Interactive CardioVascular and Thoracic Surgery 27 (2018) 153–158
doi:10.1093/icvts/ivy102 Advance Access publication 10 May 2018

Cite this article as: Papageorgiou G, Grant SW, Takkenberg JJM, Mokhles MM. Statistical primer: how to deal with missing data in scientific research? Interact CardioVasc Thorac Surg 2018;27:153–8.

Statistical primer: how to deal with missing data in scientific research?†

Grigorios Papageorgiou[a,b], Stuart W. Grant[c], Johanna J.M. Takkenberg[a] and Mostafa M. Mokhles[a,*]

[a] Department of Cardiothoracic Surgery, Erasmus University Medical Center, Rotterdam, Netherlands
[b] Department of Biostatistics, Erasmus University Medical Center, Rotterdam, Netherlands
[c] Department of Academic Surgery, University of Manchester, Manchester, UK

* Corresponding author. Department of Cardiothoracic Surgery, Erasmus University Medical Center, 's Gravendijkwal 230, 3015 CE Rotterdam, Netherlands. Tel: +31-010-7035413; fax: +31-010-7033993; e-mail: m.mokhles@erasmusmc.nl (M.M. Mokhles).

Received 10 November 2017; received in revised form 23 February 2018; accepted 27 February 2018

Abstract

Missing data are a common challenge encountered in research which can compromise the results of statistical inference when not handled appropriately. This paper aims to introduce basic concepts of missing data to a non-statistical audience, list and compare some of the most popular approaches for handling missing data in practice and provide guidelines and recommendations for dealing with and reporting missing data in scientific research. Complete case analysis and single imputation are simple approaches for handling missing data and are popular in practice, however, in most cases they are not guaranteed to provide valid inferences. Multiple imputation is a robust and general alternative which is appropriate for data missing at random, surpassing the disadvantages of the simpler approaches, but should always be conducted with care. The aforementioned approaches are illustrated and compared in an example application using Cox regression.

Keywords: Statistics • Missing data • Imputation • Research

INTRODUCTION

Missing data are a common challenge encountered by researchers while undertaking clinical research. It can occur across all types of studies including randomized controlled trials, cohort studies, case–control studies and clinical registries. The optimum approach to missing data is to ensure that strategies are devised to ensure that the amount of missing data in a study is as small as possible. Such strategies are commonly utilized in prospectively designed clinical trials as if statistical assumptions due to missing data are required, then the protection of randomization will be broken down and unbiased estimates of treatment effect will be lost. Strategies to minimize missing data in large multicentre cohort or registry studies may be employed however, data desired for research purposes may often be missing due to the retrospective nature of the study or because the data fall outside the primary purpose of the registry [1, 2].

Dealing with missing data may be low on the list of priorities for a researcher when undertaking a study but it is a vital step in data analysis as inappropriate handing of missing data can lead to a variety of problems. These included a loss of statistical power, loss of representation of key subgroups of the cohort, biased or inaccurate estimates of treatment effects and increased complexity of the statistical analysis.

To ensure that missing data are handled appropriately, there are a number of steps to follow: first, taking any necessary steps to complete or reduce the amount of missing data wherever possible; second, understanding the mechanism behind the remaining missing data; third, handling the missing data using appropriate methodology and finally, performing sensitivity analyses where appropriate. Focusing primarily on the framework of missing covariate data in non-randomized studies, this article introduces the concept behind different types of missing data and compares some of the most popular approaches for handling missing data in practice. Guidelines and recommendations for dealing with and reporting missing data in scientific research are also presented along with a simulated exercise on handling missing data.

METHODOLOGY

Missing data mechanisms

Before discussing methods for handling missing data, it is important to review the types of missingness. Commonly, these are classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [3]. An analysis of missing data patterns across contributing participants or centres,

†Presented at the 31st Annual Meeting of the European Association for Cardio-Thoracic Surgery, Vienna, Austria, 7–11 October 2017.

over time, or between key treatment groups should be performed to establish the mechanisms behind the missing data [1].

Missing completely at random. Observations of all subjects are equally likely to be missing. That is, there are no systematic differences between subjects with observed and unobserved values meaning that the observed values can be treated as a random sample of the population. For example, echocardiographic measurements might be missing due to sporadic ultrasound malfunction.

Missing at random. The likelihood of a value to be missing depends on other, observed variables. Hence, any systematic difference between missing and observed values can be attributed to observed data. That is, the relationships observed in the data at hand can be utilized to 'recover' the missing data. For example, missing echocardiographic measurements might be more normal than the observed ones because younger patients are more likely to miss an appointment.

Missing not at random. The likelihood to be missing depends on the (unobserved) value itself, and thus, systematic differences between the missing and the observed values remain, even after accounting for all other available information. In other words, there is extra information associated with the missing data that cannot be recovered by utilizing the relationships observed in the data. For example, missing echocardiographic measurements might be worse than the observed ones because patients with severe valve disease are more likely to miss a clinic visit because they are unable to visit the hospital.

Although there are a few methods proposed to test whether the data are MCAR or MAR, their practical value is dubious [4]. On the other hand, test distinguishing between MAR and MNAR always depends on data that are not observed meaning it is not possible to make this distinction based only on the observed data. Therefore, a researcher should always evaluate the plausibility of each missing data mechanism with respect to the method used to analyse the data and importantly on how the data were collected. The missing data mechanism should be regarded as an assumption that either supports an analysis or not rather than as an inherent and identifiable feature of a dataset. If that assumption is false, results may be biased. While under MCAR, most standard statistical tools will lead to valid results, that is not the case for MAR and MNAR, for which appropriate methods need to be employed. Table 1 summarizes the basic differences between the 3 missing data types and lists which of the methods discussed in the following section can be used to draw valid inference with respect to each missing data type.

Methods for handling missing data

There are various approaches for an incomplete data analysis. Two common approaches encountered in practice are complete case analysis and single imputation. Although these approaches are easily implemented, they may not be statistically valid and can result in bias when the data are not missing completely at random [5, 6]. On the other hand, multiple imputation is a more general approach that overcomes the main disadvantages of the aforementioned approaches when data are missing (completely) at random [7–9].

Complete case analysis. The easiest way to deal with missing data is to drop all cases that have one or more values missing in any of the variables required for analysis. Although under MCAR this does not lead to bias of the results, it may result in significant loss of data and associated loss of power (e.g. wider confidence intervals) because the sample size is reduced. The extent of this loss of power is associated with the amount of missing data. If the data are MAR, this approach will lead to biased results. Complete case analysis may be appropriate for missing data related to the primary outcome of the study.

Single imputation. Alternatively, missing values in any variable could be replaced with a single value that is thought to best represent the mechanism of the missing data. This could be the mean of a normally distributed continuous variable, the median/mode of a categorical variable, the predicted value from a regression equation (that is, utilizing the complete observations to predict the values of the missing observations) or the best/worst observation carried forward. There may be cases where the missing risk factor data are believed to be highly likely due to the absence of a risk factor, and in this situation, it may be reasonable to impute the absence of the risk factor.

Although this approach allows the researcher to include all subjects in the analysis, it may lead to biased results. Moreover, the uncertainty of parameter estimates of the imputed variables will not necessarily improve when compared with the complete case analysis because the imputation is not conditional on the values of the outcome variable. How large the induced bias is depends on the variability of the imputed variable and on the proportion of missing values. Single imputation is also invalid under MAR since it does not account for the inter-relationships between the variables of interest. Single imputation may, however, be used to perform sensitivity analyses for missing covariate information or primary outcome data to ensure that the reported results are valid under the worst or best-case scenario.

Table 1: Summary of missing data mechanisms

Missing data mechanism	Related to	Not related to	Probability to be missing	Valid analysis
MCAR		Observed or missing data	Equal for every data point	Complete case analysis, single and multiple imputation
MAR	Observed data	Missing data	Equal for data points within groups	Multiple imputation
MNAR	Missing data		Unequal and unknown	Sensitivity analysis

MAR: missing at random; MCAR: missing completely at random; MNAR: missing not at random.

Multiple imputation. Multiple imputation offers an alternative to overcome the disadvantages of the complete case analysis or single imputation approach. It allows the uncertainty, which is due to missing data, to be appropriately considered and can be thought of in three distinct steps: imputation, analysis and pooling of the results.

At the first step of imputation, multiple copies of the original incomplete dataset are generated. In each dataset, the missing values are replaced by values which are randomly sampled from the predictive distribution of the observed data, conditional on all other variables. The process of sampling induces variation in the imputed values which reflects the uncertainty of those imputed values.

In the analysis step, the model of interest is fitted to each imputed dataset. The results derived from each analysis will differ slightly due to the variability of the imputed values. In the third step, the results are pooled by taking the average of the estimates from the separate analyses to derive the pooled estimate and by applying Rubin's rules, which incorporate the within and between imputation uncertainty, to derive the associated standard errors. More details on Rubin's rules and the formulas that are used to obtain the pooled estimates can be found in Supplementary Material A.

In contrast to the complete case analysis, multiple imputation provides valid results when the data are MAR while avoiding the loss of power due to sample size reduction. However, loss of power may still occur when using multiple imputation if there is high uncertainty in the distributions which are used to impute the missing data. Unlike single imputation, multiple imputation provides valid results if the data are MAR. This is because systematic differences between the missing and the observed values are due to information already present in the data at hand, and this is considered in the predictive distributions utilized in the first step of the multiple imputation procedure. In addition, single imputation does not account for the between-imputation variance that leads to overestimation of accuracy (small standard errors). Table 2 briefly summarizes the advantages and disadvantages of each method.

It is important to note that generally none of the above methods will provide valid results under MNAR. Methods for dealing with MNAR data are very limited and usually complex. They are typically based on the idea of sensitivity analysis under various MNAR scenarios, for example, assuming the worst possible or best possible value for the missing data. Commonly, the goal of such sensitivity analyses is to help in assessing the robustness of the results under plausible MNAR scenarios. Multiple imputation can also potentially be used to perform sensitivity analyses if data are MNAR [10].

Multiple imputation: considerations and limitations

Multiple imputation is a general approach with numerous applications, and it is easily accessible through standard statistical software packages such as R [11], SPSS®, SAS® and STATA®. However, it should be highlighted that it is not a panacea for every incomplete data setting [12, 13]. Although multiple imputation is often considered as an out of the box method that can be easily applied in any missing data problem, this is not true. Its application requires the user to carefully consider the plausibility of each of the possible causes of missingness, thoroughly select an appropriate imputation model and choose appropriate variables to include with respect to both clinical relevance and the missing at random assumption.

Some common points and special cases to consider when performing multiple imputation are as follows:

- Missing outcome information: It should be noted that up to this point, this article has focused primarily on missing covariate information. That is because when there are missing outcome data, it has been argued that the complete case analysis is more appropriate as imputed outcome data can lead to misleading results [14, 15]. Single imputation of the worst or best-case scenario for missing outcome data may be used as sensitivity analysis to ensure the validity of trial results. Multiple imputation of missing outcome data may also be performed if there are auxiliary variables that are highly correlated with the outcome and the probability that the outcome is missing. However, this can only help in reducing the loss in accuracy of the estimates due to missing data and only if the data are at most MAR. Nevertheless, the complete case analysis should be regarded as the principle analysis in the case of missing outcome data.
- The number of imputed datasets: Although 5 imputed datasets are considered adequate, it is always advised to increase the number to improve the efficiency and the reproducibility of the results [13].
- The number of iterations: Since multiple imputation is based on an iterative algorithm, the convergence criteria should always be assessed and if necessary, the number of iterations increased [7, 10].
- Inclusion of the outcome in the imputation model: The outcome should be included in the first step of the multiple imputation procedure to take into account the association between outcome and incomplete covariates [16].
- Longitudinal studies: Common software packages usually require the transformation of long datasets (a row per measurement) to their wide (a row per subject) counterparts to perform multiple imputation. This implies that current implementation of multiple imputation in longitudinal settings works best in balanced studies (e.g. subjects are measured at the same time points).

Table 2: Comparison of incomplete data analysis methods

Methods	Pros	Cons
Complete case analysis	Simple to implement	Loss of power and efficiency and invalid under MAR
Single mean imputation	Simple to implement and avoids loss of power	Does not appropriately account for uncertainty in results and invalid under MAR
Multiple imputation	Avoids loss of power, retains efficiency and valid under MAR	Time consuming and requires more statistical knowledge

MAR: missing at random.

- Survival analysis: Because of the complex nature of the outcome variable in such cases (pairing of a binary event indicator variable with a time-to-event variable), several approaches have been proposed on how to include it in the imputation model [17–19]. The most recent research findings, however, propose to use the Nelson–Aalen estimator along with the event indicator in the imputation model rather than the event indicator along the time-to-event variable [20].
- Acceptable amount of missingness: There is no standard rule of how much missing data is too much. Theoretically, multiple imputation can handle large amounts of missingness. Nevertheless, the quality of the results is related to the complexity of the imputation model used, whether there are few or many variables with a large amount of missingness, the total sample size and the variability of the variables which are subject to missingness. For example, 50% missingness may be acceptable if the remaining 50% of the data allow accurate estimation of the predictive distribution used to draw imputed values. In settings with a small sample size, large variability and/or a heterogeneous study population, this may not be the case.

Given the potential complexities, it is clear that multiple imputation should be conducted carefully with respect to the challenges of each analysis. Advice from statistical experts is, therefore, highly recommended when considering multiple imputation to address missing data.

Reporting

Because performing analysis on incomplete data requires a lot of considerations, decisions and assumptions, it is recommended that authors provide a thorough description of the imputation procedure to ensure the transparency and reproducibility of the analysis. Often, such a description can be moved to the Supplementary Material accompanying a manuscript. Table 3 provides an extensive list of points that should be included when conducting incomplete data analysis.

Table 3: Guidelines for reporting incomplete data analysis in scientific manuscripts

Report the number/proportion of missing values per variable of interest: • Alternatively/complementary report number of complete cases • If possible, discuss potential causes for missing data
Provide comparison of complete and incomplete cases: • Table or figure comparing the distributions of variables of interest
State the methodology used for incomplete data analysis: • Which one was used: the complete case analysis, multiple imputation, etc. • Report the assumptions that were made regarding the cause of missingness: MCAR, MAR or MNAR
Indicate the software (including version number) that was used in handling missing data: • Optionally add any changes to the default settings/features of the software or/and functions which were used
Report the number of imputed datasets and number of iterations
List the variables that were included in the imputation model
Mention any higher order functions of imputed variables: • Were higher order terms such as interactions, polynomials or spline transformations of original variables used in the analysis? • If yes, were these higher order terms included in the imputation model as well?
Assess the robustness of MAR assumption by conducting sensitivity analysis for various MNAR scenarios

MAR: missing at random; MCAR: missing completely at random; MNAR: missing not at random.

Data example

To illustrate the above points with a data example, we consider a simple scenario for survival analysis. The data come from a follow-up study of patients with congenital heart disease who received a human tissue allograft in the aortic position. The aim is to investigate the association between postoperative aortic gradient (mmHg) and risk of death while accounting for baseline factors such as age at operation (years), gender, donor age (years) and allograft diameter (mm). An overview of the data for the 'all cases' scenario (before excluding any case to artificially generate missing data scenarios) is provided in the Supplementary Material B, Table S4.

To briefly illustrate a few of the points presented throughout this article for dealing with missing data, we artificially generated 40% missingness on the postoperative aortic gradient under the 3 missingness scenarios: MCAR, MAR and MNAR. Under MCAR, randomly chosen values were deleted. Under MAR, the aortic gradient measurements of younger patients (age less than the mean age in the dataset) were deleted. Finally, for MNAR, the missing values were selected to be patients with a high postoperative aortic gradient (higher than the 65th percentile of the postoperative gradient in the dataset) assuming that they are more likely to be unable to go to the hospital. We then applied Cox regression using complete cases, single mean imputation and multiple imputation under each scenario for the mechanism that generated the missing data and compared the corresponding results with those obtained when all the cases were used (no missing data). The analysis was conducted in R [11] using the packages mice [10] and mitools [21]. A sample R code for conducting multiple imputation in R is given in Supplementary Material C.

The results are summarized in Fig. 1, where the red dot and black lines represent the estimated hazard ratios and their corresponding 95% confidence intervals, respectively. As shown in this figure, under MCAR and MAR, multiple imputation provided results that were slightly closer to those of the complete data (before values were removed; 'all cases') than the results from the simpler approaches for this specific example. Nevertheless, the differences are small, and both the complete case analysis and single mean imputation are theoretically valid under MCAR. The loss in efficiency due to the reduced sample size when using only the complete cases is evident from the wider confidence intervals. Under MNAR, all approaches provided biased estimates. In this situation, further sensitivity analyses or explicit accounting of the missing data mechanism would be required [8].

DISCUSSION

Missing data are common in clinical research and should be minimized wherever possible through good study design and data collection protocols. However, in most cases, it is not possible to reduce the amount of missing data to zero. As demonstrated in the example presented in this article, inappropriate handling of missing data can potentially lead to biased results or significant loss of power. Although simpler approaches in handling missing data such as the complete case analysis or single imputation may be appropriate if the amount of missing data is small and the mechanisms behind the missing data are clearly understood, in most cases

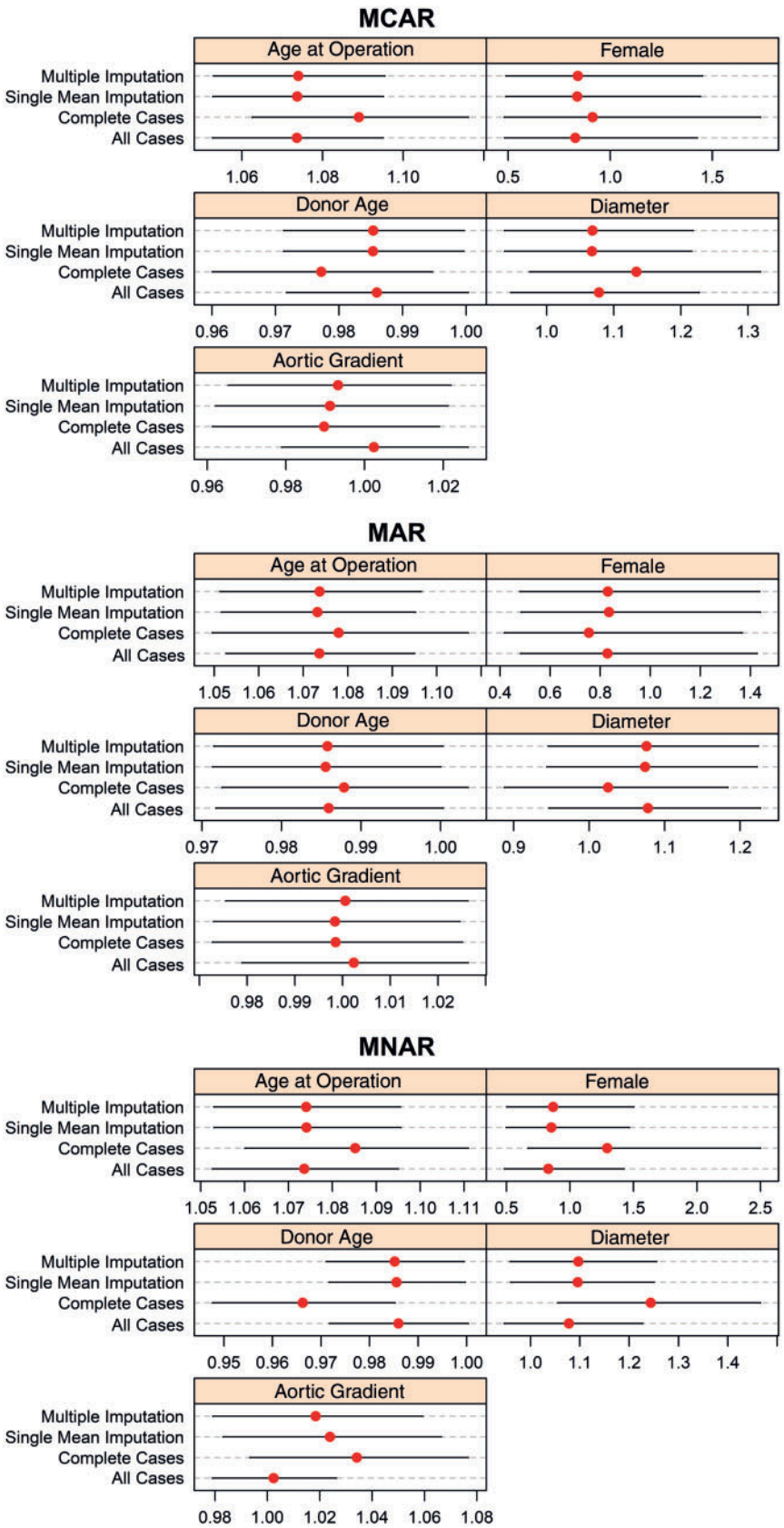

Figure 1: Hazard ratios and 95% confidence intervals of 'all cases', complete cases, single mean imputation and multiple imputation analyses under 3 missing data mechanisms. MAR: missing at random; MCAR: missing completely at random; MNAR: missing not at random.

multiple imputation is accepted as the preferred strategy for handling missing data. Although multiple imputation deals with a number of pitfalls related to complete case analysis or single imputation, it does significantly increase the complexity of the analysis and can potentially lead to bias if the data are not missing at random.

It is important to approach the handling of missing data in a systematic manner and clearly report the steps that have been undertaken in the handling of missing data as outlined in the guidelines in Table 3. Although this article is intended to give an overview for clinicians on how to handle missing data, it is strongly recommended that complex approaches to handle missing data should be performed under the guidance of a statistician.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *ICVTS* online.

ACKNOWLEDGEMENTS

We would like to thank Nicole S. Erler for the helpful discussions and valuable comments.

Funding

M.M.M. is funded by a NWO Veni grant of the Netherlands Organisation for Scientific Research (NWO 916.160.87).

Conflict of interest: none declared.

REFERENCES

[1] Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ 2013;346:e8668.
[2] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT *et al*. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367:1355–60.
[3] Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd edn, Chapter 1. Hoboken, NJ: Wiley, 2002.
[4] Enders CK. Applied Missing Data Analysis, Chapter 1. New York: Guilford Press, 2010.
[5] Carpenter J, Kenward MG. A critique of common approaches to missing data. In: Missing data in randomised controlled trials—a practical guide. Birmingham, AL: National Institute for Health Research, 2007.
[6] Vach W, Blettner M. Biased estimation of the odds ratio in case control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895–907.
[7] van Buuren S. Flexible Imputation of Missing Data, Chapters: 2, 5. Boca Raton, FL: Taylor & Francis, 2012.
[8] Carpenter JR, Kenward MG. Multiple Imputation and Its Application, Chapters: 1, 2, 10. Chichester, West Sussex, United Kingdom: John Wiley & Sons, Ltd, 2013.
[9] Schafer JL. Analysis of Incomplete Multivariate Data, Chapter 4. London: Chapman and Hall, 1997.
[10] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Soft 2011;45:1–67.
[11] R Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
[12] Erler NS, Rizopoulos D, van Rosmalen J, Jaddoe V, Franco O, Lesaffre E. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. Stat Med 2016;35:2955–74.
[13] White IR, Royston P, Wood AM. Multiple Imputation using chained equations: issues and guidance in practice. Stat Med 2011;30:377–99.
[14] Von Hippel PT. Regression with missing Ys: an improved strategy for analyzing multiply imputed data. Sociol Methodol 2007;37:83–117.
[15] Little RJA. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.
[16] Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092–101.
[17] Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol 2004;160:34–45.
[18] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18:681–94.
[19] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003;56:28–37.
[20] White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009;28:1982–98.
[21] Lumley T. mitools: Tools for Multiple Imputation of Missing Data. R Package Version 2.3. 2014. https://CRAN.R-project.org/package=mitools (20 February 2018, date last accessed).