Multiple imputation using chained equations for missing data in survival models: applied to multidrug-resistant tuberculosis and HIV data

Sizwe Vincent Mbona,[1] Principal Ndlovu,[2] Henry Mwambi,[3] Shaun Ramroop[3]

[1]Department of Statistics, Durban University of Technology, Durban; [2]Department of Statistics, University of South Africa, Pretoria; [3]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Abstract

Background. Missing data are a prevalent problem in almost all types of data analyses, such as survival data analysis. Objective. To evaluate the performance of multivariable imputation via chained equations in determining the factors that affect the survival of multidrug-resistant-tuberculosis (MDR-TB) and HIV-coinfected patients in KwaZulu-Natal. Materials and Methods. Secondary data from 1542 multidrug-resistant tuberculosis patients were used in this study. First, data from patients with some missing observations were deleted from the original data set to obtain the complete case (CC) data set. Second, missing observations in the original data set were imputed 15 times to obtain complete data sets using a multivariable imputation case (MIC). The Cox regression model was fitted to both the CC and MIC data, and the results were compared using the model goodness of fit criteria [likelihood ratio tests, Akaike information criterion (AIC), and Bayesian Information Criterion (BIC)]. Results. The Cox regression model fitted the MIC data set better (likelihood ratio test statistic =76.88 on 10 df with P<0.01, AIC =1040.90, and BIC =1099.65) than the CC data set (likelihood ratio test statistic =42.68 on 10 df with P<0.01, AIC =1186.05 and BIC =1228.47). Variables that were insignificant when the model was fitted to the CC data set became significant when the model was fitted to the MIC data set. Conclusion. Correcting missing data using multiple imputation techniques for the MDR-TB problem is recommended. This approach led to better estimates and more power in the model.

Correspondence: Sizwe Vincent Mbona, Department of Statistics, Durban University of Technology, 41-43 ML Sultan Road, Mariam BEE Building, Durban, South Africa.
Tel.: +27.313735703. Fax: +27.313735674.
E-mail: sizwem@dut.ac.za

Key words: missing data; multiple imputation; multidrug-resistance tuberculosis.

Acknowledgments: we would like to thank Dr. Marian Loveday and Prof. Glenda Matthews for allowing us to use their dataset. We also want to thank all facility-level managers, doctors, nurses and data capturers at the study sites for their assistance.

Contributions: SVM, planning of the study and writing of the initial draft of the paper and did the analysis; PN, advise on analysis and input; HW, SR, revision and editing of the paper. All the authors approved the final version to be published.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: none.

Ethics approval and consent to participate: the study protocol was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (Ref: BF052/09), and by the KwaZulu-Natal Department of Health. Only secondary data, the data routinely collected by health workers for clinical care was used in this study. To protect patient confidentiality and anonymity the databases were de-identified and access was strictly limited.

Availability of data and materials: data will be made available upon request but will be controlled.

Informed consent: informed consent was waived by the ethics committee since all patient data used were previously collected during the course of routine medical care and did not pose any additional risks to the patients.

Received for publication: 18 May 2022
Accepted for publication: 17 February 2023.

Journal of Public Health in Africa 2023; 14:2388
doi:10.4081/jphia.2023.2388

Introduction

Loveday *et al.*[1] conducted a study in KwaZulu-Natal (South Africa) whose objective was "To improve the treatment of multidrug-resistant-tuberculosis (MDR-TB) and HIV coinfected patients by investigating the relationship between health system performance and patient outcomes at 4 decentralized MDR-TB sites". Although these data are at least 10 years old, we believe they still have useful information about the survival of MDR-TB and HIV co-infected patients. Hence, we wish to use the data to investigate the factors which affect the survival of MDR-TB and HIV-coinfected patients.

The problem with the data of Loveday *et al*. is that it has missing data with an unknown missingness mechanism among Rubin:[2] i) missing at random (MAR); ii) missing completely at random (MCAR); iii) missing not at random (MNAR). This complicates the application of standard survival analysis methods to analyze the data, as the missing data mechanism determines the statistical data analysis method.[3,4] When data are MCAR, the standard statistical analysis methods applied to complete data cases obtain unbiased model parameter estimates at the cost of the loss in the precision of the estimates and the reduced power of the statistical tests about the model parameters.[5,6] The losses and reduced power are due to the reduction in the sample data size after deleting the cases with missing values. In contrast, applying standard statistical analysis methods to complete data cases after deleting cases with missing values due to either MAR or MNAR obtains biased model parameter estimates.[5,6]

Among the traditional methods developed to enable

researchers to make statistical inferences from incomplete data sets are listwise deletion or complete case analysis, pairwise deletion, mean substitution, regression imputation, and inclusion of an indicator variable.[7] The disadvantages of these methods include further loss of data, obtaining biased model parameter estimates, and underestimating their standard errors. Furthermore, if model parameter estimators are consistent in the no missing data problem, then they are also consistent in the listwise or complete case only if the missingness process is MCAR.[8,11] Better statistical methods for handling missing data include maximum likelihood estimation via the expectation-maximization algorithm and multiple imputation (MI).[9,11-17] For general missing data patterns, imputation methods such as MICE,[18-20] which is also the subject of this paper, are widely used. Simulation studies provide evidence that MICE generally yields estimates that are unbiased and provide appropriate coverage.[18,21] However, MICE is still rarely used in epidemiology, perhaps in part because relatively little practical guidance is available for implementing and evaluating this method. Only a few studies have looked at practical questions about how to implement MI in large data sets used for diverse purposes.[22-24]

The other objective of this paper is to evaluate the performance of MICE in determining the factors that affect the survival of MDR-TB and HIV-coinfected patients in KwaZulu-Natal using the Loveday *et al*. data set.[1] The paper only focuses on MICE because the method involves no variable distributional assumptions and can handle different types of variables.[25,26] MICE can also incorporate variables that are functions of other variables, and it does not require monotonic missing-data patterns.[19,26,27] Furthermore, the advantages of MI over other methods include increased efficiency and the ability to make valid inferences. The imputations are randomly drawn from the updated represented distribution of the data. Hence, the efficiency of estimation is increased by MI. By combining complete data inferences according to Rubin's rules, MI is also able to make valid inferences.[11] MI can incorporate all sources of variability and uncertainty, both within-imputation variance and between-imputation variance. By capturing the between-imputation variance, it solves the problem of standard errors that are too small.[19,26]

Materials and Methods

Data sources

The data used in this study are described in Loveday *et al*.[1] The authors were cleared to use the data by the University of KwaZulu-Natal Biomedical Research Ethics Committee (Ref: BF052/09) and by the KwaZulu-Natal Department of Health. Only secondary data, the data routinely collected by health workers for clinical care using existing records and databases, structured questionnaires, observation, and interviews, were used by Loveday *et al*. The authors report that to protect patient confidentiality and anonymity, the databases were deidentified, and access was strictly limited. Furthermore, informed consent was waived by the ethics committee since all patient data used, were previously collected during the course of routine medical care and did not pose any additional risks to the patients.

According to Loveday *et al*, their study was a prospective mixed methods case study of four decentralized MDR-TB sites in KwaZulu-Natal (South Africa) between 1 July 2008 and 30 June 2012. The authors did not include the fifth center (centralized hospital) in their study. In this study, we used data from 1542 MDR-TB patients from five TB centers (four decentralized sites and one centralized hospital) who were diagnosed with TB. The response variable of interest is the time to death of an MDR-TB patient.

The Cox proportional hazards model

One of the objectives of this study is to identify factors that affect the time (t) to death of patients with a confirmed diagnosis of MDR-TB. The Cox proportional hazards (PH) model expresses the patient hazard rates as functions of potential factors (covariates) as follows:

Let $X_i^T = (X_{i1}, X_{i12},\ldots, X_{ip})$ be a p-dimensional vector of the values of the covariates associated with the i^{th} patient. Then, the Cox proportional hazards regression model is as follows:[28-30]

$$h_i(t) = h_0(t)exp\left\{X_i^T\beta = \sum_{j=1}^{p}\beta_j X_{ij}\right\}, \quad (1)$$

where $\beta^T=(\beta_1,\beta_2,\ldots, \beta_p)$ is a p-dimensional vector of regression coefficients to be estimated from the data, and $h_0(t)$ is the unspecified baseline hazard function that does not have to be estimated.

The hazard model in model (1) makes no assumptions about the shape of the hazard function over time. A hazard function could be constant, increasing, or decreasing, or it could be a combination of two or three of these trends. Model (1) can be written in terms of the survivor function:[30,31]

$$S_i(t) = S_0(t)^{exp\{X_i^T\beta\}}. \quad (2)$$

The model (1) assumptions that may be violated by the MDR-TB data are as follows: i) the covariates X_i^T do not vary with time, and hence, the hazard rate ratios of pairs of patients do not vary with time; ii) censoring and time to cure are independent; iii) the log hazard rate is indeed a linear function of the covariates.

The parameters are estimated as values of β that maximize the Cox likelihood (also called partial likelihood) function for censored data:[30-33]

$$L(\beta) = \prod_{j=1}^{k}\left(\frac{exp(X_j^T\beta)}{\sum_{l\in R_j} exp(X_l^T\beta)}\right). \quad (3)$$

where R_j is the group of patients at risk of cure at time t_j ($0 < t_1 < t_2 < t_3 <\ldots< t_k < \infty$) be k observed death times of patients in the cohort of MDR-TB patients during the observation period).

Consider the log partial likelihood function $l(\beta) = \ln L(\beta)$ and let $l(\hat{\beta})$ evaluated at the maximum likelihood estimate of β for a reduced Cox PH model, and let $l(t)$ be $l(\beta)$ but evaluated at the maximum likelihood estimate of β of the full/saturated model. Then, the test statistic of the null hypothesis that the reduced Cox PH model fit to the data is good is:[30,34]

$$D = 2\{l(t) - l(\hat{\beta}) \sim \chi^2_{n-p} \text{ (asymptotically)}. \quad (4)$$

The null hypothesis is rejected if H_0 if H_0 if $D > \chi^2_{n-p,\alpha}$ or if the p-value < α

where

n is the number of patients

p is the number of parameters and

α is the level of significance of the test

To test the null hypothesis H_0:$\beta_j=\beta_j^0$,

The Wald test statistic $Z = \frac{\hat{\beta}_j-\beta_j^0}{se\,(\hat{\beta}_j)}$ is used, where se $(\hat{\beta}_j)$ is the

estimate of the asymptotic standard error of $\hat{\beta}_j$ (square root of the j^{th} diagonal element of $\widehat{Cov}(\hat{\beta})$).

Under the null hypothesis, the asymptotic distribution of Z is the standard normal distribution.

If the assumptions of the model (1) are not violated by the data, then developing a Cox PH model involves variable selection. To do this, the partial likelihood ratio (4) and the Wald tests are used in conjunction with information criteria such as the Akaike information criterion (AIC):[30,35]

$$AIC = -2\ln L(\hat{\beta}) + 2p, \quad (5)$$

where p is the number of model parameters. The model with the smallest *AIC* among competing Cox PH models is the best.

Proportional hazards and linearity assumptions

The null hypothesis for the linearity test is that the predictor in the Cox PH model is $X_i^T\beta$. The hypothesis may be tested by testing the null hypothesis that $\theta_2 = 0$ in the Cox PH model:[30,33]

$$h_i(t) = h_0(t)exp\{\theta_1(X_i^T\hat{\beta}) + \theta_2(X_i^T\hat{\beta})^2\}, \quad (6)$$

where $\hat{\beta}$ is from fitting the Cox PH model (1).

Rejecting the null hypothesis implies that $X_i^T\beta$ is an incorrect specification of the predictor in the Cox PH model. The proportional hazards assumption may be tested by testing the null hypothesis that $\phi_2 = 0$ in the Cox PH model:[30,33]

$$h_i(t) = h_0(t)exp\{\phi_1(X_i^T\hat{\beta}) + \phi_2(X_i^T\hat{\beta})t\}, \quad (7)$$

where $\hat{\beta}$ is from fitting the Cox PH model (1). Rejecting the null hypothesis implies that the proportional hazards assumption does not hold.

Interpretation of the estimated coefficients of the Cox proportional hazards model

The hazard ratio (HR) associated with the j^{th} covariate is:[30,33]

$$HR_j = e^{\hat{\beta}_j}, j = 1, 2, \ldots, p, \quad (8)$$

where $\hat{\beta}_j$ is the estimate of the coefficient of the j^{th} covariate (β_j).

For continuous covariates, if $\hat{\beta}_j > 0$ then a unit increase in the j^{th} covariate increases the hazard by 100 ($HR_j - 1$)%. Otherwise, a unit increase in the j^{th} covariate decreases the hazard by 100 ($HR_j - 1$)%. For categorical variables, the j^{th} covariate is actually the j^{th} category/level of the categorical variable.

Hence, if $\hat{\beta}_j > 0$, then patients in the j^{th} category/level face a hazard 100 ($HR_j - 1$)% greater than those in the specified reference category/level. Otherwise, the patients in the j^{th} category/level face a hazard 100 ($HR_j - 1$)% lower than those in the specified reference category/level.

The 100 $(1 - \alpha)$% confidence interval for the true hazard ratio associated with the j^{th} covariate (e^{β_j}) is given by:

$$CR_j = e^{\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)}, j = 1, 2, \ldots, p, \quad (9)$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $se(\hat{\beta}_j)$ is the standard error of of $\hat{\beta}_j$. If CR_j includes one, then there is no association between the hazard (and the survival) and the j^{th} covariate.

In summary, e^{β_j} indicates how large (or small) the hazard in one group or subject is with respect to the hazard in the reference group or subject.

Multiple imputation

The basic idea of MI in the context of the present study is to create a small number, m, of copies of the MDR-TB data set, each of which has the missing values suitably imputed. Traditionally, m=3 to 10. Then, the Cox PH regression model is independently fitted to each of the m complete MDR-TB data sets. Estimates of the parameters and values of other statistics from fitting the Cox PH regression models to each of the m complete MDR-TB data sets are averaged to obtain single estimates. Standard errors are computed according to the "Rubin rules".[11] The above three main steps are schematically displayed in Figure 1.[36,37]

For each of the m imputed data set point estimates, $\hat{Q}_i$, $i = 1,2,\ldots,m$, are computed for every parameter Q of interest as well as the estimate of the variance of $\hat{Q}_i$ denoted by U_i. Then, the pooled point estimate of Q, the within-imputation variance ($\bar{U}$) the between imputation variance (B) are given by:[11,16]

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i,\ \bar{U} = \frac{1}{m}\sum_{i=1}^{m}U_i \text{ and } B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})(\hat{Q}_i - \bar{Q})', \quad (10)$$

respectively. The estimate of the variance of $\bar{Q}$ is calculated as:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B, \quad (11)$$

Here, the factor $(1+m^{-1})$ multiplied by B is an adjustment to correct for the extra variance caused by using a finite number of imputations m to estimate $\bar{Q}$. This adjustment is needed to make valid inferences with low m. Otherwise, the analysis would result

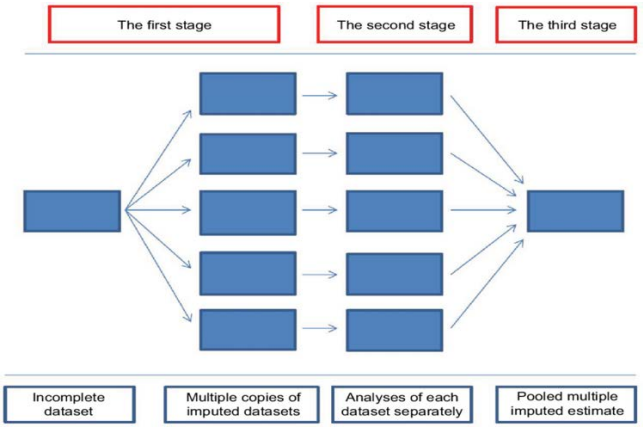

Figure 1. Schematic representation of the main steps in multiple imputation: incomplete data are the original multidrug-resistant-tuberculosis data set; imputed data are each of the *m* copies of the original multidrug-resistant-tuberculosis data set with chained equations imputed missing values; pooled results are the application of the "Rubin rules";[11] and multivariate imputation by chained equations.

in too low p values or too short confidence intervals. The procedure described above to pool the repeated-imputation results is referred to as Rubin's rules.[11,26]

Inferences about the Q (confidence intervals and hypothesis tests) are then based on Student's t approximation:[11,16,38]

$$\sqrt{T}(\bar{Q}-Q)\sim t_v, v=(m-1)\left[1+\frac{m\bar{U}}{(m+1)B}\right]^2 \quad (12)$$

where v is the degrees of freedom. The quantity $r = \frac{m\bar{U}}{(m+1)B}$ in v is the ratio of B to $\bar{U}$ and measures the relative increase in variance due to the missing data, *i.e.*, the cost due to missing data.[11] Furthermore, $\lambda = \frac{1}{1+r}$ is the rate of missing information for Q and $e = \frac{100}{m+\lambda}\%$ is the efficiency of $\bar{Q}$ based on m imputed data sets relative $\bar{Q}$ to based on an infinite number of imputed data sets.[11,16]

Multivariate imputation by chained equations

Multivariate imputation by chained equations (MICE) is one particular MI technique that is used in this study. Following Azur *et al*. and in the context of this study, the technique is implemented as follows.[19,39]

Let $Y_1, Y_2, \ldots, Y_k$ represent the variables with missing values in the MDR-TB data set.

Step 1: Perform mean imputations of the missing values for all the variables.

Step 2: Set the imputations for variable Y_i back to missing.

Step 3: Regress Y_i on all the other variables with imputed values in step 2 and all the other variables in the MDR-TB data set.

Step 4: Replace the missing values of Y_i with the predicted values from the fitted regression model in step 3.

Step 5: Repeat steps 2-4 with the other Y_i until the missing values of all the $Y_i (i = 1,2,\ldots,k)$ have been replaced with predictions from the fitted regression models.

Step 6: Repeat steps 2-5 l times until the estimated parameters of the regression models converge or become stable. Store the imputed data set.

Step 7: Repeat steps 2-6 to obtain m imputed MDR-TB data sets.

In Azur *et al*.,[39] it is suggested that $l \geq 10$ and $m \geq 10$ be used and that the precision of the pooled estimates of the parameters and the power of the tests about the parameters increases with m.

In this study, l=10 and m=15 were used to guarantee to obtain at least 95% efficient parameter estimates after assuming that the rate of missing information for all the parameters is $\lambda \leq 0.7$ (see the rates of missing data in Table 1 and Figure 2). The imputed MDR-TB data sets were generated and analyzed using the MICE in SPSS Version 25 and STATA Version 19, respectively.

Results

The median follow-up time was 26.8 months. By 30 June 2012, 56% of the patients had been cured, 15.9% were deceased, and the rest had defaulted or lost to follow-up (see Table 2). Table 2 also displays the frequency distributions of the other variables in the data whose effects on the survival time of MDR-TB patients were investigated in this study. Table 2 shows that most of the patients were between 18 and 50 years old, had no extrapulmonary TB, had no commodity diseases, and had no previous MDR-TB episodes. This suggests dropping the type of TB, previous MDR-TB episodes, and comorbidities from among the factors to be investigated for their effects on the time to cure MDR-TB patients. However, Table 1 shows that the variable comorbidities had the largest percentage of missing data (45%), followed by other variables in the table. Furthermore, Figure 2 shows that the MDR-TB data set has approximately 7.8% missing values in 50% (5) of the variables and/or in 66.7% (1028) of the data cases. Thus, the size of the complete MDR-TB data set is 514. This motivated the MICE approach of analyzing MDR-TB data to meet the objectives of this study.

Estimated HRs with 95% CIs, corresponding to the Cox proportional hazard model and imputed data sets, are given in Table 3. Age groups (41 to 50 years and 51 or more) at diagnosis and HIV-negative status were not significant in the Cox proportional model with missing data due to unavoidable loss in the power of the model. Furthermore, the female sex was significant in the Cox model but not in the imputed model. After imputing missing data, all of these variables were retained in the model.

Comparing the performance of the models in Table 4, imputation of missing data led to improvement in model goodness of fit (likelihood ratio test =76.88, AIC =1040.90, Bayesian Information Criterion (BIC) =1099.65 for the MICE versus likelihood ratio test =42.68, AIC =1186.05, BIC =1228.47 for the Cox model with missing data).

This is not shocking as the results suggest because we have seen that 45.45% of the covariates had random missing values. Recall that the MAR data are ignorable because they are derived from observed data (Y_{obs}). When missing data are ignorable, the rate of missing information can be negligible, and the chances of obtaining biased results are higher. We will interpret the results obtained using imputed data sets.

After a median follow-up of 26.8 months, the hazard ratio for baseline weight was 0.98; 95% CI: 0.97-0.99; S.E =0. We found that patients treated in decentralized sites (HR =1.72, 95% CI:

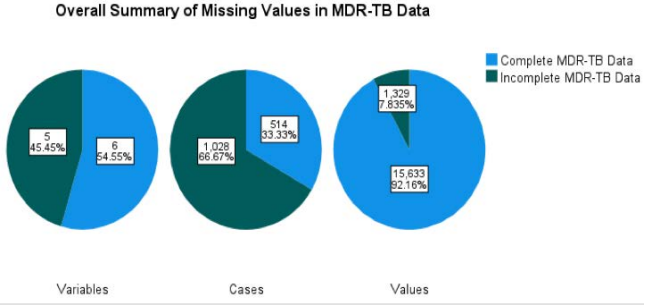

Figure 2. Frequency distributions of missing data by variable, cases, and values in the KwaZulu-Natal (RSA) multidrug-resistant-tuberculosis data set: 2008-2012.

Table 1. Frequency distribution of missing values in the KwaZulu-Natal (RSA) multidrug-resistant-tuberculosis dataset by variable: 2008-2012.

Variable	Missing	Observed	Mean
Comorbidities	694 (45.0)	848	
Time to be cured (months)	453 (29.4)	1089	18.30 [8.46]
Baseline weight (kg)	112 (7.3)	1430	48.71 [17.01]
Type of TB	1 (0.1)	1541	
HIV status	69 (4.5)	1473	

() are percentages; [] are standard deviations. TB, tuberculosis.

0.73-4.06) had higher chances of dying than those who were treated in the centralized hospital. However, this was not significant. The results also show that patients in the age group 31 to 40 years (HR =1.50, 95% CI: 1.07-2.10), 41 to 50 years (HR =1.81, 95% CI: 1.24-2.64), and more than 50 years (HR =3.21, 95% CI: 2.11-4.87) age group had higher chances of dying than those in the 18 to 30 years age group. MICE results also revealed that HIV-negative patients had lower chances of dying than HIV-positive patients (HR =0.68, 95% CI: 0.48-0.97).

It should be noted that when the missing rate is low, the results obtained using incomplete data might be similar to those of MICE. It has also been noted that with a more than 60% missing rate, the MICE model might not provide accurate estimates.[40]

Discussion

In this study, we evaluated the handling of survival data with missing covariate values by means of MICE. In general, careful modeling is required when using MICE to obtain valid statistical inferences.[27] Another important point to remember concerns the order in which the imputation models should be imputed. Imputation using chained equations does not require us to specifically order the variables that must be imputed because the software imputes by default the variables from the most observed to the least observed. Missing data are a common problem in longitudinal studies. Ignoring incompleteness or handling the data inappropriately may bias study results, reduce the power and efficiency of the study, and alter important risk/benefit relationships.

In the results presented, some variables lost their significant effect in the incomplete data analysis. For example, age is known as one of the most important prognostic variables.[41] However, this variable did not reach a significant level in the incomplete data model. Once missing data were imputed, power was increased, and variables lost their effect in the incomplete model, such as "previous MDR-TB episodes", and reached a significance level. This shows that missing data depend on other patients' characteristics and therefore can be well imputed using MI methods. The results also showed that the estimates were relatively similar to those obtained from the incomplete case analysis. However, after imputation, the standard errors were smaller and the confidence intervals were narrower.

Limitations of the study

Our work involved several limitations. We used a data set containing only eight variables. Therefore, the impact of the number of variables offered was not investigated. Furthermore, we only compared the performance of the Cox proportional model using

Table 2. Frequency distribution of the variables in the KwaZulu-Natal (RSA) multidrug-resistant-tuberculosis dataset by site (centralized, decentralized), 2008-2012.

Variable	Site Centralized hospital 812 (52.7)	 Decentralized 4 sites 730 (47.3)	 Total 1542 (100)
Age (years) at diagnosis*			
18-30	303 (19.7)	245 (15.9)	548 (35.5)
31-40	292 (18.9)	258 (16.7)	550 (35.7)
41-50	145 (9.4)	153 (9.9)	298 (19.3)
51+	72 (4.7)	74 (4.9)	146 (9.5)
Gender			
Male	399 (25.9)	346 (22.4)	745 (48.3)
Female	413 (26.8)	384 (24.9)	797 (51.7)
Previous MDR-TB episodes			
0	802 (52.0)	673 (43.6)	1475 (95.7)
1	9 (0.6)	55 (3.6)	64 (4.2)
2+	1 (0.1)	2 (0.1)	3 (0.2)
Type of TB			
Pulmonary	804 (52.1)	706 (45.8)	1510 (97.9)
Extra pulmonary	7 (0.5)	24 (1.6)	31 (2.1)
Comorbidities			
No diseases or conditions	780 (50.6)	12 (0.8)	792 (51.4)
Diabetes	10 (0.6)	10 (0.6)	20 (1.2)
Epilepsy	4 (0.3)	8 (0.5)	12 (0.8)
Hearing loss prior to treatment	1 (0.1)	10 (0.6)	11 (0.7)
Renal problems	0 (0.0)	3 (0.2)	3 (0.2)
Substance abuse	0 (0.0)	4 (0.3)	4 (0.3)
Liver problems	1 (0.1)	1 (0.1)	2 (0.2)
Psychiatric problems	4 (0.3)	0 (0.0)	4 (0.3)
HIV status			
Positive	576 (37.4)	524 (34.0)	1100 (71.4)
Negative	211 (13.7)	162 (10.5)	373 (24.2)
TB outcome**			
Cured	441 (28.6)	445 (28.9)	886 (57.5)
Died	113 (7.3)	132 (8.6)	245 (15.9)
Defaulted	229 (14.9)	105 (6.8)	334 (21.7)
Lost to follow-up	29 (1.9)	48 (3.1)	77 (5.0)

() are percentages; *median age=34 years; **median follow-up=26.8 months. MDR-TB, multidrug-resistant-tuberculosis; TB, tuberculosis.

incomplete data sets and the MICE under the MAR mechanism. It is known that the performance of models depends to a great extent on the mechanism of missing data, rate of missing data, method of imputation of missing data, and sample size.[42,43] Our work was simply a study to explain the methodological issues in the application of the MICE method and its art in the recovery of information. We showed that analyzing incomplete data decreases the power and that the MICE method recovers the data. However, at this stage, due to the limitations listed above, we cannot provide a specific guideline on how best to tackle the problem of missing data because there are many approaches to address missing data.

In our analysis, we used only the information criteria as a method of selecting variables in the Cox model. However, there are many other popular approaches that can be used, such as the boosting method, which originally evolved from the field of machine learning as an approach to classification problems and was later adapted to statistical models.[44,45] Other selection methods that can be used are stepwise regression and its variants forward selection, lasso and backward elimination.[46]

Conclusions

Procedures of MIs including the survival outcome and MIs including the imputed observed event time have been shown to perform quite adequately. From our results, we have seen that MI performs best in ordinary survival data. This procedure was found to perform well in the literature and even though more recent improvements have been suggested, this procedure is widely used by many authors. Based on our thoughts both on our own study results and on other studies, we conclude that the procedure of MI performs well and could be recommended for imputing missing covariate values with survival data because applying models to incomplete data reduces the power and efficiency of the study. Applying the MICE model may provide better and more accurate estimates and increase the power of the model.

Table 3. Comparison of the efficiency of the parameter estimates and power of the tests from fitting the Cox proportional hazards model to the original and imputed datasets: *m*=15 imputed datasets.

Variables	Original MDR-TB				Imputed MDR-TB			
	Coef.	S.E	HR (95% CI)	p-value	Coef.	S.E	HR (95% CI)	p-value
Baseline weight (kg)	-0.03	0.01	0.97 (0.96 – 0.98)	<0.01*	-0.02	0.00	0.98 (0.97 – 0.99)	<0.01*
Study sites								
Centralised hospital (Ref)								
Decentralised sites	0.45	0.49	1.57 (0.60 – 4.08)	0.36	0.54	0.43	1.72 (0.73 – 4.06)	0.21
Age group (years)								
18 – 30 (Ref)								
31 to 40	0.53	0.26	1.70 (1.02 – 2.85)	0.04*	0.41	0.17	1.50 (1.07 - 2.10)	0.02*
41 to 50	0.60	0.31	1.82 (0.99 – 3.34)	0.05	0.59	0.19	1.81 (1.24 – 2.64)	<0.01*
51 or more	0.48	0.38	1.61 (0.76 – 3.41)	0.21	1.17	0.21	3.21 (2.11 – 4.87)	<0.01*
Gender								
Male (Ref)								
Female	0.52	0.22	1.69 (1.09 – 2.61)	0.02*	0.12	0.13	1.13 (0.87 – 1.47)	0.37
Type of TB								
Pulmonary (Ref)								
Extra-pulmonary	0.25	0.72	1.28 (0.31 – 5.28)	0.73	0.28	0.39	1.33 (0.62 – 2.85)	0.47
Previous MDR-TB episodes								
No (Ref)								
Yes	-0.19	0.73	0.83 (0.20 – 3.46)	0.80	0.35	0.27	1.42 (0.84 – 2.39)	0.19
HIV status								
Positive (Ref)								
Negative	-0.06	0.26	0.94 (0.57 – 1.56)	0.82	-0.39	0.18	0.68 (0.48 – 0.97)	0.03*
Comorbidities conditions								
No (Ref)								
Yes	-0.33	0.56	0.72 (0.24 – 2.13)	0.55	-0.41	0.45	0.66 (0.27 – 1.62)	0.37
	Likelihood ratio test = 42.68 on 10 df, p-value < 0.01				**Likelihood ratio test = 72.81 on 10 df, p-value < 0.01**			

HR: Hazard Ratio, CI: Confidence Interval, N: Sample size, S.E: Standard error

Table 4. Comparison of the goodness of fit statistics from fitting the Cox proportional hazards model to the original and imputed datasets: *m*=15 imputed datasets.

Statistic	Dataset	
	Original MDR-TB	MICE MDR-TB
Sample size (n)	514	1542
Proportionality of Cox PH model (θ_2)	5.52 on df =10, p=0.854	7.96 on df =10, p=0.438
Likelihood ratio test (ϕ_2)	42.68 on df =10, p<0.01	76.88 on df =10, p<0.01
AIC of Cox PH model	1186.047	1040.902
BIC of Cox PH model	1228.469	1099.651

MDR-TB, multidrug-resistant-tuberculosis; PH, proportional hazards; AIC, Akaike information criterion; BIC, Bayesian Information Criterion; MICE, multivariate imputation by chained equations.

References

1. Loveday M, Padayatchi N, Wallengren K, et al. Association between health systems performance and treatment outcomes in patients co-infected with MDR-TB and HIV in KwaZulu-Natal, South Africa: implications for TB programmes. PLoS One 2014;9:e94016.
2. Rubin DB. Inference and missing data. Biometrika 1976;63:581-92.
3. Allison PD. Multiple imputation for missing data: A cautionary tale. Sociol Methods Res 2000;28:301-9.
4. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica neerlandica 2003;57:19-35.
5. Altman DG, Bland JM. Missing data. BMJ 2007;334:424.
6. Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. Iran J Cancer Prev 2010;3:e80700.
7. Eekhout I, de Boer RM, Twisk JW, et al. Missing data: a systematic review of how they are reported and handled. Epidemiology 2012;23:729-32.
8. Acock AC. Working with missing values. J Marriage Fam 2005;67:1012-28.
9. Little RJA, Rubin DB. Statistical analysis with missing data. 2002.
10. Peugh JL, Enders CK. Missing data in educational research: A review of reporting practices and suggestions for improvement. Rev Educ Res 2004;74:525-56.
11. Rubin DB. Multiple Imputation for Non-response in Surveys John Wiley. New York. 1987.
12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Series B (Methodol) 1977;39:1-22.
13. Enders CK. A primer on maximum likelihood algorithms available for use with missing data. Struct Equ Modeling 2001;8:128-41.
14. Rubin DB. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In Proceedings of the survey research methods section of the American Statistical Association 1978;1:20-34. Alexandria, VA, USA: American Statistical Association.
15. Rubin DB. Multiple imputation after 18+ years. J Ame Stat Assoc 1996;91:473-89.
16. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147.
17. Patrician PA. Multiple imputation for missing data. Res Nurs Health 2002;25:76-84.
18. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul 2006;76:1049-64.
19. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. J Stat Softw 2011;45:1-67.
20. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Ame J Epidemiol 2010;171:624-32.
21. Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. Emerg Themes Epidemiol 2012;9:1-0.
22. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. Ame J Epidemiol 2009;169:1133-9.
23. He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010;3:98-105.
24. Schenker N, Raghunathan TE, Chiu PL, et al. Multiple imputation of missing income data in the National Health Interview Survey. J Ame Stat Assoc 2006;101:924-33.
25. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. J Stat Softw 2011;45:1-20.
26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377-99.
27. Marchenko Y. Chained equations and more in multiple imputation in Stata 12. In 2011 Italian Stata Users Group Meeting, 2011.
28. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. New York: Springer, 2003.
29. Etikan I, Babatope G. Survival Analysis: A Major Decision Technique in Healthcare Practices. Int J Sci Res Methodol 2018;8:121-35.
30. Bengura P. Identification of factors affecting the survival lifetime of HIV+ terminal patients in Albert Luthuli municipality of South Africa. University of South Africa, 2020.
31. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. New York: Springer, 2012.
32. Lemeshow S, May S, Hosmer Jr DW. Applied survival analysis: regression modeling of time-to-event data. John Wiley & Sons, 2011.
33. Cleves M, Gould WW, Gutierrez RG, Marchenko YV. Competing risks. An Introduction to Survival Analysis Using Stata. 2010.
34. Zhang HH. Checking proportionality for Cox's regression model (Master's thesis), 2015. Available from https://www.duo.uio.no/bitstream/handle/10852/45324/HuiHongZhang_thesis.pdf?sequence=1&isAllowed=y (accessed on May 23, 2018).
35. Xu R, Vaida F, Harrington DP. Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. Statistica Sinica 2009;19:819.
36. Van Buuren S. Flexible imputation of missing data. CRC press, 2018.
37. Yang S. Flexible Imputation of Missing Data: Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xxvii+415 pp. J Ame Stat Assoc 2019;114.
38. Little RJ, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 1996:98-111.
39. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40-9.
40. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Ame J Epidemiol 2004;160:34-45.
41. Chen C, Zhu T, Wang Z, et al. High latent TB infection rate and associated risk factors in the eastern China of low TB incidence. PLOS one 2015;10:e0141511.
42. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 2007;16:277-98.
43. Musil CM, Warner CB, Yobas PK, Jones SL. A comparison of imputation techniques for handling missing data. West J Nurs Res 2002;24:815-29.
44. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Inicml 1996;96:148-56.
45. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat 2000;28:337-407.
46. Chatterjee S, Price B. Selection of variables in a regression equation. Regression analysis by example. 1977:201-3.