Handling missing data in research

Priya Ranganathan, Sally Hunsberger[1]

Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India, [1]Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, United States

Abstract Missing data are an inevitable part of research and lead to a decrease in the size of the analyzable population, and biased and imprecise estimates. In this article, we discuss the types of missing data, methods to handle missing data and suggest ways in which missing data can be minimized.

Keywords: Data collection, imputation, missing data

Address for correspondence: Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India.
E-mail: drpriyaranganathan@gmail.com

Received: 05-03-24, **Accepted:** 08-03-24, **Published:** 05-04-24.

INTRODUCTION

Missing data are an unavoidable part of research. Research data may be missing for several reasons – losses to follow-up, participant refusal to provide certain types of data, technical errors such as equipment malfunction, or inability to measure data at a particular time point (e.g., failure to capture pain scores if a patient is on mechanical ventilation). The missing data points could include the exposure, outcome, confounders, or other characteristics. In this article, we will look at the types of missing data, the impact of missing data on analysis and results, methods used to adjust for missing data, and ways to minimize missing data in research studies.

TYPES OF MISSING DATA

Three mechanisms of missing data have been described (originally attributed to Rubin).[1]

Missing completely at random

This refers to data where the pattern of missingness is completely random, and is unrelated to either observed or unobserved characteristics of the participants (or unit being measured). The probability of a value being missing is the same for all participants in the study. Examples of these include administrative censoring, accidental deletion of data, or failure of equipment used for measurements on a particular day. The remaining analyzable data are a random subset of the complete dataset.

Missing at random

This refers to situations where the missingness of data is related to participant characteristics. The probability that a value is missing is related to characteristics that have been measured in the study.

Missing not at random

Here, the probability of a value being missing is related to the value itself and to a characteristic which has not been measured by the researcher.

Let us look at the example of patients with diabetes who are asked to follow a low-carbohydrate diet and need to report to the hospital every 3 months for a blood glucose test.

1. If some patients are unable to report on a particular

Access this article online

Quick Response Code:

Website:
www.picronline.org

DOI:
10.4103/picr.picr_38_24

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Ranganathan P, Hunsberger S. Handling missing data in research. Perspect Clin Res 2024;15:99-101.

day due to bad weather, their missing values are missing completely at random (MCAR)

2. If age is associated with blood glucose and elderly patients find it more difficult to access the hospital, and therefore, have more missing values, then these data are missing at random (MAR)
3. If patients who have not complied with the diet (and are, therefore, likely to have higher blood glucose values) do not report to the hospital for testing and we have not measured compliance with the diet, then these data are missing not at random (MNAR).

IMPACT OF MISSING DATA

The percentage of missing data reflects the quality of a research study. A large extent of missing data and the use of inappropriate methods to handle the missingness can lead to bias, reduction in sample size loss of power, and imprecision in study results. Readers are referred to this article by Larkins *et al.* on how to assess the impact of missing data on the validity of the results of a research study.[2]

TECHNIQUES TO HANDLE MISSING DATA

We will briefly review some of the techniques used to handle missing data during analysis. A detailed explanation is beyond the scope of this article.

DELETION TECHNIQUES

a. Complete case analysis (also known as listwise deletion or available case analysis): Cases are dropped from the analysis if they have even one missing data point
b. Pairwise deletion: Cases are dropped from a particular analysis if they have one of the data points missing for the variables in that analysis. These cases can be included in other analyses where data points are not missing.

Deletion of cases is an easy way to account for missing data and is appropriate if data are MCAR. However, it can introduce bias by not accounting reason for missingness and leads to an ambiguous definition of sample size. These techniques are more appropriate for observational studies rather than for randomized trials where the outcome primarily depends on treatment assignment.

Missing indicator method

For regression models, where values for a predictor variable are missing, a separate code is created for the category of missing data points (and labeled as missing) so that the case remains in the regression model. This method may give biased estimates when used for nonrandomized studies.

Imputation techniques

These techniques are used to fill in the missing data with substitute values.

a. Single imputation techniques: Replace missing values with a single plausible value. Various methods of doing this include:
   i. Mean or median value for the actually observed data
   ii. Regression analysis from the observed data to predict observations for the missing data
   iii. Last observation carried forward or baseline observation carried forward
   iv. Best case/worst case analysis.

Single imputation methods treat the filled-in observations as real data so the standard deviation of the parameter estimates is not correct. Therefore, multiple imputation methods have been developed to take in the uncertainty of the filled-in observations.

b. Multiple imputation techniques: They are extended versions of simple imputation and use repeated combinations of the observed data to predict a set of plausible values (instead of a single value) for the missing data. These values are then combined to get an "average" estimate of the missing parameter.

Multiple imputation methods provide better results than single imputation methods since they account for the uncertainty of the filled-in observations. Methods such as imputing with the mean values or worst-case scenarios give biased estimates but may be useful to examine how sensitive results are to missing data. Model-based imputation that includes covariates related to the missingness structure can give unbiased results.

Maximum likelihood techniques

These use the full data set that has missing values to obtain parameter estimates. The method has a likelihood function that is factored into two components that are computed separately and then maximized together to find the parameter estimates. One component includes observations with missing data on some variables and the other component includes observations with data on all variables. If the data are MAR or MCAR and the model is specified well, the parameter estimates and standard errors are unbiased.

Generalized estimating equations

This is an analysis method used to analyze longitudinal or repeated measurement data. With this type of data, there may be some missing values for a subject. If generalized estimating equation (GEE) is used, only subjects with complete data are used, which can lead to biased estimates when data are MAR or MNAR. Multiple imputation methods and weighted methods have been developed to use with GEE that reduce bias when data is MAR.

Minimizing missing data in research

Despite the various methods to handle missing data, it is clear that there are several problems associated with incomplete research data. Researchers must adopt techniques to ensure the completeness of their data. Some of these could include:

- Careful choice of outcomes: Keep outcomes relevant and easy to measure, and collect only essential data for each outcome
- Decrease demands on participants: For example, avoid frequent follow-up visits and reduce the complexity of procedures at the follow-up visits
- Allow flexibility where it will not affect the validity of outcomes: For example, remote collection of data and window periods for study assessments
- Initiate standard operating procedures and training of the research team on study procedures and assessments
- Use a pilot study or run-in period to identify problems with compliance to the protocol
- Develop user-friendly and objective case record forms
- Prespecify the extent of missing data that will be considered acceptable, and how this will be dealt with.

REPORTING MISSING DATA IN PUBLICATIONS

To allow readers and reviewers to interpret the validity of study results, authors must clearly report the extent of missing data in their manuscripts. Frameworks such as the CONSORT checklist for randomized trials and the STROBE checklist for observational studies mandate the reporting of details of missing data.[3-5]

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Groenwold RH, Dekkers OM. Missing data: The impact of what is not there. Eur J Endocrinol 2020;183:E7-9.
2. Larkins NG, Craig JC, Teixeira-Pinto A. A guide to missing data for the pediatric nephrologist. Pediatr Nephrol 2019;34:223-31.
3. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726-32.
4. Butcher NJ, Monsour A, Mew EJ, Chan AW, Moher D, Mayo-Wilson E, *et al*. Guidelines for reporting outcomes in trial reports: The CONSORT-outcomes 2022 extension. JAMA 2022;328:2252-64.
5. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet 2007;370:1453-7.