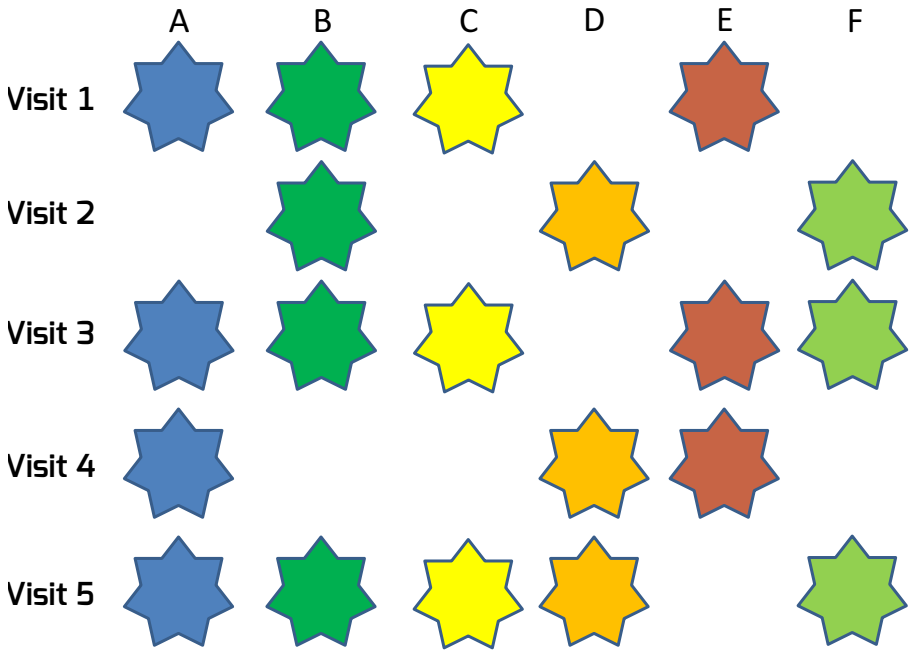
The perils of the unknown: Missing data in clinical studies

Helen Bridge[1] and **Thomas M. Schindler**[2]
1 AstraZeneca, Cambridge, UK
2 Boehringer Ingelheim Pharma, Biberach, Germany

Correspondence to:
Helen Bridge or Thomas M. Schindler
Medical Writing Europe
Boehringer Ingelheim Pharma
Birkendorfer Str. 65
88307 Biberach an der Riß, Germany
Thomas.Schindler@
boehringer-ingelheim.com

Abstract
The phenomenon of missing data is ubiquitous in clinical studies. Both the extent of missing data and the structure of missing data can introduce bias into study results and lead to wrong conclusions. Medical writers should be aware of the extent of missing data and should describe the methods used to deal with the issue. This article outlines some of the most commonly used statistical methods for handling missing data. The traditionally used last-observation-carried-forward (LOCF) method to fill data gaps is problematic in many ways. It is better to employ a method that reduces bias, such as multiple imputation (MI) or mixed-effects models for repeated measures (MMRM). Clinical study design can also help minimise the quantity of missing data.



You may think that this topic does not concern medical writers. You may think this is something for data managers or statisticians. Well, you may need to think again. Every study you write about in a publication or a study report will have dealt with the problem of missing data. Moreover, the way this problem was handled by those conducting the study can have far-reaching consequences. The complexity of clinical studies means that everything is related to everything else, so the issue of missing data is linked to many other aspects, from study design to patient retention, data analysis, and the conclusions that can be drawn. Because of this, every scientific report about a clinical study must take note of the extent to which data are missing and how this unfortunate but inevitable fact has been handled.

A version of this article previously appeared in the *AMWA Journal* 2017;32(3):116–19, 123.

Why are data missing in clinical studies?
In an ideal world and an ideal clinical trial, all patients would come to all visits, all patients would take their medication each day at the right time, and all patients would undergo all procedures as planned. No study investigators or patients would move or decide to leave the study, and nobody would have an accident, fall ill, or die during the study. Only in such a scenario could the medical writer be absolved of having to talk about missing data. But as seen from this non-exhaustive list, in the real world things are never perfect, and the issue of missing data will invariably arise.

In an ideal world and an ideal clinical trial, all patients would come to all visits, all patients would take their medication each day at the right time, and all patients would undergo all procedures as planned.

What are the issues?
We cannot assume that we will obtain all the data for all patients in a clinical study. This, however, may or may not be a problem, depending on the quantity and nature of the missing data.

There can be no doubt about it: the more data are missing, the shakier the results and conclusions become. It is very difficult to say when a critical limit of missing data has been reached because the size of the study, the indication being studied, the magnitude of difference between treatments, and the frequency and nature of the assessments must all be considered. However, if the trial is testing for a difference in outcome events (e.g. heart

attacks) then even a small number of missing data may be important. If outcome data are missing for a sizable proportion of the patients, the whole trial may become invalid.

A second issue with missing data arises when the pattern of missing data differs between the treatment groups. This is likely to introduce bias in the interpretation of results. Data can be missing for various reasons. On the one hand, it could be pure chance that values are missing. For example, a patient misses a study visit because her car broke down and she could not get to the study site. Or a patient decides to leave the study because he needs to move for his wife to take up a new job in a different region. On the other hand, the fact that data are missing could be related to the outcome that is being measured and/or the study treatment. For example, we might have a much higher dropout rate in one treatment group than in the other. This may happen for many reasons, e.g. because of adverse events, lack of efficacy, or unknown reasons. Often it is difficult to know whether data are missing by chance or because of the treatment. Consider a drug that may cause dizziness and a patient who has a traffic accident on her way to the study clinic and ends up in hospital. Is this a chance event or related to the treatment?

Statisticians have developed a theoretical framework to categorise the reasons for missing data. In brief, they distinguish data that are "missing completely at random (MCAR)" from data that are "missing at random (MAR)" and data that are "missing not at random (MNAR)". As the elaboration of these concepts is beyond the scope of this short paper, please consult the reading list.

In a randomised trial comparing two treatments, missing data because of chance events should not be much of a problem, provided they are rare. We would expect chance events to occur with similar frequency in both treatment groups, and therefore no bias is being introduced. However, "missingness" related to the treatment or outcome variable leaves us on very difficult ground.

Suppose we have a study comparing a new wonderdrug (WD) and placebo. WD may cause adverse events that lead to dropout of patients, while patients in the placebo group carry on. Conversely, WD may have good efficacy and no tolerability issues, so the patients taking it remain in the study, while patients in the placebo group drop out because they see no improvement. In these scenarios, we risk underestimating or overestimating the size of the treatment effect.

Differential withdrawal between treatment groups will result in a serious conceptual problem. The goal of randomisation is that the two treatment groups will have similar characteristics at the start of the study. If many patients in one group but not in the other withdraw from the study, the two groups may no longer be comparable at the end. If a sizeable proportion of patients in the WD group drops out because of tolerability issues, we will not only have more missing data in this group, we will also have a different group of people at study end. By exposing patients to WD for some weeks, we unintentionally "select" those patients who are able to tolerate the treatment. Hence, at study end we arrive at a comparison of the placebo group with all its initial demographic and disease characteristics and a modified WD group that consists only of those patients who have tolerated the treatment. Their demographic and baseline disease characteristics may be quite unrepresentative of the initial population. This will make it very difficult to draw any conclusions about the efficacy or safety of WD.

When reporting clinical studies, medical writers need to be alert to signs that missing data are not due to chance and therefore have the potential to cause bias.

When reporting clinical studies, medical writers need to be alert to signs that missing data are not due to chance and therefore have the potential to cause bias. Signs to watch out for include differences between treatment groups in the proportions of patients with missing values or the reasons for withdrawals. Clusterings of withdrawals or missed visits around certain points in time should also raise suspicion. A starting point could be the tables detailing the disposition of patients. If you detect any issues, it is advisable to ask the statistician to provide further information on the missing data.

Now let's look at an example of what missing data can look like for individual patients. Let's assume that we are looking at a trial in patients with type 2 diabetes. We want to find out what effect our new drug has on the long-term marker for blood sugar levels, haemoglobin A1c (HbA1c). We are looking at the change from baseline to study end as our primary endpoint for efficacy. Table 1 depicts the data of five patients.

In this example, we have all values only for patient 1, who has completed all visits. Thus only for her can we easily calculate the change from baseline. Data analysis will be more complicated for the other four patients because they have data missing for some visits. Would it therefore be a good idea to ignore the data from patients 2 to 5, i.e. concentrate the analysis only on "completers"? No, it would not. Looking at the table does not tell us the reasons why the data are missing, and this is a common situation in clinical trials. We may know the broad reasons why some patients withdrew (e.g. "adverse event" or "lack of efficacy") and the reason why a patient died, but patients who are lost to follow-up or who missed some visits may not have detailed reasons recorded. The patients who missed visits in our example may have done so because of the severity of their disease, or because they had

Table 1. Data from 5 patients in a study with the primary endpoint of change from baseline in HbA1c

	Study start / Baseline	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**	**Visit 5/ Study end**	**Comment**
Patient 1	X_0	X_1	X_2	X_3	X_4	X_5	Completer
Patient 2	X_0	X_1	X_2	–	–	–	Withdrew at Visit 2
Patient 3	–	X_1	X_2	X_3	X_4	X_5	No baseline value
Patient 4	X_0	X_1	X_2	X_3	–	–	Died after Visit 3
Patient 5	X_0	–	–	–	X_4	–	Did not attend all visits

adverse events, or because of chance events having nothing to do with their health. The patients who attended all visits could be the younger patients with fewer comorbidities who are fit and mobile enough to make it to every planned visit. If we focus on the "completers" (or "observed cases"), we may be selecting patients who are not representative of the population as a whole. Disregarding all the patients with incomplete data would not only risk bias, but would also make us lose a lot of valuable information.

What can we do about missing data?

A number of different statistical methods exist for handling missing data, and the risk of bias in a particular situation will vary depending on the method chosen.

Simple imputation methods

For both ethical and economic reasons it would be wise to use all the data we have gathered during a clinical study. Thus we need to find the best and most appropriate ways to use the data. One method that has been used for many years is the "last-observation-carried-forward" approach, or LOCF. The LOCF method is very simple as it fills in (or "imputes") the missing data items with the last observation that was obtained at a previous time point (Table 2).

After having performed LOCF, we can now easily calculate the change from baseline to study end (for patient 3 we use the data from Visit 1 as a starting point). This method looks convenient as it fulfils our aim to include all patients in the analysis and provides a mechanism for filling in the missing values. (A similar imputation method is BOCF, i.e. "baseline observation carried forward". Here a patient's baseline value is carried over.)

Although appealing in its simplicity, the LOCF method is likely to introduce bias and may even lead to wrong conclusions. Suppose, for example, we perform a study in a population of patients with depression. Typically, in a group of patients with depression some will improve spontaneously in their condition. If many patients in the active treatment group in the study drop out because of adverse events and the LOCF method were applied, this would likely result in underestimation of the treatment effect of the drug. The reason for this is that not all the spontaneous improvements in the active treatment arm would have had a chance to surface and be recorded. Conversely, suppose we perform a study in a population of patients who have a condition that worsens over time. The condition in the group of patients that received placebo would continue to worsen, resulting in a worse score at study end. If some patients in the active treatment group leave the study prematurely due to adverse events, the LOCF method would mean using an earlier, better score for these patients than the scores they would have had at study end, had they stayed on study as their condition continued to worsen over time. This would likely favour the active treatment and result in overestimation of the treatment effect. Because of its potential for introducing bias and leading to incorrect conclusions, regulators and leading statisticians urge clinical researchers to stop using the LOCF method.

> No amount of statistical expertise can make up for the absence of real data. – MG Kenward

Methods involving statistical modelling

Instead of filling in each missing value with a single "replacement" value (as with LOCF and BOCF), more sophisticated methods of handling missing data exist that use statistical modelling to minimise bias. The multiple imputation (MI) method involves using all the data collected in all patients, whether they have complete data or some missing values, to model the distribution of the missing data. This model is then used to generate a series of values (this is the "multiple") to fill in each missing observation. An overall estimate of treatment effect is derived by combining all the results.

A different approach to handling missing data is to use a model for the analysis that can take account of all the available information from patients with complete data as well as those with some missing values. This makes it unnecessary to fill in the missing values with substitute values. Such an approach, called mixed-effects models for repeated measures (MMRM), is frequently used in clinical trials where the same continuous outcome variable is measured repeatedly at different time points. In effect, these analyses combine the information available for patients who have missing data with information from the patients who have complete data to predict what the responses of the patients with missing data would have been.

Suppose a patient showed a small improvement from baseline early in the trial then withdrew after 3 weeks, while most other patients in the same treatment group had larger improvements in the first 3 weeks and then continued to improve until the end of the study. In an MMRM analysis, the pattern seen in the data collected from the patient before withdrawal will feed into the overall estimate of treatment effect, as will all of the data collected from the other patients. So in this example, the model will assume that the withdrawn patient, like the other patients, would have continued to improve after Week 3, but – based on the data from the first 3 weeks – that this patient's improvement would have been smaller than average.

By comparison with single imputation methods like LOCF and BOCF, MI and MMRM have the clear advantage of using all the available

Table 2. Data from 5 patients in a study with the primary endpoint of change from baseline in HbA1c with missing data being filled in by LOCF

	Study start / Baseline	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5 / Study end	Comment
Patient 1	X_0	X_1	X_2	X_3	X_4	X_5	Completer
Patient 2	X_0	X_1	X_2 →	→	→	X_2	Withdrew at Visit 2
Patient 3	–	X_1	X_2	X_3	X_4	X_5	No baseline value
Patient 4	X_0	X_1	X_2	X_3 →	→	X_3	Died after Visit 3
Patient 5	X_0	–	–	–	X_4 →	X_4	Did not attend all visits

information for each patient (i.e. all of the values in Table 2 instead of just one value) to arrive at an estimate of treatment effect. Both methods have also been shown to produce much less biased estimates than LOCF.

Sensitivity analyses

There is no single best solution to the missing data problem that will produce unbiased results in all circumstances. As well as choosing a method that is appropriate to the particular situation, it is important to investigate the robustness of the results by carrying out sensitivity analyses. These should include analyses using missing data handling methods that rely on different assumptions from the method that was used in the primary analysis. For example, if MMRM is used for the primary analysis, the sensitivity analysis might include MI and sophisticated modelling techniques that do not make the same assumptions as MMRM about the nature of the missing data. If the results from the primary analysis and the various sensitivity analyses are similar, then we can be confident that the results are not being unduly influenced by the method used for handling missing data. If, on the other hand, the results differ substantially, then the issue of missing data needs further investigation and discussion.

What can be done to avoid missing data?

No amount of statistical expertise can make up for the absence of real data. – MG Kenward

Preventing missing data in the first place therefore needs to be a top priority. A number of measures can be taken at the trial protocol stage to help limit the quantity of missing data. Most importantly, trials need to be designed so that they interfere only minimally with the "normal life" of the study participants. That means study visits should be scheduled at convenient times and should not take too long. If it is possible to minimise the number of visits and assessments in the trial, this is likely to help retain patients. Likewise, generous visit windows make it easier for patients to fit study visits around other commitments. The longer the follow-up period, the more patients are likely to withdraw, so using a short follow-up period, at least for the primary endpoint, can help minimise the impact of missing data. Endpoints that are difficult or time-consuming to measure, or that require invasive procedures, tend to result in a high quantity of missing data. If endpoints can be chosen that are easy to measure, this is likely to reduce the amount of missing data.

As we have seen, missing data that arise due to adverse events or lack of efficacy are especially problematic because they tend to be associated with a particular treatment and therefore risk biasing the results of a study. Withdrawals due to tolerability issues can be minimised by allowing flexible dosing. Withdrawals due to lack of efficacy are a common problem when patients receive placebo, so using an add-on design, where patients receive active treatment or placebo in addition to standard treatments, can help to avoid withdrawals for this reason. Should a patient nevertheless need to discontinue study treatment, the sponsors should ask for permission to continue to collect data from them and plan the study so that discontinuation of treatment does not necessarily mean the patient has to withdraw from the study.

During trial conduct too, precautions can be taken to limit missing data. Engaging the participants by giving clear explanations of the study purpose and the procedures will most likely reduce the number of patients who withdraw from the study.

Realistically, it will never be possible to prevent missing data altogether. In order to ensure that data are collected from enough patients to enable valid conclusions to be drawn, it is important to consider the likely number of missing values when planning the trial and to allow for them when calculating how many patients to recruit.

Disclaimers

The opinions expressed in this article are the authors' own and are not necessarily shared by their employers or EMWA.

Conflicts of interest

The authors declare no conflicts of interest.

References

Committee for Medicinal Products for Human Use (CHMP). 2010. Guideline on missing data in confirmatory clinical trials. EMA/CPMP/EWP/1776/99 Rev. 1.

Kenward MG. The handling of missing data in clinical trials. Clin Invest. 2013;3(3):241–50.

Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355–60.

Mallinckrodt CH, Sanger TM, Dubé S, DeBrota DJ, Molenberghs G, Carroll RJ, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry. 2003; 53:754–60.

National Research Council. 2010. The prevention and treatment of missing data in clinical trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther. 2012;9,3:550–4.

Author information

Helen Bridge, DPhil, was previously a university lecturer in German language and literature. She then studied life sciences with statistics at the Open University, and became a regulatory medical writer in 2012. She worked in a CRO for 6 years before moving to AstraZeneca in March 2018.

Thomas M. Schindler, PhD (Molecular Physiology). After his post-doc, he went into publishing as a popular science editor and has now gained some 20 years of experience in both medical affairs and regulatory medical writing. He is the head of the European regulatory medical writing group at Boehringer Ingelheim Pharma.