BMJ Open

Handling of missing data with multiple imputation in observational studies that address causal questions: protocol for a scoping review

Rheanna Mainzer ,[1,2] Margarita Moreno-Betancur,[1,2] Cattram Nguyen ,[1,2] Julie Simpson,[3] John Carlin,[1,3] Katherine Lee[1,2]

To cite: Mainzer R, Moreno-Betancur M, Nguyen C, *et al*. Handling of missing data with multiple imputation in observational studies that address causal questions: protocol for a scoping review. *BMJ Open* 2023;**13**:e065576. doi:10.1136/bmjopen-2022-065576

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-065576).

Received 10 June 2022
Accepted 19 January 2023

[1]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia
[2]Department of Paediatrics, The University of Melbourne, Parkville, Victoria, Australia
[3]School of Population and Global Health, University of Melbourne, Parkville, Victoria, Australia

Correspondence to
Dr Rheanna Mainzer;
rheanna.mainzer@mcri.edu.au

ABSTRACT

Introduction Observational studies in health-related research often aim to answer causal questions. Missing data are common in these studies and often occur in multiple variables, such as the exposure, outcome and/or variables used to control for confounding. The standard classification of missing data as missing completely at random, missing at random (MAR) or missing not at random does not allow for a clear assessment of missingness assumptions when missingness arises in more than one variable. This presents challenges for selecting an analytic approach and determining when a sensitivity analysis under plausible alternative missing data assumptions is required. This is particularly pertinent with multiple imputation (MI), which is often justified by assuming data are MAR. The objective of this scoping review is to examine the use of MI in observational studies that address causal questions, with a focus on if and how (a) missingness assumptions are expressed and assessed, (b) missingness assumptions are used to justify the choice of a complete case analysis and/or MI for handling missing data and (c) sensitivity analyses under alternative plausible assumptions about the missingness mechanism are conducted.

Methods and analysis We will review observational studies that aim to answer causal questions and use MI, published between January 2019 and December 2021 in five top general epidemiology journals. Studies will be identified using a full text search for the term 'multiple imputation' and then assessed for eligibility. Information extracted will include details about the study characteristics, missing data, missingness assumptions and MI implementation. Data will be summarised using descriptive statistics.

Ethics and dissemination Ethics approval is not required for this review because data will be collected only from published studies. The results will be disseminated through a peer reviewed publication and conference presentations.

Trial registration number This protocol is registered on figshare (https://doi.org/10.6084/m9.figshare.20010497.v1).

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ A targeted review of observational studies published in the five top-ranked epidemiology journals will benchmark the current state of practice for handling multivariable missingness with multiple imputation in causal analyses.
⇒ Screening, reviewing and data extraction will be performed systematically, with double data extraction for a subset of articles and any discrepancies resolved by a panel.
⇒ It is likely that some of the information sought will be ambiguously reported or not reported.
⇒ Potential challenges with data extraction have been considered and a strategy for handling these challenges has been put in place.
⇒ All extracted data and code will be made publicly available, enabling our descriptive analysis to be entirely reproducible.

INTRODUCTION

Observational studies in clinical and health-related research often aim to answer causal questions, even if this intent is only implicit.[1 2] This aim is usually addressed by estimation of a target parameter to quantify the impact of intervening on an exposure on an outcome of interest, in a given population. In observational studies, missing data are common and can occur in multiple variables, such as the exposure, the outcome and/or the variables used to control for confounding. Restricting the statistical analysis to individuals with complete data on all analysis variables, that is, conducting a 'complete case analysis' (CCA), can lead to bias and/or loss of precision in estimates of the target parameter.[3] Multiple imputation (MI) is a popular and flexible approach for estimating a target parameter in the presence of incomplete data.[4 5] In the first stage of MI, missing data are imputed multiple times with random draws from the predictive distribution of the missing values given the observed data and a specified imputation model. In the second stage, the statistical analysis of interest is applied to each imputed dataset and the results are

combined using Rubin's rules to obtain a single estimate of the target parameter with associated standard error.[4]

Standard implementations of MI are known to provide consistent estimation of target parameters under certain (unverifiable) assumptions about the mechanism leading to missing data. Assumptions about missing data are usually expressed using Rubin's classification of missing data mechanisms into missing completely at random (MCAR, where the probability of data being missing does not depend on the observed or unobserved data), missing at random (MAR, where the probability of data being missing does not depend on the unobserved data, conditional on the observed data) and missing not at random (MNAR, where the probability of data being missing depends on the unobserved data, even after conditioning on the observed data).[6] While this framework is useful if missing data occur in a single variable, it raises issues when missingness arises in more than one variable. First, what these mechanisms mean with multivariable missingness is poorly understood and does not allow for a transparent assessment of missingness assumptions.[7] Second, based on our experience researching, teaching and applying MI, these mechanisms have become widely (mis)understood as synonymous with methods. For example, researchers often use MI under the assumption that data are MAR, but this is only a sufficient and not necessary condition for standard MI to be consistent.[8] Both a CCA and an MI analysis could be unbiased under a range of multivariable missingness mechanisms (even those considered to be MNAR).[9] Likewise, there are missingness mechanisms in which neither MI nor a CCA can be used to estimate an exposure–outcome association without bias, and a different approach would be needed for unbiased estimation.

The primary analysis in a study would ideally be conducted under the missing data assumptions that the researcher believes to be most likely. However, because one cannot verify from the observed data what the true missing data mechanism is, sensitivity analyses to examine how results differ under other plausible assumptions about the missingness mechanism (hereafter, 'sensitivity analyses') are strongly recommended.[10] Such an analysis could be carried out by estimating the target parameter under the other mechanism(s) that the researcher has identified as likely. As stated by the US National Research Council, 'the usefulness of a sensitivity analysis ultimately depends on the transparency and plausibility of the unverifiable assumptions'.[10] The inherent difficulty in assessing missingness assumptions when framed in the traditional MCAR/MAR/MNAR manner is an obvious obstacle to conducting sensitivity analyses. Furthermore, from our observation, MI is routinely applied as a sensitivity analysis to a CCA. However, this practice is flawed without considering one's plausible assumptions regarding the missingness mechanism,[11] as neither of these approaches may be valid under particular assumptions regarding the missingness mechanism. If this is the case, obtaining similar results from a CCA and MI is not informative.

Most reviews of the handling and reporting of missing data, and the implementation and documentation of MI, have been carried out in the context of randomised controlled trials (RCTs).[12–18] For trials, typically only the outcome variable is incomplete, while the intervention and other key variables (typically baseline variables) are observed for all participants. In this setting where there are missing data in a single variable, the MCAR/MAR/MNAR framework is more transparent and guidance on sensitivity analyses has been well developed[15 19]. In contrast, there have been few reviews concerned with how missing data are handled in observational studies where there is the additional complication of multivariable missingness. A review by Mackinnon published in 2010 found that only 2 (4%) out of 50 non-RCT studies reviewed carried out an additional analysis that was described as a sensitivity analysis.[11] Similarly, Rezvan *et al* found that none of the 30 observational studies reviewed conducted a sensitivity analysis to departures from the missingness assumptions following MI.[20]

While the reviews by Mackinnon[11] and Rezvan *et al*[20] provide useful insight into the problem, neither focused specifically on observational studies and the issues described above. In addition, subsequent to publication of these reviews, there have been important developments in the theory and application of missingness directed acyclic graphs (m-DAGs), also known as m-graphs, a tool for the formulation of causal assumptions in the presence of multivariable missingness.[8] M-DAGs aid the depiction and assessment of missingness assumptions. Clarity regarding each plausible causal mechanism underlying the missing data then facilitates the choice of analytical approach. For example, the application of DAG theory allows one to determine whether a target parameter can be estimated without bias from the available data using an approach like CCA or MI, or whether additional assumptions and a more sophisticated analysis is required (such as a delta-adjusted MI approach, where imputations are shifted by a parameter 'delta' representing the difference between the observed and unobserved data).[9 21–23]

The aim of this scoping review is to examine the use of MI in observational studies that address causal questions relating to health. Addressing causal questions is typically the focus of epidemiological studies even when this may not be very clearly articulated.[2] These studies often face missingness in multiple variables required for analysis. We will examine (1) how missingness assumptions are expressed, (2) if and how missingness assumptions are used to justify the choice of a CCA and/or MI for handling missing data and (3) the conduct of sensitivity analyses under alternative plausible assumptions about the missingness mechanism. We will also examine how MI is implemented. This review will be used to document the current state of practice, to identify areas for improvement in the handling and reporting of missing data with MI in observational studies, and to subsequently develop guidance on these key components for researchers.

METHODS AND ANALYSIS

In this section, we provide a full description of the study design, including how articles will be selected, what information will be extracted and how extracted data will be analysed. The review described in this protocol began in June 2022 and we anticipate it will be completed by June 2023.

Search strategy

We will search five general epidemiology journals for observational studies published between January 2019 and December 2021 that aim to answer at least one causal research question using MI. The general epidemiology journals that will be included in this search are: *International Journal of Epidemiology*, *American Journal of Epidemiology*, *European Journal of Epidemiology*, *Journal of Clinical Epidemiology and Epidemiology*. These journals were chosen because they are high ranking, general journals in epidemiology that publish original research from observational studies. As such, articles from these journals should capture the current best practice in the use of MI to handle missing data when answering causal questions using observational data. They have also been used previously in a review of epidemiologic practice.[24] Original research articles will be identified using the full-text search term 'multiple imputation' on each journal's website. This search strategy is similar to that used in previous scoping reviews in this area.[11 20]

Inclusion criteria

We will include original research articles published between January 2019 and December 2021 that aim to answer at least one causal question using MI to handle missing data. We will determine that a study has aimed to answer a causal question if at least one of the following criteria is satisfied:

1. The authors explicitly stated they were estimating a causal effect.
2. The study estimated an effect that was given (at least implicitly) a causal interpretation, that is, an interpretation which suggested that intervening on the exposure could change the outcome (eg, increasing coffee consumption may be protective against stroke). This will be determined by wording in conclusions. If it is not clear from this wording alone, investigation of the following three typical signals of causal analyses will be used to aid in the determining: identification of confounders, the inclusion of a DAG to illustrate causal assumption made in the analysis, and analytical approaches incorporating adjustment for confounders (eg, estimating an effect using a regression model that was adjusted for a set of covariates).

Studies on all disease areas/medical conditions and any target population will be considered.

Exclusion criteria

Studies will be excluded from the review if they meet any of the following criteria:

- No causal question. The article did not aim to answer a causal question, for example, the aim of the study was to develop a predictive model or to estimate a disease burden.
- Unclear type of question. A clear research goal could not be identified. In other words, it was unclear whether the study aimed to answer a descriptive, predictive or causal question.
- The analysis did not use MI.
- Methodological research. The primary purpose of the article was methodological development, for example, using a simulation study to compare the performance of methods or mathematical derivations to develop a new method or model. While these articles often include comprehensive case studies, they may not be representative of empirical studies aiming primarily to answer causal research questions.
- Aggregate-level data. The analysis was based on aggregated data where MI could not be applied at the participant level, as is common in meta-analysis or interrupted time series analysis.
- Qualitative research. The article provided a commentary, review, opinion, study protocol, study profile or description only.
- Trial. The study intervention was assigned to participants by the study investigators.

Sample size

We will require at least 100 studies to estimate the percentage of studies with a particular element (eg, studies that justify their missingness assumptions) to within a maximum margin of error (two standard errors) of 10%. Assuming a prevalence of 50%, this would give a 95% CI from 40% to 60%. For a prevalence greater than or less than 50%, the 95% CI will be narrower. This sample size is similar to the sample size used in the first review of MI in medical research (n=99[11]), and many of the subsequent reviews in this area (eg, n=103,[20] 77[15] and 118[12]). We expect to identify at least 100 eligible studies given the 3-year publication time frame. All eligible studies will be included in the review.

Study selection

The search of the journal databases and selection of studies for inclusion in the review will be performed primarily by a single researcher (RM) in two steps. First, the title, abstract and date of each article will be screened to rule out studies that are clearly not eligible for the review. Second, the full text of the remaining studies will be reviewed to confirm if studies are eligible for the review. If a decision about the eligibility of an article cannot be reached by RM (eg, due to uncertainty about the inclusion criteria), a second researcher (CN) will independently review the full text. Disagreements about inclusion criteria will be resolved by discussion in meetings with at least three researchers (RM, CN and at least one of JC, JS, KL or MM-B).

Table 1 Summary of items to be extracted from each article

Category	Summary of data extraction items
Study characteristics	▶ First author's last name ▶ Publication date ▶ Journal ▶ Type of study design
Missing data	▶ Percentage of complete cases ▶ Percentage of missing values in the exposure and outcome ▶ Number of incomplete covariates
Missingness assumptions	▶ Statement of missingness data assumptions (including whether the study used m-DAGs or the MCAR/MAR/MNAR framework) ▶ Justification of missingness assumptions
Analysis methods	▶ The primary analysis method used to answer the key causal question, for example, MI or CCA ▶ Whether the primary analysis was justified on the basis of missingness assumptions ▶ If applicable, any other analyses conducted to answer the key causal question that handle the missing data differently (eg, a CCA or a delta-adjusted MI analysis) ▶ Whether the alternative analysis was justified on the basis of missingness assumptions ▶ If a delta-adjusted MI analysis was used, whether external information elicited from subject-matter experts was used to choose the value(s) of the delta parameter
MI implementation	▶ The method used for MI, for example, multivariate normal imputation or multiple imputation by chained equations ▶ The statistical software used for MI ▶ The number of imputations performed ▶ Whether all analysis variables were included in the imputation model ▶ Whether auxiliary variables (ie, variables defined as potential predictors of the variable(s) with missing data and possibly also the missingness in these variables that are not included in the target analysis) were included in the imputation model ▶ Whether interactions were included in the imputation model

CCA, complete case analysis; MAR, missing at random; MCAR, missing completely at random; m-DAGs, missingness directed acyclic graphs; MI, multiple imputation; MNAR, missing not at random.

Data extraction and management

Covidence, a web-based tool for systematic review management, will be used to perform the review.[25] The data extraction questionnaire was developed and tested for use by RM and KL using a sample of 10 articles. Data from all eligible studies will be extracted by RM. The supplementary material of all eligible studies will also be reviewed. We will use double data extraction (performed by KL) for a random selection of 10% of articles and additionally when there is uncertainty about the information being extracted. Discrepancies and uncertainties will be resolved by discussion in meetings with at least three researchers (RM, KL and at least one of JC, JS, CN or MM-B).

Outcomes measured

We will extract data pertaining to the study characteristics, the amount of missing data and in which variables it occurs, missingness assumptions, methods for handling missing data and implementation of MI. Data extraction items are summarised in table 1 and a copy of the data extraction questionnaire is provided in the online supplemental material. Because we anticipate difficulties in extracting some items (such as the percentage of complete cases), in online supplemental table 1, we list potential challenges in extracting data and any assumptions or simplifications that will be made if these challenges arise. Any post-hoc assumptions or simplifications for unanticipated challenges will be recorded and reported as part of the analysis.

Analysis

The questionnaire data will be cleaned and analysed in R. Descriptive statistics will be used to summarise the extracted data. Frequencies and percentages will be presented for categorical data, for example, the method used to obtain the primary results. Median and IQR will be presented for continuous data, for example, the percentage of complete cases in each observational study. We are also collecting free-text data on certain aspects of missing data handling to capture information that may be difficult to capture otherwise, such as the details of the justification provided for the missingness assumptions. We will examine the free-text data for themes and patterns. If possible, we will group responses into common themes and summarise these themes using frequencies and percentages. If this is not possible, we will summarise the results in text. All data and code will be made publicly available on GitHub.

Reporting

Findings from this review will be reported using the Preferred Reporting Items for Systematic reviews and

Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.[26]

Patient and public involvement

None.

ETHICS AND DISSEMINATION

Ethics approval is not required for this review because data will be collected only from published studies. The results will be disseminated through a peer-review publication and conference presentations.

DISCUSSION

Previous reviews of the handling of missing data have primarily focused on RCTs with incomplete outcome data. Observational studies that answer causal questions are common and subject to greater challenges than RCTs in terms of missing data as they often face missing data in multiple variables (exposure, outcome and/or confounders). This paper describes a protocol for a scoping review of how MI is used to handle missing data in these studies.

Strengths and limitations

There are several strengths to our study. A targeted review of observational studies in top epidemiology journals publishing general research will benchmark the current state of practice for handling multivariable missingness with MI in causal analyses. Screening, reviewing and data extraction will be performed systematically. All data and code will be made publicly available, enabling our analysis to be entirely reproducible. Results from the review will be reported according to best practice, using PRISMA-ScR.

There are also limitations. Identifying whether the aim of the research was to answer a descriptive, causal or predictive question is somewhat subjective because many researchers have not adopted this classification of research questions.[1] Although our targeted review will not include studies from all epidemiology journals, we expect that included studies (expected to be >100 studies from five major epidemiology journals) will be sufficient to provide insight and general trends on the methods of interest. It is likely that some of the information sought will be unclear or not reported. To accommodate this, we have specified how anticipated challenges with data extraction will be handled if they arise.

Implications of this research

In addition to critically appraising the current state of the literature regarding the use and reporting of causal analyses using MI to handle missing data in observational studies, this review will identify areas for improvement in the handling and reporting of missing data in these studies. The results of this review will be used to develop practical guidance for researchers and inform future research in these areas.

Contributors RM conceived the study idea, developed the methodology, designed the data extraction tool, drafted and revised the paper. KL developed the study idea, methodology, data extraction tool and revised the paper. MM-B and JS developed the study idea, methodology and revised the paper. CN developed the study idea, methodology and data extraction tool. JC developed the study idea, methodology, data extraction tool and revised the paper.

Funding This work was supported by an Australian National Health and Medical Research Council (NHMRC) Career Development Fellowship (CDF) Level 2 Grant (grant 1127984 awarded to KL), an NHMRC Investigator Grant Leadership Level 1 (grant 1196068 awarded to JS), an NHMRC Investigator Grant Emerging Leadership Level 2 (grant 2009572 awarded to MM-B) and an NHMRC Project Grant (grant 1166023). Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs

Rheanna Mainzer http://orcid.org/0000-0002-5933-8917
Cattram Nguyen http://orcid.org/0000-0002-0599-8645

REFERENCES

1 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE* 2019;32:42–9.
2 Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health* 2018;108:616–9.
3 Little RJ, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, 2019.
4 Rubin DB. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004.
5 van Buuren S. *Flexible imputation of missing data*. Boca Raton, FL: CRC press, 2019.
6 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.
7 Seaman S, Galati J, Jackson D, *et al*. What is meant by " missing at random "? *Statist Sci* 2013;28:257–68.
8 Mohan K, Pearl J. Graphical models for processing missing data. *J Am Statist Assoc* 2021;116:1023–37.
9 Moreno-Betancur M, Lee KJ, Leacy FP, *et al*. Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *Am J Epidemiol* 2018;187:2705–15.
10 National Research Council. *The prevention and treatment of missing data in clinical trials*. 2010.
11 Mackinnon A. The use and reporting of multiple imputation in medical research-a review. *J Intern Med* 2010;268:586–93.
12 Tan P-T, Cro S, Van Vogt E, *et al*. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. *BMC Med Res Methodol* 2021;21:72.
13 Rabe BA, Day S, Fiero MH, *et al*. Missing data handling in non-inferiority and equivalence trials: a systematic review. *Pharm Stat* 2018;17:477–88.

14 Fiero MH, Huang S, Oren E, *et al*. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016;17:72.
15 Bell ML, Fiero M, Horton NJ, *et al*. Handling missing data in rcts; a review of the top medical journals. *BMC Med Res Methodol* 2014;14:118.
16 Powney M, Williamson P, Kirkham J, *et al*. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials* 2014;15:1–11.
17 Ibrahim F, Tom BDM, Scott DL, *et al*. A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. *Trials* 2016;17:272.
18 Rombach I, Rivero-Arias O, Gray AM, *et al*. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Qual Life Res* 2016;25:1613–23.
19 White IR, Horton NJ, Carpenter J, *et al*. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2011;342:d40.
20 Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol* 2015;15:30.
21 Tompsett DM, Leacy F, Moreno-Betancur M, *et al*. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Stat Med* 2018;37:2338–53.
22 Cro S, Morris TP, Kenward MG, *et al*. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Stat Med* 2020;39:2815–42.
23 Hayati Rezvan P, Lee KJ, Simpson JA. Sensitivity analysis within multiple imputation framework using delta-adjustment: application to longitudinal study of australian children. *LLCS* 2018;9:259–78.
24 Penning de Vries BBL, van Smeden M, Rosendaal FR, *et al*. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. *J Clin Epidemiol* 2020;121:55–61.
25 Veritas Health Innovation. *Covidence systematic review software*. Melbourne, Australia,
26 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-scr): checklist and explanation. *Ann Intern Med* 2018;169:467–73.