AOSPINE NORTH AMERICA

Global Spine Journal
2018, Vol. 8(8) 892-894

DOI: 10.1177/2192568218811922
journals.sagepub.com/home/gsj
SAGE

The Sin of Missing Data: Is All Forgiven by Way of Imputation?

Joseph R. Dettori, PhD[1], Daniel C. Norvell, PhD[1], and Jens R. Chapman, MD[2]

> . . . in this world nothing can be said to be certain, except death and taxes.
>
> —Benjamin Franklin

> In clinical research, nothing can be said to be certain, except measurement error and *missing data*.
>
> —various clinical researchers

Missing data can seriously compromise inferences from clinical research studies. While great effort should be undertaken to limit the likelihood of missing data through careful study design and conduct, it is inevitable that some values for various reasons will be missing. Missing data are problematic for the data analyst because most statistical procedures require a value for each variable. Therefore, when a dataset is not complete, the analyst needs to decide on how to best deal with the data.

The most common method used for handling missing data is complete case analysis. In complete case analysis, one analyzes only the cases with complete data. Individuals with missing data on any of the included variables are dropped from the analysis. The advantages of this method are its simplicity, its ease of use, and that it is the default for most statistical programs. However, it has its disadvantages, which include lower sample size leading to reduced study power, and potentially biased results, depending on the reason for the missing data. One alternative method for handling missing data is to substitute each missing value with a reasonable guess, and then carry out the analysis as if there were no missing values. This method is termed imputation and is the subject of this article.

Single and Multiple Imputation

Single Imputation

In single imputation, a replacement value is determined by a certain rule. There are many forms of single imputation. Some common examples include the following:

- Last observation carried forward. In this form of simple imputation, a participant's missing values are replaced by the study participant's last observed value.
- Worst observation carried forward. This form of simple imputation provides a conservative estimate by replacing the participant's missing values with the worst observed value.
- Simple mean imputation. Missing values are replaced with the mean of the nonmissing values for that variable.

In general, single imputation methods are not recommended as they depend on assumptions that are often unrealistic and frequently result in an underestimation of the variability and a spuriously low P value.[1,2]

Multiple Imputation

Multiple imputation is a statistical technique that narrows the uncertainty around missing values by calculating several different options or imputations. It is characterized by 3 steps.

Step 1. Create multiple copies of the dataset where the missing values are replaced by imputed values. The imputed values are derived statistically from the observed data, those values not missing. Each imputed dataset is different.

Step 2. Analyze each of the completed imputed datasets. This step results in a separate result for each dataset.

Step 3. Pool the results from step 2 into a final result.

[1] Spectrum Research, Inc, Steilacoom, WA, USA
[2] Swedish Neuroscience Institute, Swedish Medical Center, Seattle, WA, USA

Corresponding Author:
Joseph R. Dettori, Spectrum Research, Inc, Box 88998, Steilacoom, WA 98388, USA.
Email: joe@specri.com

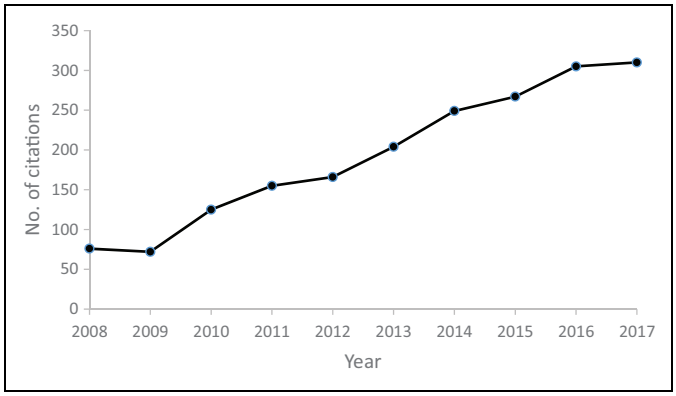

Figure 1. Annual number of citations in PubMed with "multiple imputation" in the title or abstract over a 10-year period.

The use of multiple imputation in the medical literature has become more popular of late, in part because it is readily available in standard statistical software. A search on the term in PubMed from 2008 through 2017 reveals a steady increase in the number of citations from 76 in 2008 to 310 in 2017 (Figure 1). But is multiple imputation the solution for missing data? The short answer is, "it depends on the reason why the data are missing."

Missing Data Mechanisms

Missing data can be placed into 1 of 3 categories based on the process that generated the missing data, which is called the missing data mechanism. These categories are important to understand because the problems caused by missing data and the solutions to these problems are different for the 3 categories. Comprehending the differences among the categories is complicated by the fact that 2 of these mechanisms have confusing names: *missing completely at random* (MCAR) and *missing at random* (MAR).

Missing Completely at Random

Data is said to be MCAR if the missing data value is unrelated to any observed or missing data. In this case, the tendency for the data point to be missing is completely random. In other words, there is no systematic reason that makes some data more likely to be missing than others. For example, missing data as a result of a laboratory technician dropping a blood sample, or data missing from surveys lost in the mail likely occur randomly. This assumption can be tested by separating the missing and the complete cases and examine the group characteristics. If characteristics are not similar for both groups, the MCAR assumption does not hold. Unfortunately, most missing data are not MCAR.

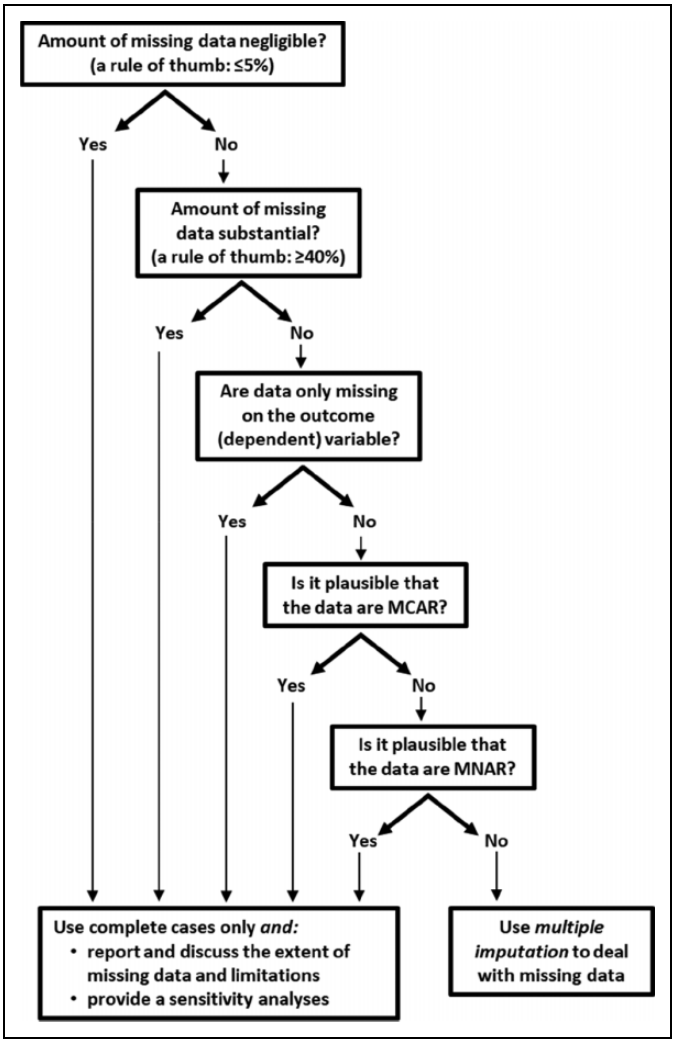

Figure 2. Flowchart to determine when multiple imputation should be used for missing data in clinical research studies. Figure adapted from Jakobsen et al.[2]

Missing at Random

Data is considered MAR if the reason for the missing data is unrelated to the missing values but are related to some of the observed data. The tendency for the data point to be missing under this assumption is systematically associated with the observed data, but not the missing data. For example, if men are more likely to correctly report weight than women, the weight variable is considered MAR.

Missing Not at Random

In data missing not at random (MNAR), missing values *do* depend on unobserved values. Examples include people who are overweight and, as a result, less likely to report their weight, or patients with more comorbidities who tend to drop out of a study more readily than those with less comorbidities.

When Is It Appropriate to Perform Multiple Imputation?[1-3]

Multiple imputation *is not appropriate* when:

- The proportion of missing data is small ($\leq 5\%$ as a general rule). In this case, the potential impact of the missing data is likely small.
- Only the outcome variable has missing values, and not covariate (independent) variables.
- The data are MCAR (rare).
- The proportion of missing data is quite large ($\geq 40\%$ as a general rule).

In all these situations, complete cases analysis can be performed. However, in each situation, an appropriate sensitivity analysis should be conducted and the potential effect of the missing values discussed.

Multiple imputation *is appropriate* when the MAR assumption is reasonable, based on the characteristics of the missing data. When used properly, multiple imputation has been shown to be a valid method for handling missing data in clinical research studies.[4-6] Figure 2 is a flowchart adapted from Jakobsen et al[2] to help determine if multiple imputation is appropriately used in a study.

Summary

- Missing data can seriously compromise inferences from clinical research studies.
- One method for handling missing data is to substitute each missing value with a reasonable guess, and then carry out the analysis as if there were no missing values (imputation).
- Single imputation replaces a missing value based on a predefined rule and includes last observation carried forward, worst observation carried forward, and simple mean imputation. In general, single imputation methods are not recommended.
- Multiple imputation is a statistical technique that creates multiple complete datasets, substituting missing with imputed values. The datasets are analyzed separately, and the results pooled into a final result.
- Whether to use multiple imputation for missing data depends on the missing data mechanism (reason for missing data). To properly use multiple imputation, the missing data should be MAR.

References

1. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012; 367:1355-1360
2. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17:162
3. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test (Madr)*. 2009;18:1-43.
4. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med*. 2013;86:343-358.
5. Jorgensen AW, Lundstrom LH, Wetterslev J, Astrup A, Gotzsche PC. Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. *PLoS One*. 2014;9: e111964.
6. Zhang Y, Alyass A, Vanniyasingam T, et al. A systematic survey of the methods literature on the reporting quality and optimal methods of handling participants with missing outcome data for continuous outcomes in randomized controlled trials. *J Clin Epidemiol*. 2017; 88:67-80.