Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Peter C. Austin, PhD, Ian R. White, PhD, Douglas S. Lee, MD PhD, Stef van Buuren, PhD

PII: S0828-282X(20)31111-9

DOI: https://doi.org/10.1016/j.cjca.2020.11.010

Reference: CJCA 3911

To appear in: *Canadian Journal of Cardiology*

Received Date: 22 October 2020

Revised Date: 20 November 2020

Accepted Date: 24 November 2020

Please cite this article as: Austin PC, White IR, Lee DS, van Buuren S, Missing Data in Clinical Research: A Tutorial on Multiple Imputation, *Canadian Journal of Cardiology* (2021), doi: https://doi.org/10.1016/j.cjca.2020.11.010.

Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Peter C. Austin, PhD (a,b,c)

Ian R. White, PhD (d)

Douglas S. Lee, MD PhD (a,b,e,f)

Stef van Buuren, PhD (g,h)

Short title: Multiple imputation in clinical research

Word count: 5,485

Author affiliations:
(a) ICES, Toronto, Ontario, Canada.
(b) Institute of Health Policy, Management and Evaluation, University of Toronto, Ontario, Canada.
(c) Sunnybrook Research Institute, Toronto, Ontario, Canada.
(d) Medical Research Council Clinical Trials Unit, University College London, London, United Kingdom.
(e) Department of Medicine, University of Toronto, Toronto, Ontario, Canada.
(f) Peter Munk Cardiac Centre and University Health Network, Toronto, Ontario, Canada.
(g) University of Utrecht, Utrecht, The Netherlands.
(h) Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands.

Address for correspondence
Peter Austin
ICES
G106, 2075 Bayview Avenue
Toronto, Ontario
M4N 3M5
Canada

E-mail: peter.austin@ices.on.ca

Phone: (416) 480-6131 Fax: (416) 480-6048

Brief summary

Missing data is a common occurrence in clinical research. We briefly discuss common approaches to addressing missing data and highlight their limitations. We introduce multiple imputation (MI), a popular approach for addressing the presence of missing data. With MI, multiple plausible values of a given variable are imputed or filled-in for each subject who has missing data for that variable. We describe the steps that should be conducted when conducting an analysis using MI.

Abstract (word count: 242)

Missing data is a common occurrence in clinical research. Missing data occurs when the value of the variables of interest are not measured or recorded for all subjects in the sample. Common approaches to addressing the presence of missing data include complete-case analyses, in which subjects with missing data are excluded, or mean-value imputation, where missing values are replaced with the mean value of that variable in those subjects for whom it is not missing. However, in many settings, these approaches can lead to biased estimates of statistics (e.g., of regression coefficients) and/or to confidence intervals that are artificially narrow. Multiple imputation (MI) is a popular approach for addressing the presence of missing data. With MI, multiple plausible values of a given variable are imputed or filled-in for each subject who has missing data for that variable. This results in the creation of multiple completed datasets. Identical statistical analyses are conducted in each of these complete datasets and the results are pooled across complete datasets. We provide an introduction to MI and discuss issues in its implementation, including developing the imputation model, how many imputed datasets to create, and addressing derived variables. We illustrate the application of MI through an analysis of data on patients hospitalized with heart failure. We focus on developing a model to estimate the probability of one-year mortality in the presence of missing data. Statistical software code for conducting multiple imputation in R, SAS, and Stata are provided.

Keywords: Missing data, multiple imputation, tutorial.

1. Introduction

Missing data are a common occurrence in clinical research. Missing data occurs when the value of the variables of interest are not measured or recorded for all subjects in the sample. Data can be missing for several reasons, including: (i) patient refusal to respond to specific questions (e.g., patient does not report data on income); (ii) loss of patient to follow-up; (iii) investigator or mechanical error (e.g., sphygmomanometer failure); (iv) physicians not ordering certain investigations for some patients (e.g., cholesterol test not ordered for some patients).

Before discussing different ways of addressing the presence of missing data, it is important to understand the conditions under which data are subject to being missing. Rubin developed a framework for addressing missing data and described three different missing-data mechanisms [1, 2]. Data are said to be 'missing completely at random' (MCAR) if the probability of a variable being missing for a given subject is independent of both the observed and unobserved variables for that subject [3] (a list of abbreviations is provided in Table 1). If data are MCAR, then the sub-sample consisting of subjects with complete (or non-missing) data is a representative sub-sample of the overall sample. An example of MCAR is laboratory values that are missing because the sample was lost or damaged in the laboratory. The occurrence of such events in the laboratory is unlikely to be related to characteristics of the subject. Data are said to be 'missing at random' (MAR) if, after accounting for all the observed variables, the probability of a variable being missing is independent of the unobserved data. If physicians were less likely to order laboratory tests for older patients and that was the only factor influencing whether or not a test was ordered and recorded, then missing laboratory data would be MAR (assuming that age was recorded for all patients). Finally, data are said to be 'missing not at random' (MNAR) if they are neither MAR nor MCAR. Thus, data are MNAR if the probability of a variable being missing, even after

accounting for all the observed variables, is dependent on the value of missing variable. An example of data that are MNAR could be income, in which more affluent subjects, even after accounting for other characteristics, are less likely to report their income in surveys than are less affluent subjects. Unfortunately, one cannot test whether the data are MAR vs. MNAR, so one must judge what is plausible using clinical knowledge [4, 5].

Historically, a popular approach when faced with missing data was to exclude all subjects with missing data on any necessary variables and to conduct subsequent statistical analyses using only those subjects who have complete data (accordingly, this approach is often referred to as a 'complete case' analysis). When only the outcome variable is incomplete, this approach is valid under MAR and often appropriate [6]. With incomplete covariates, there are disadvantages to this approach [2, 4, 7]. First, unless data are MAR, the estimated statistics and regression coefficients may be biased [4]. Second, even if data are MCAR, with the reduction in sample size there is a corresponding reduction in precision with which statistics and regression coefficients are estimated. Accordingly, estimated confidence intervals will be wider by using complete case analysis than if all the data were used. Moreover, different analyses may use different subsets of the overall sample, so that it is difficult to compare results even within the same paper.

An approach to circumvent the limitations of a complete case analysis is to replace the missing values of variables with plausible values. Such an approach is called 'imputation', as one is imputing a value of the variable for those subjects with missing data on that variable. Historically, a common approach to imputation was 'mean value imputation', in which subjects for whom a given variable is missing have the missing value replaced with the mean value of that variable amongst all subjects for whom the variable is present. Thus, subjects who are missing blood pressure have the missing value replaced with the average value of blood pressure

amongst those subjects for whom blood pressure was measured and recorded. A limitation of mean value imputation is that it artificially reduces the variation in the dataset. For example, mean imputation will artificially lower the estimated standard deviation of the variable which was imputed [2]. Furthermore, mean imputation ignores multivariate relations between different variables in the sample. For instance, older subjects may have, on average, higher blood pressure than younger subjects. This correlation between age and blood pressure is not taken into account by mean imputation.

An alternative to mean value imputation is 'conditional mean imputation' in which a regression model is used to impute a single value for each missing value [2]. From the fitted regression model, the mean or expected value, conditional on the observed covariates, is imputed for those subjects with missing data. Thus, assuming that the imputation model regressed blood pressure on age and sex, the same value of blood pressure would be imputed for all subjects of the same age and sex. A modification of conditional mean imputation draws the imputed value from a conditional distribution whose parameters are determined from the fitted regression model. However, both of these latter approaches artificially amplify the multivariate relations in the data. Another limitation is that the imputed values are treated as known with certainty and treated on an equal footing with the values for the same variable for other subjects for whom the variable was observed and recorded and not imputed. Mean imputation and conditional mean imputation are recommended for handling missing values of baseline covariates in randomized trials only [6, 8, 9].

A popular approach for addressing the issue of missing data is multiple imputation (MI) [1, 10]. MI imputes multiple values for each missing value. This results in the creation of multiple complete data sets in which the missing values have been filled in with plausible values. The

analysis of scientific interest is then conducted separately in each of these complete datasets and the results are pooled across the imputed datasets. In this way, multiple imputation allows the user to explicitly incorporate the uncertainty about the true value of imputed variables.

The current paper provides an introduction to MI and illustrates its application using a cardiovascular example. The paper is structured as follows. In Section 2 we introduce MI and discuss several issues related to its implementation. In Section 3 we illustrate its application using an example of logistic regression to model mortality in patients with heart failure. Finally, in Section 4 we summarize our brief tutorial and direct the interested reader to more detailed and comprehensive discussions of MI.

2. Multiple imputation for missing data

In this section we provide an introduction to MI and discuss issues related to its use.

2.1 Multiple imputation using Multivariate Imputation by Chained Equations (MICE)

Fully conditional specification (FCS) is a strategy for specifying multivariate models through conditional distributions. A specific implementation of this strategy in which every variable is imputed conditional on all other variables is now known as the Multivariate Imputation by Chained Equations (MICE) [10-13] algorithm. In our description of the algorithm we assume that there are *p* variables, of which *k* are subject to missing data and *p-k* that are complete. The algorithm is summarized in Table 2. The process described in Steps 3 and 4 is repeated for several cycles to create one imputed dataset. Standard software uses 5 to 20 cycles by default, and it is rarely necessary to increase these values [10, 11]. The imputed values obtained

after the last cycle are used as the imputed values for the first imputed dataset. The entire process is then repeated M times in order to produce M imputed datasets.

2.2 Multiple imputation for continuous variable using predictive mean matching

The imputation process described above uses linear regression and takes the imputed values as random draws from a normal distribution. This has problems if the residuals from the regressions are not normally distributed (e.g. if data are skewed), or if relations are non-linear (e.g. height and age). For example, a variable that can have only positive values (e.g., counts) may have imputed values that are negative. One option to address such problems is to transform the variable prior to imputation so that the transformed variable is approximately normally-distributed. For instance, the logarithmic transformation, when applied to a positively-skewed distribution, can result in a distribution that is more normally-distributed. As a last step, one may wish to back-transform imputations into the original scale. A second option to is to draw imputations from the observed values by a technique called predictive mean matching (PMM) [11]. For a given subject with missing data on the variable in question, PMM identifies those subjects with no missing data on the variable in question whose linear predictors (created using the regression coefficients from the fitted imputation model) are close to the linear predictor of the given subject (created using the regression coefficients sampled from the appropriate posterior distribution, as described above). Of those subjects who are close, one subject is selected at random and the observed value of the given variable for that randomly-selected subject is used as the imputed value of the variable for the subject with missing data. Morris et al. suggest that identifying the ten closest subjects without missing data performed well [14]. Using the terminology of Morris et al., we refer to the method described in Section 2.2 as parametric

imputation, as the imputed variables are drawn from a parametric distribution [14]. This is in contrast to PMM, where the imputed variables are drawn from an observed empirical distribution.

2.3 Analyses in the M imputed datasets

Once M complete datasets have been constructed using multiple imputation, the statistical analysis of scientific interest is conducted in each of the M complete datasets. That analysis would be the exact analysis that would be conducted in the absence of missing data. Thus, if the analysis model is a logistic regression model in which a binary outcome variable is regressed on a set of predictor variables, this model is fit in each of the M imputed datasets. The statistics of interest (e.g., estimated regression coefficients and their standard errors) are extracted from the analysis conducted in each of the M imputed datasets.

2.4 Rubin's Rules for combining estimates and standard errors across imputed datasets

Once the statistics of interest have been estimated in the M imputed datasets, they are combined using Rubin's Rules [1]. Let $\theta^{(i)}$ denote the estimated statistic of interest (e.g., a regression coefficient) obtained from the analysis in the *i*th imputed dataset ($i = 1,\ldots, M$). The pooled estimated of the statistic of interest is $\theta = \frac{1}{M}\sum_{i=1}^{M}\theta^{(i)}$. The MI estimate of the statistic is simply the average value of the estimated statistic across the M imputed datasets.

Computing the variance of the estimated statistic is more complex, as it requires accounting for the within-imputation uncertainty in the estimated statistic and the between-imputation variation in the estimated statistic. Let $W^{(i)}$ denote the estimated variance (e.g., the square of the

estimated standard error) of $\theta^{(i)}$. The average within-imputation variance is defined as $W = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} W^{(i)}$. This is simply the mean estimated variance of the estimated statistic across the M imputed datasets. The between-imputation variance of the estimated statistic is $\mathrm{B} = \frac{1}{M-1}\sum_{i=1}^{M}(\theta^{(i)} - \theta)^2$. This quantity reflects the degree to which the estimated statistic varies across the M imputed datasets. The MI estimate of the variance of θ obtained using Rubin's Rules is $\mathrm{var}(\theta) = \mathrm{W} + \left(1 + \frac{1}{\mathrm{M}}\right)\mathrm{B}$. This quantity reflects both the average within-imputation variation in θ as well as the between-imputation variation in θ. Note that when using single imputation, there is no estimate of B, so we are unable to estimate the true variation in the statistic.

2.5 How many imputations: how large should M be?

An important question is how many imputed datasets should be created. Early recommendations were that three to five imputed datasets were sufficient as long as the amount of missing information was not very high [1, 3], while others suggested that often 5-10 imputations were sufficient [7]. These early recommendations were based on the accuracy with which the regression coefficient was estimated compared to its accuracy had it been estimated with an infinite number of imputed datasets. However, analysts are interested not only in estimated regression coefficients (e.g., log-odds ratios or log-hazard ratios), but also in their associated standard errors (which are used in deriving confidence intervals and significance tests). Thus, one wants to estimate not only regression coefficients accurately, but also standard errors.

Ideally, one would select M such that the pooled estimated regression coefficients and standard errors would not vary meaningfully across repeated applications of MI (i.e., if the entire process was repeated with M new imputed datasets, one would obtain estimates comparable to those obtained using the initial M imputed datasets). The term Monte Carlo error in a given statistic (e.g., a regression coefficient or a standard error) refers to the standard deviation of that statistic across repeated applications of MI. When focusing on a single statistic, the Monte Carlo error can be computed as $\sqrt{B/M}$ [11]. White et al. suggested that, as a rule of thumb, the number of imputed datasets should be at least as large as the percentage of subjects with any missing data [11]. They suggest that this will result in estimates of regression coefficients, test statistics (regression coefficients divided by the standard error) and p-values with minor variability across repeated MI analyses (i.e., the Monte Carlo error will be low). A more advanced method for determining the number of imputations was developed by Von Hippel [15]. Nowadays computation is cheap and the use of between 20 and 100 imputed datasets is common.

2.6 Which variables to include in the imputation model?

Investigators need to distinguish between two different statistical models: the imputation model and the analysis model. The imputation model is used for imputing missing data. It is not of direct interest and is only used to provide reasonable imputations. The analysis model holds the quantities that are ultimately of scientific interest and is the focus of the research question. *The rules for building imputation and analysis models are very different.* It is important to include in the imputation model all the variables that will be included in the analysis model. Failure to include these variables in the imputation model usually results in estimates in the analysis model being biased. The variables must also be included in the imputation model in the

right way: for example, Schafer noted that if interactions are omitted from the imputation model, then the estimated interactions in the analysis model will be biased towards the null [7].

It is especially important to include in the imputation model the outcome variable for the analysis model [5, 11]. Failure to do so usually results in estimated regression coefficients for the analysis model being biased toward the null. When the outcome in the analysis model is a survival or time-to-event outcome (e.g., the outcome model is a Cox proportional hazards model) then there are two components to the outcome: a time-to-event variable denoting the time to the occurrence of the event or the time to censoring, and a binary indicator variable denoting whether the subject experienced the event or was censored. The recommended approach is to include both in the imputation model, with the time-to-event variable transformed using the cumulative survivor function [16]. In addition, the imputation model is improved by including variables that are related to the missingness and variables that are correlated with variables of interest. In longitudinal data, when imputing a variable for a specific measurement occasion (e.g., on the second clinic visit), one also needs to include in the imputation model future values of that variable (e.g. the value of that variable at the third clinic visit).

2.7 Imputing derived variables

The analysis model may include variables that are derived from other variables. Examples include body mass index (BMI, which is derived from height and weight), quadratic terms for continuous variables (e.g., age^2), and interactions between variables (i.e., products of variables). When the component variables required to create the derived variable are missing (and therefore the derived variable is also missing), there are two main options for imputing the derived variables. The first option imputes the missing component variables and creates the derived

variable after all variables have been imputed. Thus, if height was missing, height would first be imputed and then combined with weight to create BMI. Von Hippel refers to this approach as 'impute, then transform'. This approach is appealing as it leads to derived variables that are consistent with the derivation rule. The obvious problem with the approach is that the derived variable is not part of the imputation model, hence it may lead to bias, as explained in section 2.6. The second option is to treat the derived variable as simply another variable and to impute this variable directly. Thus, if height were missing (and thus BMI were also missing), height and BMI would be imputed for those subjects for whom they were missing. This approach is known as 'transform, then impute' [17] or 'just another variable' (or JAV)[11]. Note that the JAV approach incorporates the components as well as the derived variable in the imputation model. This approach is appealing as it incorporates all necessary variables into the imputation model. However, it can lead to quadratic variables with negative values or BMI values that are inconsistent with the height and weight of the subject. It has been shown that in some settings the approach leads to accurate estimates of regression coefficients in the analysis model, though it can fail in others [18, 19]. Van Buuren (Sec 6.4) describes some alternative strategies for specific types of dependencies [10]. Since no strategy performs uniformly better, we may need some tailoring to the type of derived variable.

2.8 Missing outcome variables

Multiple imputation is blind to which variables are outcomes and which variables are predictors in the final analysis model. When developing the imputation models, the important issue is to include in the imputation models all the variables from the analysis model. This suggests that one can impute values of the outcome variable (for the analysis model) for those

subjects for whom it is missing. However, von Hippel provided evidence that excluding subjects who are missing the outcome variable (for the analysis model) when fitting the outcome model will tend to be a better strategy [20]. He proposed a strategy that he referred to as 'multiple imputation, then deletion' (MID). Under MID, all subjects are used in the imputation process. Values are imputed for all missing data, including for those subjects who are missing the outcome variable. However, subjects for whom the outcome variable was imputed are then excluded when the analysis model is fit in each imputed dataset. The MID approach will tend to result in estimated regression coefficients for the analysis model that are more efficient (have smaller variability) than those obtained when fitting the analysis model in all subjects. In addition, the method is robust against bad imputation in the outcome. The MID procedure should not be used if there are auxiliary variables that are strongly related to the outcome (and not included in the analysis model), or if the scientific interest extends to parameters other than regression coefficients [11].

3. Case study

We use data on patients hospitalized with heart failure in the province of Ontario to provide a case study illustrating the application of MI. The analysis model of interest is a logistic regression model in which death within one year of hospital admission is regressed on ten patient characteristics.

3.1 Data sources

We used data from the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, which was an initiative to improve the quality of care for patients with cardiovascular

disease in Ontario [21]. We used data on 8,338 patients hospitalized with congestive heart failure between April 1, 2004 and March 31, 2005 at 81 Ontario hospital corporations. Data on patient demographics, vital signs and physical examination at presentation, medical history, and results of laboratory tests were collected on these patients by retrospective chart review. Subjects were linked to administrative health care data to determine vital status.

For the purposes of this case study, we considered ten baseline covariates: age, respiratory rate at admission, glucose level, urea level, low density lipoprotein (LDL) cholesterol level, sex, S3 (third heart sound) on admission, S4 (fourth heart sound) on admission, neck vein distension on admission, and cardiomegaly on chest X-ray. The first five were continuous while the last five were binary. The outcome was a binary outcome denoting whether the patient died within 365 days of hospital admission. Logistic regression models for 30-day and 1-year mortality are often used in cardiovascular research [22-24]. Our purpose in using these data was to illustrate the application of statistical methods and not to draw clinical conclusions. Accurate estimation of the association of variables with cardiovascular outcomes in current patients may require the use of more recent data and of a more comprehensive set of predictor variables. Furthermore, depending on the objective of the intended study, a different regression model may be more appropriate.

3.2 Descriptive statistics

Means and percentages are reported for the continuous and binary variables, respectively, in Table 3. We also report the percentage of subjects with missing data for each of the variables. The percentage of missing data ranged from a low of 0% (age and sex) to a high of 73% (LDL cholesterol). Overall, 78% of subjects had missing data on at least one variable.

3.3 Comparison of subjects with and without missing data

We conducted univariate comparisons of those with and without missing data. There are at least two reasons for these comparisons. First, as noted above, the imputation model is improved by including variables that are related to the missingness. Thus, these comparisons will help identify variables that should be included in the imputation model. Second, these analyses provide evidence as to the plausibility of the MAR assumption. If those with and without missing data differ on many observed variables, then it is plausible that they may also differ on unobserved variables. Note that a lack of significant univariate associations does not provide proof that the data are MCAR or MAR.

There were meaningful differences in age, sex and mortality (the three variables that were not subject to missingness) between those with complete data and those with missing data. The average age of those with complete data was 73.7 years while it was 77.5 years for those with missing data. Of those with complete data, 43.4% were female, while 53.0% of those with missing data were female. Of those with complete data, 23.7% died within one year of admission, while 33.9% of those with missing data died within one year of admission. Patients with missing data tended to be older, were more likely to be female, and more likely to die compared to those with complete data.

3.4 Complete case analysis

We conducted a complete case analysis that was restricted is to the 1,806 subjects with complete data. The reason for doing this is that complete case analysis is less prone to user error than MI (as it does not rely on an imputation model) and we should be able to explain any

differences between the complete case analysis and the MI analysis [5]. We used logistic regression to regress death within one year of hospital admission on the 10 baseline covariates. The logarithm of the estimated odds ratios and associated 95% confidence intervals are reported in Figure 1 (log-odds ratios are reported so that the confidence intervals are symmetric). Increasing age and urea were associated with an increased odds of death within one year and had 95% confidence intervals that excluded the null value. None of the binary variables had odds ratios whose associated 95% confidence interval excluded the null value. Note that the odds ratios for the five continuous variables are not directly comparable with one another as they are measured on different scales.

3.5 Multiple imputation

Imputation was conducted using the MICE algorithm using PROC MI in SAS (SAS/STAT version 14.1). Logistic regression models were used as the imputation models for the binary variables, while linear regression models were used as the imputation models for the continuous variables. All variables (including the binary outcome variable) were included in each imputation model (with the obvious exception of the variable that was being imputed). Using the rule of thumb suggested by White et al., we created 78 imputed datasets since 78% of subjects had any missing data. For comparative purposes, we used von Hippell's two-stage algorithm with 10 imputed datasets in the first stage with the criterion that the standard errors of the estimated regression coefficients be estimated accurately to two decimal places. The algorithm suggested 80 imputed datasets were necessary to estimate the standard error of the intercept term with the desired precision and that at most 15 imputed datasets were necessary to estimate the standard errors of the 10 covariates with the desired precision.

As a sensitivity analysis we used predictive mean matching when imputing missing values for the continuous variables. Software code for conducting these analyses are provided in Supplementary Appendix S1 (SAS code), Supplementary Appendix S2 (R code), and Supplementary Appendix S3 (Stata code).

3.6 Descriptive statistics in the imputed datasets

Non-parametric density plots were used to describe the distribution of the four continuous variables that were subject to missing data in the complete cases and in those subjects who were missing data for the given continuous variable. The latter was done separately in each of the imputed datasets. These are described in Figure 2 (parametric imputation) and Figure 3 (predictive mean matching). The density function in the complete cases is described using a solid black line, while the density function of the imputed variable in each of the imputed datasets is described using a dashed red line. When using parametric imputation, the distribution of imputed respiratory rate, glucose, and urea failed to display the skewness seen in subjects for whom the variable was observed. However, the distribution of imputed values of LDL was comparable to the empirical distribution in subjects for whom LDL was measured. When using predictive mean matching, the distribution of the imputed values tended to be very similar to that of the observed values of the variable.

3.7 Logistic regression in the imputed datasets

In each imputed dataset, we regressed the binary outcome denoting death within one year of hospital admission on the 10 covariates described in Table 3. The regression coefficients and their standard errors were pooled using Rubin's Rules. The estimate of the Monte Carlo error for

the ten estimated regression coefficients ranged from 0.000042 for age to 0.005502 for LDL cholesterol. Thus, if we repeated the entire imputation process multiple times, we would expect to see only minor variation in the estimated regression coefficients.

The log-odds ratios and their associated 95% confidence intervals obtained using parametric imputation are reported in Figure 1. Three continuous variables (age, respiratory rate, and urea) had a positive association with 1-year mortality, while females had a lower risk of death than males. The odds ratios and associated 95% confidence intervals obtained using predictive mean matching imputation are also reported in Figure 1. The estimated odds ratios and associated confidence intervals obtained using predictive mean matching imputation were essentially identical to those obtained using parametric imputation. In comparing the results of the three regression analyses, one observes that the confidence intervals obtained from the imputation-based analyses were narrower than those obtained in the complete case analysis. For some variables (e.g., age, S3 and S4), the confidence intervals obtained using the complete case analysis were substantially wider than those obtained using multiple imputation.

4. Discussion

Missing data occurs frequently in clinical research. MI is a statistical tool that allows the researcher to replace missing values with multiple plausible values of the variable in question. The use of MI allows the researcher to analyze complete datasets while incorporating the uncertainty in the imputed values of the variable. We provided a brief introduction to MI and provided guidance regarding its implementation. We illustrated the application of MI through the analysis of data on patients hospitalized with heart failure.

When applying MI, researchers should explore differences between the observed and imputed distributions and between the complete case analyses and the MI analyses. We refer readers to previously-published guidelines for reporting analyses affected by missing data [5, 25].

The current introduction to MI was not intended to be exhaustive. We refer the interested reader to several excellent texts on MI [1-3, 10] as well as to more detailed overview articles [7, 11]. We have focused our attention on multiple imputation in observational studies in which clustering of subjects or a multilevel structure is absent. Other works describe methods for using multiple imputation with multilevel data [10, 26-29]. Similarly, we have focused on the use of parametric models (e.g., logistic regression models or linear regression models) for the imputation models. An area of current research is on the use of machine learning methods for multiple imputation [30]. We have focused on the use of MI when data are either MCAR or MAR. The described methods must be modified if it is thought that the data are MNAR. Van Buuren summarizes different methods to address data that are MNAR [10]. The simplest approach is to assume that the distribution of a variable in those with missing is shifted compared to the distribution in those with complete data. Sensitivity analyses can be conducted in which the magnitude of the shift parameter is allowed to vary.

We have focused on the MICE algorithm for multiple imputation, along with a modification, predictive mean matching. This is not the only method to impute missing data. An earlier method has been described as 'joint modeling' [10], of which MI under a normal model is a specific implementation [4]. This approach assumes that the set of variables follow a joint multivariate distribution. The multivariate normal distribution is widely used in applications [10]. Under this implementation, the variables are assumed to follow a multivariate normal distribution. Once the parameters of this distribution have been estimated, missing values can be

imputed by random draws from this multivariate distribution. In theory, this approach requires that all the variables be continuous. In practice, binary or categorical variables occur frequently (e.g., presence or absence of diabetes). Schafer and Graham suggest that despite this theoretical limitation, they have found the multivariate normal distribution to be useful in a wide range of settings [4]. Furthermore, they provide suggestions for incorporating binary and categorical variables as well as non-normally distributed continuous variables. However, others have suggested that these methods of incorporating non-continuous variables may not perform as desired [10]. Given the flexibility of the MICE algorithm and its ability to explicitly incorporate different types of variables, its use may be attractive to researchers in biomedical research.

In our case study, we obtained similar parameter estimates when using parametric imputation as when using predictive mean matching imputation. This is to be expected for estimates that depend on the middle of the distribution, such as means or regression coefficients. In practice, it may be difficult to provide examples where PMM imputation beats a well-crafted parametric imputation model. However, in practice, analysts often prefer PMM imputation because it preserves typical features in the raw data. For example, it accounts for discreteness of data, avoids impossible values, preserves location of quantiles, and is highly robust to imputation model misspecification. All this costs no additional work on the part of the analyst. If the complete-data model depends on such features, then the inference will also be better when using PMM imputation.

In this tutorial article we have focused on the use of multiple imputation in observational studies. In randomized controlled trials (RCTs), multiple imputation is not always the optimal approach [6]. When a univariate outcome is MAR, a complete case analysis using an adjusted analysis is unbiased and efficient [6]. With a multivariate outcome (e.g., an outcome measured at

multiple occasions over the course of follow-up), the use of a linear mixed model with missing data in the outcome only will tend to result in estimates with smaller standard errors compared to the use of multiple imputation [6]. If multiple imputation is used, it is suggested that imputation be conducted separately in the different arms of the trial [6].

In summary, MI replaces missing values with plausible values. By creating multiple imputed datasets, the analyst can explicitly account for the uncertainty inherent in the imputed values. Historical approaches such as complete case analysis, mean imputation or single imputation potentially result in bias, incorrect estimates of standard errors, and consequently incorrect tests of statistical significance. Researchers are encouraged to consider MI as an important tool to address the problems associated with missing data in clinical research.

Funding sources: This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. The data sets used for this study were held securely in a linked, de-identified form and analysed at ICES. While data sharing agreements prohibit ICES from making the data set publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR). The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research (Grant numbers CTP79847 and CRT43823). Drs. Austin and Lee are supported in part by Mid-Career Investigator awards from the Heart and Stroke Foundation. Dr. Lee is supported by the Ted Rogers Chair in Heart Function Outcomes. Ian White was supported by the Medical Research Council Programme MC_UU_12023/21.

Disclosures: None.

References

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.

2. Little RJA and Rubin DB. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons; 2002.

3. Carpenter JR and Kenward MG. *Multiple Imputation and its Application*. Chichester, UK: John Wiley & Sons; 2013.

4. Schafer JL and Graham JW. Missing Data: Our View of the State of the Art. *Psychological Methods*. 2002;7:147-1777.

5. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM and Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.

6. Sullivan TR, White IR, Salter AB, Ryan P and Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat Methods Med Res*. 2018;27:2610-2626.

7. Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research*. 1999;8:3-15.

8. White IR and Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med*. 2005;24:993-1007.

9. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG and Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*. 2012;184:1265-9.

10. van Buuren S. *Flexible Imputation of Missing Data, Second Edition*. Boca Raton, FL: CRC Press; 2018.

11. White IR, Royston P and Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *StatMed*. 2011;30:377-399.

12. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16:219-42.

13. van Buuren S and Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45.

14. Morris TP, White IR and Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14:75.

15. von Hippell PT. How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*. 2018:1-20.

16. White IR and Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009;28:1982-98.

17. von Hippell PT. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*. 2009;39:265-291.

18. Seaman SR, Bartlett JW and White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12:46.

19. Vink G and van Buuren S. Multiple imputation of squared terms. *Sociological Methods & Research*. 2013;42:598-607.

20. von Hippell PT. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*. 2007;37:83-117.

21. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA and Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association*. 2009;302:2330-2337.

22. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D and Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association*. 2003;290:2581-2587.

23. Tu JV, Austin PC, Walld R, Roos L, Agras J and McDonald KM. Development and validation of the Ontario acute myocardial infarction mortality prediction rules. *Journal of the American College of Cardiology*. 2001;37:992-997.

24. Lee DS, Lee JS, Schull MJ, Borgundvaag B, Edmonds ML, Ivankovic M, McLeod SL, Dreyer JF, Sabbah S, Levy PD, O'Neill T, Chong A, Stukel TA, Austin PC and Tu JV. Prospective Validation of the Emergency Heart Failure Mortality Risk Grade for Acute Heart Failure. *Circulation*. 2019;139:1146-1156.

25. Akl EA, Shawwa K, Kahale LA, Agoritsas T, Brignardello-Petersen R, Busse JW, Carrasco-Labra A, Ebrahim S, Johnston BC, Neumann I, Sola I, Sun X, Vandvik P, Zhang Y, Alonso-Coello P and Guyatt GH. Reporting missing participant data in randomised trials: systematic survey of the methodological literature and a proposed guide. *BMJ Open*. 2015;5:e008431.

26. Longford NT. Missing Data. In: J. de Leeuw and E. Meijer, eds. *Handbook of Multilevel Analysis* New York, NY: Springer; 2008: 377-399.

27. van Buuren S. Multiple Imputation of Multilevel Data. In: J. J. Hox and J. K. Roberts, eds. *Handbook of Advanced Multilevel Analysis* New York, NY: Routledge; 2011: 173-196.

28. Molenberghs G and Verbeke G. Missing Data. In: M. A. Scott, J. S. Simonoff and B. D. Marx, eds. *The SAGE Handbook of Multilevel Modeling* London: SAGE; 2013: 403-424.

29. Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter JR, Van Buuren S and Resche-Rigon M. Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*. 2018;33:160-183.

30. Richman M, Trafalis T and Adrianto I. Multiple imputation through machine learning algorithms. *87th AMS Annual Meeting*. 2007.

Table 1. List of abbreviations

Abbreviation	Full term
MI	Multiple imputation
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
FCS	Fully conditional specification
MICE	Multivariate Imputation by Chained Equations
PMM	Predictive mean matching
JAV	Just another variable
MID	Multiple imputation, then deletion

Table 2. Multivariate Imputation by Chained Equations (MICE) algorithm for multiple imputation

1. Specify an imputation model for each of the *k* variables that are subject to missing data.
2. For each of the *k* variables that are subject to missing data, fill in the missing values with random draws from those subjects with observed values for the variable in question. Note that these initial imputed values do not respect the multivariate relations in the data and will be overwritten by better imputed values in later stages of the algorithm.
3. For the first variable that is subject to missing data:
    a. Regress this first variable on all the other variables using those subjects with complete data on the first variable and observed or currently imputed values of the other variables.
    b. The estimated regression coefficients and their variance-covariance matrix (and the estimated variance of the residual distribution if a linear regression model was fit for a continuous variable) are extracted from the regression model estimated in (a).
    c. Using the quantities obtained in (b), randomly perturb the estimated regression coefficients in a way that reflects the degree of uncertainty arising from the data.
    d. Using the set of perturbed regression coefficients obtained in (c), the conditional distribution of the first variable is determined for each subject with missing data on that variable.
    e. A value of the variable is drawn from this conditional distribution for each subject with missing data on the first variable.
4. Repeat Step 3 for each of the variables that is subject to missing data. Step 3 and Step 4 form one cycle of the imputation process for creating one imputed dataset.
5. Repeat Steps 3 and 4 the desired number of times (suggested values: 5 to 20 cycles). The final imputed values are used as the imputed values in first imputed dataset.
6. Repeat Steps 2 to 5 M times to produce M imputed datasets (the choice of M, the number of imputed datasets, is discussed in Section 2.5).

Table 3. Descriptive statistics of case study data

Variable	Mean (SD)/ Percentage	Number of subjects with observed data	Number of subjects with missing data	Percentage of subjects with missing data
Continuous variables				
Age (years)	76.7 (11.6)	8338	0	0%
Respiratory rate at admission (breaths per minute)	24.5 (7.0)	8138	200	2.4%
Glucose (initial lab test) (mmol/L)	8.6 (4.1)	8051	287	3.4%
Urea (initial lab test) (mmol/L)	10.3 (6.6)	8028	310	3.7%
LDL cholesterol (mmol/L)	2.2 (0.9)	2272	6066	72.8%
Binary variables				
Female	50.9%	8338	0	0%
S3	6.2%	8126	212	2.5%
S4	2.7%	8135	203	2.4%
Neck vein distension	66.1%	7586	752	9.0%
Cardiomegaly on chest X-ray	47.7%	7711	627	7.5%
Outcome				
Death within one year	31.7%	8338	0	0%

Note: SD: standard deviation

Figure legends

Figure 1:

Title: Distribution of continuous variables in complete cases and in those with imputed data when using parametric imputation.

Caption: The solid black line denotes the distribution of the given continuous variable in those subjects for whom that variable was not missing. The red lines denote the distribution of the imputed value for that variable in those subjects for whom the variable was missing. There is one red line for each of the imputed datasets.

Figure 2:

Title: Distribution of continuous variables in complete cases and in those with imputed data when using predictive mean matching (PMM).

Caption: The solid black line denotes the distribution of the given continuous variable in those subjects for whom that variable was not missing. The red lines denote the distribution of the imputed value for that variable in those subjects for whom the variable was missing. There is one red line for each of the imputed datasets.

Figure 3.

Title: Estimated odds ratios and 95% confidence intervals for variables in the logistic regression model fit in the case study.

Caption: There are three estimates/confidence intervals for each of the ten variables: (i) using complete cases; (ii) multiple imputation analyses when using parametric imputation; (iii) multiple imputation analyses when using predictive mean matching (PMM).

Supplementary Appendix S1. SAS code for multiple imputation

```
* This code is provided for illustrative purposes and comes
  with absolutely no warranty;
* The dataset used for these analyses cannot be publicly
  Distributed. Please do not contact the authors requesting
  the dataset;

* MI using the default parametric imputation for the continuous
variables;

* In the dataset are the following variables:
   age: patient age (years)
   resp: Respiratory rate
   glucose: Glucose
   urea: Urea
   ldl: LDL cholesterol level
   female: Binary variable (female=1/male=0)
   s3: S3 (third heart sound)
   s4: S4 (fourth heart sound)
   neckvdis: Neck vein distension
   cmg: Cardiomegaly on chest X-ray
   mort1yr: Binary variable denoting death within one year;

proc mi data=cohort seed=2122019 nimpute=pctmissing (max=99)
  out=tutorial_mi;
  class female s3 s4 neckvdis cmg mort1yr;
  fcs plots=trace nbiter=20 logistic(mort1yr female s3 s4 neckvdis
cmg);
  var mort1yr age female resp s4 s3 glucose urea cmg neckvdis ldl;
run;

* MI using PMM for the continuous variables;

proc mi data=cohort seed=2122019 nimpute=pctmissing (max=99)
  out=tutorial_mi_pmm;
  class female s3 s4 neckvdis cmg mort1yr;
  fcs plots=trace nbiter=20 logistic(mort1yr female s3 s4 neckvdis
cmg);
  fcs regpmm(age resp glucose urea ldl);
  var mort1yr age female resp s4 s3 glucose urea cmg neckvdis ldl;
run;


* Analyses in the imputed datasets;

Proc sort data=tutorial_mi; by _imputation_; run;

Proc logistic data=tutorial_mi descending;
   Model mort1yr = age resp glucose urea ldl female s3 s4 neckvdis cmg
   /covb;
```

```
   By _imputation_;
   Ods output ParameterEstimates=lgsparms covB=lgscovb;
Run;

* Pooling results using Rubin’s Rules;

Proc mianalyze parms=lgsparms covb(effectvar=stacking)=lgscovb;
  Modeleffects Intercept age resp glucose urea ldl female s3 s4
    neckvdis cmg;
  ods output ParameterEstimates=MI;
run;

proc print data=MI;
run;
```

Supplementary Appendix S2. R code for multiple imputation

```
# This code is provided for illustrative purposes and comes
  with absolutely no warranty;

#################################################################################
# Read in data.
#################################################################################

zlist<- list(age=0,resp=0,glucose=0,urea=0,ldl=0,female=0,s3=0,s4=0,neckvdis=0,
  cmg=0,mort1yr=0)

cohort <- data.frame(scan("mi_tutorial.txt",zlist))

data <- cohort[,c("mort1yr","age","female","resp","s4","s3",
                  "glucose","urea","cmg","neckvdis","ldl")]

#################################################################################
# MI using the parametric imputation for the continuous variables
#################################################################################

meth <- make.method(data)
meth[meth == "pmm"] <- "norm"
nimp <- 100 * nic(data) / nrow(data)

imp.parm <- mice(data,m=nimp,method=meth,maxit=20,seed=2122019)
plot(imp.parm)

# Analyses in the imputed datasets
fit.parm <- with(imp.parm,glm(mort1yr ~ age + resp + glucose +
  urea + ldl + female + s3 + s4 + neckvdis + cmg,family = "binomial"))

# Pooling results using Rubin's Rules
summary(pool(fit.parm), confint = TRUE, exponentiate = TRUE)

#################################################################################
# MI using the default PMM for the continuous variables
#################################################################################

imp.pmm <- mice(data,m=nimp,maxit=20,seed=2122019)
plot(imp.pmm)

# Analyses in the imputed datasets
fit.pmm <- with(imp.pmm,glm(mort1yr ~ age + resp + glucose +
  urea + ldl + female + s3 + s4 + neckvdis + cmg,family = "binomial"))

# Pooling results using Rubin's Rules
summary(pool(fit.pmm), confint = TRUE, exponentiate = TRUE)
```

Supplementary Appendix S3. Stata code for multiple imputation

```
* This code is provided for illustrative purposes and comes
  with absolutely no warranty;

infile age resp glucose urea ldl female s3 s4 neckvdis cmg mort1yr
using "mi_tutorial.txt"

set seed 2122019
mi set flong
mi register imputed resp glucose urea ldl // continuous and incomplete
mi register imputed s3 s4 neckvdis cmg // binary and incomplete
mi register regular mort1yr age female // complete

* MI using parametric imputation for the continuous variables;
*mi impute chained (logit) s3 s4 neckvdis cmg (regress) resp glucose
urea ldl = mort1yr age female, add(20)

* alternative MI using PMM for the continuous variables;
mi impute chained (logit) s3 s4 neckvdis cmg (pmm, knn(10)) resp
glucose urea ldl = mort1yr age female, add(20)

mi estimate: logistic mort1yr age resp glucose urea ldl female s3 s4
neckvdis cmg
```

Figure 1 Estimated log-odds ratios and 95% confidence intervals

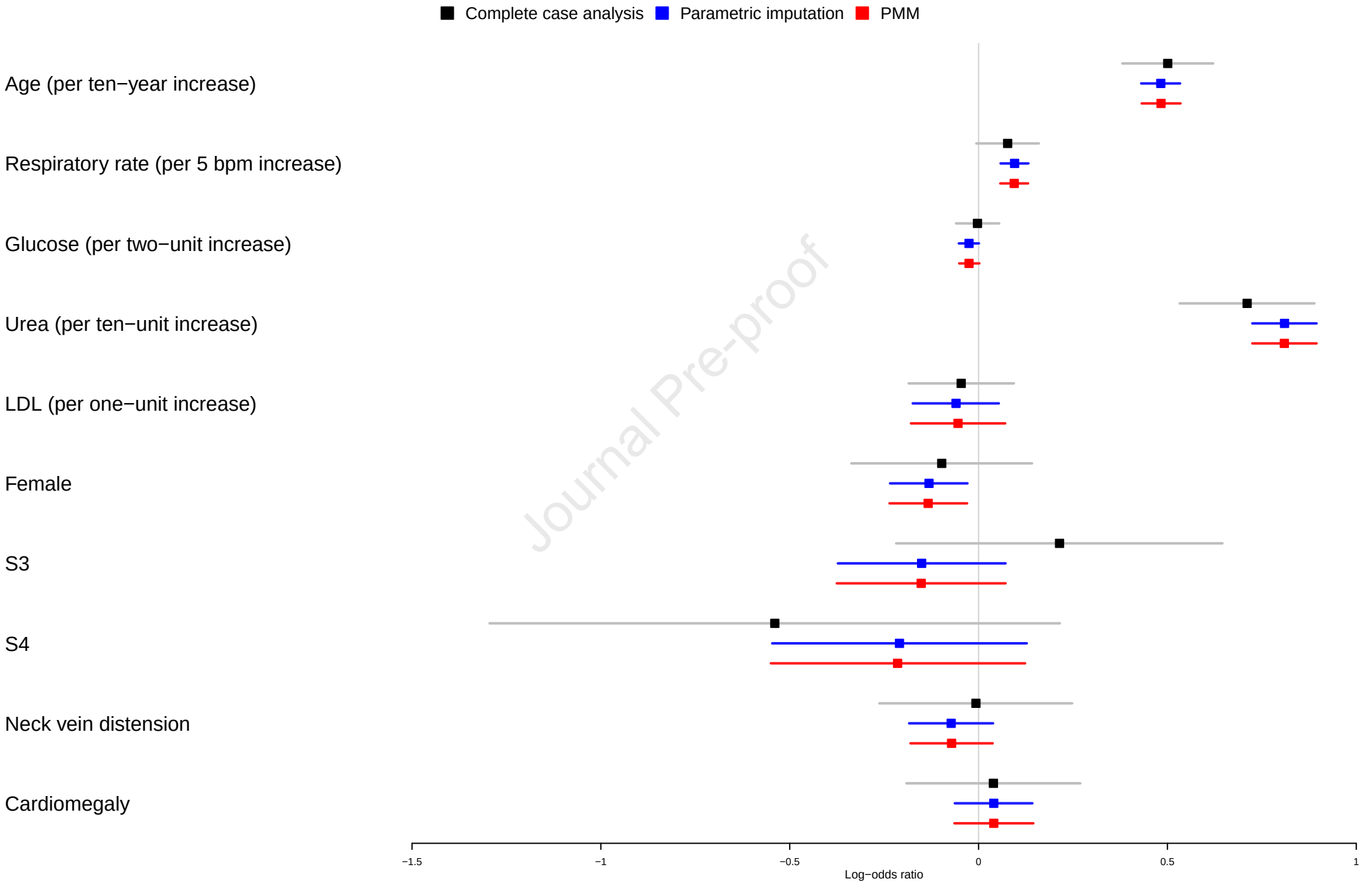

Figure 2. Distribution of continuous variables in complete cases and in those with imputed data

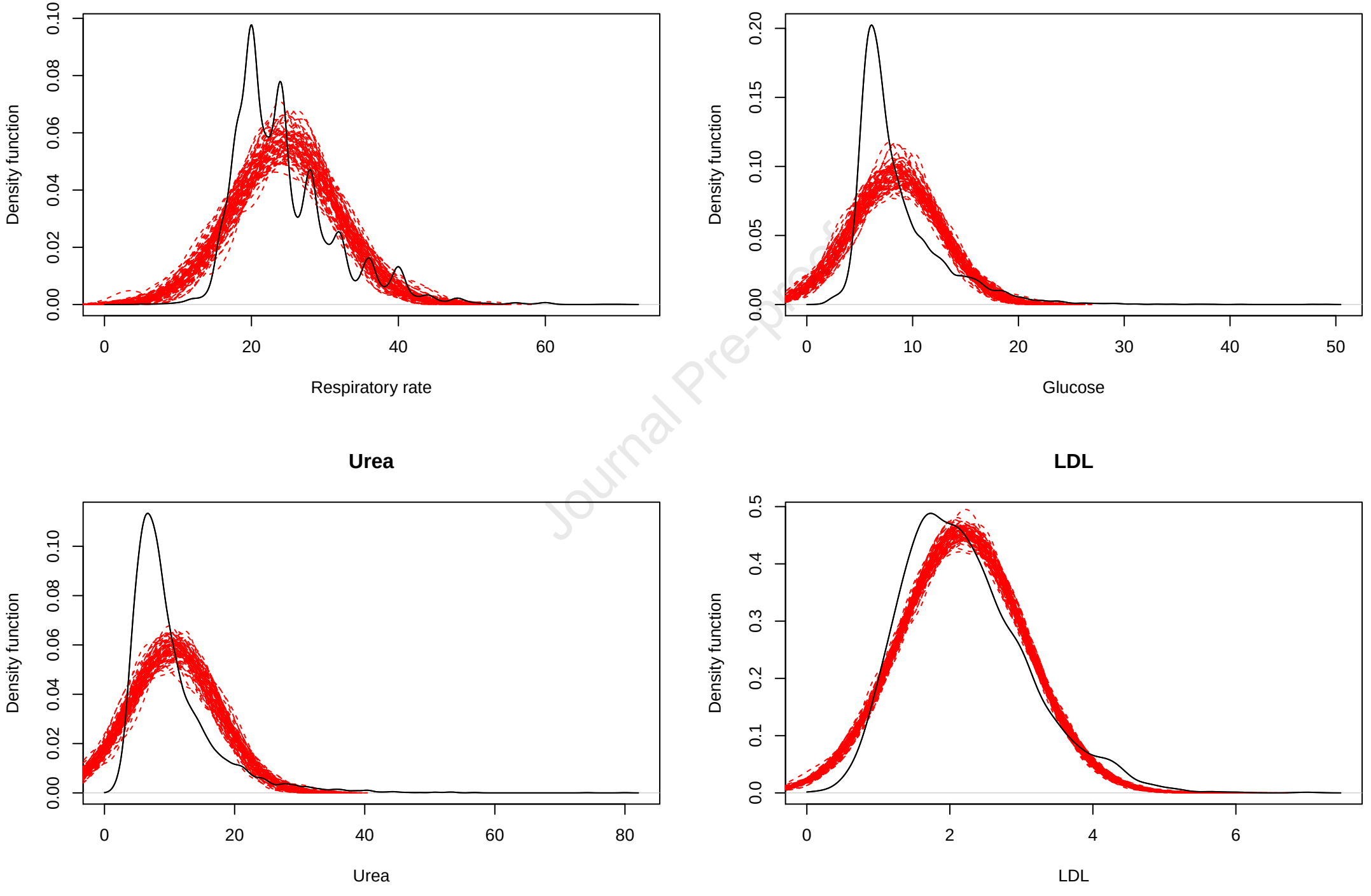

Figure 3. Distribution of continuous variables in complete cases and in those with imputed data (PMM)