Missing data 6. Choices in doing multiple imputation

Ian R White, Nikolaos Pandis, Tra My Pham

Earlier articles in this series have discussed why missing data present a problem and why multiple imputation (MI) is a popular solution ***[add refs at proof stage – adjust the refs section accordingly]****.* The previous article described the principles of MI ***[add ref at proof stage – adjust the refs section accordingly]***. However, a number of choices have to be made in implementing MI, and if we want to get valid results, then we have to make suitable choices. This article describes the main steps in doing this. We assume that the reader has in mind a particular analysis that they would perform if the data were complete: for example, a linear regression of a recession score on age, gender, and other variables, which we term the "Analysis Model". We describe the steps required to allow the analysis model to be fitted after MI.

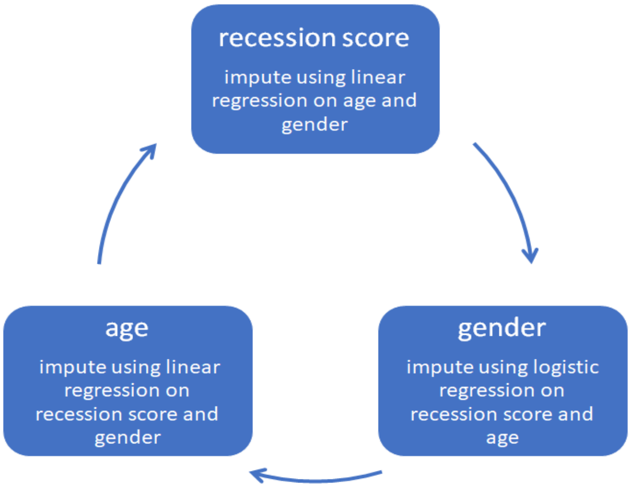
Most data sets have missing data in a number of different variables, which presents further difficulties. The most popular approach is multivariate imputation by chained equations (MICE),[1] and we focus on that here.

MICE involves setting up a suitable imputation model for each variable that is to be imputed, using the ideas in article 5 ***[add ref at proof stage – adjust the refs section accordingly]***. Usually the imputation model is some form of regression model.[2] For example, a continuous variable like age might be imputed using a linear regression model, while a binary variable like gender might be imputed using a logistic regression model. The form of the imputation model can be tailored to the variable type. Categorical variables with more than 2 levels can be handled either assuming the levels are unordered or assuming they are ordered: the latter is more robust, if it is appropriate for the variable concerned. Quantitative variables can be handled by linear regression or by predictive mean matching,[3] a technique which imputes the observed value of a near neighbour.

An important choice in MICE is which variables to include in the procedure. Any incomplete variables that are needed in the analysis model clearly need to have their missing values imputed. However, we also need to consider which variables will be used in order to construct good imputations. It turns out that all variables in the analysis model should be included in the imputation models, whether they are complete or incomplete.[2] This ensures that all the inter-relationships seen in the observed data are carried over into the imputed data.

Figure 1 shows an illustrative example created using data from a cohort study conducted to assess whether gingival recession is more likely in individuals who had orthodontic treatment compared to those without orthodontic treatment.[4] Here we consider three incomplete variables, recession score, age at the end of treatment and gender, and one complete variable, treatment group. The MICE procedure is initialised by imputing arbitrary values and then each variable in turn is re-imputed, with the new imputed values overwriting the old ones. Usually fewer than 5 cycles of this process are needed to achieve stable imputations, and statistical software typically performs at least 10 cycles.

Figure 1. Illustration of MICE procedure for variables recession score, age (recorded at the end of treatment), gender and treatment group. All imputations are done separately by treatment group


recession score
impute using linear regression on age and gender
gender
impute using logistic regression on recession score and age
age
impute using linear regression on recession score and gender

Readers will note in the figure that values of recession score, the outcome in the analysis model, are used in imputing age, a covariate in the analysis model. This use of the future to predict the past may seem jarring. However, our aim here is not to construct a clinical prediction model, but only to get plausible representations of what the missing data might have been, and we must use all relevant variables to do this.[5] In fact, if we omit recession score from the imputation model for age, then we get imputations that do not respect the associations between variables, and so we tend to under-estimate the association between age and recession score; this may have consequent effects on other associations.

We have said that the imputation models should include all variables in the analysis model. They should also include any variables that are not in the analysis model but are needed to make the data missing at random (MAR) and thus to make the analyses appropriate (see articles 2 and 3) ***[add refs at proof stage – adjust the refs section accordingly]***.

A further possibility is to include other variables that are not in the analysis model but are highly correlated with variables being imputed. Such variables can improve the quality of the imputations and are called “auxiliary variables”. For example, in a repeated measures study, the analysis model might involve only the outcome measured at 6 months, but there may be individuals with the outcome measured at 3 months but not at 6 months. Then including the outcome measured at 3 months as an auxiliary variable will give better imputations at 6 months. Note the difference between MI, which uses the observed relationship between 3- and 6-month outcomes to impute 6-month outcomes, and last observation carried forward, which simply imputes the missing 6-month outcome as equal to the observed 3-month outcome; the latter procedure is hard to justify.[6]

Next, we need to choose the form of each imputation model. Usually each incomplete variable is imputed using a regression model (linear, logistic, etc.) with all the other variables included in their natural form: for example, in imputing age, a four-level ethnicity variable would be included as a factor variable (three dummies), and a clinical score would be included as a linear effect. Sometimes it makes sense to enter variables in a transformed way, for example using a log transformation. Often this is most easily done by transforming the variable outside the MI procedure, so that the procedure imputes the transformed values, which then have to be transformed back to the original scale for analysis. However, imputing and analysing on different scales can be problematic,[7] and statistical advice should be sought.

A particular issue arises if the analysis model explores interactions or non-linear effects, because then the imputation model needs to be complex enough to respect these complexities. In general, this requires statistical advice. An exception is in a randomised trial, where imputing separately in each randomised group automatically respects any treatment-by-covariate interactions. In other data sets it may also make sense to impute separately by a key variable, but this is only possible if that key variable is complete.

A final issue is how many imputations are needed. The main consideration here is to be confident that the results from applying the analysis model to the imputed data are unlikely to be changed in any important way if more imputations were done. A rule of thumb is that the number of imputations should be at least equal to the percentage of incomplete cases[2]: for example, in the example above, if 66% of records are complete for all the variables in the analysis model and 34% of records are incomplete, we could choose 34 imputations. More important is to use one's statistical software to report the Monte Carlo errors on the imputation results. Monte Carlo errors reflect the uncertainty in the results compared to a MI analysis with very many imputations. The analyst should check that varying the results by one or two Monte Carlo errors would not change their interpretation of the results.

Table 1 shows an example using the recommended 34 imputed data sets. Monte Carlo errors are given in square brackets after each result. It is clear that changes of one or two Monte Carlo errors would not change the interpretation of our results. For example, the largest Monte Carlo error is for the confidence interval for gender, and the upper confidence interval which is reported as -0.09 might in fact be as low as -0.13 or as high as -0.05, but in each case the 95% confidence interval for gender lies wholly below zero, indicating statistically significant evidence of a lower recession score for female individuals. Therefore, we could be confident that using fewer imputed data sets than recommended has not affected our interpretation in this instance.

Table 1. Results of regressing recession score on group, gender and age at impression using multiple imputation with 34 imputed data sets. Monte Carlo errors in square brackets express the likely error in the multiple imputation coefficient and 95% confidence limits

Variable	Coefficient	95% confidence interval
Treatment group	-0.89 [<0.01]	(-1.39 [<0.01], -0.39 [<0.01])
Gender	-0.63 [0.01]	(-1.17 [0.02], -0.09 [0.02])
Age	0.03 [<0.01]	(-0.07 [<0.01], 0.12 [<0.01])

This article has explained the choices that need to be made in planning or performing a MI analysis. The next article takes a different perspective: it describes the pitfalls in doing MI.

References

1. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; 18: 681–694.

2. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30: 377–399.

3. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014; 14: 1–13.

4. Gebistorf M, Mijuskovic M, Pandis N, et al. Gingival recession in orthodontic patients 10 to 15 years posttreatment: A retrospective cohort study. *Am J Orthod Dentofac Orthop* 2018; 153: 645–655.

5. Moons KGM, Donders RART, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; 59: 1092–1101.

6. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials* 2012; 9: 396–407.

7. Seaman SR, Bartlett JW, 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.