Missing data: the impact of what is not there

Rolf H H Groenwold[1,2] and **Olaf M Dekkers**[1,3]

Departments of [1]Clinical Epidemiology, [2]Biomedical Data Sciences, and [3]Endocrinology, Leiden University Medical Center, Leiden, the Netherlands

Correspondence should be addressed to R H H Groenwold
Email
R.H.H.Groenwold@lumc.nl

Abstract

The validity of clinical research is potentially threatened by missing data. Any variable measured in a study can have missing values, including the exposure, the outcome, and confounders. When missing values are ignored in the analysis, only those subjects with complete records will be included in the analysis. This may lead to biased results and loss of power. We explain why missing data may lead to bias and discuss a commonly used classification of missing data.

European Journal of Endocrinology
(2020) **183**, E7–E9

Introduction

In almost all clinical research, one or more of the measured variables have missing values. For example, in a study of daily exercise and the risk of type 2 diabetes mellitus, measurements of exercise levels may not be available for all subjects, and if smoking is a confounder in this study, smoking status may be unknown for some of the subjects. This situation is commonly referred to as missing data. Standard statistical approaches ignore missing data, meaning that subjects with a missing value will not contribute to the analysis. This is called complete (or available) case analysis. Importantly, this applies to all variables in the model, not only outcomes. There are two potential problems with missing data: loss of power and bias. Here, we discuss why missing data can lead to bias and argue that claims about the possible impact of missing data should neither be based on the study design nor on the percentage of missing data.

Bias due to missing data

Consider a randomised trial among elderly with subclinical hypothyroidism comparing levothyroxine against placebo; 120 patients receive levothyroxine, while another 120 receive placebo. The results of this hypothetical study are summarized in Table 1.

In scenario A, no missing data, the outcome is observed for all patients in the trial, and the trial suggests a 40% risk reduction (relative risk 0.60). We consider scenario A to be the reference. In scenarios B and C, only 2% of the patients have missing outcome values, while in scenario D this is as much as 50%. In a complete case-analysis, in scenarios B, C, and D, the data of 235, 235, and 120 patients, respectively, contribute to the analysis. Note that the effect estimates in scenarios B and C differ from the reference value; the effect estimate in scenario D does not, although it is much less precise owing to the smaller sample size. Apparently, the percentage of missing data is not very informative about the risk of bias. Note also that missing data can lead to an overestimation as well as to an underestimation of the treatment effect.

The effect estimate in scenarios B differs from the reference value, because the risk of the outcome among those with an observed outcome value ($n=115$, risk 17%) in the levothyroxine arm does not correspond to the risk of the outcome among all subjects in that treatment arm (risk 20%, unobserved). Hence, the calculated risk ratio will be biased too. Likewise, for scenario C the observed risk in the placebo arm ($n=115$, risk 30%) does not represent the true (yet unobserved) risk among the placebo treated subjects ($n=120$, risk 33%). The effect estimate in scenario D does not differ from the reference value, because in each treatment arm the risk of the outcome among those with

https://eje.bioscientifica.com
https://doi.org/10.1530/EJE-20-0732

Published by Bioscientifica Ltd.

Table 1 Numerical examples of the possible impact of missing data in a hypothetical trial of levothyroxine.

Scenario	Percentage missing data	Levothyroxine treatment	Placebo	RR (95% CI)
		$n = 120$	$n = 120$	
A	0%	24/120 (20%)	40/120 (33%)	0.60 (0.93; 0.93)
B	2%	19/115 (17%)	40/120 (33%)	0.50 (0.31; 0.80)
C	2%	24/120 (20%)	35/115 (30%)	0.66 (0.42; 1.03)
D	50%	12/60 (20%)	20/60 (33%)	0.60 (0.32; 1.12)
E	17%	15/100 (15%)	25/100 (25%)	0.60 (0.34; 1.07)

RR, risk ratio.

an observed outcome value equals the risk of the outcome among all subject in those treatment arms.

In scenario E, in each treatment arm the risk of the outcome among those with an observed outcome value differs from the true risks in both groups (reference). Nevertheless, the *risk ratio* that is calculated based on these risks corresponds to the true value.

In observational studies with the need to adjust for confounding (1), the proportion of missing values can be considerably larger than in randomised trials. Think of an observational cohort to study the effect of glucose levels on cardiovascular events; there are many potential confounders for this association (age, BMI, lifestyle, amount of salat eating, etc). Although a single confounder may have only 5% missing values, with 10 potential confounders (not unlikely for the association between glucose and cardiovascular events) this could mean that only for 50% of the subjects information is available about all confounders. Even if a complete case analysis does not lead to biased results, it would still be very inefficient (i.e. low power and wider CIs) than a situation without missing data. As the default option in statistical software is to include only subjects without missing values, missing data are overlooked easily. Therefore, for each analysis that is conducted, the actual number of included subject should be reported. Researchers could have a look at the final adjusted statistical model, where the output displays the number of subjects included. Preferably, a comparison is made of subjects with and without missing values, because this may also provide insight in the possible reasons for missing data (the missing data mechanism) and guidance about choosing the optimal statistical approach.

Classification of missing data

A commonly used classification of missing data describes the (assumed) mechanism that leads to the data being missing (see (2, 3) for an introduction to the topic and Table 2 for definitions (4, 5)). If missingness is a random process (e.g. a batch with lab tests gets lost in the lab) and no systematic difference exist between those with and those without missing values, this is referred to as missing completely at random (MCAR); scenario D could be an example of MCAR. If missingness is a random process within levels of an observed variable, it is – somewhat confusingly – called missing at random (MAR). For example, in the trial it could be that outcome values are more often missing for males than for females, but among males it is a random process whether or not the outcome is observed (and ditto for females). If missingness is not a random process (within levels of an observed variable), but depends on unobserved variables, such that systematic difference between those with and those without missing values depend on unobserved factors, this is referred to as missing not at random (MNAR). Scenarios B and C are example of MNAR. Although we know that, in scenario B, all five subjects with a missing outcome value in fact had the outcome, obviously the researchers will not know this.

Table 2 Classification of missing data.

- Missing completely at random (MCAR) means that the probability of a value being missing is the same for all subjects in a study and does not depend either on observed or on unobserved characteristics of the subjects in the study. In that case, missingness is unrelated to the specific values that are missing or observed values in the data.
- Missing at random (MAR) means that the probability of a value being missing is the same within groups of subjects, where the groups are defined based on the observed data. In that case, missingness depends on observed, but not on unobserved, characteristics of the subjects in the study, including the specific values that are missing.
- When missing data are neither MCAR nor MAR they are said to be missing not at random (MNAR), which means that the probability of a value being missing depends on the specific value that is missing in addition to observed characteristics of the subjects in the study.

The distinction between MCAR and MAR can be made based on the observed data. However, because the distinction between MCAR/MAR and MNAR relies on unobserved data, this distinction cannot be made using observed data only. Therefore, assumptions about missing data mechanisms can be supported by data analysis, but cannot ultimately be confirmed; the data will not tell which missing data mechanism is at work.

Various methods to deal with missing data have been developed to reduce the bias that can accompany complete case analysis. Multiple imputation is, nowadays, commonly used to 'impute' (i.e. fill in) the missing value using a predicted value that is based on the observed data (6). One crucial assumption underlying this method is that missing data are MAR, as in that case missing data can be filled in validly based on observed data; however, in the case of MNAR, results may still be biased. For an introductory overview of methods to handle missing data, we refer to the literature (2, 3, 4).

Whether or not results are biased depends on the missing data mechanism in combination with the method that is applied to deal with missing data and the method of data analysis (7). For example, complete case analysis might be appropriate in case of missing data that are MCAR, but perhaps not if missing data are MAR. Multiple imputation, however, may be appropriate when missing data are MAR, but not if these are MNAR. However, there are also situations in which complete case analysis is appropriate even when missing data are MNAR (7); scenario E in Table 1 is an example. It is too simplistic to say that, for example, MAR will never and MNAR will always result in a bias. To make claims about the potential impact of missing data requires assumptions about the missing data and an understanding of how missing data affect the analysis method that is applied.

Concluding remarks

Missing data can result in bias, although this need not always be the case, depending on the missing data mechanism and the applied statistical approach. In a complete case analysis, already with low percentages of missing values there can be substantial bias and with high percentages there need not be a bias. Nevertheless, the percentage of missing values may be related to the quality of the study in general and specifically the quality of the collected data. As such, the percentage of missing values may be a proxy for study quality and risk of bias, although not necessarily bias due to missing data. Even randomised trials are not immune to bias due to missing data (8, 9, 10), although the extent of missing data in trials is probably smaller than in observational studies. As default statistical methods ignore subjects with missing values, each reported analysis should be accompanied by the actual number of subjects included in that analysis. Apart from a possible impact in terms of bias, missing data reduce the precision of effect estimations. Instead of depreciating any missing data bias, because of a study being a randomised trial or because of the low percentage of missing values, researchers should discuss the possible missing data mechanism in relation to the data analysis and consider possible solutions, including imputation techniques.

Declaration of interest

R G reports no conflicts of interest. O M D is a Deputy Editor for *European Journal of Endocrinology*. He was not involved in the review or editorial process for this paper, on which he is listed as an author.

Funding

This work was supported by grants from the Netherlands Organization for Scientific Research (ZonMW-Vidi project 917.16.430) and the LUMC.

References

1 Groenwold RHH & Dekkers OM. METHODOLOGY FOR THE ENDOCRINOLOGIST: Basic aspects of confounding adjustment. *European Journal of Endocrinology* 2020 **182** E5–E7. (https://doi.org/10.1530/EJE-20-0075)

2 Donders AR, van der Heijden GJ, Stijnen T & Moons KG. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 2006 **59** 1087–1091. (https://doi.org/10.1016/j.jclinepi.2006.01.014)

3 Lee KJ, Tilling K, Cornish RP, Little RJ, Bell ML, Goetghebeur E, Hogan JW & Carpenter JR for the STRATOS initiative Framework for the treatment and reporting of missing data in observational studies: the TARMOS framework. arXiv:2004.14066 2020.

4 Van Buuren S. *Flexible Imputation of Missing Data*. CRC Press, 2018.

5 Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A & Verbeke G (eds). *Handbook of Missing Data Methodology*. CRC Press, 2014.

6 Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM & Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009 **338** b2393. (https://doi.org/10.1136/bmj.b2393)

7 Hughes RA, Heron J, Sterne JAC & Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* 2019 **48** 1294–1304. (https://doi.org/10.1093/ije/dyz032)

8 Bell ML, Kenward MG, Fairclough DL & Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ* 2013 **346** e8668. (https://doi.org/10.1136/bmj.e8668)

9 Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA *et al.* The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* 2012 **367** 1355–1360. (https://doi.org/10.1056/NEJMsr1203730)

10 Groenwold RH, Moons KG & Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. *Canadian Medical Association Journal* 2014 **186** 1153–1157. (https://doi.org/10.1503/cmaj.131353)

Received 29 June 2020
Accepted 17 July 2020