Missing values in data analysis

Shengping Yang PhD, Gilbert Berdine MD

I recently completed a randomized clinical trial and found out that there were missing values in the data. I have heard of missing value imputations and am wondering if such imputations could improve data analysis.

In clinical studies, it is common for some participants to be lost to follow-up. Epidemiological studies are also susceptible to missing values, depending on the nature of the studies. While we all understand that avoiding this situation is preferable, and that it is important to design studies and develop trial protocols to limit the amount of missing data, sometimes, missing values are unavoidable.

In this article, we will provide a brief introduction on whether we should worry about missing values, and on the strategies used to correct this situation. The issue of handling missing values is still under discussion, and more research needs to be conducted to improve our understanding.

1. Should we worry about missing values?

To better answer this question, we start with explaining the three types of missing mechanisms.

1.1 Missing complete at random (MCAR)

In general, if the probability of being missing is the same for all subjects in a study, then the data are called MCAR.[1] Although MCAR results in a reduction in statistical power as well as the precision in estimating the effect of interest due to a decrease in the number of subjects, it does not cause bias in data analysis. This is because that with MCAR, the subjects with a valid value are considered as a random sample from all the subjects enrolled in a study, and thus they unbiasedly represent all the subjects. An example of MCAR is a missing outcome measurement due to a participant's death caused by a traffic accident. Under MCAR, the probability of being missing does not depend on any observed or missing values, thus there is no systematic differences between the missing and the observed values. In general, if the proportion of subjects with missing value is small, e.g., <5%, then we do not worry too much about missing values.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swrccc.v10i44.1075

1.2 Missing at random (MAR)

If the probability of being missing is the same within groups defined by the observed data, i.e., the missing mechanism depends only on the observed data, but not the missing data, then the data are called MAR.[1] An example is that younger people are more willing to answer a question related to income than older people, and within both the younger and older groups, the probabilities of answering this question are the same across subjects. Intuitively, MAR allows for predicting missing values based on the observed values because the missing mechanism can be modeled. The analysis of MAR data may cause bias in estimating the effect(s) of interest if the missing mechanism is not considered, and thus methods have been developed aiming at providing unbiased results, as well as recovering efficiency (see section 2).[2] Note that, the definition of MAR is more realistic and is broader than MCAR, and many methods make assumptions based on MAR.

1.3 Missing not at random (MNAR)

Distinctly different from both MCAR and MAR, if the probability of being missing depends on the missing data, and remains so even given the observed data, then the data are called MNAR.[1] Specifically, because the probability of being missing is related to unknown data, the missing-data mechanism cannot be modeled. In addition, there is a possibility that there are systematic differences between the missing and observed values, and these differences cannot be evaluated meaningfully. An example of MNAR is that higher income people are less willing to answer a question related to income than people with lower income. Note that, in the example presented in 1.2 (under MAR), income value missingness depends on age, which is observed, and missing values can be predicted based on age; on the other hand, under MNAR, missingness is related to unobserved income, which renders modeling and predicting missing value virtually impossible, and thus no existing statistical method can with certainty take account of the associated potential bias.

To obtain unbiased estimates when missing values exist, the relationship between the missing values and the probability of being missing needs to be modeled. Compared with MAR, MNAR missingness is considered a more serious problem because external information or strong assumptions are needed to model data missingness. Thus, the best strategy to deal with MNAR is to find more data about the causes, then perform sensitivity analysis and make evaluations in an exploratory manner. Note that sensitivity analysis often is a part of data analysis plan and should be well described in advance. An *ad hoc* sensitivity analysis can also be performed if necessary.[1]

With the introduced three main types of missing mechanisms, it may be tempting to determine which one of the mechanisms the data of interest fit best, and then perform data analysis accordingly. However, although it is possible to demonstrate that missing data are not MCAR,[3] the MAR and MNAR mechanisms cannot be distinguished because the missing data are unknown, and it therefore cannot be verified whether the observed data can predict the unknown data. Consistent with this reality, pure MCAR, MAR and MNAR mechanisms rarely exist, and it would be more meaningful to consider the true mechanism as a continuum of MAR and MNAR, and to develop a data analysis plan that accommodates such a complexity.

2. The strategies to handle missing values

2.1 Complete-case analysis

This is the default method used in many data analysis plans. Specifically, all the analyses are performed on only those who have completed a study (completers), and subjects with any missing data are excluded. When the typical statistical analysis methods are used, the required assumption is that the completers are a random sample of the complete study participants, i.e., MCAR. Otherwise, if the missing mechanism is not MCAR, then the effect estimation(s) might be biased. It is evident that the analysis of completers results in reduced statistical power because the total number of subjects is smaller compared with the study enrollment. To increase the number of subjects, sometimes a pairwise deletion method, which uses all the observed data, can be used. For example, to calculate the correlation between two variables, all the participants with valid data for these two variables will be included, regardless of whether there are missing values in the other variables. One potential inconvenience of these analyses is that the number of subjects might differ for different outcomes, and it is important to clearly report such differences.

2.2 The indicator method

This method is commonly used in epidemiological studies, especially when the baseline covariates are partially observed.[4] By including an indicator variable, the systematic differences between the observed and missing data can be modeled and the full dataset can be retained. For example, some participants do not have baseline smoking status data, and they can be assigned to an "unknown" category. Although this category might include a heterogeneous group of people, they share some similarities, compared to those who reported smoking status. Under certain conditions, the indicator method yields an unbiased effect estimate (van Buuren 2018).[5] However, if these conditions are not met, it might generate severely biased estimates, even under MCAR.

2.3 The imputation methods

2.3.1 Single imputation

There are various forms of single imputation methods. In general, it is required that the missing values are replaced based on certain rules.

a. Mean imputation–The mean of the observed values for each variable is computed, and the missing value is replaced by the mean. This method can yield severely biased estimates even under MCAR.[6]

b. Last observation carried forward–In longitudinal studies, a missing value is replaced by the previous observed value. While it might sound meaningful for certain situations, there are strong oppositions on applying this method in any studies because this method has never been shown to be able to generate unbiased estimates.[7]

c. Baseline observation carried forward–the missing value is replaced by the baseline observed value.

One commonly acknowledged limitation of a single imputation is that it results in underestimated data/effect variability. This is because this method directly replaces a value of uncertain, i.e., missing, with a value with certainty. It worth noting that the validity of these methods depends more on the assumptions, e.g., why the missing values should be considered the same as the previous values rather than random (MCAR). In general, most of single imputation methods are based on strong and unrealistic assumptions, and studies have demonstrated that multiple imputation methods often outperform these methods under most commonly seen conditions.

2.3.2 Multiple imputation (MI)

Multiple imputation is a flexible and attractive approach dealing with missing values and has been incorporated in several commonly used statistical software packages.[1,8,9] It tackles the uncertainty about the missing data in two steps,

a. Creating several different imputed data sets
Because we can never know what the missing values are exactly, it is a viable solution to generate several values by randomly sampling the predictive distribution derived from using the observed data, to account for the uncertainty associated with the missing values. And it has been shown that generating a small number of imputed datasets, for example 5, could substantially improve the quality of estimation.[10]

b. Combining the results obtained from datasets generated from step 'a.'
Once the imputed datasets have been generated, standard statistical methods can be applied to each of the imputed dataset. Then, all the effect estimates from these analyses can be summarized into one statistical inference. This is valid because the summarized information is obtained by averaging over the distribution of the missing data, given the observed data.

Note that to model the distribution of the missing data correctly, a wide range of variables should be included in the imputation models, including all variables required for evaluating the effect of interest, all variables predictive of the missing values, as well as those influencing the process causing the missing values.[9]

Although MI has many desirable properties for data with MCAR and MAR–for example, MI can recover efficiency when data are MCAR–it might result in biases similar or even larger than complete-case analysis if the assumptions made are not valid. Therefore, if results from complete-case and MI analyses differ substantially, then attempts should be made to understand the causes of such inconsistencies.

2.4 Model-based methods

The maximum likelihood method is one of the model-based methods for modeling data with missing values. Specifically, the joint distribution of outcome and covariates is fitted to maximize the probability of observing the values that are truly observed.[11] One potential issue with the likelihood method is that the normality assumption might not be valid. In addition, only a few model-based methods are available in commonly used statistical software.

In summary, missing values are often unavoidable in clinical and epidemiological studies. Depending on the mechanisms that lead to missing data, different approaches should be taken to avoid biased results. While complete-case analysis is often the default method to apply, predicting missing values by multiple imputation is gaining increased popularity. It worth noting that no single analysis method is definitive when missing data occur, and a combination of different

strategies and methods should be employed to minimize the risk of making incorrect conclusions.

Keywords: Missing values, multiple imputation, missing at random

Article citation: Yang S, Berdine G. Missing values in data analysis. The Southwest Respiratory and Critical Care Chronicles 2022;10(44):57–60

From: Department of Biostatistics (SY), Pennington Biomedical Research Center, Baton Rouge, LA; Department of Internal Medicine (GB), Texas Tech University Health Sciences Center, Lubbock, Texas

Submitted: 7/11/2022
Accepted: 7/13/2022
Conflicts of interest: none

References

1. Jakobsen JC, Gluud C, Wetterslev J, et al. 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 Dec 6;17(1):162. doi: 10.1186/s12874-017-0442-1.
2. Dziura JD, Post LA, Zhao Q, et al. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med 2013 Sep 20;86(3):343–58.
3. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988; 83(404):1198–202.
4. Perkins NJ, Cole SR, Harel O, et al. Principled approaches to missing data in epidemiologic studies. Am J Epidemiol 2018 Mar 1;187(3):568–575.
5. van Buuren S. Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC. 2018.
6. Jamshidian M, Bentler PM. ML estimation of mean and covariance structures with missing data using complete data routines. J Educational Behavioral Statistics 1999;24:21–41.
7. Lachin JM. Fallacies of last observation carried forward analyses. Clin Trials 2016 Apr;13(2):161–8.
8. Harel O, Mitchell EM, Perkins NJ, et al. Multiple imputation for incomplete data in epidemiologic studies. Am J Epidemiol 2018 Mar 1;187(3):576–584.
9. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393.
10. Hand DJ, Adèr HJ, Mellenbergh GJ. Advising on Research Methods: A Consultant's Companion. Huizen, Netherlands: Johannes van Kessel.2008;305–332.
11. Allison PD. Handling missing data by maximum likelihood, statistical horizons. In: SAS global forum 2012 statistics and data analysis; 2012.