Missing data 1. Why missing data are a problem

Tra My Pham, Nikolaos Pandis, Ian R White

Missing data are a common issue in medical research. In this series of articles, we seek to explain in non-technical language some of the important ideas about missing data, and how they can be addressed in practice, using examples from orthodontic studies. We start with a discussion of why missing data are problematic.

Missing data refer to values that we intended to collect in a study to answer a research question, but for some reason, we were not able to collect them.[1] Let's consider an example of how missing data might occur in practice. This example was created using data from an orthodontic randomised controlled trial comparing how probing depth evolves over time when using two types of lower lingual retainers.[2] In this example, probing depth (the outcome) is measured on six teeth (lower left canine to lower right canine) per individual at six time points. Table 1 presents an example of data on treatment, gender and age at baseline, and mean probing depth across six teeth at six time points, for five individuals. Some individuals have data observed for all variables (individual 10). Some individuals fail to attend some follow-up appointments, resulting in their outcome data being missing intermittently for some time points (e.g. mean probing depth is missing for individual 14 at time point 5). Other individuals end their participation before the end of the trial, with their outcome data being missing for time points subsequent to dropout (e.g. mean probing depth is missing for individual 11 at time points 4–6). Sometimes, the recorded probing depth might also be judged to be wrong, and deleted during the data cleaning process. Similarly, there might be missing data in the individual characteristics collected at baseline (e.g. age at baseline is missing for individuals 11 and 13). Therefore, when such data are analysed, missing values can occur in more than one variable considered in the analysis, e.g. the outcome, one or more covariates, or both.[1]

Table 1. An example of missing data in an orthodontic study comparing probing depth between two types of retainers

Individual	Baseline characteristics		Treatment	Mean probing depth across six teeth					
	Gender	Age		Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
10	F	17	B	2.83	1.50	1.67	1.75	1.75	1.50
11	F	.	A	1.08	0.83	1.08	.	.	.
12	F	31	A	1.08	0.92	.	1.50	.	.
13	F	.	B	1.17	0.92	1.00	1.17	0.92	.
14	M	17	A	2.50	1.58	1.42	1.58	.	1.83

Missing data pose several challenges to statistical analysis.[3] Since missing data lead to a loss of information, one direct consequence of having missing values in the analysis is reduced statistical power: that is, the probability of a correct conclusion is reduced in the case that the alternative hypothesis is true. A second consequence of the occurrence of missing data is that we cannot proceed with performing the analysis that is intended for complete data, but we need to first find a way to deal with the missing values.

Whenever there are missing values in our data set, any analysis that involves partially observed variables (i.e. those containing missing values) makes certain assumptions about how these values have become missing. These assumptions will be introduced in article 2. The validity of these assumptions cannot be verified from the observed data alone, as we need to know the missing values in order to check these assumptions. However, some assumptions might be more plausible than others, depending on the context of the study and the process of collecting the data.

Following from this, a more insidious pitfall of missing data is that making the wrong assumption about the missing values can lead to wrong conclusions being drawn from the study. There might be bias in the estimates of treatment effects or regression coefficients: that is, estimates might be systematically different from their true underlying values. Using the probing depth example above, if for some reason individuals with greater probing depth values are lost from one treatment arm and excluded from the analysis, this treatment arm will appear 'better' than the comparator arm just because the 'worse' outcomes were excluded. Standard errors might also be wrongly estimated, resulting in incorrect p-values and confidence intervals. Due to the loss of information from available data, analyses might also be inefficient, leading to confidence intervals that are too wide.

The data analyst should therefore consult the research staff who collect the data as well as clinical experts in order to determine a plausible assumption for the missing values. Additionally, assessing sensitivity of the results under alternative assumptions is key in any analysis with missing data (see article 7).

This article has summarised the challenges presented by missing data. The next article will explain the assumptions that can be made about missing data. Articles 3 and 4 will explain how to explore missing data and how the type of missing data influences the choice of analysis. The remaining four articles will explain multiple imputation, the most popular way to handle missing data, focussing on its basic ideas, the choices it requires, the pitfalls to be avoided, and how to report data analysed by multiple imputation.

References

1. Carpenter JR, Kenward MG. *Missing data in randomised controlled trials – a practical guide*. Birmingham: National Institute for Health Research, Publication RM03/JH17/MK, http://www.missingdata.org.uk (2008).
2. Węgrodzka E, Kornatowska K, Pandis N, et al. A comparative assessment of failures and periodontal health between 2 mandibular lingual retainers in orthodontic patients. A 2-year follow-up, single practice-based randomized trial. *Am J Orthod Dentofac Orthop* 2021; 160: 494-502.e1.
3. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004; 1: 368–376.

Funding, acknowledgment, and affiliations for all articles

Funding

TMP and IRW were supported by the Medical Research Council [grant number MC_UU_00004/07].

Acknowledgement

We would like to thank Jadbinder Seehra for reviewing and commenting on draft versions of these articles.

Affiliations

TMP, IRW: MRC Clinical Trials Unit at UCL, 90 High Holborn, London WC1V 6LJ, UK

NP: Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, University of Bern, Bern, Switzerland