BMJ Open

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-065576
Article Type:	Protocol
Date Submitted by the Author:	10-Jun-2022
Complete List of Authors:	Mainzer, Rheanna; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Moreno-Betancur , Margarita ; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Nguyen, Cattram; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Simpson, Julie; University of Melbourne School of Population and Global Health Carlin, John; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Lee, Katherine; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Corresponding author: 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 such studies and can occur in the exposure, outcome and/or variables used to control for confounding. The standard classification of all missing data as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR), 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 (i) how missingness assumptions are expressed and assessed, (ii) the connection between missingness assumptions and the use of MI or other approaches for handling missing data, and (iii) the conduct of sensitivity analyses under alternative plausible missingness mechanisms.

Methods and analysis We will systematically review observational studies that aim to answer causal questions using 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". Information extracted from eligible studies will include details about the study characteristics, missing data, missingness assumptions, analysis methods and MI implementation. Systematic review methods will be used to screen, review and extract data. 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.

Registration 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. Although our targeted review will not include all relevant studies, we expect that included studies will be sufficient to provide insight and general trends on the application and reporting of multiple imputation in observational studies.
- 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.
- All data and code will be made publicly available, enabling our analysis to be entirely reproducible.
- 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.
- Results from the review will be reported according to best practice, using the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR).

INTRODUCTION

Observational studies in clinical and health-related research often aim to answer causal questions, i.e. to estimate the effect of an exposure on an outcome.(1) In such studies missing data are common and can occur in the exposure, the outcome and/or the variables used to control for confounding. Restricting statistical analysis to individuals with available data (complete case analysis, CCA) can lead to bias and/or loss of precision in estimates of the average causal effect. (2) Multiple imputation (MI) is a popular and flexible approach for estimating target quantities in the presence of incomplete data.(3, 4) 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 with associated standard error.(3)

Standard implementations of MI are known to provide consistent estimation of target parameters under certain unverifiable assumptions about the mechanism leading to missing data. These assumptions 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).(5) While this framework is useful if missing data occur in a single variable, it is poorly understood and does not allow for a transparent assessment of missingness assumptions when missingness arises in more than one variable.(6) For example, MAR is a sufficient but not necessary condition for the validity of standard MI estimates.(7) Further, because one cannot be sure about the true missing data mechanism, sensitivity analyses to examine the robustness of results to alternative plausible missingness mechanisms (hereafter, "sensitivity analyses") are strongly recommended.(8) 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."(8) The inherent difficulty in assessing missingness assumptions when framed in the traditional MCAR/MAR/MNAR manner and their lack of one-to-one correspondence with analytic approaches in the presence of multivariable missingness leads to further complications when planning and conducting sensitivity analyses.

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) with missing outcome data.(9-15) For trials, typically only the outcome variable is incomplete, while the intervention and other key variables are observed for all participants. In this setting where there is missing data in a single variable, the MCAR/MAR/MNAR framework is more transparent and guidance on sensitivity analyses has been well-developed (see, for example, (12, 16)). 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 two (4%) out of 50 non-RCT studies reviewed carried out an additional analysis that was described as a sensitivity analysis.(17) Similarly, Rezvan et al. (2015) found that none of the 30 observational studies reviewed conducted a sensitivity analysis to departures from the missingness assumptions following MI.(18) Even when they are carried out, what is meant by a "sensitivity analysis" is often unclear. Confusion between sensitivity analyses and secondary analyses has been observed, (17, 19) and the logic behind applying MI as a sensitivity analysis to a CCA (or vice versa) is unsound.(17) While the reviews by Mackinnon and Rezvan et al. 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.(7) M-DAGs can aid the depiction and assessment of missingness assumption, which is important since transparency in the assumed causal mechanisms underlying the missing data facilitates the choice of analytical approach.(20) Although, it is currently unclear how much m-DAGs are being used in the literature.

The aim of this scoping review is to systematically review the epidemiological literature to examine the use of MI in observational studies that address causal questions, which is typically the focus of such studies even when this may not be very clearly articulated.(21) These studies often face missingness in multiple variables required for analysis. We will examine (i) how missingness assumptions are expressed, (ii) their connection to the justification for the use of MI or other approaches for handling missing data, and (iii) the conduct of sensitivity analyses to alternative plausible missingness mechanisms. 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 of reporting on the handling of missing data with MI in observational studies, and to subsequently develop guidance for researchers.

METHODS AND ANALYSIS

In this section we provide a full description of the study design, including how articles will be selected, what outcomes will be measured, and how data will be extracted and analysed. The anticipated start date of this review is 13th June 2022 and the anticipated completion date is 30th November 2022.

Search strategy

We will systematically 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 systematic review of epidemiologic practice.(22) 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.(17, 18)

Inclusion criteria

We will include original research articles that were published between January 2019 and December 2021, and aim to answer at least one causal question using MI to handle the 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, i.e., an interpretation which suggested that intervening on the exposure could change the outcome (e.g., increasing coffee consumption may be protective against stroke). This will be determined by wording in conclusions and typically signalled by the identification of confounders, the inclusion of a DAG to illustrate causal assumption made in the analysis, and/or analytical approaches incorporating adjustment for confounders (for example, estimating an effect using a regression model that was adjusted for a set of covariates).

All disease areas/medical conditions will be considered and there will be no restrictions on the study participants.

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 validate 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 published studies that aim 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 trial investigators.

Sample size

We will require at least 100 studies to estimate the percentage of studies with a particular element (e.g., 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% confidence interval from 40% to 60%. For a prevalence greater than or less than 50%, the 95% confidence interval will be narrower. This sample size is similar to the sample size used in the first review of MI in medical research (n = 99, (17)), and many of the subsequent reviews in this area (e.g., n = 103 in (18), 77 in (12) and 118 in (9)). We expect to identify at least 100 eligible studies given the three-year publication time frame. All eligible studies will be included in the review.

Study selection

The search of the journal databases will be performed by a single researcher (RM). The title, abstract and date of each article will be screened for eligibility. When a decision about the eligibility of an article cannot be reached based on the title, abstract and publication date alone, the full text will be screened for eligibility. A second researcher (CN) will independently screen articles when there is uncertainty about the inclusion criteria. 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 MMB).

Data extraction and management

Covidence, a web-based tool for systematic review management, will be used to perform the review.(23) The data extraction questionnaire was developed and tested for use by RM and KL using a sample of 10 articles. All eligible studies will be extracted and reviewed by RM. The supplementary material of all eligible studies will also be extracted and reviewed. We will use double data extraction (performed by CN) 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, CN and at least one of JC, JS, KL or MMB).

Outcomes measured

We will extract data pertaining to the study characteristics, the amount of missing data and in which variables, missingness assumptions, methods for handling missing data and implementation of multiple imputation. Data extraction items are summarised in Table 1. Because we anticipate difficulties in extracting some items (such as the percentage of complete cases), in Supplementary 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.

Table 1. Summary of items to be extracted from each article.

Category	Summary of data extraction items
Study characteristics	• Title • Authors • 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, e.g. 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 (e.g. a CCA or a delta-adjusted MI analysis, where imputations are shifted by a parameter "delta" representing the difference between the observed and unobserved data(24)) • Whether the alternative analysis was justified • 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 (i.e. variables defined as potential predictors of missingness and/or the variable(s) with missing data, but are not included in the target analysis) were included in the imputation model • Whether interactions were included in the imputation model

Analysis

The questionnaire data will be cleaned and analysed in R. Descriptive statistics will be used to summarise the data. Frequencies and percentages will be presented for categorical data, for example, the method used to obtain the primary results. Median and interquartile range will be presented for continuous data, for example, the percentage of complete cases in each observational study. 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.(25)

Patient and public involvement

There will be no patient or public involvement in this project because data will be collected only from published studies.

DISCUSSION

Previous reviews of the handling of missing data have primarily focused on RCTs with incomplete outcomes. Observational studies are 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 observational studies that answer causal questions.

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. 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 analyses using MI to handle missing data, this review will identify areas for improvement in the handling and reporting of missing data in observational studies. The results of this review will be used to develop practical guidance for researchers and promote the formulation of missingness assumptions in a clear and transparent manner.

Funding sources / sponsors

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 KJL), a NHMRC Investigator Grant Leadership Level 1 (grant 1196068 awarded to JAS), a NHMRC Investigator Grant Emerging Leadership Level 2 (grant 2009572 awarded to MMB) and a NHMRC Project Grant (grant 1166023). Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Authors' contributions

RM conceived the study and wrote the first draft of the manuscript. All authors contributed to the design of the study, revision of the manuscript and take public responsibility for its content.

Competing interests statement

None declared.

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(1):42-9.
2. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2019.
3. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.
4. Van Buuren S. Flexible imputation of missing data: CRC press; 2018.
5. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
6. Seaman S, Galati J, Jackson D, Carlin J. What is meant by "missing at random"? Statistical Science. 2013;28(2):257-68.
7. Mohan K, Pearl J. Graphical models for processing missing data. Journal of the American Statistical Association. 2021;116(534):1023-37.
8. National Research Council. The prevention and treatment of missing data in clinical trials. 2010.
9. Tan P-T, Cro S, Van Vogt E, Szigeti M, Cornelius VR. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. BMC medical research methodology. 2021;21(1):1-17.
10. Rabe BA, Day S, Fiero MH, Bell ML. Missing data handling in non-inferiority and equivalence trials: A systematic review. Pharmaceutical statistics. 2018;17(5):477-88.
11. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.

12. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology. 2014;14(1):1-8.
13. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014;15(1):1-11.
14. Ibrahim F, Tom BD, Scott DL, Prevost AT. A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. Trials. 2016;17(1):1-8.
15. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. Quality of Life Research. 2016;25(7):1613-23.
16. White IR, Horton NJ, Carpenter J, statistics rim, social, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ. 2011;342:d40.
17. Mackinnon A. The use and reporting of multiple imputation in medical research–a review. Journal of internal medicine. 2010;268(6):586-93.
18. 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 medical research methodology. 2015;15(1):1-14.
19. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PP. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ open. 2016;6(10):e012594.
20. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. American Journal of Epidemiology. 2018;187(12):2705-15.
21. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. American journal of public health. 2018;108(5):616-9.
22. de Vries BBP, van Smeden M, Rosendaal FR, Groenwold RH. , abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. Journal of Clinical Epidemiology. 2020;121:55-61.
23. Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia.
24. 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. Statistics in medicine. 2020;39(21):2815-42.
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018;169(7):467-73.

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Corresponding author: Rheanna Mainzer; rheanna.mainzer@mcri.edu.au

Supplementary Table 1. Anticipated challenges with data extraction and how they will be handled.

Challenge for data extraction	Category of items affected	How challenge will be handled
Articles may have more than one publication date, for example, the date the article first appeared online and when it was published in-print.	Inclusion criteria	Only one publication date is required to be between January 2019 and December 2021. If two or more publication dates are between January 2019 and December 2021, the earlier date will be recorded.
There are multiple causal questions, exposures or outcomes.	Missing data	We will identify the primary causal question based on the research aims and conclusion. The proportion of missing data in the exposure, outcome and confounders used to answer this primary question will be recorded. This is expected to be acceptable in most cases. If the primary causal question cannot be identified due to multiple outcomes, we will report the missing data details for the first outcome listed in the methods section. (This is comparable to the strategy taken by Fiero et al. (1)) Similarly, if the primary causal question cannot be identified due to multiple exposures, we will report the missing data details for the first exposure listed in the methods section.
Multiple sets of covariates are used for adjustment.	Missing data	The largest adjustment set will be considered. The number of incomplete covariates will be recorded categorically (no incomplete covariates, 1 incomplete covariate, 2 or more incomplete covariates, not stated or unable to establish). This categorisation has been chosen to enable determination of multivariable missingness.
Not clear whether all variables in the target analysis were	MI implementation	If some (but not all) analysis variables were reported as being included in the imputation model then we will assume

included in the imputation model.		that the analysis variables not explicitly mentioned were excluded from the imputation model. If there was no description of the imputation model, then we will categorise this as “unclear”.
Not clear whether auxiliary variables or interactions were included in the imputation model.	MI implementation	If it is not explicitly stated that these were included in the imputation model, we will assume they were excluded. If there was no mention of the imputation model then we will categorise this as “unclear”.
Imputation method used not explicitly stated.	MI implementation	If the imputation method used (e.g. multivariate normal imputation or multiple imputation by chained equations) is not provided, we will infer the method used, where possible, from the statistical software procedures listed in the main paper or supplementary material. If the method is unable to be inferred, we will categorise this as “unclear”.

REFERENCE

1. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.

BMJ Open

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-065576.R1
Article Type:	Protocol
Date Submitted by the Author:	15-Nov-2022
Complete List of Authors:	Mainzer, Rheanna; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Moreno-Betancur , Margarita ; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Nguyen, Cattram; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Simpson, Julie; University of Melbourne School of Population and Global Health Carlin, John; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Lee, Katherine; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Public health, Research methods
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Corresponding author: 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 (MCAR), missing at random (MAR) or missing not at random (MNAR), 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 (i) missingness assumptions are expressed and assessed, (ii) missingness assumptions are used to justify the choice of a complete case analysis and/or MI for handling missing data, and (iii) sensitivity analyses under alternative plausible assumptions about the missingness mechanism are conducted.

Methods and analysis We will systematically 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.

Registration 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 statistical analysis to individuals with complete data on all analysis variables, i.e., 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 a 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.

Because one cannot verify from the observed data what the true missing data mechanism is, sensitivity analyses to examine the robustness of results to alternative plausible assumptions about the missingness mechanism (hereafter, "sensitivity analyses") are strongly recommended.(10) However, 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 this. Furthermore, the mistakenly assumed one-to-one correspondence with analytic approaches in the presence of multivariable missingness leads to misguided practices. For example, from our observation, MI is routinely applied as a sensitivity analysis to a CCA. However, the logic behind applying MI as a sensitivity analysis to a CCA (or vice versa) without first considering one's assumptions about the missingness mechanism is unsound.(11) Obtaining similar or different estimates from these analyses does not provide insight into the impact of alternative plausible assumptions about the missingness mechanism on the study results unless one has first made their missingness assumptions explicit and identified these two approaches as appropriate for estimating the target parameter under those explicit assumptions.

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 (see, for example, (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 two (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. (2015) 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 and Rezvan et al. 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 systematically review the epidemiological literature to examine the use of MI in observational studies that address causal questions, which is typically the focus of such 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 (i) how missingness assumptions are expressed, (ii) if and how missingness assumptions are used to justify the choice of a CCA and/or MI for handling missing data, and (iii) the conduct of sensitivity analyses to 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, which variables will be extracted, and how 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 systematically 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 systematic 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, i.e., an interpretation which suggested that intervening on the exposure could change the outcome (e.g., 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 (for example, 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 (e.g., 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% confidence interval from 40% to 60%. For a prevalence greater than or less than 50%, the 95% confidence interval 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 (e.g., n = 103 in (20), 77 in (15) and 118 in (12)). We expect to identify at least 100 eligible studies given the three-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 (for example, 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 MMB).

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 MMB).

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 multiple imputation. Data extraction items are summarised in Table 1. Because we anticipate difficulties in extracting some items (such as the percentage of complete cases), in Supplementary 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.

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, e.g. 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 (e.g. 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 (i.e. variables defined as potential predictors of the variable(s) with missing data that are not included in the target analysis) were included in the imputation model • Whether interactions were included in the imputation model

Analysis

The questionnaire data will be cleaned and analysed in R. Descriptive statistics will be used to summarise the data. Frequencies and percentages will be presented for categorical data, for example, the method used to obtain the primary results. Median and interquartile range will be presented for continuous data, for example, the percentage of complete cases in each observational study. 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

There will be no patient or public involvement in this project because data will be collected only from published studies.

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.

Funding sources / sponsors

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 KJL), a NHMRC Investigator Grant Leadership Level 1 (grant 1196068 awarded to JAS), a NHMRC Investigator Grant Emerging Leadership Level 2 (grant 2009572 awarded to MMB) and a NHMRC Project Grant (grant 1166023). Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Authors' contributions

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. MMB 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.

Competing interests statement

None declared.

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(1):42-9.
2. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. American journal of public health. 2018;108(5):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: CRC press; 2018.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, 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, Carlin J. What is meant by "missing at random"? Statistical Science. 2013;28(2):257-68.
8. Mohan K, Pearl J. Graphical models for processing missing data. Journal of the American Statistical Association. 2021;116(534):1023-37.
9. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. American Journal of Epidemiology. 2018;187(12):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. Journal of internal medicine. 2010;268(6):586-93.
12. Tan P-T, Cro S, Van Vogt E, Szigeti M, Cornelius VR. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. BMC medical research methodology. 2021;21(1):1-17.
13. Rabe BA, Day S, Fiero MH, Bell ML. Missing data handling in non-inferiority and equivalence trials: A systematic review. Pharmaceutical statistics. 2018;17(5):477-88.
14. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.
15. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology. 2014;14(1):1-8.
16. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014;15(1):1-11.
17. Ibrahim F, Tom BD, Scott DL, Prevost AT. A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. Trials. 2016;17(1):1-8.
18. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. Quality of Life Research. 2016;25(7):1613-23.
19. White IR, Horton NJ, Carpenter J, statistics rim, social, Pocock SJ. 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 medical research methodology. 2015;15(1):1-14.
21. Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Statistics in medicine. 2018;37(15):2338-53.
22. 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. Statistics in medicine. 2020;39(21):2815-42.
23. Rezvan PH, Lee KJ, Simpson JA. Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to Longitudinal Study of Australian Children. Longitudinal and Life Course Studies. 2018;9(3):259-78.

24. de Vries BBP, van Smeden M, Rosendaal FR, Groenwold RH. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. Journal of Clinical Epidemiology. 2020;121:55-61.
25. Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia.
26. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018;169(7):467-73.

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Corresponding author: Rheanna Mainzer; rheanna.mainzer@mcri.edu.au

Supplementary Table 1. Anticipated challenges with data extraction and how they will be handled.

Challenge for data extraction	Category of items affected	How challenge will be handled
Articles may have more than one publication date, for example, the date the article first appeared online and when it was published in-print.	Inclusion criteria	Only one publication date is required to be between January 2019 and December 2021. If two or more publication dates are between January 2019 and December 2021, the earlier date will be recorded.
There are multiple causal questions, exposures or outcomes.	Missing data	We will identify the primary causal question based on the research aims and conclusion. The proportion of missing data in the exposure, outcome and confounders used to answer this primary question will be recorded. This is expected to be acceptable in most cases. If the primary causal question cannot be identified due to multiple outcomes, we will report the missing data details for the first outcome listed in the methods section. (This is comparable to the strategy taken by Fiero et al. (1)) Similarly, if the primary causal question cannot be identified due to multiple exposures, we will report the missing data details for the first exposure listed in the methods section.
Multiple sets of covariates are used for adjustment.	Missing data	The largest adjustment set will be considered. The number of incomplete covariates will be recorded categorically (no incomplete covariates, 1 incomplete covariate, 2 or more incomplete covariates, not stated or unable to establish). This categorisation has been chosen to enable determination of multivariable missingness.
Not clear whether all variables in the target analysis were	MI implementation	If some (but not all) analysis variables were reported as being included in the imputation model then we will assume

included in the imputation model.		that the analysis variables not explicitly mentioned were excluded from the imputation model. If there was no description of the imputation model, then we will categorise this as “unclear”.
Not clear whether auxiliary variables or interactions were included in the imputation model.	MI implementation	If it is not explicitly stated that these were included in the imputation model, we will assume they were excluded. If there was no mention of the imputation model then we will categorise this as “unclear”.
Imputation method used not explicitly stated.	MI implementation	If the imputation method used (e.g. multivariate normal imputation or multiple imputation by chained equations) is not provided, we will infer the method used, where possible, from the statistical software procedures listed in the main paper or supplementary material. If the method is unable to be inferred, we will categorise this as “unclear”.

REFERENCE

1. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-065576.R2
Article Type:	Protocol
Date Submitted by the Author:	03-Jan-2023
Complete List of Authors:	Mainzer, Rheanna; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Moreno-Betancur , Margarita ; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Nguyen, Cattram; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Simpson, Julie; University of Melbourne School of Population and Global Health Carlin, John; Murdoch Childrens Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics Lee, Katherine; Murdoch Children's Research Institute, Clinical Epidemiology and Biostatistics Unit; The University of Melbourne, Department of Paediatrics
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Public health, Research methods
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

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

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Correspondence to:

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 (MCAR), missing at random (MAR) or missing not at random (MNAR), 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 (i) missingness assumptions are expressed and assessed, (ii) missingness assumptions are used to justify the choice of a complete case analysis and/or MI for handling missing data, and (iii) 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.

Study registration: 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, i.e., 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 a 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 (see, for example, (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 two (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. (2015) 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 and Rezvan et al. 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 (i) how missingness assumptions are expressed, (ii) if and how missingness assumptions are used to justify the choice of a CCA and/or MI for handling missing data, and (iii) 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, i.e., an interpretation which suggested that intervening on the exposure could change the outcome (e.g., 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 (for example, 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 (e.g., 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% confidence interval from 40% to 60%. For a prevalence greater than or less than 50%, the 95% confidence interval 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 (e.g., n = 103 in (20), 77 in (15) and 118 in (12)). We expect to identify at least 100 eligible studies given the three-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 (for example, 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 MMB).

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 MMB).

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 multiple imputation. Data extraction items are summarised in Table 1 and a copy of the data extraction questionnaire is provided in the Supplementary Material. Because we anticipate difficulties in extracting some items (such as the percentage of complete cases), in Supplementary 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.

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, e.g. 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 (e.g. 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 (i.e. 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

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 interquartile range 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.

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 KJL), a NHMRC Investigator Grant Leadership Level 1 (grant 1196068 awarded to JAS), a NHMRC Investigator Grant Emerging Leadership Level 2 (grant 2009572 awarded to MMB) and a NHMRC Project Grant (grant 1166023). Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

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. MMB 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.

Competing interests

None declared.

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(1):42-9.
2. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. American journal of public health. 2018;108(5):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: CRC press; 2018.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, 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, Carlin J. What is meant by "missing at random"? Statistical Science. 2013;28(2):257-68.
8. Mohan K, Pearl J. Graphical models for processing missing data. Journal of the American Statistical Association. 2021;116(534):1023-37.
9. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. American Journal of Epidemiology. 2018;187(12):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. Journal of internal medicine. 2010;268(6):586-93.
12. Tan P-T, Cro S, Van Vogt E, Szigeti M, Cornelius VR. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. BMC medical research methodology. 2021;21(1):1-17.
13. Rabe BA, Day S, Fiero MH, Bell ML. Missing data handling in non-inferiority and equivalence trials: A systematic review. Pharmaceutical statistics. 2018;17(5):477-88.
14. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.
15. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology. 2014;14(1):1-8.
16. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014;15(1):1-11.
17. Ibrahim F, Tom BD, Scott DL, Prevost AT. A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. Trials. 2016;17(1):1-8.
18. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. Quality of Life Research. 2016;25(7):1613-23.

19. White IR, Horton NJ, Carpenter J, statistics rim, social, Pocock SJ. 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 medical research methodology. 2015;15(1):1-14.
21. Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Statistics in medicine. 2018;37(15):2338-53.
22. 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. Statistics in medicine. 2020;39(21):2815-42.
23. Rezvan PH, Lee KJ, Simpson JA. Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to Longitudinal Study of Australian Children. Longitudinal and Life Course Studies. 2018;9(3):259-78.
24. de Vries BBP, van Smeden M, Rosendaal FR, Groenwold RH. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. Journal of Clinical Epidemiology. 2020;121:55-61.
25. Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia.
26. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018;169(7):467-73.

The handling of missing data with multiple imputation in observational studies that address causal questions: Protocol for a scoping review

Supplementary Material

Rheanna M. Mainzer*[1,2], Margarita Moreno-Betancur[1,2], Cattram D. Nguyen[1,2], Julie A. Simpson[3], John B. Carlin[1,2,3], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
3. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

*Corresponding author: Rheanna Mainzer; rheanna.mainzer@mcri.edu.au

Supplementary Table 1. Anticipated challenges with data extraction and how they will be handled.

Challenge for data extraction	Category of items affected	How challenge will be handled
Articles may have more than one publication date, for example, the date the article first appeared online and when it was published in-print.	Inclusion criteria	Only one publication date is required to be between January 2019 and December 2021. If two or more publication dates are between January 2019 and December 2021, the earlier date will be recorded.
There are multiple causal questions, exposures or outcomes.	Missing data	We will identify the primary causal question based on the research aims and conclusion. The proportion of missing data in the exposure, outcome and confounders used to answer this primary question will be recorded. This is expected to be acceptable in most cases. If the primary causal question cannot be identified due to multiple outcomes, we will report the missing data details for the first outcome listed in the methods section. (This is comparable to the strategy taken by Fiero et al. (1)) Similarly, if the primary causal question cannot be identified due to multiple exposures, we will report the missing data details for the first exposure listed in the methods section.
Multiple sets of covariates are used for adjustment.	Missing data	The largest adjustment set will be considered. The number of incomplete covariates will be recorded categorically (no incomplete covariates, 1 incomplete covariate, 2 or more incomplete covariates, not stated or unable to establish). This categorisation has been chosen to enable determination of multivariable missingness.

Not clear whether all variables in the target analysis were included in the imputation model.	MI implementation	If some (but not all) analysis variables were reported as being included in the imputation model then we will assume that the analysis variables not explicitly mentioned were excluded from the imputation model. If there was no description of the imputation model, then we will categorise this as "unclear".
Not clear whether auxiliary variables or interactions were included in the imputation model.	MI implementation	If it is not explicitly stated that these were included in the imputation model, we will assume they were excluded. If there was no mention of the imputation model then we will categorise this as "unclear".
Imputation method used not explicitly stated.	MI implementation	If the imputation method used (e.g. multivariate normal imputation or multiple imputation by chained equations) is not provided, we will infer the method used, where possible, from the statistical software procedures listed in the main paper or supplementary material. If the method is unable to be inferred, we will categorise this as "unclear".

REFERENCE

1. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1-10.

Data extraction questionnaire.

Study characteristics

Authors

First author last name, e.g., Mainzer

Publication date

Publication date (mm-yyyy).

Journal

Journal in which paper was published

1. ○ International Journal of Epidemiology
2. ○ American Journal of Epidemiology
3. ○ European Journal of Epidemiology
4. ○ Journal of Clinical Epidemiology
5. ○ Epidemiology

Inclusion criteria

Select all that apply

1. □ Study authors stated they were estimated a causal effect
2. □ Study authors estimated an effect of an exposure on an outcome that was given (at least implicitly) a causal interpretation

Did the study use any of the following approaches (typical signals of a causal question)?

Select all that apply

1. □ Study used a directed acyclic graph (DAG) or m-DAG to illustrate causal assumptions made in the analysis
2. □ Study identified a set of variables that were used to control for confounding
3. □ Study estimated an effect of an exposure on an outcome using a regression model that was adjusted for a set of covariates

Causal interpretation

If the study estimated an effect that was given (at least implicitly) a causal interpretation, provide details of the text indicating this. (Copy and paste)

Type of study design

1. ○ Prospective longitudinal study

2. ○ Individual patient data (IPD) meta-analysis / pooled cohort analysis
3. ○ Retrospective analysis of routinely collected data (e.g., administrative or EMR data)
4. ○ Interrupted time series (ITS)
5. ○ Case-control study
6. ○ Case-cohort study
7. ○ Cross-sectional study
8. ○ Other

Missing data

Was the size of the inception sample* for the research question of interest available or able to be established?

*Inception sample: Participants who met eligibility criteria for inclusion in the study to answer the research question of interest, where eligibility criteria does not include any requirements for variables to be complete.

1. ○ Yes
2. ○ No, eligibility criteria required one or more variables to be complete
3. ○ Other

What was the size of the inception sample?

Number or NA

Was there a reduction in participants from the inception sample to the analysis sample* due to non-response or missing data in a variable used in the analysis (exposure, outcome, covariates)?

*Analysis sample: participants who were included in the study to address the research question of interest, who may or may not having missing data for analysis variables

1. ○ Yes
2. ○ No
3. ○ NA
4. ○ Other

What was the size of the analysis sample?

Number of NA

Was the percentage of complete cases* available or able to be established?

*Cases with observed data for each variable included in the analysis that was used to answer the research question of interest. The denominator is the size of the analysis sample.

1. ○ Yes
2. ○ Able to establish an upper bound only
3. ○ No

Percentage of complete cases / upper bound on the percentage of complete cases

Give number to nearest percent, e.g. 64, or NA. Use the size of the analysis sample as the denominator.

What was the exposure?

What/which exposure was considered for this review?

If there are multiple exposures: Identify the primary causal questions based on the research aims and conclusion and use the exposure in this question. If the primary causal question can not be identified due to multiple exposures, use the first exposure listed in the methods section.

Were there missing values in the exposure?

1. ○ Yes
2. ○ Yes, but only able to establish a lower bound on the percentage of missing values
3. ○ Yes, but unable to establish the percentage of missing values
4. ○ No
5. ○ Unclear

Percentage of missing values in the exposure / lower bound on the percentage of missing values in the exposure

Give number to nearest percent, e.g. 64, or NA. Use the size of the analysis sample as the denominator.

What/which outcome was considered for this review?

If there are multiple outcomes: Identify the primary causal question based on the research aims and conclusion and use the outcome in this question. If the primary causal question can not be identified due to multiple outcomes, use the first outcome listed in the methods section.

Were there missing values in the outcome?

1. ○ Yes
2. ○ Yes, but only able to establish a lower bound on the percentage of missing values
3. ○ Yes, but unable to establish the percentage of missing values
4. ○ No
5. ○ Unclear

Percentage of missing values in the outcome / lower bound on the percentage of missing values in the outcome

Give number to nearest percent, e.g. 64, or NA. Use the size of the analysis sample as the denominator.

Were there missing values in the covariates?

If multiple sets of covariates are used for adjustment, consider the largest adjustment set.

1. ○ Yes, in 2 or more covariates
2. ○ Yes, in 1 covariate only
3. ○ No
4. ○ Unable to establish

Missingness assumptions

Was a statement provided about what missingness assumptions were made?

1. ○ No
2. ○ Yes, authors invoked (either explicitly or implicitly) the missing at random assumption
3. ○ Yes, authors provided a comprehensive description of assumptions made about the missingness process for all variables subject to missing data, for example, using a m-DAG or a more simplified causal diagram
4. ○ Other

Were missingness assumptions justified?

For example, comparison of baseline data between responders and non-responders (to rule out MCAR) or a substantive assessment using expert knowledge. Note, no analysis of data can rule out MNAR.

1. ○ Yes
2. ○ No

Details of justification for missingness assumptions

For example, comparison of baseline data between responders and non-responders (to rule out MCAR) or a substantive assessment using expert knowledge. Note, no analysis of data can rule out MNAR. If missingness assumptions were not justified, enter NA.

Did authors address the potential for data to be MNAR?

1. ○ Yes, using external evidence such as expert knowledge
2. ○ Yes, but only as a study limitation
3. ○ No, the possibility that data were MNAR was not addressed
4. ○ Other

Analysis methods

What method was used to obtain the primary results?

1. ○ MI using the full analysis sample
2. ○ MI using a reduced analysis sample
3. ○ CCA, weighted (e.g. using IPW)
4. ○ CCA, unweighted
5. ○ delta-adjusted MI
6. ○ Other

Was the primary analysis justified on the basis of missingness assumptions?

1. ○ Yes
2. ○ No

Details of justification for primary analysis on the basis of missingness assumptions.

Examples include: (i) CCA was used because there was a small proportion of missing data that was unlikely to influence the results; (ii) CCA was used because a comparison of responders and non-responders did not rule out data being MCAR; (iii) MI was used because it was assumed that data were MAR; (iv) MI was used because comparison of responders and non-responders ruled out data being MCAR.

If the primary analysis was not justified on the basis of missingness assumptions, write "NA".

Was a secondary analysis that handles missing data differently used to answer the same causal question?

Select all that apply.

1. □ Yes, MI using the full analysis sample
2. □ Yes, MI using a reduced analysis sample
3. □ Yes, weighted CCA (e.g. using IPW)
4. □ Yes, unweighted CCA
5. □ Yes, delta-adjusted MI
6. □ No
7. □ Other

Was the secondary analysis justified?

1. ○ No
2. ○ Yes, as a sensitivity analysis (without further justification)
3. ○ Yes, as a sensitivity analysis to examine the influence of missing data
4. ○ Yes, as a sensitivity analysis to parametric modelling assumptions
5. ○ Yes, as a sensitivity analysis to causal assumptions made about the missing data mechanism
6. ○ NA
7. ○ Other

If a delta-adjusted analysis was used, was external information incorporated in the analysis?

If not delta-adjusted analysis select NA

1. ○ Yes
2. ○ No or not stated
3. ○ NA

If a delta-adjusted analysis was used, provide details of the delta-adjusted analysis

How was external information incorporated? What values of delta were considered? How was the analysis implemented? Etc. If no delta-adjusted analysis was used, enter NA.

MI implementation

What method was used for multiple imputation?

If the imputation method used (e.g. multivariate normal imputation or multiple imputation by chained equations) is not provided, we will infer the method used, where possible, from the statistical software procedures listed in the main paper or supplementary material. If the method is unable to inferred, we will categorise this as "unclear".

1. ○ MICE
2. ○ MVNI
3. ○ Unclear
4. ○ Other

What software was used for multiple imputation?

1. ○ R
2. ○ SAS
3. ○ SPSS
4. ○ Stata
5. ○ Unclear
6. ○ Other

Number of imputations used in the multiple imputation procedure

Were all analysis variables included in the imputation model?

If some (but not all) analysis variables were reported as being included in the imputation model then we will assume that the analysis variables not explicitly mentioned were excluded from the imputation model. If there was not description of the imputation model, then we will categorise this as "unclear".

1. ○ Yes
2. ○ No
3. ○ Unclear

Were auxiliary variables included in the imputation model?

If it is not explicitly stated that these were included in the imputation model, we will assume they were excluded. If there was no mention of the imputation model, then we will categorise this as "unclear".

1. ○ Yes
2. ○ No
3. ○ Unclear

Were interactions included in the imputation model?

If it is not explicitly stated that these were included in the imputation model, we will assume they were excluded. If there was no mention of the imputation model, then we will categorise this as "unclear".

1. ○ Yes
2. ○ No
3. ○ Unclear

Reported results

If results were obtained using both a CCA and MI, did the authors observe any substantial difference between these?

Substantial difference: a difference that the authors acknowledged as important or significant (for example, based on a clinical cut-off or a P values)

1. ○ Yes
2. ○ No
3. ○ NA

If results were obtained using both a CCA and MI, AND no substantial difference between these two sets of results was observed, was any interpretation or explanation provided for the similarities between the two sets of results? If so, what was the interpretation or explanation.

If yes, add details. Otherwise: no or NA.

Other

Funding

How was the study funded?

Any other comments?