STATISTICS IN MEDICINE
*Statist. Med.* 2009; **28**:462–475
Published online 7 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3480

# Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation

Chiu-Hsieh Hsu[1,*,†] and Jeremy M. G. Taylor[2]

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

## SUMMARY

When the event time of interest depends on the censoring time, conventional two-sample test methods, such as the log-rank and Wilcoxon tests, can produce an invalid test result. We extend our previous work on estimation using auxiliary variables to adjust for dependent censoring via multiple imputation, to the comparison of two survival distributions. To conduct the imputation, we use two working models to define a set of similar observations called the imputing risk set. One model is for the event times and the other for the censoring times. Based on the imputing risk set, a nonparametric multiple imputation method, Kaplan–Meier imputation, is used to impute a future event or censoring time for each censored observation. After imputation, the conventional nonparametric two-sample tests can be easily implemented on the augmented data sets. Simulation studies show that the sizes of the log-rank and Wilcoxon tests constructed on the imputed data sets are comparable to the nominal level and the powers are much higher compared with the tests based on the unimputed data in the presence of dependent censoring if either one of the two working models is correctly specified. The method is illustrated using AIDS clinical trial data comparing ZDV and placebo, in which CD4 count is the time-dependent auxiliary variable. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: auxiliary variables; dependent censoring; log-rank test; multiple imputation; Wilcoxon test

## 1. INTRODUCTION

In a situation that the censoring time is informative of the time to disease occurrence (dependent censoring), the marginal survival distribution is not identifiable without certain assumptions and,

[*]Correspondence to: Chiu-Hsieh Hsu, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, Arizona Cancer Center, University of Arizona, Tucson, AZ 85724, U.S.A.
[†]E-mail: phsu@azcc.arizona.edu

*Received 22 July 2007*
*Accepted 15 September 2008*

furthermore, dependent censoring will create difficulties in performing nonparametric comparisons between two groups. As a result, the conventional tests, e.g. the log-rank and Wilcoxon tests, could be invalid. In many studies, there is other information provided about each subject, and such data are often informative about both the health condition of the subjects and the censoring mechanism. Some examples of this are CD4 count and viral load in studies of AIDS. These markers are often associated with both the event and censoring times and, therefore, may be treated as auxiliary variables that can help recover some of the lost information for censored subjects and can also be either directly or indirectly incorporated into comparison of two survival distributions to provide a valid test.

There are an increasing number of statistical methods incorporating information from auxiliary variables [1–9] into survival analysis. Most of them focus on estimating the survival function or comparing the treatment effect through a proportional hazards (PH) regression model, but only few focus on nonparametrically comparing two survival distributions. For the use of auxiliary variables, these methods either directly incorporate the information from auxiliary variables into estimation of the survival function [1–5, 8, 9] or use auxiliary variables to modify the redistribution to the right algorithm [10] to derive the survival function [6, 7]. Furthermore, some of them [3–7] focus mainly on using the information in the auxiliary variables for prediction of the event time. In a case with dependent censoring, to produce reasonable survival estimates those methods will strongly rely on the assumption of independent censoring conditional on the auxiliary variables, which are used to predict the event time. In addition, of the methods that use auxiliary variables to handle censored subjects, most of them [1, 2, 8, 9] have adopted approaches using either parametric or partially parametric models to estimate censoring probabilities using auxiliary variables. In this paper, our focus will be on using a direct approach to handling censored observations, while comparing two survival distributions in a nonparametric way in the presence of dependent censoring. We will be using the auxiliary variables for prediction of both event and censoring times. Our approach is a less parametric method in the sense that the models containing the auxiliary variables are not directly used to estimate survival or probabilities of censoring. Furthermore, the approach also allows complex auxiliary variable structures.

In our previous work [11, 12], we treated censored observations as missing event times [13] and then used multiple imputation [14–18] to impute event times [19–21] for the censored observations. The idea is analogous to the redistribution to the right algorithm. We showed that imputation schemes without using auxiliary variables can reproduce the standard Kaplan–Meier (KM) estimates [11], which provides a theoretical foundation for nonparametric imputation of event times. This also provides the theoretical foundation to generalize the imputation approach to handle censored observations in a situation with auxiliary variables [12], where we proposed using two risk scores to define a neighborhood to impute event times for each censored case. The idea is similar to predictive mean matching [22] and propensity score matching [23] in the missing data literature. The two risk scores were derived from two working PH models, one for the failure time and one for the censoring time. We showed that the use of two working risk scores can induce a form of double robustness, where the double robustness property [24] is defined as meaning if one of the two working models is correctly specified, then the estimate is consistent. We showed that by incorporating the auxiliary variables into the multiple imputation method using the two risk scores we can both reduce the bias due to dependent censoring of the marginal survival distribution and increase the efficiency, compared with standard estimates [12], and the approach compared well with other estimation methods for handling dependent censoring. In this paper, we extend our previous work on estimation using auxiliary variables to adjust for dependent censoring via multiple imputation to the case of the comparison of two survival distributions. Robins and Finkelstein [8] proposed using the inverse probability of censoring weighted (IPCW) method, where the weight is derived from a treatment-specific PH model with auxiliary variables as the covariates, to incorporate auxiliary variables into the log-rank test. The IPCW log-rank test can be considered as a semi-parametric approach and can also induce a property of double robustness locally after further modifications on the test statistic and the estimated treatment-specific survival functions [1]. We will compare log-rank test derived from the multiple imputation method with the IPCW method in simulations.

This paper is organized as follows. In Section 2, we briefly describe the imputation procedures and how to construct the test statistic based on the multiply imputed data sets. In Section 3, we study properties of the imputation procedures in finite sample sizes through simulation. In Section 4, we apply the multiple imputation techniques to data from an AIDS study. A discussion follows in Section 5.

## 2. IMPUTATION PROCEDURES

In this section, we briefly describe the multiple imputation approach in [12], including defining two working PH models for deriving risk scores and imputation schemes, and how to conduct two-sample tests on a multiply imputed data set.

### 2.1. Two working PH models and risk scores

Let $T$ denote time to the outcome of interest and $C$ the potential censoring time. The observable random variables are $X = \min(T, C)$ and $\Delta = I(T \leqslant C)$. Let $\bar{\mathbf{V}}(\mathbf{t}) \equiv \{V_1(x), \ldots, V_p(x); 0 \leqslant x \leqslant t\}$ be a vector of auxiliary variables. The observable data for a subject is $\mathbf{O} = (X, \Delta, \bar{\mathbf{V}})$. We assume that we observe $n$ subjects that come from a random sample and are independent.

In order to use auxiliary variables to define an imputing risk set for each censored observation, Hsu *et al.* [12] proposed to reduce the auxiliary variables into two risk scores by fitting two working PH models. One working PH model is for the event times

$$\lambda_{w\mathrm{f}}(t|\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{f}}) = \lambda_{0\mathrm{f}}(t) \exp\{\beta_{\mathrm{f}} \bar{\mathbf{V}}(\mathbf{t})_{\mathrm{f}}\}$$

where $\lambda_{w\mathrm{f}}(t|\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{f}})$ is the conditional cause-specific hazard of failure at time $t$ given $\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{f}}$, $\lambda_{0\mathrm{f}}(t)$ is the baseline hazard function (unknown), $\bar{\mathbf{V}}_{\mathrm{f}}$ are the auxiliary variables in the model, and $\beta_{\mathrm{f}}$ are the corresponding regression coefficients. The other working PH model is for the censoring times, which is

$$\lambda_{w\mathrm{c}}(t|\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}}) = \lambda_{0\mathrm{c}}(t) \exp\{\beta_{\mathrm{c}} \bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}}\}$$

where $\lambda_{w\mathrm{c}}(t|\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}})$ is the conditional cause-specific hazard of censoring at time $t$ given $\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}}$, $\lambda_{0\mathrm{c}}(t)$ is the baseline hazard function (unknown), $\bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}}$ are the auxiliary variables in the model, and $\beta_{\mathrm{c}}$ are the corresponding regression coefficients.

Each risk score is then a linear combination of auxiliary variables. They are defined as $\hat{RS}_{\mathrm{f}} = \hat{\beta}_{\mathrm{f}} \bar{\mathbf{V}}(\mathbf{t})_{\mathrm{f}}$ and $\hat{RS}_{\mathrm{c}} = \hat{\beta}_{\mathrm{c}} \bar{\mathbf{V}}(\mathbf{t})_{\mathrm{c}}$, respectively, where $\hat{\beta}_{\mathrm{f}}$ denotes the estimates of the parameters of the PH model for the event times and $\hat{\beta}_{\mathrm{c}}$ denotes the estimates of the parameters of the PH model

DOI: 10.1002/sim

for the censoring times. Each risk score is centered and scaled by subtracting the mean and dividing by the standard deviation of the risk scores (denoted as $\hat{RS}_{\mathrm{f}}^{*}$ and $\hat{RS}_{\mathrm{c}}^{*}$). This strategy summarizes the multi-dimensional structure of the auxiliary variables into a two-dimensional structure. In a situation with time-independent auxiliary variables, these two working models are fitted once to each treatment group separately to allow potential different associations between the event and censoring times and the auxiliary variables across the treatment groups. However, in a situation with time-dependent auxiliary variables, for every censored observation these two time-independent PH models are fitted to the data of those at risk at the censoring time using the currently available auxiliary variables as fixed covariates. We note that in a case with only one time-independent auxiliary variable, the risk score is the covariate itself. Therefore, there is no need to fit the two working models.

The working model could be misspecified in either of the link function (i.e. the true model is not from a PH model) or the covariates included in the working model not the same as the ones in the true model (e.g. $\bar{\mathbf{V}}_{\mathrm{f}} \neq \bar{\mathbf{V}}$, where $\bar{\mathbf{V}}_{\mathrm{f}}$ are the covariates in the working failure time model and $\bar{\mathbf{V}}$ are the covariates in the true model). In this paper, we assume the true model is from a PH model and mainly focus on misspecification of the covariates.

As shown in [12], if one of these two working models is correctly specified, i.e. the true event (censoring) time model is from a PH model and $\bar{\mathbf{V}} \equiv \bar{\mathbf{V}}_{\mathrm{f}}(\bar{\mathbf{V}}_{\mathrm{c}})$, then conditional on these two risk scores calculated from the two working models, the event times are independent of the censored times. Hence, within an imputing risk set that is defined using two risk scores, the event times are independent of the censoring times. This property will be useful in constructing the imputes.

### 2.2. *Imputation schemes*

The two scale-free risk scores are used to select an imputing risk set for each censored observation by defining the distance between subjects. The distance, based on the original data, between subject $j$ and $k$ is defined as

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

where $w_{\mathrm{f}}$ and $w_{\mathrm{c}}$ are nonnegative weights that sum to 1 and $w_{\mathrm{c}}$ is used to adjust for depending censoring. The imputing risk set, $R(j^{+},\mathrm{NN})$, for the censored subject $j$ consists of NN subjects who have longer survival time than the censoring time of subject $j$ and the NN smallest distances from the censored subject $j$. When the number of individuals still at risk ($\mathrm{NN}_r$) is less than NN, then we use $\mathrm{NN}=\mathrm{NN}_r$. For each censored subject, a nonparametric multiple imputation scheme, called Kaplan–Meier imputation (KMI), is used. In this method an event time is drawn from a KM estimator of the distribution of event times calculated from the observations in the imputing risk set [11, 12]. In this paper, we only impute event times from the subjects in the same treatment group as the censored subject for the same reason we mention earlier. Once the new data set is created, the procedure can be independently repeated $M$ times to obtain multiple imputed data sets for use in testing. We typically use a value of $M$ of 10 or higher, and we found that NN in the range 5–20 gave reasonable results, but there is clearly the potential to optimize the choice of NN. In addition, we also notice that there are alternatives for calculating the distance, such as a rectangular distance. However, we expect that the weight will play a more critical role in selecting a nearest neighborhood for each censored subject compared with the way the distance is defined

DOI: 10.1002/sim

and, therefore, will mainly focus on the effect of weights on the two-sample test and explore it in a simulation study.

As mentioned in [11], the KMI procedure by itself does not incorporate the full uncertainty in the imputes, because it does not include a first stage of an initial parameter draw. Therefore, it would not be viewed as a proper multiple imputation scheme [25]. The KMI procedure can be enhanced by including a bootstrap stage in the procedure [25]. This can be achieved by performing the above imputation procedures on a bootstrap sample selected with replacement from the original data. The KMI method incorporating bootstrap methods is denoted as KMIB and, specifically, KMIB using the weights of $w_{\mathrm{f}}$ and $w_{\mathrm{c}}$ to define distance between subjects to select a neighborhood size of NN is denoted as $\mathrm{KMIB}_{(\mathrm{NN};w_{\mathrm{f}},w_{\mathrm{c}})}$. Multiple imputations are created by independently repeating the bootstrap stage for each of the $M$ data sets. The inclusion of a bootstrap stage has been shown to improve the properties of multiple imputation procedures [25, 26]. In [11], it was shown that when imputing event times the inclusion of the bootstrap stage improved the coverage rate of confidence intervals.

In addition, based on the property of independent censoring conditional on two risk scores if one of the two working models is correctly specified, Hsu *et al.* [12, 27] have shown that the nearest neighbor-based multiple imputation approach can induce a double robustness property.

### 2.3. Test statistics for a multiply imputed data set

In Taylor *et al.* [11], we considered and studied two approaches to comparing two survival curves with multiply imputed data. In this paper, we will use the same approaches and briefly describe them below. The first approach (denoted as Meth1) can be considered as a direct application of the procedure in Li *et al.* [28] for testing a null hypothesis ($\theta=\theta_0$) with multiply imputed data, which applies the estimation rules for multiply imputed data in [25] to obtain the point estimate ($\bar{\theta}$), equal to the average of the $M$ point estimates and the associated variance ($V_1$), equal to $U_1+(1+M^{-1})B_1$ ($B_1$ is the sample variance of the $M$ point estimates and $U_1$ is the average of the $M$ variance estimates). The test statistic ($D$) is equal to $(\bar{\theta}-\theta_0)' V_1^{-1}(\bar{\theta}-\theta_0)$ and the associated distribution, $F_{1,v_1}$ distribution, where $v_1$ is equal to $4+(t-4)(1+(1-2t^{-1})/r)$ [28], $t=M-1$ and $r=(1+M^{-1})B_1U_1^{-1}$. The second approach (denoted as Meth2) using the complete data method to derive the test statistic ($Z_m$) and associated variance (i.e. 1) for each of the $M$ data sets. The rules in [25] are then used to obtain the average ($\bar{Z}$) of the $M$ test statistics ($Z_m, m=1,\ldots,M$) and the associated variance ($V_2$), which is equal to $1+(1+M^{-1})B_2$ ($B_2$ is the sample variance of the $M$ test statistics, i.e. $Z_m$). A $t$-based test with a degree of $v_2$ is then used to compare $\bar{Z}$ to its standard error, $V_2$, where $v_2$ is equal to $[1+(M/(M+1))1/B_2]^2(M-1)$.

In this paper, we are interested in exploring the performance of log-rank and Wilcoxon tests based on multiply imputed data sets derived by using information from auxiliary variables when there is dependent censoring. For each completed data set denote the test statistic (log-rank or Wilcoxon) by $R_m$ and its standard error by $E_m$ and let $Z_m=R_m/E_m$. The first approach estimates $\theta$ by $R_m$, and the second approach estimates $\theta$ by $Z_m$. In both cases $\theta_0=0$ under the null hypothesis that the two curves are equal. The performance of the two approaches will be studied in simulation. In previous work [11], for nondependent censoring, we found that the second method was slightly preferred, although the differences were small. As mentioned in [11], $M=10$ makes the estimated degrees of freedom very large for both ways of calculating it, such that the $t$ distribution is negligibly different from a normal distribution.

## 3. SIMULATION STUDY

We perform several simulation studies to investigate the properties of the multiple imputation-based procedures. We consider situations with multiple auxiliary variables, binary and continuous. In all situations, we investigate size and how these are affected by the sample size, misspecification of the working PH models for calculating the risk scores and the weights of the nearest neighborhood for selecting the imputing risk set.

### *3.1. Data generation*

For each of 1000 independent simulated data sets, there are five hypothetical auxiliary variables, including three binary variables $(Z_1, Z_3, Z_5)$ independently generated from a $Bernoulli(0.5)$ distribution and two continuous variables $(Z_2, Z_4)$ independently generated from a *Uniform*$(0,1)$ distribution. The event time is generated from the model $\lambda_{\mathrm{f}}(t)=t^4*\exp(\psi\mathrm{Trt}-2.0Z_1+0.5Z_2-2.0Z_3+2.0Z_4+2.0Z_5)$, where $\psi$ is set equal to 0 for the study of size and 0.75 for the study of power and Trt is the treatment indicator. In a situation with independent censoring, the censoring time is generated from *Exponential*$(0.6)$. To induce dependent censoring, the censoring time is generated from the model $\lambda_{\mathrm{c}}(t)=t^3*\exp(-3(\mathrm{Trt}+0.1)Z_1+0.5Z_2-2(\mathrm{Trt}+0.1)Z_3+1.5Z_4+2(\mathrm{Trt}+0.1)Z_5)$. We primarily focus on model misspecification of the two working PH models. We consider situations where either both of the two working PH models are correctly specified or one of them is misspecified. Two working models are fitted to each treatment group separately. Specifically, for each treatment group the working failure time model is either correctly specified or misspecified as $\lambda_{w\mathrm{f}}(t)=\lambda_{0\mathrm{f}}(t)*\exp(\beta_1Z_1+\beta_2Z_2+\beta_3Z_3)$, and the censoring time model is either correctly specified or misspecified as $\lambda_{w\mathrm{c}}(t)=\lambda_{0\mathrm{c}}(t)*\exp(\gamma_1Z_1+\gamma_2Z_2+\gamma_3Z_3)$. The sample sizes $(n)$ are 200/400 subjects, 100/200 receiving placebo and 100/200 receiving treatment. The imputation procedures are conducted on each treatment group separately. The computer program written in R for the proposed multiple imputation approach can be requested via email at phsu@azcc.arizona.edu.

### *3.2. Imputation and analysis*

For the ‘fully observed’ (FO) analysis, treated as the gold standard, we apply the log-rank and Wilcoxon tests to each generated data set before any censoring is applied. For the ‘partially observed’ (PO) analysis, we apply the log-rank and Wilcoxon test to each data set (unimputed) with random censoring. For the IPCW method, only the log-rank test is conducted where the weight is derived from a correctly specified treatment-specific PH model for censoring time and the standard error is estimated from 500 bootstrap samples. For the multiple imputation methods, for each simulated data set, we multiply impute times for each observed censored time using the auxiliary variables as described in Section 2. We compute the log-rank and Wilcoxon test statistics for each augmented data set and perform the multiple imputation analysis.

### *3.3. Results*

Table I provides the sizes of the log-rank test in a situation with independent censoring. Both the IPCW method and the PO analysis generate a size comparable to the FO analysis. For the KMIB method, in all situations the sizes are slightly higher than that of the FO analysis, especially in a

Table I. Monte Carlo results: size (per cent) of log-rank tests with independent censoring and time-independent auxiliary variables; $M=10$; censoring rate: 41 per cent for both placebo and treated groups.

| Method | $n=200$ | |
|---|---|---|
| FO | 4.6 | |
| PO | 5.8 | |
| IPCW | 3.5 | |
| | Approach | |
| | Meth1 | Meth2 |
| *Both working models correctly specified* | | |
| KMIB$_{(5;1.0,0.0)}$ | 6.4 | 6.3 |
| KMIB$_{(5;0.8,0.2)}$ | 7.3 | 7.2 |
| KMIB$_{(5;0.5,0.5)}$ | 7.5 | 7.4 |
| KMIB$_{(5;0.2,0.8)}$ | 6.1 | 5.9 |
| *Working failure time model misspecified* | | |
| KMIB$_{(5;1.0,0.0)}$ | 7.8 | 7.7 |
| KMIB$_{(5;0.8,0.2)}$ | 6.5 | 6.4 |
| KMIB$_{(5;0.5,0.5)}$ | 7.1 | 6.7 |
| KMIB$_{(5;0.2,0.8)}$ | 6.1 | 6.0 |
| *Working censoring time model misspecified* | | |
| KMIB$_{(5;1.0,0.0)}$ | 7.5 | 7.5 |
| KMIB$_{(5;0.8,0.2)}$ | 6.9 | 6.6 |
| KMIB$_{(5;0.5,0.5)}$ | 7.7 | 7.2 |
| KMIB$_{(5;0.2,0.8)}$ | 7.3 | 7.1 |

situation that the working failure time model is misspecified and the whole weight is put on the risk score from the working failure time model.

Table II provides the sizes of the log-rank test in a situation with dependent censoring. The results indicate that the sizes based on the IPCW method and the PO analysis are both well above that of the FO analysis. The magnitude of bias for the PO analysis increases with sample size but stays the same for the IPCW method. For the KMIB method, in a situation with both of the working models correctly specified, all combinations of the weights (i.e. $w_{\mathrm{f}}$ and $w_{\mathrm{c}}$) produce similar sizes for a given sample size. As sample size increases from 200 to 400, the sizes of the KMIB method are all comparable to the nominal level (5 per cent) for a NN of 5. However, as the NN increases from 5 to 10, the size increases, as well. This indicates a sufficient sample size and a closeness of the nearest neighbors given by small NN are needed for the KMIB to produce a comparable size to the nominal level. In a situation with only the working failure time model misspecified, the sizes of the KMIB method in general are greater than that of the FO analysis. By putting a small weight (e.g. $w_{\mathrm{c}}=0.2$) on the risk score derived from the working censoring time model to define a NN, the size gets closer to that of the FO analysis and even comparable to the nominal level when the sample size is 400 and NN is 5. This indicates that when the working failure time model is misspecified, incorporating the information from the working censoring time has the potential to adjust for the bias. In a situation with only the working censoring time model misspecified, when only the risk score from the working failure time model (i.e. $w_{\mathrm{f}}=1.0$ and $w_{\mathrm{c}}=0.0$) is used to select a NN, the size of the KMIB is greater than that of the FO analysis,

Table II. Monte Carlo results: size (per cent) of log-rank tests with dependent censoring and time-independent auxiliary variables; $M=10$; censoring rate: 66 per cent (placebo) and 35 per cent (treated).

| Method | $n=200$ | | $n=400$ | |
|---|---|---|---|---|
| FO | 5.0 | | 4.6 | |
| PO | 15.7 | | 25.2 | |
| IPCW | 18.0 | | 18.5 | |
| | Approach | | | |
| | Meth1 | Meth2 | Meth1 | Meth2 |
| *Both working models correctly specified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 7.8 | 7.6 | 6.1 | 5.8 |
| KMIB$_{(5;0.8,0.2)}$ | 7.1 | 7.2 | 5.4 | 5.3 |
| KMIB$_{(5;0.5,0.5)}$ | 7.5 | 7.3 | 6.0 | 5.9 |
| KMIB$_{(5;0.2,0.8)}$ | 7.1 | 7.1 | 6.0 | 6.1 |
| KMIB$_{(10;1.0,0.0)}$ | 7.4 | 7.3 | 6.8 | 6.7 |
| KMIB$_{(10;0.8,0.2)}$ | 7.7 | 7.7 | 7.0 | 6.9 |
| KMIB$_{(10;0.5,0.5)}$ | 7.4 | 7.4 | 7.6 | 7.3 |
| KMIB$_{(10;0.2,0.8)}$ | 7.4 | 7.1 | 6.5 | 6.3 |
| *Working failure time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 8.2 | 8.2 | 6.8 | 6.9 |
| KMIB$_{(5;0.8,0.2)}$ | 7.0 | 6.9 | 4.9 | 5.0 |
| KMIB$_{(5;0.5,0.5)}$ | 6.8 | 6.7 | 6.3 | 6.4 |
| KMIB$_{(5;0.2,0.8)}$ | 6.6 | 6.6 | 6.5 | 6.6 |
| KMIB$_{(10;1.0,0.0)}$ | 8.1 | 8.1 | 7.8 | 7.8 |
| KMIB$_{(10;0.8,0.2)}$ | 8.3 | 8.3 | 6.5 | 6.6 |
| KMIB$_{(10;0.5,0.5)}$ | 6.9 | 6.9 | 6.3 | 6.4 |
| KMIB$_{(10;0.2,0.8)}$ | 6.9 | 6.8 | 7.1 | 7.2 |
| *Working censoring time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 7.1 | 7.0 | 6.4 | 6.3 |
| KMIB$_{(5;0.8,0.2)}$ | 7.2 | 7.3 | 6.6 | 6.7 |
| KMIB$_{(5;0.5,0.5)}$ | 6.9 | 7.1 | 5.9 | 6.0 |
| KMIB$_{(5;0.2,0.8)}$ | 7.4 | 7.4 | 6.9 | 6.9 |
| KMIB$_{(10;1.0,0.0)}$ | 8.5 | 8.3 | 7.3 | 7.1 |
| KMIB$_{(10;0.8,0.2)}$ | 9.1 | 8.8 | 6.8 | 6.8 |
| KMIB$_{(10;0.5,0.5)}$ | 8.2 | 8.5 | 7.6 | 7.5 |
| KMIB$_{(10;0.2,0.8)}$ | 7.9 | 8.0 | 7.7 | 7.6 |

especially for a NN of 10. As sample size increases to 400, the size gets closer to that of the FO analysis and comparable to the nominal level for a NN of 5. Even if the working censoring time model is misspecified, putting a weight of 0.5 (i.e. $w_c=0.5$) on the risk score derived from the working censoring time model to select a NN of 5 for imputing could obtain a size comparable to the nominal level. However, as the weight increases to 0.8, the size increases, as well, and becomes significantly greater than the nominal level. In all situations, two approaches (Meth1 and Meth2) of performing two-sample tests for multiply imputed data produce similar sizes.

Table III provides the sizes of the Wilcoxon test in a situation with dependent censoring. The results indicate that the size based on the PO analysis is well above that of the FO analysis and

Table III. Monte Carlo results: size (per cent) of Wilcoxon tests with dependent censoring and time-independent auxiliary variables; $M=10$; censoring rate: 66 per cent (placebo) and 35 per cent (treated).

| Method | $n=200$ | | $n=400$ | |
|---|---|---|---|---|
| FO | 4.6 | | 5.1 | |
| PO | 19.7 | | 29.5 | |
| | Approach | | | |
| | Meth1 | Meth2 | Meth1 | Meth2 |
| *Both working models correctly specified* | | | | |
| $\text{KMIB}_{(5;1.0,0.0)}$ | 5.7 | 5.7 | 4.4 | 4.2 |
| $\text{KMIB}_{(5;0.8,0.2)}$ | 5.7 | 5.6 | 4.8 | 4.8 |
| $\text{KMIB}_{(5;0.5,0.5)}$ | 6.0 | 6.0 | 4.7 | 4.4 |
| $\text{KMIB}_{(5;0.2,0.8)}$ | 6.5 | 6.5 | 5.3 | 5.3 |
| $\text{KMIB}_{(10;1.0,0.0)}$ | 5.8 | 5.8 | 4.6 | 4.6 |
| $\text{KMIB}_{(10;0.8,0.2)}$ | 7.4 | 7.3 | 4.6 | 4.6 |
| $\text{KMIB}_{(10;0.5,0.5)}$ | 6.2 | 6.2 | 4.6 | 4.6 |
| $\text{KMIB}_{(10;0.2,0.8)}$ | 6.0 | 6.0 | 4.9 | 4.9 |
| *Working failure time model misspecified* | | | | |
| $\text{KMIB}_{(5;1.0,0.0)}$ | 7.0 | 7.0 | 6.4 | 6.4 |
| $\text{KMIB}_{(5;0.8,0.2)}$ | 5.4 | 5.4 | 4.9 | 4.9 |
| $\text{KMIB}_{(5;0.5,0.5)}$ | 6.0 | 6.0 | 4.6 | 4.6 |
| $\text{KMIB}_{(5;0.2,0.8)}$ | 5.9 | 5.8 | 5.6 | 5.4 |
| $\text{KMIB}_{(10;1.0,0.0)}$ | 7.4 | 7.2 | 5.8 | 5.8 |
| $\text{KMIB}_{(10;0.8,0.2)}$ | 6.5 | 6.5 | 4.5 | 4.5 |
| $\text{KMIB}_{(10;0.5,0.5)}$ | 5.8 | 5.8 | 4.8 | 4.7 |
| $\text{KMIB}_{(10;0.2,0.8)}$ | 5.7 | 5.7 | 5.2 | 5.2 |
| *Working censoring time model misspecified* | | | | |
| $\text{KMIB}_{(5;1.0,0.0)}$ | 6.0 | 6.0 | 4.8 | 4.8 |
| $\text{KMIB}_{(5;0.8,0.2)}$ | 6.2 | 6.2 | 5.4 | 5.4 |
| $\text{KMIB}_{(5;0.5,0.5)}$ | 6.6 | 6.4 | 4.8 | 4.8 |
| $\text{KMIB}_{(5;0.2,0.8)}$ | 6.1 | 6.0 | 5.4 | 5.3 |
| $\text{KMIB}_{(10;1.0,0.0)}$ | 6.2 | 6.2 | 4.2 | 4.2 |
| $\text{KMIB}_{(10;0.8,0.2)}$ | 6.4 | 6.4 | 5.0 | 5.0 |
| $\text{KMIB}_{(10;0.5,0.5)}$ | 6.7 | 6.6 | 4.7 | 4.7 |
| $\text{KMIB}_{(10;0.2,0.8)}$ | 7.1 | 7.1 | 5.0 | 5.0 |

the magnitude of bias increases with sample size. For the KMIB method, when the sample size is large ($n=400$), in all situations the size is comparable to the nominal level. In addition, the results also indicates that the performance of the KMIB method, with the Wilcoxon test, highly depends on the sample size and is less affected by misspecification of one of the two working models, the weights of $w_{\text{f}}$ and $w_{\text{c}}$ or the size of NN compared with the log-rank test. This could be because the Wilcoxon test puts more weight on the early-on event times. In that situation, the problem of lack of available donors in the imputing risk set at long follow-up times for multiple imputation approaches has less impact on the Wilcoxon test. In all situations, two approaches (Meth1 and Meth2) of performing two-sample tests for multiply imputed data produce similar sizes.

Table IV. Monte Carlo results: power (per cent) analysis with dependent censoring and time-independent auxiliary variables; $M=10$; censoring rate: 66 per cent (placebo) and 24 per cent (treated).

| | Log-rank test | | | |
|---|---|---|---|---|
| Method | $n=200$ | | $n=400$ | |
| FO | 61.2 | | 89.1 | |
| PO | 29.1 | | 50.0 | |
| IPCW | 47.5 | | 75.0 | |
| | Approach | | | |
| | Meth1 | Meth2 | Meth1 | Meth2 |
| *Both working models correctly specified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 51.3 | 50.3 | 78.3 | 77.8 |
| KMIB$_{(5;0.8,0.2)}$ | 50.0 | 49.3 | 78.1 | 77.6 |
| KMIB$_{(5;0.5,0.5)}$ | 50.2 | 49.9 | 79.2 | 78.9 |
| KMIB$_{(5;0.2,0.8)}$ | 49.1 | 48.4 | 76.9 | 76.7 |
| *Working failure time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 50.7 | 50.2 | 76.6 | 76.2 |
| KMIB$_{(5;0.8,0.2)}$ | 50.5 | 50.0 | 76.4 | 75.9 |
| KMIB$_{(5;0.5,0.5)}$ | 49.4 | 48.5 | 77.1 | 76.9 |
| KMIB$_{(5;0.2,0.8)}$ | 48.8 | 47.8 | 77.2 | 76.8 |
| *Working censoring time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 51.7 | 50.8 | 78.0 | 77.6 |
| KMIB$_{(5;0.8,0.2)}$ | 49.7 | 49.2 | 78.3 | 78.0 |
| KMIB$_{(5;0.5,0.5)}$ | 51.6 | 50.8 | 77.8 | 77.1 |
| KMIB$_{(5;0.2,0.8)}$ | 51.6 | 51.4 | 77.0 | 76.2 |
| | Wilcoxon test | | | |
| | $n=200$ | | $n=400$ | |
| FO | 61.6 | | 89.1 | |
| PO | 25.8 | | 50.0 | |
| | Approach | | | |
| | Meth1 | Meth2 | Meth1 | Meth2 |
| *Both working models correctly specified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 51.2 | 51.0 | 80.3 | 80.4 |
| KMIB$_{(5;0.8,0.2)}$ | 50.2 | 49.9 | 80.3 | 80.2 |
| KMIB$_{(5;0.5,0.5)}$ | 49.7 | 49.7 | 80.3 | 80.3 |
| KMIB$_{(5;0.2,0.8)}$ | 49.3 | 49.1 | 78.9 | 78.9 |
| *Working failure time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 49.4 | 49.4 | 76.7 | 76.6 |
| KMIB$_{(5;0.8,0.2)}$ | 48.3 | 48.4 | 77.6 | 77.6 |
| KMIB$_{(5;0.5,0.5)}$ | 47.3 | 47.0 | 78.3 | 78.1 |
| KMIB$_{(5;0.2,0.8)}$ | 48.1 | 47.6 | 78.4 | 78.3 |
| *Working censoring time model misspecified* | | | | |
| KMIB$_{(5;1.0,0.0)}$ | 51.2 | 51.1 | 79.5 | 79.7 |
| KMIB$_{(5;0.8,0.2)}$ | 50.8 | 50.6 | 79.4 | 79.4 |
| KMIB$_{(5;0.5,0.5)}$ | 49.8 | 49.8 | 80.1 | 80.1 |
| KMIB$_{(5;0.2,0.8)}$ | 51.5 | 51.5 | 79.8 | 79.7 |

Table IV provides the powers of the log-rank and Wilcoxon tests in a situation with dependent censoring. The results indicate that the power based on the PO analysis is much lower than that of the FO analysis. For the log-rank test, the IPCW method produces a power about 14 per cent lower than the FO analysis. On the average the KMIB method produces a power about 10 per cent lower than the FO analysis and at least 20 per cent higher than the PO analysis. Of the three model specification situations, the power is the highest when both working models are correctly specified. The power is the lowest when the working failure time model is misspecified. For the log-rank test, the power is less affected by the weights compared with the size. As expected, the power increases with the sample size for all methods and situations. In all situations, two approaches (Meth1 and Meth2) of performing two-sample tests for multiply imputed data produce similar powers.

In summary, the PO analysis tends to produce a higher size and a lower power compared with the FO analysis. With appropriate choice of NN and weights, the KMIB method could produce a size comparable to the FO analysis even if one of the two working models is misspecified. The KMIB method consistently produces a power much higher than the PO analysis even if the working failure time model is misspecified. Both Meth1 and Meth2 approaches produce similar results in all situations.

## 4. APPLICATION TO AIDS DATA

We apply the nonparametric multiple imputation schemes to AIDS data from the ACTG-019 clinical trial [12, 29, 30]. 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. There were 25 events in the treatment group and 33 events in the placebo group. The median follow-up time is 50 weeks. For each subject, besides several baseline characteristics and the treatment indicator, CD4 counts were measured at several time points. We are interested in comparing survival functions between placebo and treated groups. Since CD4 count is a critical aspect of the immune system and is the only time-dependent variable measured in this AIDS trial, we use CD4 counts as an auxiliary variable to fit the two working PH models to derive two risk scores to select a risk set for imputing. We use both baseline CD4 count and the latest observed CD4 count before each censored time as covariates in these working models. For each observed censored time we use individuals who survived longer than the censored subject and who shared the same treatment group to fit two working PH models for the censoring and failure time distributions. Based on the results in the simulation study, we choose $\mathrm{NN}=5$, $w_{\mathrm{f}}=0.8$ and $w_{\mathrm{c}}=0.2$ to define the imputing risk set for each censored subject.

The results for method 2 are provided in Table V. This table displays the log-rank and Wilcoxon tests from the PO analysis, where the statistics are already divided by the estimated standard errors, that is, the analysis of the observed censored event time data and from the multiple imputation analysis. The results indicate that the multiple imputation method, KMIB, yields larger test statistics and smaller $p$-values compared with the PO analysis. Thus, the multiple imputation analysis could provide an adjustment for dependent censoring in the two-sample test by choosing appropriate auxiliary variables. In addition, for both the PO analysis and the KMIB method the statistic derived from a log-rank test is similar to that derived from a Wilcoxon test. This is probably due to a high percentage of censored observations that they were administratively censored at the end of the follow-up time, 90 per cent in the placebo group and 95 per cent in the treated group.

Table V. AIDS study using baseline CD4 count and latest observed CD4 count before each censored time as auxiliary variables with $M = 10$.

| Method | Statistics | se | $z$ | $p$-Value |
|---|---|---|---|---|
| | | Log-rank test | | |
| PO | 3.082 | 1.000 | 3.082 | 0.0021 |
| $\mathrm{KMIB}_{(5;0.8,0.2)^*}$ | 3.852 | 1.165 | 3.307 | 0.0012 |
| | | Wilcoxon test | | |
| PO | 3.100 | 1.000 | 3.100 | 0.0019 |
| $\mathrm{KMIB}_{(5;0.8,0.2)}$ | 3.918 | 1.162 | 3.372 | 0.0010 |

*$p$-Value calculated from $t_{z,v}$.

## 5. DISCUSSION

The research in this paper provides a direct way to test for the difference between two survival curves in the case of dependent censoring, using the information in auxiliary variables. The approach is nonparametric multiple imputation of the event times for censored observations. The simulation study shows that the use of this nonparametric multiple imputation method can lead to a valid log-rank or Wilcoxon test even in the presence of dependent censoring. In general, the sizes of the tests for the multiple imputation methods are much closer to the nominal level than the sizes produced by analyzing the observed data without using the auxiliary variables. In addition, with sufficient sample size and a good choice of NN the imputation method could produce a size comparable to the nominal level, and a power higher than the power produced by analyzing the observed data without using the auxiliary variables even if one of the two working models is misspecified. In contrast, the IPCW method produces a size much higher than the nominal level in a situation with dependent censoring. We suspect this is due to unstable weights in the IPCW at large follow-up times.

An attractive aspect of the nonparametric multiple imputation procedure in this paper is that the reliance on specific parametric statistical models is weak, because the two working models are only used to identify a neighborhood of similar observations from which an appropriate nonparametric distribution for imputing the censored observations is developed. After the imputation the analysis is based on the original data, augmented by the imputed data. This indicates that this multiple imputation method indirectly incorporates the information from the auxiliary variables into estimation of the survival function. In this sense the properties of the two-sample tests are derived mainly from the data, rather than from the assumptions in the working models. In contrast, most of the methods in the literature directly incorporate the information from auxiliary variables into estimation of the survival function and, therefore, their performance will highly depend on the assumptions in the models.

In this paper, we fix the size of the nearest neighborhood. We could employ a dynamic scheme to select the size of the nearest neighborhood dependent on the time of the censored observation we want to impute in the future research. Furthermore, the adequacy of the 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 lack of availability of possible donor observations will increase the variation of the imputation and also create bias in estimation in the tails. One way to minimize the influence from the tails is by choosing a two-sample test that assigns smaller weight to the tails, such as Gehan's and Tarone-Ware's tests.

To study the performance of the proposed multiple imputation approach, we considered in the simulations a simple situation with time-independent auxiliary variables, which were known at baseline and can be treated as covariates and directly incorporated into the two working models. The two working models are only fitted once to all of the data. In contrast, in a situation with time-dependent auxiliary variables, as we demonstrate in the application to AIDS data, the two working models need to be fitted at every censored observation to the data of those at risk at the censoring time using the currently available auxiliary variables as fixed covariates. The simulation conclusion is based on the results from a situation with time-independent auxiliary variables. Time-dependent auxiliary variables are often more predictive of event times compared with time-independent auxiliary variables. We expect the multiple imputation approach could have more power in testing two survival functions using the information from time-dependent auxiliary variables compared with only using the information from time-independent auxiliary variables.

## REFERENCES

1. 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.
2. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmaceutical Section*, *American Statistician Association*, San Francisco, CA, U.S.A., 1993; 24–33.
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. 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.
9. Satten GA, Datta S, Robins JM. An estimator for the survival function when data are subject to dependent censoring. *Statistics and Probability Letters* 2001; **54**:397–403.
10. Efron B. The two sample problem with censored data. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4. University of California Press: Berkeley, 1967; 831–853.
11. Taylor JMG, Murray S, Hsu C-H. Survival estimation and testing via multiple imputation. *Statistics and Probability Letters* 2002; **58**:221–232.
12. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* 2006; **25**:3503–3517.
13. Heitjan DF. Ignorability in general incomplete-data models. *Biometrika* 1994; **81**:701–707.
14. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
15. Kenward MG, Carpenter JR. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007; **16**:199–218.
16. Carpenter JR, Goldstein H. Multiple imputation in MLwiN. *Multilevel Modelling Newsletter* 2004; **16**:918.
17. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society*, *Series A* 2006; **169**:571–584.
18. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007; **16**:259–275.

19. Pan W. A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine* 2000; **19**:1–11.
20. Pan W. A multiple imputation approach to Cox regression with interval censored data. *Biometrics* 2000; **56**: 192–203.
21. Pan W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 2001; **57**:1245–1250.
22. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* 1986; **4**:87–94.
23. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 1985; **39**:33–38.
24. Robins JM, Rotnitzky A, van der Laan M. Comment on 'On profile likelihood'. *Journal of the American Statistical Association* 2000; **95**:477–482.
25. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**:585–598.
26. Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991; **40**:13–29.
27. 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.
28. Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association* 1991; **86**: 1065–1073.
29. Volberding PA, Lagakos SW, Koch MA, Pettinelli C, Myers MW, Booth DK, Balfour HH, Reichman RC, Bartlett JA, Hirsch MS *et al.* Zidovudine in asymptomatic human immunodeficiency virus infection. A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *New England Journal of Medicine* 1990; **322**:941–949.
30. 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.

DOI: 10.1002/sim