ISSN: 2317-0840

A comparison of multiple imputation methods for the analysis of survival data with outcome related missing covariate values

José Luiz P. Silva[†]

Federal University of Paraná (UFPR).

Resumo: *O modelo de taxas de falha proporcionais de Cox é comumente usado na área médica para investigar a associação entre o tempo de sobrevivência e covariáveis. No entanto, é bastante comum que a análise envolva covariáveis com valores ausentes. Uma suposição razoável é que os dados são* censoring-ignorable MAR*, no sentido de que o mecanismo de perda não depende do tempo de censura, mas pode depender do tempo de falha. Nesse caso, uma análise de casos completos produz estimativas viesadas para os coeficientes de regressão. Através de um estudo de simulação, comparamos três abordagens de imputação múltipla para uma covariável parcialmente observada quando o mecanismo de perda envolve o tempo de sobrevivência: (i) o método proposto por White & Royston (2009) que usa a função de taxa de falha acumulada em uma aproximação do modelo de imputação, (ii) o método descrito por Bartlett* et al. *(2015) que incorpora o modelo de Cox no processo de imputação, e (iii) a abordagem CART, um método conhecido por lidar com distribuições assimétricas, interações e relações não lineares. Os resultados da simulação mostraram que o método de White & Royston (2009) pode produzir estimativas severamente viesadas enquanto a abordagem CART subestima a incerteza da imputação resultando em baixas taxas de cobertura. O método de Bartlett* et al. *(2015) apresentou o melhor desempenho geral, com pequeno viés de pequenas amostras e taxas de cobertura próximas aos valores nominais. Os métodos de imputação são aplicados a um conjunto de dados de sobrevida de pacientes com doença de Chagas.*

Palavras-chave: *Covariáveis ausentes, modelo de Cox, imputação múltipla, estudo de simulação, MAR.*

Abstract: *The Cox proportional hazards model is commonly used in medical research for investigating the association between the survival time and covariates. However, it is quite common for the analysis to involve missing covariate values. It is reasonable to assume that the data are censoring-ignorable MAR in the sense that missingness does not depend on censoring time but may depend on failure time. In this case, a complete cases analysis produce biased regression coefficient estimates. Through a simulation study, we compare three multiple imputation approaches for a missing covariate when missingness is survival time-dependent: (i) the method proposed by White & Royston (2009) that uses the cumulative hazard in an approximation to the imputation model, (ii) the method described by Bartlett* et al. *(2015) that incorporates the Cox model in the imputation process, and (iii) the CART approach, a method known to deal with skewed distributions, interaction and nonlinear relations. Simulation results show that the method of White & Royston (2009) may produce very biased estimates while the CART approach underestimates the imputation uncertainty resulting in low coverage rates. The method of Bartlett* et al. *(2015) had the best performance overall, with small finite sample bias and coverage rates close to nominal values. We apply the imputation approaches to a Chagas disease dataset.*

Keywords: *Missing covariates, Cox regression, multiple imputation, simulation study, censoring-ignorable MAR.*

[†] Correspondent author: jlpadilha@ufpr.br.

Introduction

Analysis of time-to-event data often involves missing covariates and the Cox proportional hazards model (Cox, 1972) is generally adopted for analysis. A reasonable assumption is that data are censoring-ignorable MAR in the sense that missingness does not depend on censoring time but may depend on failure time. In this case, the so-called *complete-case analysis* (that is, discarding missing data in covariates) leads to loss of efficiency and results in bias in the estimates of the regression parameters (Hsu & Yu, 2019). When data are censoring-ignorable MAR, we cannot directly use the Cox partial likelihood since we need to model the failure time and the covariates jointly (Chen *et al.*, 2009). Common approaches to the missing data problem are inverse probability weighting (IPW) (Robins *et al.*, 1994) and multiple imputation (MI) (Little & Rubin, 2019). IPW methods require a model for the probability that an individual has complete data and uses estimated weights to rebalance the complete cases so that the complete data are representative of the whole sample (Seaman *et al.*, 2012). MI, on the other hand, needs a model for the joint distribution of the missing data (a multivariate outcome) given the observed data and is generally more efficient than IPW (Seaman & White, 2013). Qi *et al.* (2010) presented a comparison of multiple imputation and fully augmented weighted estimators. Hsu & Yu (2019) proposed a nonparametric multiple imputation approach and compared it with existing augmented IPW methods. Yi *et al.* (2020) developed a method based on inverse probability weighting with the propensity estimated by nonparametric kernel regression.

The focus of this work is on multiple imputation. The method was initially proposed by Rubin (1987) and is now a well-established technique for analyzing data sets where some units have incomplete observations (Carpenter *et al.*, 2006). The problem with developing the imputation model for survival data is that, excluding some very special cases, the conditional distribution of covariates given survival time does not follow any common distribution (Carpenter & Kenward, 2012). White & Royston (2009) developed approximations to the imputation model which are valid for small covariate effects and/or small cumulative incidence White & Royston (2009). In their approach, the cumulative hazard is used in the imputation model replacing the observed survival time. Bartlett *et al.* (2015) presented an imputation model that incorporates the substantive model – the Cox model – in the derivation of imputations. Because the conditional distribution of covariates given survival times is unknown, a rejection algorithm is used to simulate draws from this predictive distribution. The acceptance probability depends on whether the missing observations refer or not to a censored individual. Unfortunately, Bartlett *et al.* (2015) simulations considered only missing completely at random (MCAR) data. There is a growing interest in the use of machine learning techniques for multiple imputation (Van Buuren, 2018). A popular class of algorithms is Classification and Regression Trees (CART) (Breiman *et al.*, 2017) which have been promoted as strong tools for prediction modeling (Steyerberg *et al.*, 2019). This nonparametric technique uses recursive partitioning, a statistical method to construct binary trees. CART methods are robust against outliers, can deal with multicollinearity and skewed distributions, and are flexible to fit interactions and non-linear relations. There is, to our knowledge, no comparison of these imputation methods for the case of censoring-ignorable MAR data.

Our goal is to present a comparison of MI methods for imputing missing covariates data when missingness of the covariate is outcome related. Specifically, we compare the performance of three multiple imputation approaches: (i) the method proposed by White & Royston (2009), (ii) the method described by Bartlett *et al.* (2015), and (iii) the CART approach. The imputation methods are applied to the Chagas disease study which involved 619 patients between the years of 1999 and 2019. The aim of the study was to identify factors that influence the risk of death in patients with heart disease due to Chagas. A covariate of particular interest, the right ventricular Tei index, was missing for 182 (29.4%) of the patients for which we believe the missingness propensity could depend on the survival outcome.

This paper is organized as follows. In Section , we discuss the multiple imputation methods in the context of missing covariates in the survival model. In Section , we give results from a simulation study. In Section , we apply the methods to the Chagas data. A discussion follows in Section . Finally, Section concludes the paper.

Methods

Let T_i the time at which follow-up of individual i ends, with $D_i = 1$ if the T_i is the survival time and $D_i = 0$ if T_i is the censoring time, for $i = 1, \ldots, n$. When $D_i = 0$ the event occurs at some $\tilde{T}_i > T_i$. Assume that we have two covariates Y_{i1} and Y_{i2}, so our data consists of the triple $(T_i, D_i, \boldsymbol{Y}_i)$ for the ith individual, where the covariate vector $\boldsymbol{Y}_i$ is measured at baseline. We also assume the survival data are CAR (censored at random), that is, conditional on covariates in the survival model, the censoring process is independent of the survival times.

For a survival time distribution $f(t),\ t \geq 0$ with cumulative distribution function $F(t)$, the survival function is $S(t) = Pr(T > t) = 1 - F(t)$, the hazard function is $h(t) = f(t)/S(t)$ and the cumulative hazard is $H(t) = \int_0^t h(s)ds = -\log\{S(t)\}$.

Under the Cox proportional hazards model (Cox, 1972, 1975) we have $h(T_i|Y_{i1}, Y_{i2}) = h_0(T_i)\exp(\beta_1 Y_{i1} + \beta_2 Y_{i2})$ and $H(T_i|Y_{i1}, Y_{i2}) = H_0(T_i)\exp(\beta_1 Y_{i1} + \beta_2 Y_{i2})$. With complete data (i.e. no data are missing), the estimation of $\boldsymbol{\beta}$ is accomplished by solving the following estimating equations

$$\boldsymbol{U}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} D_i \left[\boldsymbol{Y}_i - \frac{\sum_{j \in R(T_i)} \boldsymbol{Y}_i \exp(\beta_1 Y_{i1} + \beta_1 Y_{i2})}{\sum_{j \in R(T_i)} \exp(\beta_1 Y_{i1} + \beta_1 Y_{i2})} \right] = \boldsymbol{0}, \tag{1}$$

where $R(T_i)$ is the set of all individuals who are still under study at a time just prior to t_i. The maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$, from (1) is consistent and asymptotically normal under certain regularity conditions (Tsiatis, 1981).

We turn now to the problem of estimation with missing covariates. Suppose that Y_{i1} is subject to missing data and Y_{i2} is always observed and let $R_i = 1$ if Y_{i1} is observed and 0 if Y_{i1} is missing. Denote by $\boldsymbol{Y}_{obs}$ the observed component of $\boldsymbol{Y}$ and by $\boldsymbol{Y}_{mis}$ the missing counterpart. The missing value mechanism relates the probability of observing unit i's data giver their potentially unseen values (Carpenter & Kenward, 2012). The complete-case analysis of $\boldsymbol{\beta}$ is based on the solution to (1) using only those individuals with $R_i = 1$. This analysis can be quite inefficient if there are appreciable missing values and will be biased if missingness depends on the outcome (Yi *et al.*, 2020; Paik & Tsai, 1997; Kalbfleisch & Prentice, 2011; Rathouz, 2007). With missing data, a consistent estimator of $\boldsymbol{\beta}$ can be obtained by taking the expectation over the conditional predictive distribution $f(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{R})$, that is, solving the estimating equation

$$E_{f(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{R})} \left\{ \boldsymbol{U}(\hat{\boldsymbol{\beta}}) \right\} = \boldsymbol{0}. \tag{2}$$

The MI solution reverses the order of expectations and solution in (2). The main idea is repeatedly draw missing values $\tilde{\boldsymbol{Y}}_{mis}$ from the (Bayesian) conditional predictive distribution of the missing observations, $f(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{R})$, and solve $\boldsymbol{U}(\hat{\boldsymbol{\beta}}, \boldsymbol{Y}) = \boldsymbol{0}$. Then, combine the results in a single inference taking into account the imputation uncertainty. Multiple imputation is generally more efficient than complete-case analysis (Van Buuren, 2018). Imputation of partially observed covariates under the Cox regression model is complicated because, excluding some special cases, the conditional distribution of covariates given survival times will not follow any common distribution (Carpenter & Kenward, 2012; White & Royston, 2009).

In this paper, we compare three ways to overcome this problem: (i) the approximation of White & Royston (2009) that uses the cumulative hazard in the imputation model, (ii) the method of Bartlett *et al.* (2015) that incorporates de Cox model in the imputation process factoring $f(T_i, Y_{i1}, Y_{i2}) = f(T_i|Y_{i1}, Y_{i2}) f_{12}(Y_{i1}, Y_{i2})$, and (iii) the CART approach to impute missing values. Because we are not using full Bayesian framework, the draw of missing values from the predictive distribution is approximate.

All these imputation models can be used in an approach commonly referred to as full conditional specification (FCS) which imputes missing data on a variable-by-variable basis (Carpenter *et al.*, 2006; Van Buuren, 2018; White *et al.*, 2011). FCS has the ability to handle different variable types because each variable is imputed using its own imputation model.

Using the cumulative hazard

White & Royston (2009) proposed an imputation approach from the consideration of the conditional distribution of covariates given the survival time. Under the proportional hazards model, the log conditional distribution of Y_1 given Y_2 and the survival time is, up to a constant of proportionality,

$$\begin{aligned}\log\{f(Y_1|T,D,Y_2)\} = &\log\{f(Y_1|Y_2)\} + D\left[\log\{h_0(T)\} + (\beta_1 Y_1 + \beta_2 Y_2)\right] - \\ &H_0(T)\exp(\beta_1 Y_i + \beta_2 Y_2).\end{aligned} \tag{3}$$

When Y_1 is binary in (3), White & Royston (2009), using Taylor series approximation valid when $Var(Y_2)$ is small, showed that we can write

$$\begin{aligned}\operatorname{logit}\{Pr(Y_1=1|T,D,Y_2)\} &= \operatorname{logit}\{Pr(Y_1=1|Y_2)\} + D\beta_1 - H_0(T)(e^{\beta_1}-1)e^{\beta_2 Y_2} \\ &\approx \zeta_0 + \zeta_1 Y_2 + \zeta_2 D + \zeta_3 H_0(T) + \zeta_4 H_0(T)\times Y_2,\end{aligned} \tag{4}$$

for constants $\zeta_0, \ldots, \zeta_4$. That is, (4) is approximately the logistic regression of Y_1 on Y_2, D, $H_0(T)$. The result is exact when Y_2 is not present; otherwise, it is approximated and the approximation gets worse for larger $Var(Y_2)$. When Y_1 is continuous, in particular $Y_1|Y_2 \sim N(\alpha_0+\alpha_1 Y_2, \sigma^2)$, we have from (3),

$$\log\{f(Y_1|T,D,Y_2)\} \approx \frac{(Y_1-\alpha_0-\alpha_1 Y_2)^2}{2\sigma^2} + D\beta_1 Y_1 - H_0(T)e^{(\beta_1 Y_1+\beta_2 Y_2)}. \tag{5}$$

Following a fuller Taylor series approximation for $e^{(\beta_1 Y_1+\beta_2 Y_2)}$, valid when $Var(Y_1)$ and $Var(Y_2)$ are small, White & Royston (2009) concluded that the imputation model resulting from (5) is approximately a linear regression on D, $H_0(T)$ and Y_2. The authors also pointed that the addition of an interaction term $Y_2 \times H_0(T)$ could improve the accuracy of the approximation. In a simulation study comparing various strategies for estimating the cumulative hazard $H_0(T)$, the authors concluded that the method that approximates $H_0(t)$ by $H(t)$, the Nelson-Aalen estimate of the cumulative hazard, the censoring indicator and Y_2 as covariates was the best method in general.

Incorporating the substantive model

Now we describe the modified FCS approach of Bartlett *et al.* (2015) that accommodates the substantive model in the imputation process. This is known as congenial imputation model (Carpenter & Kenward, 2012). Because the conditional distribution of covariates given survival times does not belong to a standard parametric family, a rejection algorithm is used to simulate draws from this predictive distribution.

The first step is to fit the substantive model (the Cox model) to the observed data and currently imputed values. Then, the maximum partial likelihood estimators $\hat{\beta}$ and its associated covariance matrix $\widehat{\Omega}$ are used to approximate a draw from the Bayesian posterior by drawing β from $N(\hat{\beta}, \widehat{\Omega})$ and extract the current estimate of $H_0(t)$. Define a proposal distribution $f(\cdot)$ for missing Y_1 given Y_2 and considers the proposal Y_{i1}^* for the ith individual missing Y_1. The acceptance probability depends on whether or not it refers to a censored individual. If $D_i = 0$, a censored observation, the acceptance probability is

$$S(T_i|Y_{i1}^*, Y_{i2}; \beta) \tag{6}$$

while for $D_i = 1$, an uncensored observation, we accept Y_{i1}^* with probability

$$H_0(t)\exp\left[1 + (Y_{i1}^*\beta_1 + Y_{i2}\beta_2) - H_0(t)\exp(Y_{i1}^*\beta_1 + Y_{i2}\beta_2)\right]. \tag{7}$$

The proposal distribution $f(\cdot)$ depends on the nature of the missing covariate. That is, for Y_1 binary we consider a logistic regression while a linear regression can be used when Y_1 is continuous. If Y_2 is also subject to missingness we should consider a proposal distribution for missing Y_2 given Y_1 and simply add another step to the algorithm.

Classification and regression trees

Classification and regression trees (CART) (Breiman *et al.*, 2017) are a popular class of machine learning algorithms. CART models seek predictors and cut points in the predictors that are used to split the sample. The method is based on statistically optimal splitting of the individuals into pairs of smaller subgroups. Splits are based on cut-off levels of the predictors, which produce maximum separation among two subgroups and a minimum variability with these subgroups with respect to the outcome (Steyerberg *et al.*, 2019). The target variable can be discrete (classification tree) or continuous (regression tree).

CART methods have properties that make them attractive for imputation: they are robust against outliers, can deal with multicollinearity and skewed distributions, and are flexible enough to fit interactions and nonlinear relations. The idea is to form a donor pool of all observed cases at the terminal node of the fitted tree, and then randomly draw a case from the donor group to be used as the imputed value. Parameter uncertainty is incorporated by fitting the tree on a bootstrap sample.

Assume again we have two covariates Y_1 and Y_2 where Y_1 is subject to missingness. We have n_1 observed values and n_0 missing values for Y_1. Based on Van Buuren (2018), the major steps of the proposed algorithm for creating imputations is as follows.

1. Draw a bootstrap sample of size n_1 from the observed data.
2. Fit a tree model $f(Y_1)$ to the bootstrap sample using Y_2, T and D as predictors.
3. Predict the n_0 terminal nodes g_j from $f(Y_{1,miss})$.
4. Construct n_0 sets Z_j of all cases at node g_j, each containing d_j candidate donors.
5. Draw one donor i_j from Z_j randomly for $j = 1, \ldots, n_0$.
6. Calculate the imputations $\widetilde{Y}_{j1} = Y_{i_j1}$ for $j = 1, \ldots, n_0$.

The composition of the donor groups will vary over different bootstrap replications, thus incorporating sampling uncertainty about the tree (Van Buuren, 2018). The algorithm is repeated to produce M imputed datasets. Again, the FCS framework can be used in the case of multiple missing covariates.

The multiple imputation estimator

Once M multiply imputed datasets are obtained, the results are combined using the rules established in Rubin (1987). Specifically, let $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{V}}_k$ denote the estimate of $\boldsymbol{\beta}$ and its covariate matrix from the fit of the Cox regression model to the kth completed dataset, $i \in (1, \ldots, M)$. The MI estimate of $\boldsymbol{\beta}$ is the simple average

$$\widehat{\boldsymbol{\beta}}_{MI} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\boldsymbol{\beta}}_i. \tag{8}$$

Define the average within-imputation covariance matrix as

$$\widehat{\boldsymbol{W}} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\boldsymbol{V}}_i, \tag{9}$$

and the between-imputation covariance matrix,

$$\widehat{\boldsymbol{B}} = \frac{1}{M-1} \sum_{i=1}^{M} (\widehat{\boldsymbol{\beta}_i} - \widehat{\boldsymbol{\beta}}_{MI})(\widehat{\boldsymbol{\beta}_i} - \widehat{\boldsymbol{\beta}}_{MI})^T. \tag{10}$$

Then, an estimate of the covariance of $\widehat{\boldsymbol{\beta}}_{MI}$ is given by

$$\widehat{\boldsymbol{V}}_{MI} = \widehat{\boldsymbol{W}} + \left(1 + \frac{1}{M}\right) \widehat{\boldsymbol{B}}. \tag{11}$$

So, the precision for $\widehat{\boldsymbol{\beta}}_{MI}$, given by (11) involves three sources of variation: the between- and within-imputation variability plus the extra variance caused by the fact that a finite number of imputations is used for estimating $\boldsymbol{\beta}$. Traditional choices for M are $M = 3$, $M = 5$ and $M = 10$. The larger M gets, the smaller the effect of simulation error on the total variance (Van Buuren, 2018). For a scalar β and large sample sizes, the reference distribution for interval estimates and significance tests is a t distribution,

$$(\beta - \hat{\beta}_{MI})\widehat{V}_{MI}^{-1/2} \sim t_\nu,$$

where the degrees of freedom, obtained from a Satterthwaite approximation (Little & Rubin, 2019), are given by

$$\nu = (M-1)\left\{1 + \frac{1}{M+1}\frac{\widehat{W}}{\widehat{B}}\right\}^2. \tag{12}$$

Simulation Study

In what precedes, various imputation approaches to overcome the bias occurring in the complete-case analysis have been presented. It is of interest to quantify the bias and precision under various scenarios of missing data and censoring. To this end, a simulation study was conducted. The following estimators were compared: the estimator computed without missing values (FULL), the complete-case (CC) estimator, the MI estimator using the congenial model (CONG), the MI estimator using the approximation of White and Royston (WR) and the CART approach (CART).

Data generation

The data generation process was based on the work of Yi *et al.* (2020). Three different settings were considered with varying censoring and missing mechanisms. Tables 1-3 provide details about data generation. The following is a summary of the simulation scheme.

- *Covariate vector:* The time-independent covariate vector is (Y_1, Y_2) in settings 1-2 and (Y_1, Y_2, Y_3) in setting 3.
- *True hazard of T:* The survival times follows a Cox proportional hazards model with baseline hazard $h_0(t)$, which is equal to 1 in settings 1-2 and $t/2$ in setting 3.
- *True propensity:* The probability of observing Y_1 is logistic depending only on T in setting 1, on T and Y_2 in settings 2-3.
- *True censoring:* The censoring time follows another Cox proportional hazards model, which depends on no covariate in setting 1, on Y_1 in setting 1, and on (Y_1, Y_2) in setting 3.
- *Censoring and missing rate:* The censoring rate varied from 37% to 82% and the missing rate varied from 29% to 47%.

In each setting, the survival times follow a Cox proportional hazards model with baseline hazard and the censoring times follow another Cox proportional hazards models which may or may not depend on covariates. That is, censoring and survival times are independent conditional on covariates.

Computation

All computations were done in R 4.1.1, using the packages survival, mice and tidyverse. Samples of size $n = 500$ and $n = 1000$ were generated according to the settings given in Tables 1-3. A total of $S = 1000$ such samples were generated. For each sample, the five estimators under comparison were obtained. $M = 10$ multiple imputations were considered for all imputation methods and the function mice::pool() was used to combine the inferences. For the WR and CONG methods, the cumulative baseline hazard $H_0(t)$ was estimated via mice::nelsonaalen() and survival::basehaz(), respectively, and entered the model in a smoothed version obtained through stats::loess() function. Multiple imputations for these methods were obtained using the function mice::mice() with default arguments. The CONG method was implemented taking a single imputation from the CART algorithm as starting values for missing observations. The algorithm was run for another 100 steps and the imputations was taken as imputed values in steps multiples of 10.

Several measures were computed to measure the relative performance of the various methods. Bias was defined as the difference between the estimate and the true value of the parameter, SE was defined as the asymptotic standard error of the estimator. In addition, the empirical standard deviation SD of the estimator over $S = 1000$ simulations and empirical coverage probability CP of 95% confidence intervals were calculated. For both Full and CC methods, the confidence interval was constructed based on the normal approximation, and for imputation methods, it was constructed based on the t distribution.

Tables 1-3 present a summary of the simulation results.

Results for the first setting

Table 1 shows the results for the first setting with 1000 simulation runs and sample sizes $n = 500$ and $n = 1000$. Mean proportion of missing data and censored observations were 44.8% and 47.6%, respectively.

In summary, CC showed considerable bias for both parameter estimates, with low empirical coverage rates increasing for the larger sample size. In terms of coverage and bias, WR performed as poorly as CC analysis. CART showed small bias but underestimated the variance, implying empirical coverage rates under the nominal level. CONG method produced the best results, with negligible small-sample bias and good empirical coverage. In general, MI recovered information, as can be seen from the standard error of estimators between those of FULL and CC analysis.

Tabela 1: Bias, SE, SD and CP based on 1000 runs under setting 1

		Estimation of β_1				Estimation of β_2			
n	Method	Bias	SE	SD	CP	Bias	SE	SD	CP
500	FULL	0.002	0.078	0.079	0.952	0.000	0.133	0.133	0.954
	CC	0.150	0.116	0.120	0.756	0.145	0.196	0.199	0.892
	CONG	0.010	0.102	0.109	0.940	-0.015	0.153	0.156	0.941
	WR	-0.176	0.111	0.080	0.655	-0.136	0.160	0.134	0.902
	CART	0.012	0.094	0.116	0.886	-0.053	0.146	0.168	0.880
1000	FULL	0.004	0.055	0.056	0.944	0.002	0.094	0.093	0.948
	CC	0.150	0.081	0.084	0.544	0.150	0.137	0.138	0.810
	CONG	0.007	0.071	0.077	0.937	-0.016	0.107	0.107	0.948
	WR	-0.175	0.078	0.055	0.354	-0.138	0.112	0.094	0.795
	CART	0.019	0.065	0.082	0.873	-0.027	0.102	0.117	0.901

Covariate vector: (Y_1, Y_2), $Y_1 \sim N(0,1)$, $Y_2 \sim \text{Ber}(0.5)$, $Y_1 \perp Y_2$
True hazard of T: $h(T) = \exp(\beta_1 Y_1 + \beta_2 Y_2)$, $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1, 1)$
True propensity: $Pr(R = 1) = 1 - \{1 + \exp(T - 0.5)\}^{-1}$
True censoring: $S_C(T) = \exp(-T^{1/2})$
Source: Author.

Results for the second setting

Table 2 shows the results for the second setting with 1000 simulation runs and sample sizes $n = 500$ and $n = 1000$. Mean proportion of missing data and censored observations were 29.4% and 37.0%, respectively.

Tabela 2: Bias, SE, SD and CP based on 1000 runs under setting 2

		Estimation of β_1				Estimation of β_2			
n	Method	Bias	SE	SD	CP	Bias	SE	SD	CP
500	FULL	0.002	0.122	0.124	0.947	0.006	0.123	0.125	0.950
	CC	0.080	0.148	0.150	0.914	0.160	0.154	0.160	0.836
	CONG	0.002	0.138	0.141	0.947	-0.003	0.128	0.131	0.948
	WR	0.001	0.144	0.141	0.956	-0.024	0.129	0.126	0.954
	CART	0.014	0.136	0.148	0.938	-0.008	0.127	0.133	0.937
1000	FULL	0.002	0.086	0.086	0.946	0.002	0.086	0.090	0.942
	CC	0.079	0.104	0.108	0.881	0.156	0.108	0.113	0.705
	CONG	0.003	0.097	0.101	0.935	-0.007	0.090	0.093	0.940
	WR	0.004	0.101	0.100	0.947	-0.027	0.090	0.089	0.931
	CART	0.016	0.095	0.105	0.913	-0.006	0.089	0.096	0.936

Covariate vector: (Y_1, Y_2), $Y_1 \sim \text{Ber}(0.5)$, $Y_2 \sim \text{Ber}(0.5)$, $Y_1 \perp Y_2$
True hazard of T: $h(T) = \exp(\beta_1 Y_1 + \beta_2 Y_2)$, $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1, 1)$
True propensity: $Pr(R = 1) = 1 - \{1 + \exp(T + Y_2)\}^{-1}$
True censoring: $S_C(T) = \exp\left\{-T^{1/2} \exp(-Y_1/4)\right\}$
Source: Author.

In contrast to the first scenario, the bias of CC analysis is now larger for Y_2 than Y_1. Note that Y_2 is included in the propensity for missing Y_1. The performance of WR was good, with little bias and acceptable empirical coverage rates. As in the first setting, the CART method presented a small bias but underestimated uncertainty. CONG was the best imputation method again, with negligible bias and empirical coverage very close to the nominal level. Even though MI can recover information, the reduction in standard errors was smaller than in the first setting, which can be explained by the uncertainty of modeling a binary covariate.

Results for the third setting

Table 3 shows the results for the second setting with 1000 simulation runs and sample sizes $n = 1000$. Mean proportion of missing data was 46.9% while the censoring rate was 60.2% and 81.5% for the two scenarios considered.

Tabela 3: Bias, SE, SD and CP based on 1000 runs under setting 3

		Estimation of β_1				Estimation of β_2				Estimation of β_3			
α	Method	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
1	FULL	0.006	0.109	0.111	0.948	0.004	0.109	0.102	0.953	0.007	0.354	0.359	0.946
	CC	0.210	0.176	0.178	0.781	0.206	0.174	0.175	0.795	-0.036	0.575	0.597	0.947
	CONG	0.021	0.144	0.146	0.946	0.020	0.113	0.110	0.957	0.062	0.374	0.395	0.940
	WR	-0.001	0.165	0.155	0.955	-0.028	0.114	0.103	0.958	-0.020	0.376	0.365	0.952
	CART	0.053	0.129	0.188	0.809	-0.017	0.112	0.105	0.958	0.003	0.367	0.384	0.948
2	FULL	0.005	0.161	0.168	0.940	0.004	0.161	0.159	0.953	0.015	0.521	0.543	0.948
	CC	0.265	0.294	0.309	0.854	0.273	0.293	0.302	0.857	-0.026	0.962	0.973	0.949
	CONG	0.107	0.235	0.248	0.914	0.015	0.168	0.166	0.960	0.098	0.551	0.570	0.943
	WR	0.088	0.290	0.273	0.954	-0.030	0.169	0.160	0.958	0.020	0.552	0.538	0.963
	CART	0.055	0.199	0.361	0.692	-0.007	0.165	0.162	0.959	0.016	0.541	0.567	0.943

Covariate vector: (Y_1, Y_2, Y_3), $Y_1|Y_2, Y_3 \sim \text{Ber}(\text{logit}(0.5\text{Y}_2 - \text{Y}_3))$, $Y_2 \sim \text{Ber}(0.5)$, $Y_3 \sim \text{Unif}(0, 0.5)$, $Y_2 \perp Y_3$
True hazard of T: $h(T) = 0.5T \exp(\beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_2) = (1, 1, 1)$
True propensity: $Pr(R = 1) = 1 - \{1 + \exp(2T - Y_3 - 1.5)\}^{-1}$
True censoring: $S_C(T) = \exp\left\{-(\alpha T)^{1/2} \exp(0.2Y_1 + 0.1Y_2)\right\}$
Source: Author.

In this extreme scenario, large bias occurred in the estimation of the first two parameters. Note that the missing rate is higher than the previous settings and, although the true propensity for missing Y_1 involves Y_3, little bias was observed in the estimation of the third parameter.

- For $\alpha = 1$, WR gave the best results overall. CONG and CART were similar in terms of bias, although serious undercoverage was observed for the effect of Y_1 when the CART method was used.

- For $\alpha = 2$, CART had the best performance in terms of bias, followed by WR and CONG. However, empirical coverage for CART for the first parameter was about 69%. The underestimation of uncertainty is evident when SD and SE are compared.

Real Data Example

In this Section we apply the imputation methods to a dataset of Chagas patients. Chagas disease is a neglected tropical disease, with the majority of the individuals affected living in Latin America. It is caused by the protozoan Trypanosoma cruzi, with risk factors strongly related to low socioeconomic status. Chagas disease, is an important cause of heart failure, stroke, arrhythmia, and sudden death (Nunes *et al.*, 2018).

The study involved patients who were referred for treatment at the Hospital das Clínicas at the Federal University of Minas Gerais, Brazil, in the years of 1999 to 2019. A total of 619 patients with Chagas and idiopathic cardiomyopathy were included. The survival outcome is time to death from any cause, which occurred for 209 patients. Several patient characteristics were measured at baseline. The following variables were considered in this application: Chagas serology group (Chagas or idiopathic cardiomyopathy); ejection fraction (EF), a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction; the New York Heart Association (NYHA) Functional Classification, a categorization of cardiac symptoms on a patient's daily activities, varying from 1 (no limitation of physical activity) to 4 (unable to carry on any physical activity without discomfort); and right ventricular (RV) Tei index (RVTei), an echocardiographic measurement of right ventricular function, which is defined as the sum of the isovolumetric contraction and relaxation time divided by the ejection time of the RV. For 182 patients (29.4%), however, RVTei could not be obtained because one or both of the quantities used in the calculation could not be evaluated. RV dysfunction, assessed by the Tei index, is known to be a strong indicator of poor prognosis (Nunes *et al.*, 2008). Missingness propensity for this variable is related to the presence of arrhythmias, especially atrial fibrillation or ventricular arrhythmias, which can alter isovolumetric contraction and relaxation times, and right ventricular ejection time. In addition, the presence of arrhythmias expresses greater severity of myocardial involvement and greater risk of death in patients with heart disease due to Chagas. Therefore, we believe missingness in RVTei is outcome related and we further assume missingness does not depend on censoring times.

The analysis model of interest for these data is a Cox proportional hazards model including covariates RVTei, Chagas group, ejection fraction and NYHA. We fitted the three imputation strategies as previously described in Session and compared them to the CC analysis. MI was conducted with the same configuration as the simulation study so $M = 10$ imputations were considered. Table 4 summarizes the results of estimating the coefficients of covariates in Cox regression.

Tabela 4: Chagas data: results of Cox regression by complete cases and three imputation methods

	CC			CONG			WR			CART		
	Est.	SE	P value	Est.	SE	P value	Est.	SE	P value	Est.	SE	P value
RVTei	1.334	0.271	0.000	1.173	0.242	0.000	1.256	0.272	0.000	1.263	0.285	0.000
Chagas	0.763	0.208	0.000	0.966	0.190	0.000	0.926	0.185	0.000	0.916	0.189	0.000
EF	-0.052	0.009	0.000	-0.065	0.007	0.000	-0.060	0.008	0.000	-0.060	0.008	0.000
NYHA2	0.180	0.205	0.378	0.237	0.190	0.216	0.238	0.186	0.203	0.239	0.185	0.198
NYHA3	0.441	0.256	0.085	0.489	0.233	0.037	0.481	0.234	0.042	0.476	0.232	0.042
NYHA4	1.082	0.310	0.000	1.075	0.270	0.000	0.859	0.281	0.003	0.926	0.271	0.001

Source: Author.

Based on these results, the following conclusions can be drawn.

1. All variables were found to be important to explain time to death. In particular, the risk is higher among patients in the Chagas group, with lower ejection fraction values, greater functional impairment and higher Tei index values.

2. WR, CONG, and CART all yielded similar conclusions. Compared with CC analysis, imputation methods gave higher estimated effects for the coefficients associated with RVTei and NYHA class 4, but lower effects for the others. Setting a significance level of 5%, the crucial change in inferences occurs for the NYHA class 3 effect, whose comparison with class I changes from non-statistically significant in the CC analysis to significant in all models. There is no evidence that NYHA class II patients differ from those in class I. The remaining effects are highly significant.

3. As occurred in the simulation study, the standard errors oF the CC estimator was generally higher compared to MI, which explains the change in significance of the NYHA class II effect. In general, the CART method had the lowest SE. The inference using this method should be viewed with caution, due to the underestimation of the uncertainty of the imputation that resulted in anti-conservative coverage probabilities.

Discussion

Multiple imputation has become the dominant approach in medical research to deal with missing values (Steyerberg *et al.*, 2019). MI is attractive because it is both practical and widely applicable. To be valid, each imputation must be a *Bayesian* draw from the conditional predictive distribution of the missing observations, in a way that the multiply imputed values take into account the uncertainty about the imputed value. Carpenter & Kenward (2012) provide the

Bayesian justification of MI and discuss the conditions under which the approximation provided by MI has the expected frequentist properties.

MI methods are not robust to misspecification of the conditional distribution of missing values given the observed quantities. Bartlett *et al.* (2015) note that when the Cox proportional hazards model is the substantive model, standard software implementations of the MI may impute values from models that are incompatible with the substantive model. This is the case when the true response model involves interactions and/or non-linear terms (Carpenter & Kenward, 2012). In this sense, the approximation by White & Royston (2009) is expected to be valid only for small covariate effects and/or small cumulative incidence. Because imputations are drawn conditional on a chosen statistical model, if the imputation model is wrong, then parameter estimates are inconsistent (Carpenter *et al.*, 2006). To overcome the dependence on the specified parametric model, we seek to investigate the performance of the CART approach. This statistical learning method is known for its predictive performance and the ability to fit interactions and nonlinear relations. As mentioned in Van Buuren (2018), unfortunately nearly all implementations of tree methods produce single imputations. Furthermore, we are not aware of any comparisons of this method for outcome related missing covariate missingness.

Simulation results confirmed that CC analysis is severely biased when missingness is related to the outcome. The performance of the WR method was highly dependent on the simulation setting. In the first setting, the estimator returned bias even larger than CC analysis although the performance was good in the other settings. This result emphasizes the point that when imputation models are wrong, resulting estimates are not consistent and inferences are invalid. The predictive performance of CART has also been confirmed. The variables can enter the imputation model without any transformation, therefore reducing additional work by the data analyst. For the WR and CONG methods, on the other hand, estimates of the cumulative baseline function are required for the imputation model. However, underestimation of imputation uncertainty have been found in all settings for CART. As a result, confidence intervals presented lower than nominal coverage probabilities. CART imputes values from a donor pool comprised of all observed cases at the terminal node. Depending on the missingness propensity, the donor pool may not carry the correct sampling uncertainty. It appears to be a problem especially when sample sizes are small or the imputation model contains variables strongly related to the missingness. CONG presented good performance in all cases, with negligible bias and good CP. In settings 1-2, the estimates were very close to the FULL analysis in terms of Bias and CP. In setting 3, however, due to the high proportion of censored observations, the performance distinction between the imputation methods were not so clear. A more comprehensive investigation of the relative merits of these imputation methods in the case of multiple missing covariates is lacking.

Even though we can never recover the true missing data, MI recovers information from the observed data (Carpenter & Kenward, 2012). As a consequence, MI results in narrower confidence intervals as compared to CC analysis. In addition, improvement can be obtained through adding additional variables in the imputation model variables that are predictive of missing data and/or missing data mechanism (Collins *et al.*, 2001; Rubin, 1996).

An alternative to the MI approach, not pursued in this paper, is the IPW method (Yi *et al.*, 2020). IPW estimators include only individuals who were fully observed and are sensitive to the choice of weighting model. The inverse probability-weighted estimates are known to be less efficient compared to MI (Seaman *et al.*, 2012). Methods for gaining efficiency and robustness to the weighting model have been proposed (Hsu & Yu, 2019; Qi *et al.*, 2010). A comparison between them will be the focus of future research.

Final Remarks

Missing data are a pervasive problem in the analysis of epidemiological and clinical data. In this paper, we discussed the problem of missing covariate values and their impact on inferences for the Cox model.

As shown in a simulation study, analysis of complete data results in loss of efficiency and causes biased estimates of regression model parameters when the missing data mechanism involves the response. Multiple imputation is an elegant and powerful technique that can be used to solve the problem. However, care must be taken when choosing the imputation model, as the inferences are not robust to the misspecification of the model. In this sense, our simulation study demonstrated that the White and Royston approximation can perform worse than the complete-cases analysis, whereas the CART method, although it is recognized as a powerful predictive tool, can considerably underestimate the imputation uncertainty, resulting in very low coverage. The MI strategy that accommodates the analysis model in the imputation was, in general, the best method.

In simulations, we have assumed that only one covariate is subject to missing data and that the proportional hazards assumption holds. Real-world situations often involve multiple covariates with missing values. Missingness in these covariates, which are often time-dependent, pose additional challenges to the analysis. Other practical problems may involve non-proportional hazards, cure fraction and informative censoring. Further research on the imputation methods is still required.

Acknowledgements

We thank Maria do Carmo Pereira Nunes (School of Medicine, Federal University of Minas Gerais) for permission to use the Chagas data.

Disclosure statement

No potential conflict of interest was reported by the author.

Data availability statement

The R code and data that support the findings of this study are available from the corresponding author on request.

Referências

Bartlett, Jonathan W, Seaman, Shaun R, White, Ian R, Carpenter, James R, & Initiative*, Alzheimer's Disease Neuroimaging. 2015. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, **24**(4), 462–487.

Breiman, Leo, Friedman, Jerome H, Olshen, Richard A, & Stone, Charles J. 2017. *Classification and regression trees*. Routledge.

Carpenter, James, & Kenward, Michael. 2012. *Multiple imputation and its application*. John Wiley & Sons.

Carpenter, James R., Kenward, Michael G., & Vansteelandt, Stijn. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(3), 571–584.

Chen, Ming-Hui, Ibrahim, Joseph G, & Shao, Qi-Man. 2009. Maximum likelihood inference for the Cox regression model with applications to missing covariates. *Journal of multivariate analysis*, **100**(9), 2018–2030.

Collins, Linda M, Schafer, Joseph L, & Kam, Chi-Ming. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, **6**(4), 330.

Cox, David R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Cox, David R. 1975. Partial likelihood. *Biometrika*, **62**(2), 269–276.

Hsu, Chiu-Hsieh, & Yu, Mandi. 2019. Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical methods in medical research*, **28**(6), 1676–1688.

Kalbfleisch, John D, & Prentice, Ross L. 2011. *The statistical analysis of failure time data.* Vol. 360. John Wiley & Sons.

Little, Roderick JA, & Rubin, Donald B. 2019. *Statistical analysis with missing data.* Vol. 793. John Wiley & Sons.

Nunes, Maria Carmo Pereira, Beaton, Andrea, Acquatella, Harry, Bern, Caryn, Bolger, Ann F, Echeverria, Luis E, Dutra, Walderez O, Gascon, Joaquim, Morillo, Carlos A, Oliveira-Filho, Jamary, *et al.* 2018. Chagas cardiomyopathy: an update of current clinical knowledge and management: a scientific statement from the American Heart Association. *Circulation*, **138**(12), e169–e209.

Nunes, Maria do Carmo Pereira, Rocha, Manoel Otávio C, Ribeiro, Antônio Luiz P, Colosimo, Enrico A, Rezende, Renato A, Carmo, Guilherme Augusto A, & Barbosa, Marcia M. 2008. Right ventricular dysfunction is an independent predictor of survival in patients with dilated chronic Chagas' cardiomyopathy. *International journal of cardiology*, **127**(3), 372–379.

Paik, Myunghee Cho, & Tsai, Wei-Yann. 1997. On using the Cox proportional hazards model with missing covariates. *Biometrika*, **84**(3), 579–593.

Qi, Lihong, Wang, Ying-Fang, & He, Yulei. 2010. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in medicine*, **29**(25), 2592–2604.

Rathouz, Paul J. 2007. Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics*, **8**(2), 345–356.

Robins, James M, Rotnitzky, Andrea, & Zhao, Lue Ping. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, **89**(427), 846–866.

Rubin, Donald B. 1987. *Multiple imputation for survey nonresponse.*

Rubin, Donald B. 1996. Multiple imputation after 18+ years. *Journal of the American statistical Association*, **91**(434), 473–489.

Seaman, Shaun R, & White, Ian R. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, **22**(3), 278–295.

Seaman, Shaun R, White, Ian R, Copas, Andrew J, & Li, Leah. 2012. Combining multiple imputation and inverse-probability weighting. *Biometrics*, **68**(1), 129–137.

Steyerberg, Ewout W, *et al.* 2019. *Clinical prediction models.* Springer.

Tsiatis, Anastasios A. 1981. A large sample study of Cox's regression model. *The Annals of Statistics*, **9**(1), 93–108.

Van Buuren, Stef. 2018. *Flexible imputation of missing data.* CRC press.

White, Ian R., & Royston, Patrick. 2009. Imputing missing covariate values for the Cox model. *Statistics in medicine*, **28**(15), 1982–1998.

White, Ian R, Royston, Patrick, & Wood, Angela M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, **30**(4), 377–399.

Yi, Yanyao, Ye, Ting, Yu, Menggang, & Shao, Jun. 2020. Cox regression with survival-time-dependent missing covariate values. *Biometrics*, **76**(2), 460–471.