

Missing data 7. Pitfalls in doing multiple imputation

Ian R White, Nikolaos Pandis, Tra My Pham

Previous articles have explained the principles of multiple Imputation (MI) and how to make the right choices in MI in order to get valid inferences from the data **[add refs at proof stage – adjust the refs section accordingly]**. However, there are still pitfalls – things that can go wrong that can lead to wrong answers. This article explains what these pitfalls are and how to spot them.

The main way MI can go wrong is if the imputations are poorly constructed. Sometimes poor imputations can be hard to spot, but sometimes they are obvious. A very useful technique is to compare the distribution of the imputed values with the distribution of the observed values. Again, we use as illustration the 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.¹ We have data on the outcome variable (recession score), exposure (treatment group), the individual's gender and age at the end of treatment. Figure 1 compares the distribution of the imputed values with the distribution of the observed values for the variable age. The figure shows box and whisker plots for the observed data (on the left) and for the imputed values in the first five imputed data sets. The imputed values are not identically distributed to the observed data, but they are broadly similar.

Figure 1. Box and whisker plots showing observed values of age at baseline and imputed values in the first five imputed data sets, when a suitable imputation model was used

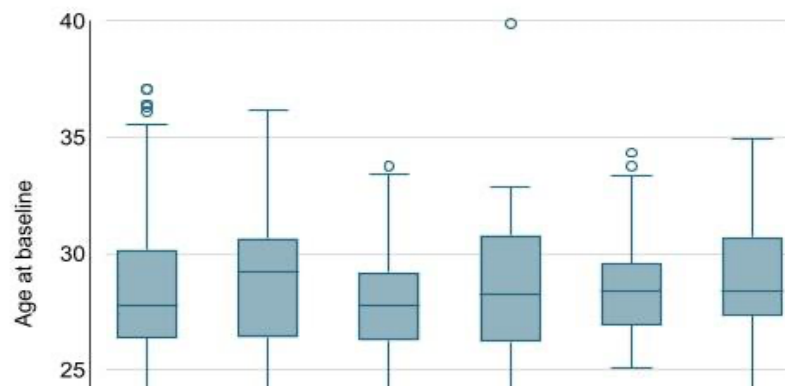
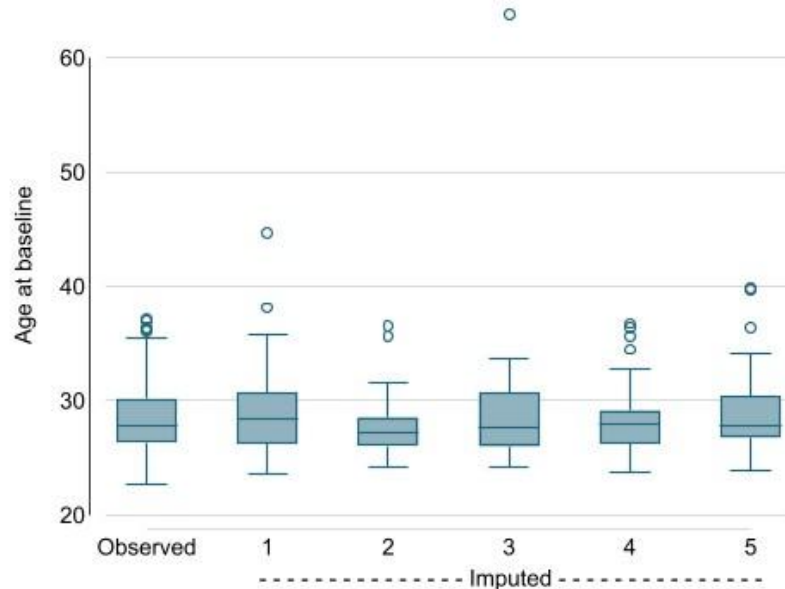
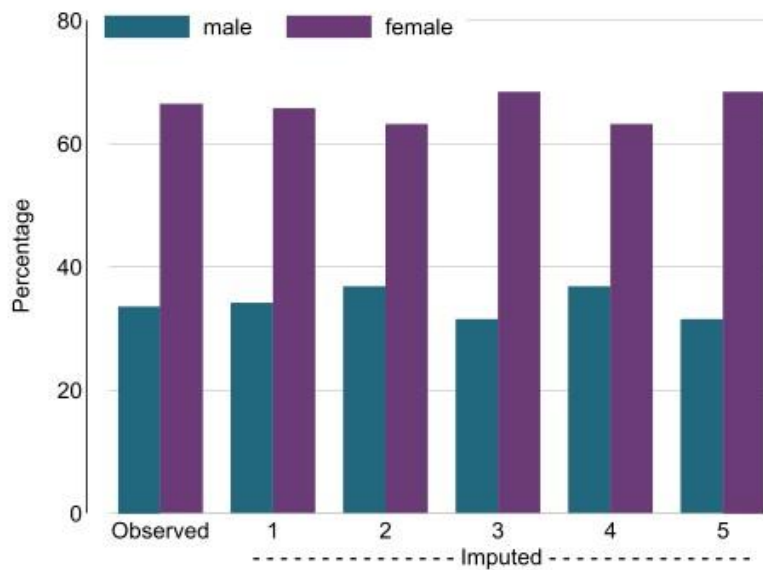


Figure 2. Box and whisker plots showing observed values of age at baseline and imputed values in the first five imputed data sets, when an unsuitable imputation procedure was used. Note the implausible imputed value of age in the 3rd imputed data set



A different graphical approach is required for categorical variables. A bar graph for gender is shown in Figure 3. Again we see that the distribution of the imputed results is similar to, but not identical to, that of the observed results. In some cases when the data are not missing completely at random (MCAR – see article 2 **[add ref at proof stage – adjust the refs section accordingly]**), imputed and observed values may be substantially different: in this case the analyst needs to explain the differences in terms of a different variable. For example, maybe one study centre had more missing data and more women, and therefore including centre in the imputation model rightly led missing data to be more likely to be imputed as women.

Figure 3. Bar graph showing observed distribution of gender and distribution of imputed values in the first five imputed data sets



A useful way to avoid pitfalls is to compare results of MI analysis with complete case analysis.² They are not expected to be identical, since we know that MI makes better use of the data, but they should not be hugely different; if they are substantially different, then we should be able to explain the differences. Table 1 gives an example. First, note that the 95% confidence intervals are narrower in the MI analysis: for example, that for treatment group has width 1.25 in complete case analysis and 1.00 in MI. This is more easily seen by comparing the standard errors (0.32 versus 0.25). The gain in precision is what we hope to achieve by using MI; in these data it is equivalent to increasing the size of the data set by some 60% (the square of 0.32/0.25 is about 1.60). Next, note that the coefficients are somewhat different between complete cases and MI analyses. Some differences are to be expected, since the MI analysis uses more data and makes a different assumption. Here, the differences are less than a standard error, and therefore are not a concern.

Table 1. Results of regressing recession score on treatment group, gender and age at impression using the complete cases and using multiple imputation with 5 imputed data sets

| Variable | Complete case analysis (N=125) | | | Multiple imputation (N=190) | | |
|--------------------|--------------------------------|-------------------|----------------------------|-----------------------------|-------------------|----------------------------|
| | Coeff- icient | Standard error | 95% confidence interval | Coeff- icient | Standard error | 95% confidence interval |
| Treatment group | -0.75 | 0.32 | (-1.38, -0.13) | -0.89 | 0.25 | (-1.39, -0.39) |
| Gender | -0.98 | 0.33 | (-1.64, -0.32) | -0.63 | 0.28 | (-1.17, -0.09) |
| Age | 0.01 | 0.05 | (-0.09, 0.12) | 0.03 | 0.05 | (-0.07, 0.12) |

Another pitfall is failure of the missing at random (MAR) assumption. If data are missing not at random (MNAR), for example if an individual is less likely to attend a dental check-up when their teeth are healthy, then it is hard to find a good technique. Our best suggestion is to consider doing a sensitivity analysis in which imputed values are varied away from the values imputed under MAR. Examples of how to do this are given in White et al (2010).³

A final issue to bear in mind is that there may be a method of analysis that is both simpler and more suitable than MI; some examples are described in article 4 **[add ref at proof stage – adjust the refs section accordingly]**.

References

1. 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.
2. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; 338: b2393.
3. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30: 377–399.