A Simulation Study Comparing Two Methods of Handling Missing Covariate Values when Fitting a Cox Proportional-Hazards Regression Model

Ali Satty[1] | *Elneelain University, Khartoum, Sudan*

Abstract

Missing covariate values is a common problem in a survival data research. The aim of this study is to compare the use of the multiple imputation (MI) and last observation carried forward (LOCF) methods for handling missing covariate values in the Cox proportional hazards (PH) regression model. The comparisons between the methods are based on simulated data. The missingness mechanism is assumed to be missing at random (MAR). Missing covariate values are generated under different missingness rates. The results from both methods are compared by assessing the bias, efficiency and coverage. The simulation results in general revealed that MI is likely to be the best under the MAR mechanism.

Keywords

Missing covariate values, multiple imputation (MI), Cox proportional hazard model , Last observation carried forward (LOCF), missing at random (MAR)

JEL code

C15

INTRODUCTION

One of the challenges in modeling practice is missing data. A problem occurs when some data on covariates are missing in survival analysis, where the Cox proportional-hazards (PH) model (Cox, 1972) is usually used for analysis. Covariate observations may be missing for some individuals, for whatever reason. An important concept with missing data, specifically where there are multiple covariates with missing values, relates to the mechanism of missing data. Rubin (1976, 1987) classified these mechanisms into three basic categories: missing completely at random (MCAR), meaning that the missingness process does not depend on the observed responses, missing at random (MAR), when the missingness process depends on the observed responses and probably on measured covariates but not on the unob-

[1] School of Statistics and Actuarial Science, Elneelain University, Khartoum, Sudan. E-mail: Alisatty1981@gmail.com.

served responses, and missing not at random (MNAR) which allows the missingness process to depend on the unobserved responses as well as on the observed responses.

Given the problems that can arise in the Cox PH model when there are missing covariate values, the following question is forced upon researchers. What methods can be utilized to handle these potential pitfalls? The goal is to use approaches that better avoid the generation of biased results. There are several ways to deal with missing covariate values in Cox PH model. One recommendation is to discard subjects with incomplete sequences, and then analyze only the units with complete data. Method that uses this solution is called complete case analysis (CC) (Little and Rubin, 1987). However, this method has numerous disadvantages leading to reduction in the sample size, which reduces the precision of estimates and therefore can lead to biased results (Schafer and Graham, 2002).

In contrast to the CC analysis, there are other ways that can help to tackle the problem of missing covariate values in Cox PH model. There are methods that do generate possible values for the missing covariates. These methods are called imputation methods, where one fills-in (imputes) the missing covariate values to obtain a full dataset, and the resultant data are then analyzed by standard statistical methods without concern as if the set represented the true and complete dataset (Rubin, 1987; Little and Rubin, 1987). This is the key idea behind commonly used procedures for imputation which include, simple and multiple imputations (Little and Rubin, 1987). There are different simple imputation methods. In this study however we restrict ourselves to outlining one of them, which is called last observation carried forward (LOCF). LOCF substitutes one value for every missing covariate value in the dataset (Little and Rubin, 1987, 2002). Under certain restrictive circumstances, LOCF can produce unbiased results. In addition, in some situations, LOCF does not produce conservative results. However, this approach can still provide conservative results, under some specific circumstances. The method will be readdressed in detail in the following section. In contrast to the LOCF method, MI fills in more than one value for each missing covariates item and carries out the analysis as if the imputed values were observed data to allow for the appropriate evaluation of imputation uncertainty (Rubin, 1987; Little and Rubin, 1987). MI was proposed by Rubin (1978) and described in detail by Little and Rubin (1987). Considerable research has focused on MI for handling missing covariate data in Cox PH model (See, Paik, 1997; van Buuren et al.1999; Brazi and Wooward, 2004; White and Royston, 2009).

This study deals with the problem of missing covariate values in the Cox regression model. It is devoted to a comparison of two imputation techniques or methods. The methods that were compared include multiple imputation (MI) and last observation carried forward (LOCF). The main objective of this paper is to study imputation techniques and compare them with others to estimate Cox PH model parameters with missing covariates values. The missing data mechanism is assumed to be MAR. The comparisons are based on a simulation study. The comparisons are made through the evaluation of bias, efficiency, and coverage. The rest of this paper is organized as follow: Section 1 describes the notation and model assumptions. An overview of methods for analyzing missing covariate values is also given. Section 2 presents the simulation study scheme including the study design, data generation and the evaluation criteria used in the analysis. The results from the simulations of the two methods are presented in section 3. Finally, a brief discussion and concluding remarks are provided in the last section.

1 METHODS

1.1 Notation and model assumptions

Assume there are n independent individuals. For each individual, $i = 1,..., n$. Let c and T be the censoring and failure, respectively. Now, we assume the hazard for individual i follows a Cox proportional hazards regression model:

$$\lambda\ (t \mid xi) = \lambda_0\ (t)\ exp\ (\beta' x_i), \qquad (1)$$

where $\lambda 0(t)$ is an unspecified baseline hazard function, x represent a set of independent covariates that may be categorical or continuous, and β is a $p \times 1$ parameter vector. In this study however an application will be confined to the continuous covariates case (i.e., x are continuous covariates). The vector of observed time to follow-up was obtained by $T = \min\ (T,\ c)$, and failure indicator vector δ by $\delta = 1$ if $T \leq c$ and $\delta = 0$ if censored. We suppose that x, T and c are independent. We restrict ourselves to consider that the survival time is fully observed, while some of the covariates x_i contains missing values. Now, partition the covariate vector x_i into its observed covariates and missing covariates, such that $x_i = (x^{obs}, x^{mis})$. Let R be a vector that represents the missing covariate process, with $R = 1$ if the covariate is observed (i.e. x^{obs}), and $R_{ij} = 0$ if the covariate is missing (i.e. x^{mis}). When MAR holds, the missing covariate mechanism is determined by the conditional distribution of R conditional upon (Z, δ, x^{obs}), which is Bernoulli with probability $\hbar = P\ (R = 1 \mid Z, \delta, x^{obs})$, where Z denotes the survival outcome. For each individual, let $(Z_i, \delta_i, x_i^{obs}, x_i^{mis}, R_i)$ denote *i.i.d* copies of $(Z, \delta, x^{obs}, x^{mis}, R)$. Thus the observed covariate data being analyzed are $(Z_i, \delta_i, x_i^{obs}, x_i^{mis})$ if $R = 1$, and $(Z_i, \delta_i, x_i^{obs})$ if R=0. There are a variety of methods that can be used to deal with missing covariate values (x^{mis}).The subsections that follow provide a review of the methods that are used in this study.

1.2 Multiple imputation (MI)

Following is a brief description of MI and its application. According to Rubin (1987), MI consists of three steps. First, each missing value is replaced by $M \geq 2$ simulated values. Each of these sets of plausible values can be used to fill-in the missing values and create a completed dataset. This method is valid under the MAR mechanism (Little and Rubin, 1987). Further, when MAR holds, for univariate x^{mis} and given the observed data (Z, δ, x^{obs}), sets of plausible values for missing observations (x^{mis}) can be created to reflect uncertainty about the stochastic non-response model. This can be done using an appropriate imputation model $P\ (x^{mis} | \ Z, \delta, x^{obs})$. In doing so, SAS PROC MI can be used. PROC MI fills in the missing covariate values and therefore the above univariate method can be conducted to each missing covariate x^{mis} in turn. This can be achieved using all the imputed values of the other missing covariates in case of creating new values of x^{mis}. This process is repeated until a suitable convergence criterion is satisfied. Second, each of the M complete datasets are analyzed using standard statistical methods, such as Cox proportional regression model. The use of the number of imputations M needs not be very large since, in practice, 3-10 imputations often provided satisfactory results (Schafer, 1997; Schafer and Olsen, 1998). Finally, the M results are combined using methods that allow for uncertainty regarding the imputation to be taken into account. The steps described earlier are repeated independently M times, resulting in β^*_m, where β^*_m is the parameter estimate of interest from imputation $m = 1,\ldots, M$. Steps 1 and 2 are referred to as the imputation task, and step 3 is the estimation task. Finally, we combine the estimates obtained after M imputations. The results of the M separate analyses (e.g. parameter estimates) are then combined into a single value as:

$$\beta^*_m = \frac{1}{M} \sum\nolimits_m^M \beta^*_m \ , \tag{2}$$

where β^*_m is the parameter estimates of interest from imputation $m=1, 2..., M$. The variance for these estimates is composed of two parts: the between imputation variance and within imputation variance. Between imputation variance takes the form:

$$B = \sum\nolimits_{m=1}^{M} \frac{(\beta^*_m - \beta^*_m)\ (\beta^*_m - \beta^*_m)'}{M-1} \ . \tag{3}$$

The within imputation variance, U^{-}, is the mean of estimated variances across the M imputations. The total variance for MI is then calculated as:

$$T = \bar{U} + \left(1 + \frac{1}{M}\right)B\,, \tag{4}$$

where:

$$\bar{U} = \sum\nolimits_{m=1}^{M} \frac{T_m}{M}\,. \tag{5}$$

The MI inference assumes that the analysis model is the same as the model used to impute missing values (the imputation model). Practically, the two models might not be the same (Meng, 1994; Schafer, 1997). The quality of the imputation model influences the quality of the analysis model results and therefore it is important to carefully consider the design of the imputation model. In this study, the imputation model is based on the Cox proportional hazards regression model (1). However, the imputation model for missing covariates requires a valid characterization of the conditional distribution of missing covariates conditional upon the observed data. This problem of the conditional distribution poses a major complication under a Cox PH model. White and Royston (2009) stated that such conditional distribution did not have standard and closed forms for Cox PH model. Thus, one recommendation is to use some of the common regression models to approximate the covariate distribution (Lihong et al., 2009). Following van Buuren et al. (1999) and White and Royston (2009), we used the linear regression model to impute the continuous covariate data. The linear regression model provides an appropriate imputation model for a continuous x_i^{mis}, that is x^{mis}~β0 + β1 Z + β2 δ + Δ_3^{T} x^{mis}. This model includes the following variables as predictors: the survival outcome Z, censoring δ, and the observed covariate x^{obs}. This means we used all the available data (including the outcome variable - survival time) to predict the missing covariate values to make the MAR assumption more plausible as well as to improve the accuracy and efficiency of the imputation. The survival time variable was included in the analysis as the outcome should be included in the imputation model (Moons et al., 2006). This was done to avoid the outcome-covariate association that might be biased toward null using the imputed data (Collins et al., 2001).

1.3 Last observation carried forward (LOCF)

The simplest imputation approach is the LOCF method in which every missing covariates value is replaced by the last observed covariates value from the subject or time series, i.e. it is a method that assumes that the outcomes would not have changed from the last observed value. We refer to Siddiqui and Ali (1998) and Satty and Mwambi (2012) for more details, and where insightful illustrations of the issues of this method are provided in Kenward, and Molenberghs (2009). It is a general and flexible technique for handling missing data, and can be implemented quickly in several statistical softwares. However, with respect to accurately reproducing known population results (parameter estimates and standard errors), the LOCF method has been found to be inadequate (Schafer and Graham, 2002). It shares with other single imputation methods that it tends to create inflated artificial values than truly expected, since imputed values are treated as observed values (Kenward and Molenberghs, 2009). Hence, the variability of the estimators is also underestimated. The problems linked with LOCF include: (1) the performance of this method is poor even when the ignorable missing data mechanism (MCAR or MAR) holds, a situation that limits their suitability to quite a restricted set of assumptions (Allison, 2002); (2) it produces seriously biased results that may or may not be predictable; (3) when using this technique, the standard errors and standard deviations tend to be underestimated, and, therefore, there is a greater likelihood of committing type-I error (see, Schafer and Graham, 2002). However, despite these shortcomings, the LOCF method has been recognized as a popular technique in dealing with missing data for the following reasons: its simplicity, in that the method can be quite effective and may be satisfactorily used with small amounts of missing data (Unnebrink and Jurgen, 2001), it is easy to carry out in most statistical software packages but it has varying details of implementations, and in some applications it makes sense to use this technique. LOCF does well when the missingness mechanism is assumed to be MCAR (Unnebrink and Jurgen, 2001). However, because such circumstance is rare, Kenward and Molenberghs (2009) advise that one should avoid this method whenever possible. In general, LOCF might become attractive under specific circumstances.

2 SIMULATION STUDY

We carried out a simulation study to compare the performance of the MI and LOCF methods. The simulations were conducted with 100 replications and sample size n = 1 000 for each replication. We simulated the survival time z_i from an exponential distribution using the following hazard:

$$\eta z = \exp(\beta_0 \ + \beta_1 x_{i1} + \beta_2 \, x_{i2} + \beta_3 x_{i3}), \tag{6}$$

where $(\beta 0, \beta 1, \beta 2, \beta 3) = (-1.5, 0.5, 1.0, 1.0)$. That is, the survival time for each individual was distributed according to equation (1). The covariates x_1, x_2 and x_3 were generated from the multivariate normal distributions, i.e., x_1~N (10, 0.25), x_2~N (10, 0.20) and x_3~N (10, 0.20). We assume that z_i observations are randomly censored with probability 0.20. Let R_i be a vector that represents the missing data process, with $R_{ij} = 1$ if the jth covariate is observed for individual i, and $R_{ij} = 0$ otherwise, where $i = 1, ..., n$, $j = 1, ...,$ p. Let, too, Ri2 and Ri3 represent whether x_{i2} and x_{i3} are unobserved. We created missing covariate mechanism according to the following models:

$$p(R_{i2=0} \mid h_i(z), x_{i,obs}, \theta_2) = \frac{\exp\,(\theta_{20}+\theta_{21h_i(z)}+\theta_{22x_{i1}})}{1+\exp\,(\theta_{20}+\theta_{21h_i(z)}+\theta_{22x_{i1}})}, \tag{7}$$

$$p(R_{i3=0} \mid h_i(z), x_{i,obs}, x_{i2}, \theta_3) = \frac{\exp\,(\theta_{30}+\theta_{31h_i(z)}+\theta_{32x_{i1}}+\theta_{33x_{i2}})}{1+\exp\,(\theta_{30}+\theta_{31h_i(z)}+\theta_{32x_{i1}}+\theta_{33x_{i2}})}, \tag{8}$$

where θ denotes the parameters of the missingness distribution, $h(z)$ is the observed event times, $\theta_2 = (-2, 1.5, 2.5)$ and $\theta_3 = (-1, 0.5, 0.5, 0.5)$. We created missing covariate observations under MAR mechanism. Namely, the probability of having a missing covariate values depends on an observed covariate values. The MAR mechanism was generated with the fraction of missing covariates set to 10%, 20% and 30%. Now, after the missing covariate values had been generated, MI was carried out using SAS PROC MI. With PROC MI, we considered the linear regression (Little, 1988) as an imputation model for continuous missing covariates data. PROC MI was applied to generate M = 5 complete datasets. These 5 imputations are often sufficient to obtain satisfactory results (Rubin, 1987; Schafer, 1997). Note that the choice of M =5 was considered adequate and the efficiency of the parameter estimate based on imputation given by $(1+{}^{\mathrm{v}}\!/_{\mathrm{M}})^{-1}$ here ν is the rate of missing data (Rubin, 1987). This formula shows that the relative efficiency of the MI inference is related to the missingness rate (ν) in combination with the number of imputations (M). For 10%, 20% and 30% rates of missing data and estimates based on M = 5 implies we achieve at least 98%, 96% and 94% efficiency, respectively. A Cox PH model was then fitted to each completed dataset using SAS procedure PHREG to estimate the overall parameters. A Cox PH model that we considered is based on (6). Thereafter, results of the analysis from these 5 completed (imputed) datasets were combined into a single inference using SAS PROC MIANALYZE.

The simpler LOCF technique replaced the missing covariate values by the last available observed values, and once the dataset has been completed in this way, it is analyzed as if it were fully observed. LOCF was conducted by using a macro in SAS software. After applying LOCF, as above introduced, the same model (6) as before being fitted is analyzed. In model (6), if x_i has missing covariate value, it will be filled in by the previous observed covariate value x_{i-1}. Comparisons of MI and LOCF were assessed using criteria recommended in Schafer and Graham (2002): (1) Bias of the estimates: the difference between the average of the 1000 coefficient estimates and the corresponding true coefficient. Thus a better approach that does on the average presents the population value with less bias. (2) The efficiency: the variability of the estimates around the true population coefficient. It was measured in this study by the average width of the 95% confidence interval. Thus, a wider interval implies a less efficient technique. (3) The coverage of the confidence interval: the percentage of 95% confidence intervals estimates across 1000 replicates. If a method is working well, the actual coverage should be close to the nominal rate (95%).

3 RESULTS

The results obtained from a Cox PH model (6) for the bias, efficiency and coverage of the MI and LOCF methods, under different missing covariate values rates are presented in Tables 1, 2 and 3. Note that the largest bias and less efficiency for each given estimate appear in bold.

Table 1 Bias, Efficiency and Coverage of MI and LOCF, under 10% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.006	1.158	0.971
		$\beta 2$	**0.018**	**1.113**	0.974
		$\beta 3$	0.017	1.116	0.967
10%					
	LOCF	$\beta 1$	**0.051**	**1.176**	0.902
		$\beta 2$	0.011	1.112	0.911
		$\beta 3$	**0.022**	**1.171**	0.908

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

Under 10% missing covariates rate, the results of MI and LOCF in terms of bias, efficiency and coverage, are displayed in Table 1. By looking at this table we find the following. With respect to biasdness of the estimates, the performance of MI was unsurprisingly, better than that for LOCF. However, the LOCF based estimates were closer to those based on MI, and only slightly less biased in estimating x2. Efficiency estimates associated with LOCF were slightly elevated when compared to those with MI. The MI method was more efficient in most cases, except for x2. For coverage criterion, according to Schafer and Graham (2002), the performance of a method can be regarded to be poor if its coverage drops below 90%, and hence leads to substantially increased Type-I error rate. By this rationale, both approaches yielded acceptable coverage of parameters. Their coverage rates were consistently above 90%.

An examination of Table 2, for 20% missing covariate rate, reveals that among the methodologies examined here, LOCF was notable for consistently producing the most biased estimates vis-a-vis those in the MI method. Namely, treating the data with MI appears to have resulted in fairly minor bias. MI yielded equally acceptable performance across all covariates. Comparing the efficiency results, just as was the case in Table 1, efficiency by LOCF appeared to be independent of the missing covariate rates, meaning the MI method yielded more efficient estimates under 20%. MI resulted in smaller estimates than estimates of LOCF. Differences in efficiency estimates between the 10% and 20% missing covariate rates were more pronounced for LOCF than for MI. Coverage rates obtained by the LOCF method in all cases were unsatisfactory, as its coverage rates were less than 90%.

Table 2 Bias, Efficiency and Coverage of MI and LOCF, under 20% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.011	1.177	0.960
		$\beta 2$	0.064	1.181	0.966
		$\beta 3$	0.051	1.152	0.957
20%					
	LOCF	$\beta 1$	**0.067**	**1.801**	0.891
		$\beta 2$	**0.089**	**1.811**	0.881
		$\beta 3$	**1.030**	**1.852**	0.889

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

Considering the 30% missing covariate values, the results shown in Table 3 reveal that in nearly all cases, LOCF consistently produced the most biased estimates. The efficiency performance was acceptable for MI but low for all parameters under LOCF. In general, the MI method tends to have the smallest estimates for efficiency condition. Thereby, it was more efficient than LOCF. With respect to coverage condition investigated, similar to the findings obtained under 10% and 20% missing covariate values, MI produced uniformly acceptable coverage; none was less than 90%. The LOCF's coverage at 95% was consistently lower than 90%. This coverage was indicative a seriously low level of coverage as 90% corresponds to a doubling of the nominal rate of error (0.05). As can be seen in the results, the low coverage rates by LOCF can also be attributed to its large biases.

Generally speaking, across all missing covariate rates, the worst performance for analyses run with LOCF occurred for the highest missing covariate rate, and declined in relative magnitude as the missingness rate decreased. In other words, when the missing covariate rate decreased to 10%, the results from LOCF became nearly closer to those of MI, but for 20% and 30%, it has seriously less efficient estimates.

Table 3 Bias, Efficiency and Coverage of MI and LOCF, under 30% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.054	1.801	0.951
		$\beta 2$	0.098	1.826	0.942
		$\beta 3$	0.102	1.900	0.938
30%					
	LOCF	$\beta 1$	**1.124**	**2.522**	0.862
		$\beta 2$	**1.021**	**2.091**	0.859
		$\beta 3$	**1.205**	**2.511**	0.849

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

DISCUSSIONS AND CONCLUSION

This study has discussed the performance of using the MI and LOCF methods for handling missing covariate values in survival analysis. The main objective was to address and compare the use of these methods when there are missing covariate values in Cox PH regression model. The methods were compared on simulated data. Missing covariate values were generated under three missingness rates. The missing data mechanism was assumed to be MAR. The comparisons between the two methods were made through the evaluation of bias, efficiency and coverage. Based on the simulation results, we reached the following conclusions:

- The results in general revealed that MI is likely to be the best under the MAR mechanism. MI consistently outperformed LOCF in terms of bias, efficiency and coverage. This advantage for the MI method is well documented in terms of the MAR mechanism (Little and Rubin, 1987; Schafer, 1997).
- The findings further suggested the inappropriateness of LOCF analysis. LOCF can lead loss in power of the covariates and imprecise parameter estimates. To avoid this problem, an application of MI can be utilized to handle this potential pitfall. Moreover, it appeared that no strong differences were seen between MI's results and those for LOCF when the missing data rate was low (10%). This indicates that the LOCF method can be applied if the proportion of missing covariate values is low. This LOCF situation is well stated in Unnebrink and Jurgen (2001) and Halabi et al. (2003). It would appear that Kenward and Molenberghs's (2009) recommendation to avoid the LOCF analysis whenever possible is supported by the current analysis.
- As missingness mechanism was simulated to be MAR, the current simulation results has shown clearly that the LOCF's performance was unsatisfactory under this assumption. This situation can be justified by some previous studies which show that LOCF is more widely used under MCAR than under MAR (See, Siddiqui and Ali, 1998; Halabi et al., 2003; Kenward and Molenberghs, 2009). Therefore, the better ways of dealing with missing covariate values in Cox PH model and the best method should be dependent on the nature of the missing covariate values mechanism. Conse quently, one needs to know why are there missing covariate values, and under which mechanism they are missing.
- In conclusion, we recommend that some techniques or methods use different approaches to address missing covariates in Cox PH model. The literature presents various techniques that can be used to deal with missing covariate values in Cox PH model, and these range from simple classical ad hoc methods to model-based methods. These methods should be fully understood and appropriately characterized in relation to missing data and should be theoretically proved before they are used practically. Additionally, each method is based on a specific missingness mechanism, but one needs to realize that at the heart of the missingness problem it is impossible to identify the missing data mechanism.

References

ALLISON, P. D. *Missing data*. Thousand Oaks, CA: Sage, 2002.

BARZI, F., WOODWARD, M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 2004, 160, pp. 34–45.

COLLINS, L. M., SCHAFER, J. L., KAM, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 2001, 6, pp. 330–351.

COX, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B, 1972, 34, pp. 187–220.

HALABI, S., WUN, C., DAVIS, B. R. Analysis of survival data with missing measurements of a time-dependent binary covariate. *Journal of Biopharmaceutical Statistics*, 2003, 13, pp. 253–270.

KENWARD, M. MOLENBERGHS, G. Last observation carried forward: A crystal ball. *Journal of Biopharmaceutical Statistics*, 2009, 872, pp. 872–888.

LIHONG, QI., YING-FANG, W., YULEI, HE. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*, 2009, 29, pp. 2592–2604.

LITTLE, R., RUBIN, D. B. *Statistical analysis with missing data*. New York: John Wiley, 1987.

LITTLE, R. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 1988, 6, pp. 287–296.

LITTLE, R., RUBIN, D. B. *Statistical analysis with missing data* (2nd ed.). New York: John Wiley and Sons, 2002.

MENG, X. L. Multiple imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 1994, 10, pp. 538–573.

MOONS, K. G., DONDERS, R. A., STIJNEN, T., HARRELL, JR. F. E. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 2006, 59, pp. 1092–1101.

PAIK, M. C. Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis*, 1997, 3, pp. 289–298.

RUBIN, D. B. Inference and missing data. *Biometrika*, 1976, 63, pp. 581–592.

RUBIN, D. B. Multiple imputation in sample survey. *Proc. Survey Res. Meth. Sec.*, Am. Statist. Assoc., 1978, pp. 20–34.

RUBIN, D. B. *Multiple imputation for non-response in surveys.* Wiley: New York, 1987.

SATTY, A. MWAMBI, H. Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, 2012, 46, pp. 327–356.

SCHAFER, J. L. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.

SCHAFER, J. L., GRAHAM, J. W. Missing data: Our view of the state of the art. *Psychological Methods*, 2002, 7, pp. 147–177.

SIDDIQUI, O., ALI, M. W. A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, 1998, 8, pp. 545–563.

UNNEBRINK, K., JURGEN, W. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in Medicine*, 2001, 20, pp. 3931–3946.

VAN BUUREN, S., BOSHUIZEN, H. C., KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 1999, 18, pp. 681–694.

WHITE, I. R., ROYSTON, P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 2009, 28, pp. 1982–1998.