Missing data 8. Reporting analyses with missing data

Ian R White, Nikolaos Pandis, Tra My Pham

We end this series by discussing how to report one's missing data in a research paper. As with other aspects of reporting, the key is to be transparent about what one has done. Relevant checklists include CONSORT for randomised trials and STROBE for observational studies.[1,2] We give a short discussion; a detailed framework for observational studies is given by STRATOS.[3]

Following all the steps described below is more complicated than simply excluding any records with missing values, but it has three key advantages: it acknowledges the problem of missing data, it should alleviate any bias due to missing data, and it avoids allowing the missing data to lead to an inefficient analysis.

Methods section

Researchers should describe, in the methods section, how missing data were handled. This should be done in enough detail for the analysis to be reproduced. Multiple imputation (MI) analysis involves a number of choices (article 6) ***[add ref at proof stage – adjust the refs section accordingly]***, so details of MI procedures may be best reported in supplementary materials.

Researchers should also say what the assumptions were for the methods used, and why they were reasonable. This is often not done. In the case of a complete cases analysis, a suitable justification could be that the complete cases were judged to be representative of all cases, and any loss of precision was unlikely to affect conclusions. For a MI analysis, the justification should focus on why the missing at random (MAR) assumption was thought to be plausible.

Results section: descriptive analyses

Researchers should describe the amount of missing data in key variables. The CONSORT diagram does this for main trial outcomes.

It is also useful to report key comparisons (see article 3 ***[add ref at proof stage – adjust the refs section accordingly]***) such as comparisons of one key variable between those with observed and missing values for another key variable. If, for example, individuals with missing outcome had worse dental health at baseline than those with observed outcome, this may suggest they also had worse dental health at follow-up times, which suggests a missing not at random mechanism (MNAR, article 2 ***[add ref at proof stage – adjust the refs section accordingly]***).

In a randomised trial, it is useful to report a comparison of baseline variables between randomised groups, restricted to individuals with observed outcomes. This assesses to what extent missing outcome data may have destroyed the baseline balance between groups that was achieved by randomisation.

Where MI is used, comparisons of imputed with observed values (article 7 ***[add ref at proof stage – adjust the refs section accordingly]***) should be done by the analysts, but are not usually reported.

Results section: main analyses

It is always helpful to report the number of individuals included in an analysis. Where MI has been used, this is typically the whole sample. It is usually helpful to report a complete cases analysis alongside the MI analysis, as a check for serious failures of the imputation model.[4] For example, if an MI analysis were substantially less precise than the complete case analysis, or if the results were very different from each other, then we would question whether the MI had been performed suitably.

Discussion section

A key issue in the discussion section is whether any further bias could arise from the missing data. For example, in a randomised trial with 30% missing outcome data, it is important to consider whether plausible systematic differences between missing and observed data (i.e. departures from MAR) could have introduced bias. Reasons for missing data, if collected, should be used to inform this discussion. The existence of bias may be tackled qualitatively, but a formal sensitivity analysis is preferred.

Example

The Box gives an example of how missing data might be reported in the example described in articles 5–7 ***[add refs at proof stage – adjust the refs section accordingly]***. The example was created using data from a cohort study conducted to assess whether gingival recession is more likely in individuals who had orthodontic treatment compared to those without orthodontic treatment.[5] The results of the MI analysis would be as shown in article 7 of this series ***[add ref at proof stage – adjust the refs section accordingly]***. We have presented our artificial deletion of data as an administrative error, though in practice to lose data on gender and sex might be seen as poor research standards that would cast doubt on other aspects of the study quality.

Box. An example of how missing data might be reported in the recession data

In Methods section

Missing data in age and gender were handled by multivariate imputation by chained equations. These variables were imputed by logistic and linear regression respectively. Recession score was included in all imputation models, and imputation was done separately by treatment group. 34 imputed data sets were created. Analyses of imputed data used Rubin's rules. This analysis assumes that the data are missing at random, which is reasonable since data were missing due to administrative oversights.

In Results section

Age and gender each had 38 (20%) missing values. Missing values were similarly distributed across treatment groups, and recession score was similarly distributed across those with missing and observed values (results not shown).

Regression of recession score on treatment group, age and gender gave the results shown (see article 7).

In Discussion section

A possible limitation of this study is the occurrence of 20% missing data on age and gender. We believe that failure to collect these variables was due to an administrative error which, while undesirable, is unlikely to relate to any characteristics of the individuals, and is therefore unlikely to introduce any bias into our analysis which assumed the data were missing at random.

References

1. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340: 698–702.
2. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ* 2007; 335: 806--808.
3. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol* 2021; 134: 79–88.
4. 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.
5. Gebistorf M, Mijuskovic M, Pandis N, et al. Gingival recession in orthodontic patients 10 to 15 years posttreatment: A retrospective cohort study. *Am J Orthod Dentofac Orthop* 2018; 153: 645–655.