Missing data: Issues, concepts, methods

Tra My Pham [a,*], Nikolaos Pandis [b], Ian R White [a]

[a] *MRC Clinical Trials Unit at UCL, 90 High Holborn, London WC1V 6LJ, UK*
[b] *Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine/Medical Faculty, University of Bern, Bern, Switzerland*

ARTICLE INFO

ABSTRACT

Missing data are a common issue in medical research. We aim to explain in non-technical language the issues and concepts around missing data, as well as discuss common methods for handling missing data. Specifically, our objectives are to answer the following questions: (1) What are missing data and why should we care about them? (2) What are the missingness mechanisms and how do they impact statistical analysis? (3) How can we explore missing values in our datasets? (4) What are ad-hoc methods for dealing with missing values and are they valid? (5) What is multiple imputation? (6) What should we consider when conducting a multiple imputation analysis? (7) Is multiple imputation always needed? (8) How should we report an analysis with missing data? We illustrate discussions with examples from an orthodontic study.

Introduction

Missing data remain a common issue in medical research, despite researchers' best efforts to prevent their occurrence through careful design and conduct of studies.[1,2] The literature on missing data is well-developed, with various discussions on different methods for handling missing data and their applications in both randomised trials and observational studies.[3−7]

Here we aim to explain in a non-technical manner key issues and concepts around missing data in biomedical research, and some common methods for handling missing data. This paper follows a similar structure to a series of articles recently published by the authors.[8−15] We will illustrate the discussions using different examples from an orthodontic study.

What are missing data and why should we care about them?

Missing data are data that we planned to collect to answer a research question, such as participant characteristics at the start of the study or their health outcomes after receiving some treatments, but for some reason we were not able to. In practice there are various ways in which missing data can arise. Table 1 describes an example created based on a randomised trial comparing probing depth (a sign of periodontal health) between 2 types of treatments (retainer A/B).[16] The outcome, mean probing depth over 6 lower teeth, was measured at baseline (pd1) and 5 follow-up time points (pd2−pd6). Data collected on other participant characteristics at baseline included age (years) and sex (female/male).

Sometimes, a mean probing depth value might be recorded, later judged to be wrong, and deleted from the dataset, which also gives rise

to missing data. Missing data can occur in the outcomes, one or more covariates, or both the outcomes and covariates.

A participant's probing depth could also be unobserved if the participant had lost their teeth or died during the study, after which their data no longer existed and would be considered 'truncated'.[17,18] Data truncation is conceptually different from data that are missing due to e.g. failure to attend a follow-up visit, and is not the focus of this paper.

Since missing data represent a loss of information, having missing values reduces statistical power of our study, which is the likelihood of a hypothesis test concluding an effect if there is one. The occurrence of missing data complicates statistical analysis, because we cannot perform the analysis originally intended for complete data without having to handle the missing values first. A direct consequence of this is that inappropriate handling of missing values can lead to bias and incorrect conclusions.[4]

What are the missingness mechanisms and how do they impact statistical analysis?

Any analysis that uses variables containing missing values (i.e. 'partially observed' or 'incomplete' variables) makes some *untestable* assumptions about how these values have become missing. These assumptions are called missingness mechanisms.[19]

The missingness mechanism describes how the likelihood of data being observed or missing is associated with the values of the variables included in our analysis. In our previous example, consider probing depth data at the end of the trial for 5 participants (pd6, Table 2). Conceptually, missingness in pd6 is represented by a binary indicator r6.

Table 1
An example of missing data in an orthodontic randomised trial.

ID	Age	Sex	Randomised treatment	pd1	pd2	pd3	pd4	pd5	pd6
1	37	Male	B	1.67	1.50	1.5	1.17	1.42	1.5
11	34	Female	A	1.08	0.83	1.08	.	.	.
32	36	Male	B	1.83	1.42	.	1.5	1.75	1.42
90	18	Female	A	1.58	.	.	.	.	.

Note: ID, participant identification; pd1−pd6, mean probing depth over 6 lower teeth measured at baseline and 5 follow-up time points, respectively; dot, missing value.

Table 2
Missing values and missingness indicators.

ID	pd1	pd6	r6	pd6*
107	1.17	0.92	1	0.92
108	2.25	0.83	1	0.83
109	1.58	.	0	**1.17**
110	1.42	.	0	**1.08**
111	1.25	0.75	1	0.75

Note: ID, participant identification; pd6, probing depth measured at time point 6; r6, indicator of whether pd6 is observed or missing; pd6*, an example of what pd6 could be had it been observed.

This variable takes value 1 if the participant's probing depth is observed and 0 if the participant's probing depth is missing. In case probing depth is missing (denoted by a dot in pd6), there exists an underlying measurement of pd6 (e.g. possibly the values in **bold** in pd6*) that we were not able to observe.

There are 3 broad missingness mechanisms, missing completely at random, missing at random, and missing not at random. Continuing with our previous example, suppose probing depth was measured for all participants at baseline (i.e. pd1 is fully observed), while some participants did not have their probing depth measured at the final time point (pd6 is partially observed, missingness is governed by indicator r6, such as in Table 2).

Here, the missingness mechanism describes how the likelihood of pd6 being observed or missing depends on pd1 as well as the (possibly missing) values of pd6. Directed acyclic graphs (DAGs) could be used to describe these relationships, with nodes representing the variables and arrows representing their relationships and directions of effects.[20]

If the likelihood of pd6 being observed is independent of both pd1 and the values of pd6 (i.e. no arrow from either pd1 or pd6 to r6, Fig. 1a), then pd6 is said to be missing completely at random (MCAR). The missing pd6 values are a random subset of and are fully comparable to the observed values.

If the likelihood of pd6 being observed depends on pd1 (i.e. arrow from pd1 to r6, Fig. 1b), but among participants with the same pd1

value, the likelihood of pd6 being observed is the same (i.e. no arrow from pd1 to r6, Fig. 1b), then pd6 is said to be missing at random (MAR) *conditional on pd1*. MAR means that given the same value of pd1, the missing values of pd6 are fully comparable with the observed values of pd6.

Finally, if even among participants with the same pd1 value, pd6 is more likely to be missing for e.g. participants with higher values of pd6 (i.e. arrows from both pd1 and pd6 to r6, Fig. 1c), then pd6 is considered missing not at random (MNAR). MNAR means that given the same value of pd1, the likelihood of observing pd6 may still vary with the values of pd6. This assumption means that missing pd6 values are not fully comparable with the observed values, even among participants with the same pd1 value.

As will be shown in the next section, assumptions about the missingness mechanism are untestable and often we cannot conclude whether missing values are MCAR, MAR or MNAR from just looking at the observed data. MCAR is the most restrictive and least likely to be plausible in medical research, since there is often some information related to the data collection that can be used to partly explain how the missingness occurs.[21] The least restrictive mechanism is MNAR, which is also the hardest to handle because in reality we will never truly know how the missingness may have arisen. In practice, MAR is often a good starting point for analysis, and standard implementation of methods such as multiple imputation (see 'What is multiple imputation?') is based on the MAR assumption.[22] In our example, the MAR mechanism for pd6 is described using a single variable pd1, but an incomplete variable can be MAR conditional on several variables. Therefore, the MAR assumption can be made more plausible by collecting data on variables that are associated with both the missingness as well as the values of the incomplete variable (see next section).

How can we explore missing values in our datasets?

Before doing any statistical analysis, the first step is to understand the extent of missingness (e.g. how much and in which variables) in our dataset, as well as whether some missingness mechanisms are more plausible than others. This is important because the consequence of incorrect handling of missing values (e.g. assuming the wrong missingness mechanisms) has a greater effect with more missing data.

Simple statistics such as percentages of missing values in relevant variables should be calculated. They are useful to assess the reliability of the data collection process, especially during the conduct of a study. We could also consider omitting from analysis some less important covariates with very high percentages of missing values.

The missingness patterns describe the location of the missing values (Table 3). These are categorised into univariate or multivariate, and the latter is categorised as monotone or non-monotone. When missingness occurs in a single variable, the missingness pattern is said to be univariate (Table 3a). When there are more than one incomplete variable in the dataset, the missingness pattern is said to be multivariate. If the incomplete variables could be arranged such that when a variable is missing for a participant then all subsequent variables are also missing for that participant, then the missingness pattern is monotone (Table 3b). Monotone missingness patterns often occur in studies where participants
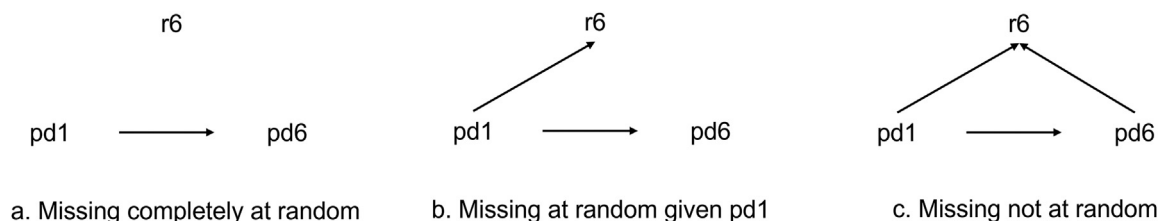

Fig. 1. Directed acyclic graphs of missingness mechanisms in probing depth at time point 6. Note: pd1 and pd6, probing depth measured at baseline and time point 6 (end of trial); r6, binary indicator of whether pd6 is observed or missing.

Table 3

An example of what different missingness patterns of probing depth measured at baseline and 5 follow-up time points might look like in a dataset.

Group of participants	pd1	pd2	pd3	pd4	pd5	pd6
	a. Univariate					
1	x	x	x	x	x	x
2	x	x	x	x	x	x
3	x	x	x	x	x	x
4	x	x	x	x	x	x
5	x	x	x	x	x	x
6	x	x	x	x	x	.

Group of participants	pd1	pd2	pd3	pd4	pd5	pd6
	b. Multivariate monotone					
1	x	x	x	x	x	x
2	x	x	x	x	x	x
3	x	x	x	x	x	.
4	x	x	x	x	.	.
5	x	x	x	.	.	.
6	x	x	.	.	.	.

Group of participants	pd1	pd2	pd3	pd4	pd5	pd6
	c. Multivariate non-monotone					
1	x	x	x	x	.	x
2	.	x	x	x	x	x
3	x	x	x	x	x	x
4	x	x	.	x	x	.
5	x	.	.	.	x	.
6	x	x	x	.	x	.

Note: pd1—pd6, probing depth measured at 6 time points; the dots and 'x's represent probing depth being missing and observed for the time points, respectively.

might attend all follow-up visits up to a point and then drop out, after which point all of their subsequent data become missing. When we cannot order variables according to their missingness in this way, e.g. some participants might attend all follow-up visits except the first one, while some participants do not have data at baseline and the last time point, then the missingness pattern is non-monotone (Table 3c). A tabulation of missingness patterns, or graphical tools such as UpSet plots (Fig. 2),[23]
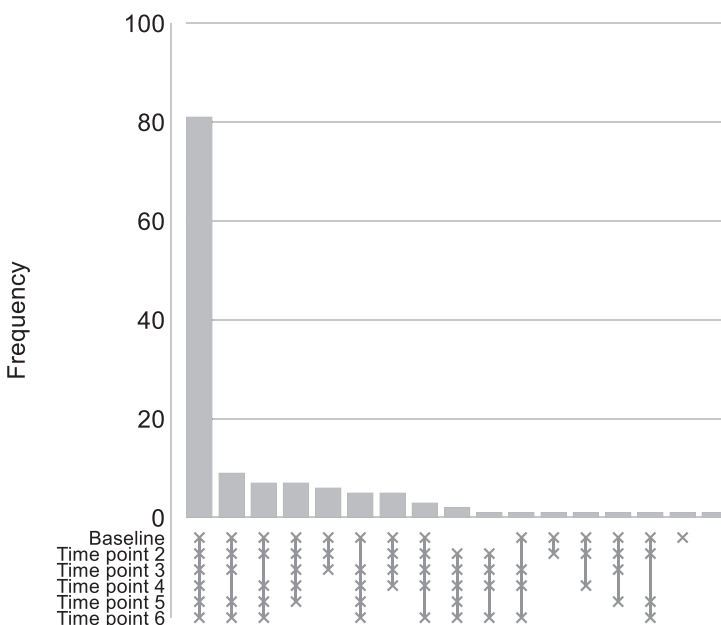
could be used to assess possible errors in the data collection and processing, e.g. when a variable is only observed if another variable is observed and exceeds a threshold.

Observed data can be used to assess, to some extent, the potential missingness mechanism. Suppose that in the probing depth randomised trial, we were able to measure everyone's probing depth at baseline, but probing depth at the end of the trial was missing for some participants. We could check whether probing depth at baseline was predictive of missingness in probing depth at the end of the trial, e.g. by fitting a logistic regression model with the missingness indicator of probing depth at time 6 (i.e. r6) as the dependent variable, and baseline probing depth (pd1) as the explanatory variable.[24] If there are other variables that might explain missingness in pd6, such as baseline age and sex, they could be included in this logistic regression model as additional explanatory variables. The choice of variables that might explain missingness depends on the context of each study and should be defined a-priori where possible (e.g. in the study protocol or statistical analysis plan). Identifying these variables may require input from clinical members of the research team as well as those involved in data collection.

If the estimated odds ratio and 95% confidence interval from fitting this logistic regression model provide evidence of an association, e.g. pd6 was more likely to be observed in participants with large baseline probing depth pd1, this indicates that our data are not likely to be MCAR, and that we need to adjust for pd1 in the analysis. From a study conduct point of view, this information could be used to improve completeness of probing depth data collection at time point 6.

The logistic regression model described above is a useful tool for checking whether our data are likely to be MCAR or not MCAR. If we have evidence against MCAR, we cannot proceed to distinguish between MAR and MNAR, as doing so would require e.g. fitting a logistic regression of r6 on pd6, which we cannot do because of the missing values in pd6. Therefore, as mentioned in section 'What are the missingness mechanisms and how do they impact statistical analysis?', we cannot verify whether the assumed MAR mechanism holds from the observed data alone. In some longitudinal studies where the outcome is measured over time (e.g. the probing depth trial), we could cross-tabulate missingness in the outcome measured at a given time point against the outcome measured at a previous time point. If participants with poor outcome tend to having missing values at the next time, this could suggest that their outcome might even be poorer over time, which is consistent with a MNAR mechanism.

Fig. 2. UpSet plot of an example multivariate non-monotone missingness pattern of probing depth measured at baseline and 5 follow-up time points for $N = 133$ participants. Note: the 'x' at each time point indicates that the participant had a probing depth measurement at that time point; the bars represent the frequency of participants with a particular pattern, e.g. the first bar indicates that $n = 81$ participants had a probing depth measurement at all time points including baseline; the second bar indicates that $n = 9$ participants had a probing depth at all time points except time point 4.

The plausibility of the MAR assumption can be improved by adjusting for variables that are predictive of both the missingness and the values of the incomplete variables.[4] In our example above, if probing depth at time point 6 was more likely to be missing for younger participants with low probing depth at baseline, then including age and baseline probing depth in the analysis will make the MAR assumption more plausible. Because of the untestable nature of these assumptions, it is important in any analysis with missing data to perform sensitivity analyses to assess the results under plausible alternative MNAR assumptions.[25−27]

What are ad-hoc methods for dealing with missing values and are they valid?

There are simple methods for handling missing values, such as creating a 'missing' category for categorical variables, replacing missing values of a continuous variable with a summary measure (e.g. mean) of the observed values, or replacing missing values with the last observed values in longitudinal studies (also known as last observation carried forward). These methods were proposed mainly for their computational convenience, and except in some specific situations[28] they are almost never valid. There have been extensive discussions on the limitations of these methods in the missing data literature.[21,29,30]

When an analysis is performed on a dataset that contains some missing values, the default option in most statistical software packages is to exclude from analysis cases with missing values in any of the variables of interest. This method is known as complete case or complete record analysis. The validity of complete case analysis might depend on whether data are missing in the outcome, or in the covariates.[31,32] It is important to state clearly the assumptions being made for missing data when a complete case analysis is performed.

In randomised trials, participant data can be missing in several ways. Participants may stop taking part in some or all aspects of the trial, after which point they may not provide further data on the outcome of interest. They may also be unable to attend some follow-up visits, resulting in missing outcome data for those visits. Some baseline data might occasionally not be collected, leading to missing values in the covariates. In observational studies, the same issues may occur, but missing values in the covariates occur more frequently.

It is well known that complete case analysis is valid when missing values are MCAR, such that the complete cases are a random subset of the whole sample. However, a complete case analysis is also valid in regression analyses where missingness does not depend on the outcome.[31,33]

In our previous probing depth example, suppose that the only variable containing missing values was probing depth at the end of the trial, pd6, and missingness in pd6 was explained by pd1, e.g. such that participants with low probing depth at baseline tended to be more likely to have missing probing depth at the end of the trial. Assume that we want to model the effect of treatment on probing depth at the end of the trial using a linear regression, i.e. pd6 is regressed on randomised treatment. When this analysis is performed among participants whose outcome was observed, the analysis is valid provided that the model adjusts for baseline probing depth, i.e. a linear regression of pd6 on randomised treatment and pd1. Adjusting for baseline probing depth would also be desirable to improve the precision of treatment effect estimation.[34]

Suppose instead that pd6 was observed for everyone, but some pd1 values were missing, and missingness in pd1 depended on its (possibly missing) values. Here, the linear regression of pd6 on randomised treatment and pd1 among the complete cases is unbiased, even though the assumption for missingness in baseline probing depth is consistent with a MNAR mechanism. In settings such as missing baseline data in a trial, this MNAR assumption might be more easily justified than a MAR assumption where baseline data (e.g. pd1) are missing conditional on the outcome (e.g. pd6) and the outcome is measured in the future. However, there are 2 issues with handling missing baseline covariates with a complete case analysis in randomised trials. First, the approach does not comply with the intention to treat principle that all randomised participants should be included in the analysis. Second, complete case analysis is inefficient because outcome data from participants who have missing baseline covariate values are excluded from analysis. This is a particular setting where using a simple method such as mean imputation for handling missing baseline covariates is appropriate and preferred to a complete case analysis.[28]

What is multiple imputation?

Multiple imputation (MI) is a popular approach for handling missing values in medical research. The basis of imputation is to replace missing values with some guesses. A simple imputation approach mentioned in the previous section is to replace missing values with the mean (or mode) of the observed data (mean or mode imputation) for a quantitative (or categorical) variable. Apart from some specific settings (see 'What are ad-hoc methods for dealing with missing values and are they valid?' and 'Is multiple imputation always needed'), simple imputation is generally poor, due to a few reasons.

The main drawback of simple imputation is that it fails to account for associations between variables in the dataset. For example, missing probing depth values at time point 6 should be imputed in a way that accounts for their correlation with other time points 1−5. Another downside of simple imputation is that there is no distinction between the imputed and observed data. For example, when imputing missing values in pd6, we need to acknowledge that there is uncertainty in predicting what the missing values could be. Mean imputation is a form of single imputation since each missing value is replaced with a single imputed value. When the analysis is performed, it does not distinguish between the values that have been imputed and the values that were observed, and imputed values are treated as actually observed values. Analyses following single imputation are therefore likely to be overconfident.

MI overcomes this issue by recognising that we do not know the true values of the missing data, and instead we provide some best guesses of what these values might have been with some uncertainty. This is done by creating several imputed datasets, each one representing a guess of what the complete data might have looked like.

While the implementation of MI is readily available in common statistical software packages, it is important to understand the steps involved in creating an imputed dataset using MI. Continuing with our probing depth example, suppose we want to use MI to impute missing values in pd6 based on pd1, and the imputation process is carried out separately in each randomised arm. Fig. 3 shows a scatterplot of pd6 (to be imputed) against pd1 (assumed complete). Each filled circle represents one observed data point. The solid lines are lines of best fit among the observed data, showing a positive association between baseline probing depth and probing depth measured at the end of the trial. The dashed lines are a perturbed version of the lines of best fit, representing a possibility of what the true relationship might look like. Each '+' represents a participant whose pd1 was observed and pd6 was missing. The hollow circles are values drawn from a distribution around the predicted pd6 for participants with missing pd6. For example, consider the participant who was randomised to receiving retainer A, whose pd1 was 3.5 and pd6 was missing. Their missing pd6 measurement has been imputed with a value above the line of best fit (2.42); however, in another imputed dataset this value could also be imputed to be below the predicted line.

This process is repeated several (M) times, resulting in M completed datasets. For example, if we create $M=5$ imputations of pd6, we could obtain the 5 completed datasets shown in Fig. 4, where the missing values in pd6 have been replaced with similar but not identical imputed values.

After M completed datasets have been created, the analysis planned for complete data is carried out in each completed dataset separately. For example, to compare probing depth between 2 treatments we could regress pd6 on randomised treatment (retainer), and adjust for baseline
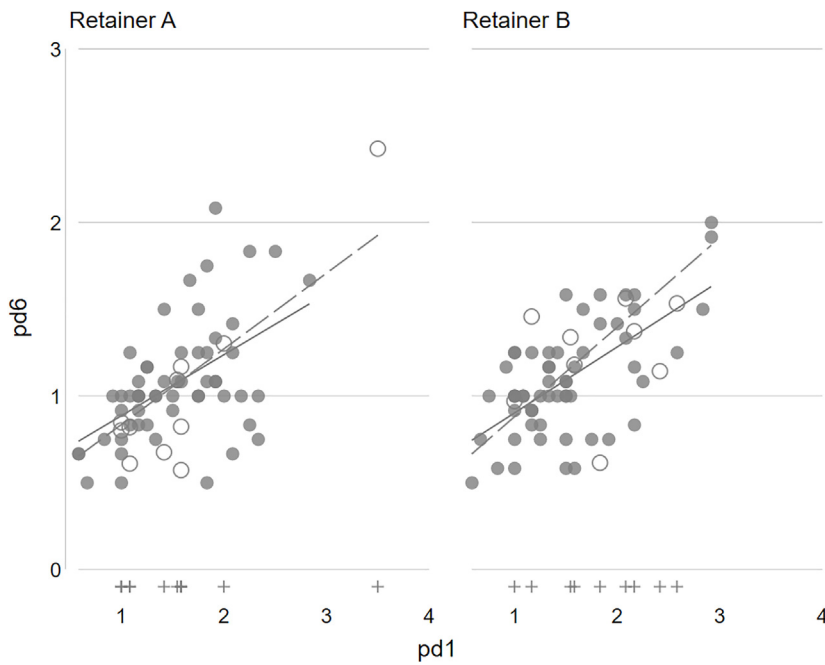
Fig. 3. An illustration of how an incomplete variable (pd6) may be imputed using a complete variable (pd1) separately in groups randomised to retainer A or retainer B.

factors such as age, sex and pd1. Doing this separately in each of M completed datasets will give us M estimates of the treatment effect together with M estimates of the uncertainty (Table 4). Finally, these M results are pooled together using a procedure called Rubin's rules.[22] First, the overall estimate of the treatment effect is simply an average of the M estimated treatment effects. Second, the overall measure of uncertainty associated with this treatment effect (e.g. standard error and confidence interval) is obtained by pooling together (i) the uncertainty in analysing each of the M completed datasets and (ii) the variability between the M completed datasets. The second source of variation is what is lacking in single imputation methods and represents our uncertainty about the true values of the missing data.

MI using Rubin's rules provides valid results (good estimates and valid confidence intervals) as long as the imputation procedure is carried out appropriately. Standard implementation of MI is based on the assumption of data being MAR. MI under MNAR is possible but often more complex and requires further (usually untestable) assumptions about the missing values.
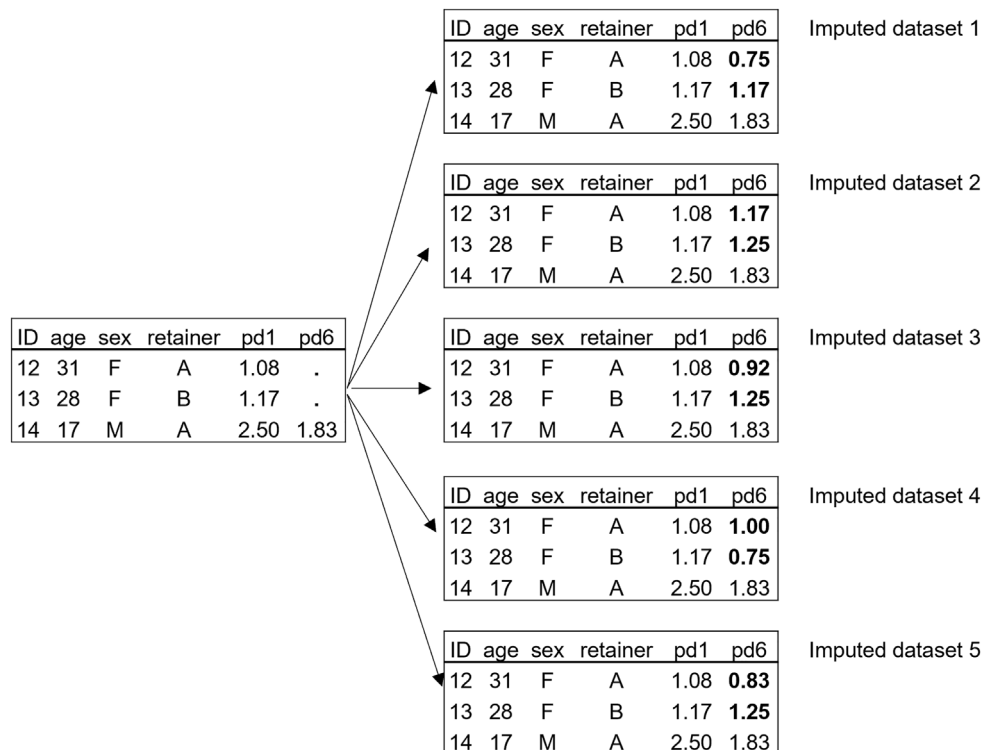

Fig. 4. An example of $M = 5$ imputations of pd6, where pd6 is imputed using age, sex, randomised retainer, and pd1.

Table 4
An example of combining coefficient and standard error estimates from 5 completed datasets using Rubin's rules.

Completed dataset	Estimated treatment effect on pd6	Standard error	Overall treatment effect on pd6	Overall standard error
1	0.047	0.048	0.040	0.050
2	0.029	0.048		
3	0.048	0.048		
4	0.020	0.046		
5	0.055	0.047		

Note: the coefficient and standard error estimates are obtained by fitting, separately in each of the 5 completed dataset, a linear regression model of pd6 on randomised treatment and baseline factors including age, sex, and pd1. The overall estimates are obtained using Rubin's rules.

What should we consider when conducting a multiple imputation analysis?

The previous section illustrated the principles of MI using an example of a single incomplete variable (univariate missingness). In practice, most datasets have missing data in several variables (multivariate missingness), and MI can also be used to impute more than one variable.[3] Multivariate missingness poses further difficulties and requires some considerations on how MI should be performed.

The most common MI approach when there are several incomplete variables is multivariate imputation by chained equations (MICE).[35] In MICE, we set up an imputation model for each variable to be imputed, based on the principles described in the previous section. A suitable regression model would typically be used as the imputation model. In our probing depth example, continuous variables like probing depth or age might be imputed using a linear regression model, while a binary variable like sex might be imputed using a logistic regression model. In general, quantitative variables could be imputed by either linear regression, as described in Fig. 3, or predictive mean matching where each missing value is imputed with the observed value of another participant with similar characteristics.[36] Categorical variables could be imputed with ordered or multinomial logistic regression models, depending on whether or not there is an order to the categories.

MI requires careful consideration of which variables to be included. It is essential that all variables present in the analysis model are included in the imputation models, regardless of whether they are complete or incomplete. By doing this, we ensure that all relationships present in the observed data are reflected in the imputed data.

Continuing with our probing depth example, suppose our analysis model is a linear regression of pd6 on randomised treatment, adjusting for baseline factors including age, sex, and pd1. Here we assume pd6, age, and pd1 are incomplete, while randomised treatment and sex are fully observed. The MICE procedure often starts by imputing arbitrary values, then each incomplete variable is imputed in turn using an appropriate imputation model, with new imputed values replacing the previously imputed ones (Fig. 5). Stable imputations are usually achieved after this process has completed a small number of cycles, typically less than 10.

In Fig. 5, we can see that pd6 is used to impute missing values in pd1 despite being measured later than pd1. This is important.[37] Using pd6 to predict the missing values in pd1 ensures that the relationship between probing depth measured at the 2 different time points are respected, and the imputations are plausible representations of what the missing values could have been.

In addition to variables that are in the analysis model, it is also useful to include in MI variables that are not in the analysis model but are (i) needed to make the MAR assumption more plausible, or (ii) predictive of values of the incomplete variables. Variables that fulfil the latter criterion are also known as 'auxiliary variables'. For example, in the probing
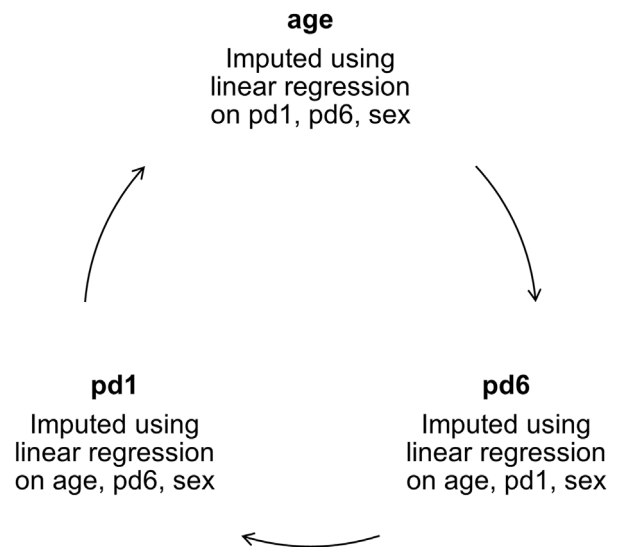

Fig. 5. An illustration of the MICE procedure for pd6, pd1, age, sex, and treatment group.Note: pd1, pd6, age are partially observed while sex and randomised treatment are fully observed; MI is performed in each treatment group separately.

depth example, we could use probing depth measurements at other time points (pd2−pd5) to inform imputations of pd1 and pd6, since repeated measurements of the outcome are often correlated. Selecting good candidate auxiliary variables and the number of auxiliary variables to be included also requires careful consideration (e.g. with regards to small study samples or complex analysis models); this issue has been discussed elsewhere.[38−40]

Above we emphasised that all variables present in the analysis model must be included in the imputation procedure to generate plausible imputations. When the analysis model contains structures such as interactions or non-linear effects, the imputation also needs to account for such structures, which could quickly complicate the MI procedure, especially in observational studies.[41] In randomised trials a straightforward way to account for any potential treatment-covariate interactions is to perform MI separately in each randomised arm. Similarly, in other datasets MI could be performed separately by a key variable, but this variable needs to be fully observed.

Another implementation issue is how many imputations are needed. We want to perform enough imputations so that we are confident that the results obtained from MI are unlikely to be different substantially if more imputations were created. A practical rule of thumb is to create as many imputations as the percentage of incomplete cases. For example, if 80% of participants have complete data in all variables considered in the analysis model, and the remaining 20% of participant have some missing values, then at least 20 imputations are needed in MI.

Common statistical software packages can also report the Monte Carlo errors of the imputation results. An example is given in Table 5. The analysis performed was a linear regression of pd6 on randomised treatment, adjusting for baseline factors including age, sex, and pd1. Missing values in pd1 were handled with mean imputation; missing values in pd6 were handled with MI using predictive mean matching to 5 nearest neighbours, conditional on age, sex, and (observed and mean imputed) pd1. MI of missing values in pd6 used $M = 20$ imputations. The upper confidence limit is estimated to be 0.138, which might in fact be as low as 0.131 and as high as 0.145 ($0.138 \pm 1.96 \times$ Monte Carlo error). In either case, the 95% confidence interval covers 0, indicating no evidence of a difference in probing depth between the 2 treatments.

A common question is: how much missing data can multiple imputation handle? In principle, multiple imputation can handle a large amount of missing values, but the multivariate imputation by chained

Table 5
An example of estimated treatment effect on probing depth (pd6, incomplete) from a linear regression model adjusted for baseline factors, using $M = 20$ imputations.

Variable	MI coefficient	Standard error	95% confidence interval	
Randomised treatment	0.036 [0.004]	0.051 [0.001]	−0.067 [0.006]	0.138 [0.004]

Note: MI coefficient and standard error were obtained using Rubin's rules; Monte Carlo errors in square brackets represent the likely error in the MI coefficient, standard error, and 95% confidence interval; baseline factors included age (complete), sex (complete), and pd1 (incomplete, mean imputed).

equations algorithm might converge slowly, and bias arising from incorrect specification of the imputation model (e.g. incorrectly diluting a relationship by omitting a variable in the analysis model) will be greater with more missing data. The choice of whether to perform multiple imputation given the extent of missing data is also context dependent. For example, in a study involving a very rare outcome, the analyst might choose to perform multiple imputation over a complete case analysis even if there is only a very small percentage of missing values. This should also be accompanied by relevant sensitivity analyses.

Is multiple imputation always needed?

Although MI is a popular method for handling missing data, we have illustrated in previous sections that doing MI well requires a lot of careful considerations. In addition, when faced with the issue of missing data, we should always ask whether MI is needed.

It is useful to consider whether simpler alternatives are valid given our assumed missingness mechanism. As shown in section 'What are ad-hoc methods for dealing with missing values and are they valid?', in several situations including missing outcome data in randomised trials, a complete case analysis may be valid. The method is also valid under certain MNAR assumptions (e.g. missingness in baseline covariates not dependent on outcome) that are arguably more plausible than a MAR assumption. In addition, in randomised trials, for missing values in baseline covariates, simple methods such as mean imputation or missing indicator can be used.[28]

When analysis involves repeated measurements of an outcome which are analysed using a linear mixed model, then missing data in the repeated outcome are implicitly 'handled' under the assumption of data being MAR. Here MI is only preferred to a mixed model when e.g. there are auxiliary variables that could be used to improve the imputation.

Since standard implementation of MI is based on the MAR assumption, which is untestable, it is important to conduct sensitivity analyses to explore alternative plausible MNAR assumptions. Such sensitivity analyses could be based on some information about the incomplete variable that is available externally,[42] or information provided by experts about the potential difference in outcome between participants who remained in the study versus those who were lost to follow-up.[43,44] In randomised trials, plausible MNAR assumptions about the missing data could also be constructed using information internal to the trial, e.g. observed information in one randomised arm could be used to inform missing values in the other arm.[25]

How should we report an analysis with missing data?

As with other aspects of reporting, analysis with missing data should be reported clearly and with sufficient details to ensure transparency. There are extensive guidelines and checklists for randomised trials and observational studies.[45,46]

Here we provide an example of an analysis of data from the probing depth trial ($N = 133$), together with how results could be reported.

The aim of this analysis was to compare the effect of 2 treatments on mean probing depth over 6 lower teeth measured at the end of the trial. The analysis model intended for complete data was a linear regression

Table 6
An example of how an analysis of the probing depth data might be reported.

In the Methods section

Missing data in probing depth measured at baseline were handled by mean imputation. Missing data in probing depth measured at 5 follow-up time points were handled by multivariate imputation by chained equations. These variables were imputed by predictive mean matching to 5 nearest neighbours. All imputation model included as covariates baseline factors (age, sex, and baseline probing depth). Imputation was performed separately by randomised treatment. M = 50 imputed datasets were created. Analyses of imputed data used Rubin's rules. This analysis assumes that data were missing at random.

In the Results section

Missing values occurred in probing depth measured at baseline and 5 follow-up time points. There were 81 (61%) participants with observed probing depth at all time points, and 111 (84%) with observed probing depth at both baseline and the end of the trial. Probing depth was missing more frequently in the group randomised to retainer A at later time points [*present a table of missingness pattern by randomised treatment*].

Estimated treatment effects from a complete case analysis and a MI analysis are shown in **Table A**. In both analyses, there was no evidence of a difference in probing depth between 2 treatments at the end of the trial.

Table A. Estimated treatment effect from a complete case analysis and a MI analysis.

Method	Estimated treatment effect	Standard error	95% confidence interval	
Complete case analysis (N = 111)	0.036	0.052	−0.067	0.138
MI (N = 133)	0.027 [<0.001]	0.049 [<0.001]	−0.071 [<0.001]	0.125 [<0.001]

Note: Monte Carlo errors are presented in square brackets for MI estimates.

In the Discussion section

The MI analysis was performed under the assumption of data being missing at random. Further sensitivity analyses could be conducted to explore alternative assumptions to missing at random, for example, where mean probing depth of participants who were lost to follow-up is assumed to be *x* points lower or higher than mean probing depth of those who remained in the trial.

of probing depth at time point 6 on randomised treatment, adjusting for baseline factors including age, sex, and baseline probing depth. Probing depth was measured at baseline and 5 follow-up time points, and there were missing values at all time points. Other variables were fully observed. We performed MI of missing probing depth values at all time points using MICE. This analysis, which improves slightly on that shown in Table 5, is reported in Table 6.

Funding

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tra My Pham reports financial support was provided by UKRI Medical

CRediT authorship contribution statement

Tra My Pham: Writing − original draft, Methodology, Formal analysis, Conceptualization. **Nikolaos Pandis:** Writing − review & editing, Data curation, Conceptualization. **Ian R White:** Writing − review & editing, Validation, Methodology, Conceptualization.

References

1. Hussain J, Peters TJ, Johnson M, Byrne A, Preston N, Haines A, et al. Missing data in palliative and end of life care trials. Guidance on how to reduce, handle and report incomplete data [Internet]. 2022. Available from: https://www.mariecurie.org.uk/globalassets/media/documents/research/Missing-Guidance-Report-final-June-2022.pdf.
2. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Eng J Med*. 2012;367(14):1355–1360. https://doi.org/10.1056/nejmsr1203730.
3. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–399. https://doi.org/10.1002/sim.4067.
4. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. 1st ed. John Wiley & Sons, Ltd.; 2013.
5. van Buuren S. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC; 2012.
6. Quartagno M, Grund S, Carpenter J. jomo: a flexible package for two-level joint modelling multiple imputation. *R Journal*. 2019;11(2):205–228. https://doi.org/10.32614/rj-2019-028.
7. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. https://doi.org/10.1136/bmj.b2393.
8. Pham TM, Pandis N, White IR. Missing data, part 1. Why missing data are a problem. *Am J Orthod Dentofacial Orthop*. 2022;161(6):888–889. https://doi.org/10.1016/j.ajodo.2022.02.006.
9. Pham TM, Pandis N, White IR. Missing data, part 2. Missing data mechanisms: missing completely at random, missing at random, missing not at random, and why they matter. *Am J Orthod Dentofacial Orthop*. 2022;162:138–139. https://doi.org/10.1016/j.ajodo.2022.04.001.
10. Pham TM, Pandis N, White IR. Missing data, part 3. How to explore missing data. *Am J Orthod Dentofacial Orthop*. 2022;162(2):282–283. https://doi.org/10.1016/j.ajodo.2022.05.002.
11. Pham TM, Pandis N, White IR. Missing data, part 4. Missing data in the outcomes vs in the covariates. *Am J Orthod Dentofacial Orthop*. 2022;162(3):432–433. https://doi.org/10.1016/j.ajodo.2022.06.001.
12. White IR, Pandis N, Pham TM. Missing data, part 5. Introduction to multiple imputation. *Am J Orthod Dentofacial Orthop*. 2022;162(4):581–583. https://doi.org/10.1016/j.ajodo.2022.07.004.
13. White IR, Pandis N, Pham TM. Missing data, part 6. Choices in doing multiple imputation. *Am J Orthod Dentofacial Orthop*. 2022;162(5):795–797. https://doi.org/10.1016/j.ajodo.2022.08.004.
14. White IR, Pandis N, Pham TM. Missing data, part 7. Pitfalls in doing multiple imputation. *Am J Orthod Dentofacial Orthop*. 2022;162(5):975–977. https://doi.org/10.1016/j.ajodo.2022.08.013.
15. White IR, Pandis N, Pham TM. Missing data, part 8. Reporting analyses with missing data. *Am J Orthod Dentofacial Orthop*. 2023;163:139–140. https://doi.org/10.1016/j.ajodo.2022.10.001.
16. Węgrodzka E, Kornatowska K, Pandis N, Fudalej PS. 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 Dentofacial Orthop*. 2021;160(4):494–502. https://doi.org/10.1016/j.ajodo.2021.02.015.e1..
17. Colantuoni E, Scharfstein DO, Wang C, et al. Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*. 2018;360:j5748. https://doi.org/10.1136/bmj.j5748.
18. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH harmonised guideline addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials ICH E9(R1); 2019.
19. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592. https://doi.org/10.1093/biomet/63.3.581.
20. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
21. Carpenter JR, Kenward MG. *Missing Data in Randomised Controlled Trials − A Practical Guide*. Birmingham: National Institute for Health Research; 2008... Publication RM03/JH17/MK; http://www.missingdata.org.uk.
22. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley; 1987.
23. Morris TP, Cox NJ. UPSETPLOT: Stata module to show Euler or Venn diagrams mapped to bar charts, upsetplot style. Published online 2023. https://ideas.repec.org/c/boc/bocode/s459161.html.
24. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Stat Med*. 2002;21(8):1043–1066. https://doi.org/10.1002/sim.1065.
25. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Stat Med*. 2020;39(21):2815–2842. https://doi.org/10.1002/sim.8569.
26. Smuk M, Carpenter JR, Morris TP. What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry. *BMC Med Res Methodol*. 2017;17(1):21. https://doi.org/10.1186/s12874-017-0301-0.
27. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. *BMC Med Res Methodol*. 2014;14:11. https://doi.org/10.1186/1471-2288-14-11.
28. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med*. 2005;24:993–1007. https://doi.org/10.1002/sim.1981.
29. Siddiqui O, Hung HMJ, O'Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. *J Biopharm Stat*. 2009;19(2):227–246. https://doi.org/10.1080/10543400802609797.
30. White I, Carpenter J, Horton N. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials*. 2012;9(4):396–407. https://doi.org/10.1177/1740774512450098.
31. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. *Biom J*. 2021;63(5):915–947. https://doi.org/10.1002/bimj.202000196.
32. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Am J Epidemiol*. 2015;182(8):730–736. https://doi.org/10.1093/aje/kwv114.
33. Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics*. 2014;15(4):719–730. https://doi.org/10.1093/biostatistics/kxu023.
34. Morris TP, Walker AS, Williamson EJ, White IR. Planning a method for covariate adjustment in individually randomised trials: a practical guide. *Trials*. 2022;23(1):1–17. https://doi.org/10.1186/s13063-022-06097-z.
35. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18:681–694.
36. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14(1):1–13. https://doi.org/10.1186/1471-2288-14-75.
37. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59:1092–1101. https://doi.org/10.1016/j.jclinepi.2006.01.009.
38. Thoemmes F, Rose N. A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behav Res*. 2014;49(5):443–459. https://doi.org/10.1080/00273171.2014.931799.
39. Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med Res Methodol*. 2012;12(184). https://doi.org/10.1186/1471-2288-12-184.
40. Curnow E, Cornish RP, Heron JE, Carpenter JR, Tilling K. Multiple imputation assuming missing at random: auxiliary imputation variables that only predict missingness can increase bias due to data missing not at random. *medRxiv*. 2023. https://doi.org/10.1101/2023.10.17.23297137.
41. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462–487. https://doi.org/10.1177/0962280214521348.
42. Pham TM, Carpenter JR, Morris TP, Wood AM, Petersen I. Population-calibrated multiple imputation for a binary/categorical covariate in categorical regression models. *Stat Med*. 2019;38(5):792–808. https://doi.org/10.1002/sim.8004.
43. Mason AJ, Grieve RD, Richards-Belle A, Mouncey PR, Harrison DA, Carpenter JR. A framework for extending trial design to facilitate missing data sensitivity analyses. *BMC Med Res Methodol*. 2020;20(1):1–12. https://doi.org/10.1186/s12874-020-00930-2.
44. Mason AJ, Gomes M, Grieve R, Ulug P, Powell JT, Carpenter J. Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: application to the IMPROVE trial. *Clin Trials*. 2017;14(4):357–367. https://doi.org/10.1177/1740774517711442.
45. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ (Online)*. 2010;340(7748):698–702. https://doi.org/10.1136/bmj.c332.
46. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806–808. https://doi.org/10.1136/bmj.39335.541782.AD.