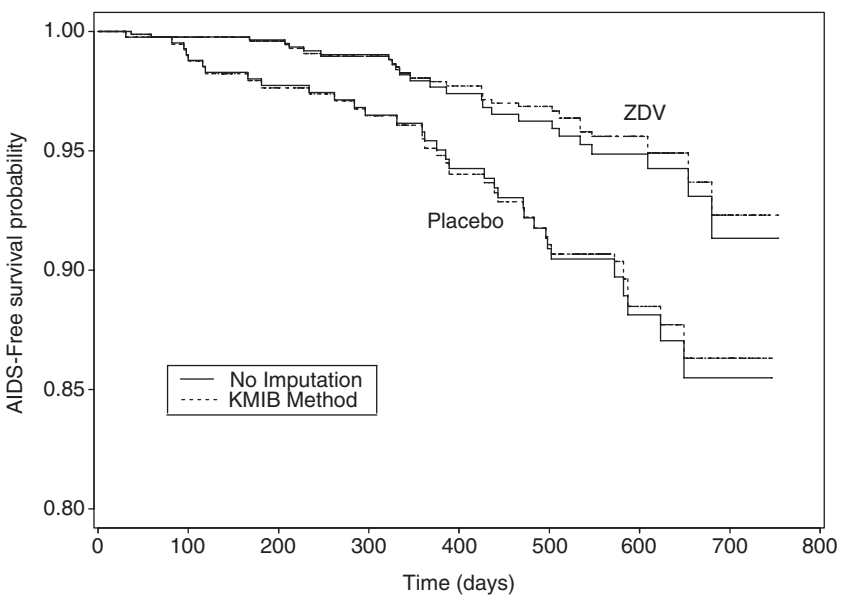

Figure 1. Comparison of KM curves based on the partially-observed data (no imputation) and based on KMIB method using the latest CD4 count as the auxiliary variable.

Table II. Monte Carlo results for a binary covariate: survival estimates.

Method	True value	Average	SD*	SE†	CR‡
Independent censoring; *censoring rate*: 0.47					
Event times $\sim$Exp(1.0/0.1); *censoring times* $\sim$Exp(0.28)					
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 $\sim$Exp(1.0/0.1); *censoring times* $\sim$Exp(0.5/0.2)					
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

*Empirical standard deviation.
†Estimated standard error based on Greenwood's formula.
‡Coverage rate of 95 per cent confidence interval calculated as estimate $\pm t_v^{(0.975)}$ standard error.
The total sample size is 80, 40 for each group. Results based on 500 replications and $M = 50$.

5. SIMULATION STUDY

We perform several simulation studies to investigate the properties of the multiple imputation procedures and to compare with IPCW methods. We consider a binary and multiple time-independent auxiliary variables, and time-dependent auxiliary variables. We investigate the effects of the censoring mechanism, model mis-specification for calculating the risk scores, and the weights and the size of the nearest neighbourhood (NN) on survival estimates.

5.1. Data generation

With a binary covariate (Table II), the event time and censoring time are both generated from an exponential distribution. For more complex situations (Tables III–V), the event and censoring times are both generated from hypothetical PH models conditional on auxiliary variables.

For the time-independent situation with the failure time working model potentially mis-specified (Tables III–IV), five hypothetical auxiliary variables $(Z_1, \ldots, Z_5)$ are independently generated from a U(0, 1) distribution. The event time is generated from $\lambda(t) = t^4 \times \exp(-2.0Z_1 + 0.5Z_2 - 2.0Z_3 + 2.0Z_4 + 2.0Z_5)$. The censoring time is generated from $\lambda_c(t) = t^3 \times \exp(-3.0Z_1 + 0.5Z_2 - 2.0Z_3 + 1.5Z_4 + 2.0Z_5)$ for dependent censoring or from $\lambda_c(t) = 0.6$

Table III. Monte Carlo results for five time-independent covariates with dependent censoring: the effects of the size of the nearest neighbourhood and weights (w_f, w_c) on survival estimates (true value = 0.5).

Method	NN	Average	SD	SE	CR	SDR‡
*Failure**: *correct*; *censoring*†: *correct*						
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
Method	Weights	Average	SD	SE	CR	SDR
Failure: *correct*; *censoring*: *correct*, NN = 10						
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

*Working failure time model.
†Working censoring time model.
‡Squared difference ratio = $\sum(PO - FO)^2 / \sum(\text{est} - FO)^2$.
Censoring rate: 0.51. Sample size 200, 500 replications, $M = 10$.

for independent censoring. When the working failure time or working censoring models are mis-specified only the terms for Z_1, Z_2 and Z_3 are included.

For time-dependent covariates two models are considered (Table V), either a random effects (RE) model or a Brownian motion (BM) model. For the RE model, an auxiliary

Table IV. Monte Carlo results with five 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).

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

Censoring rate: 0.51. Results based on 500 replications, $(w_f = 0.8, w_c = 0.2)$, NN = 10, and $M = 10$.

Table V. Monte Carlo results for time-dependent covariates. Results based on sample size 300, 500 replications, NN = 10, and $M = 10$.

Method	Average	SD	SE	CR*
(a) *Random effects model*				
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*				
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

(a) Random effects model, survival estimates at 1 year (b) Brownian motion model, survival estimates at 6 months.

variable (Z_{ij}) is generated from $Z_i(t) = b_{0i} + b_{1i} \times t$, where t has units of days, $b_{0i} \sim \mathrm{N}(0, 9)$, and b_{1i} distributed as $\mathrm{N}(0, 0.005^2)$. For the BM model Z_i is generated from $Z_i(t) = b_{0i} + b_1 \times t + \sigma \mathrm{BM}_i(t)$, where $b_{0i} \sim \mathrm{N}(5.1696, 0.3)$, $b_1 = -0.2/365$, $\sigma^2 = 0.05/365$ and $\mathrm{BM}_i(t)$ is a Brownian motion stochastic process generated using an iterative algorithm (i.e. $\mathrm{BM}_i(0) = 0$, $\mathrm{BM}_i(t+1) = \mathrm{BM}_i(t) + e_{it}$, where $e_{it} \sim \mathrm{N}(0, 1)$). The auxiliary variable is generated at time 0 and every 3 months for a total duration of 2 years and a maximum of eight measurements. The event time T_i is generated from $\lambda_{\mathrm{f}}(t) = \phi_0 \exp\{\phi_1 l_i(t)\}$ and the random censoring time C_i is generated from $\lambda_{\mathrm{c}}(t) = \psi_0 \exp\{\psi_1 l_i(t)\}$, 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, $\phi_0 = 0.0008 \times \mathrm{e}^{-0.15}$, $\phi_1 = -1.5$, $\psi_0 = \mathrm{e}^{-6}$ and $\psi_1 = 0.0$ for the RE model and $\phi_0 = 5 \times 10^8$, $\phi_1 = -1.0$, $\psi_0 = 0.0025$ and $\psi_1 = 0.0$ for the BM model. With dependent censoring, $\phi_0 = 0.0008 \times \mathrm{e}^{-0.15}$, $\phi_1 = -1.5$, $\psi_0 = \mathrm{e}^{-6}$ 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. The latest observed covariate before each censored time is the only auxiliary variable used to define the nearest neighbourhood.

We note that for the BM model the conditions necessary for the KMI method to eliminate bias 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. For the RE model the conditions are not satisfied, in particular knowledge of the current value and slope are needed, and this does not reduce to a single linear combination of current and past covariate values.

5.2. *Imputation and analysis*

For the 'fully-observed' (FO) analysis (the gold standard), we derive the empirical cumulative distribution function for 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 and then compute KM estimates for each augmented data set and perform the multiple imputation analysis.

5.3. *Results*

5.3.1. Binary time-independent covariate. In Table II, 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. By comparing with the SD values for FO and PO, we see that the KMIB method recovers about $\frac{1}{3}$ of the information lost due to censoring. In addition, the KMI and RSI methods yield almost identical standard deviations as the WKM method and the IPCW method. All 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. All the methods work well in the case of independent censoring, thus we focus on the case of dependent censoring and show only the results for KMIB and IPCW because we find that RSI and RSIB are less good at

reducing the bias. In Table III we show the impact of the size of the NN. The best choice of NN appears to be 10 in this setting. The KMIB method substantially reduces the bias, but does not eliminate it. Table III 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 mis-specified and increases it when the censoring time model is mis-specified. According to Table III, a reasonable choice for w_f would be 0.8 or 0.5 in this setting. 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 mis-specified.

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 data set, 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 when the failure time model is correctly specified and more weight is placed on this model. Table IV shows that as the sample size increases, the bias for KMIB slowly decreases and the coverage rate improves.

5.3.3. Time-dependent covariates. Table V shows the results for KMIB and IPCW for both the RE and the BM generation scheme. For the RE 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 BM model with independent censoring the KMIB has no bias, but the IPCW introduces bias. With dependent censoring, for both the RE and BM models the KMIB substantially reduces but does not eliminate the bias. The IPCW method is less successful at reducing the bias. 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.

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.

All the above results where the SE's of the imputation methods were close to the SD's indicate that the variance formula in Reference [11] is working well in our complex situations.

6. DISCUSSION

The research in this paper provides a direct, simple and transparent approach, non-parametric multiple imputation, to using auxiliary variables to recover information for censored observations. This approach is expected to have weak reliance on a statistical model, because the model is only used to identify a nearest neighbourhood. Once this neighbourhood is defined, the imputation is conducted on a non-parametrically estimated residual time distribution 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

analysing the observed data without using the auxiliary variables. Of the imputation schemes, 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 neighbourhood contains some observations that are not close enough to the target value, so some remnants of dependent censoring remain within the neighbourhood. This is likely to be more of a problem with high dimensional covariates compared to cases with less than say five 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 neighbourhood that is truly 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 mis-specified the bias is greater than when it is correctly specified. In addition, 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 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 neighbourhood. Future research could employ a dynamic scheme to select the size of the nearest neighbourhood dependent on the time of the censored observation. There are other possible adaptations that might improve the KMIB method. For example, rather than equally weighting all the observations in the nearest neighbourhood, one could give more weight to close observations. Instead of using the KM 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

1. Heitjan DF. Ignorability in general incomplete-data models. *Biometrika* 1994; **81**:701–708.
2. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*: *Methodological Issues*, Jewell N, Dietz K, Farewell V (eds). Birkhauser: Boston, 1992; 297–331.
3. Finkelstein DM, Schoenfeld DA. Analysing survival in the presence of an auxiliary variable. *Statistics in Medicine* 1994; **13**:1747–1754.
4. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* 1994; **13**:955–968.
5. Gray RJ. A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* 1994; **81**:527–539.
6. Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 1995; **82**:515–526.

7. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; **52**:137–151.
8. Hubbard A, van der Laan MJ, Robins JM. 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*, Halloran E, Berry D (eds). Springer: New York, 2000; 135–177.
9. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
10. Jewell NP, Kalbfleisch J. Marker processes in survival analysis. *Lifetime Data Analysis* 1996; **2**:15–29.
11. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
12. Taylor JMG, Murray S, Hsu C-H. Survival estimation and testing via multiple imputation. *Statistics and Probability Letters* 2002; **58**:221–232.
13. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* 1996; **22**:425–446.
14. Faucett CL, Schenker N, Taylor JMG. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* 2002; **58**:37–47.
15. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* 1986; **4**:87–94.
16. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 1985; **39**:33–38.
17. Robins JM, Rotnitzky A, van der Laan M. Comment on 'On profile likelihood'. *Journal of the American Statistical Association* 2000; **95**:477–482.
18. Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* 1991; **10**:585–598.
19. Fay R. When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*. American Statistical Association: Washington DC, 1992; 227–232.
20. Meng XL. Multiple imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**: 538–573 (with Discussion).
21. Rubin DB. Multiple imputation after 18 years. *Journal of the American Statistical Association* 1996; **91**: 473–490.
22. Robins J, Wang N. Inference in imputation estimators. *Biometrika* 2000; **87**:113–124.
23. Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991; **40**:13–29.
24. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Unpublished Technical Report*, University of Michigan, 2004.
25. Dabrowska DM. Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Annals of Statistics* 1989; **17**:1157–1167.
26. Pepe MS. Inference using surrogate outcome data and a validation sample. *Biometrika* 1992; **79**:355–365.
27. Volberding PA, Lagakos SW, Koch MA *et al.* Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* 1990; **322**:941–949.

