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.

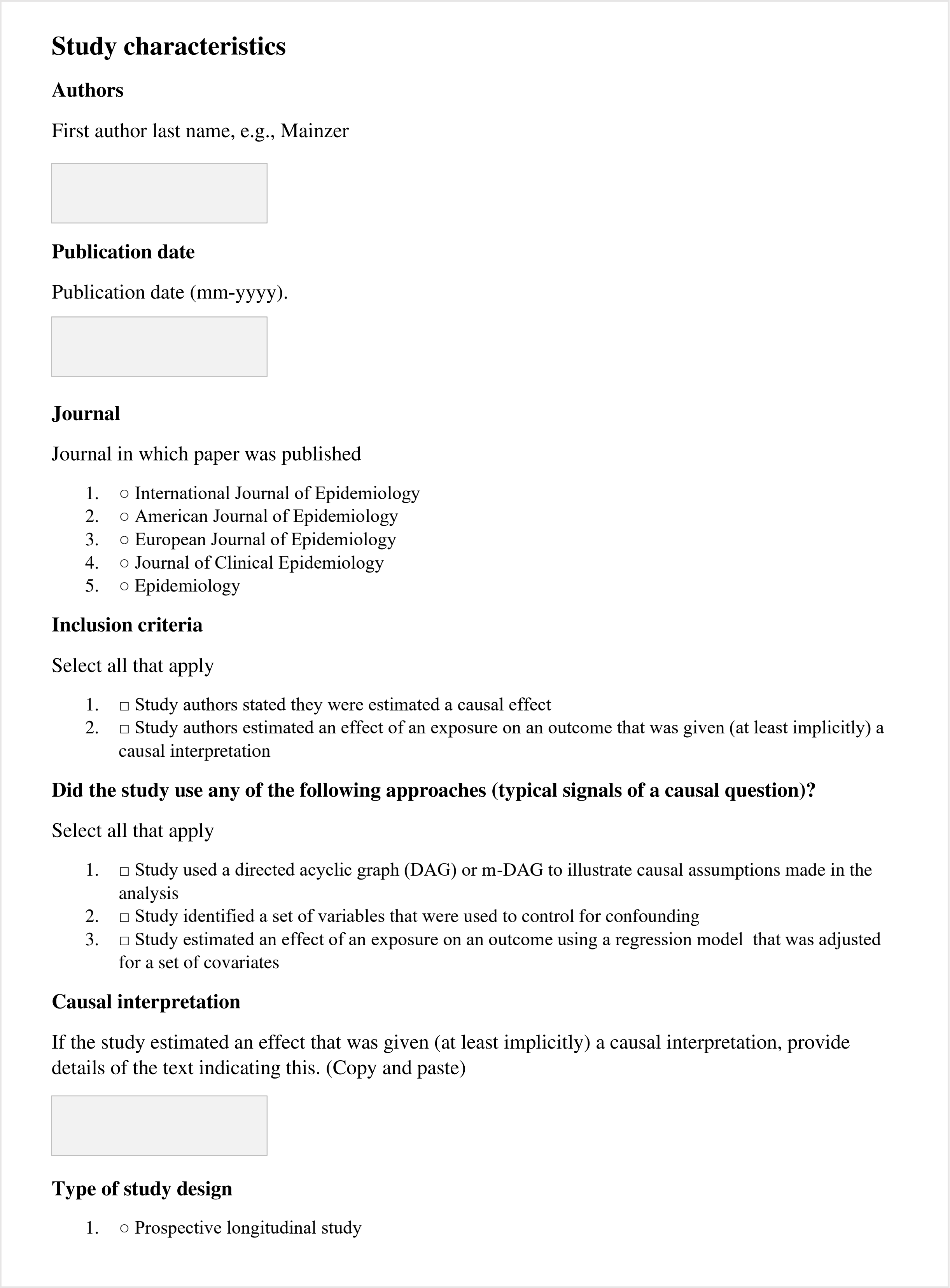
Challenge for data extraction	**Category of items affected**	**How challenge will be handled**
Articles may have more than one publication date, for example, the date the article first appeared online and when it was published in-print.	Inclusion criteria	Only one publication date is required to be between 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

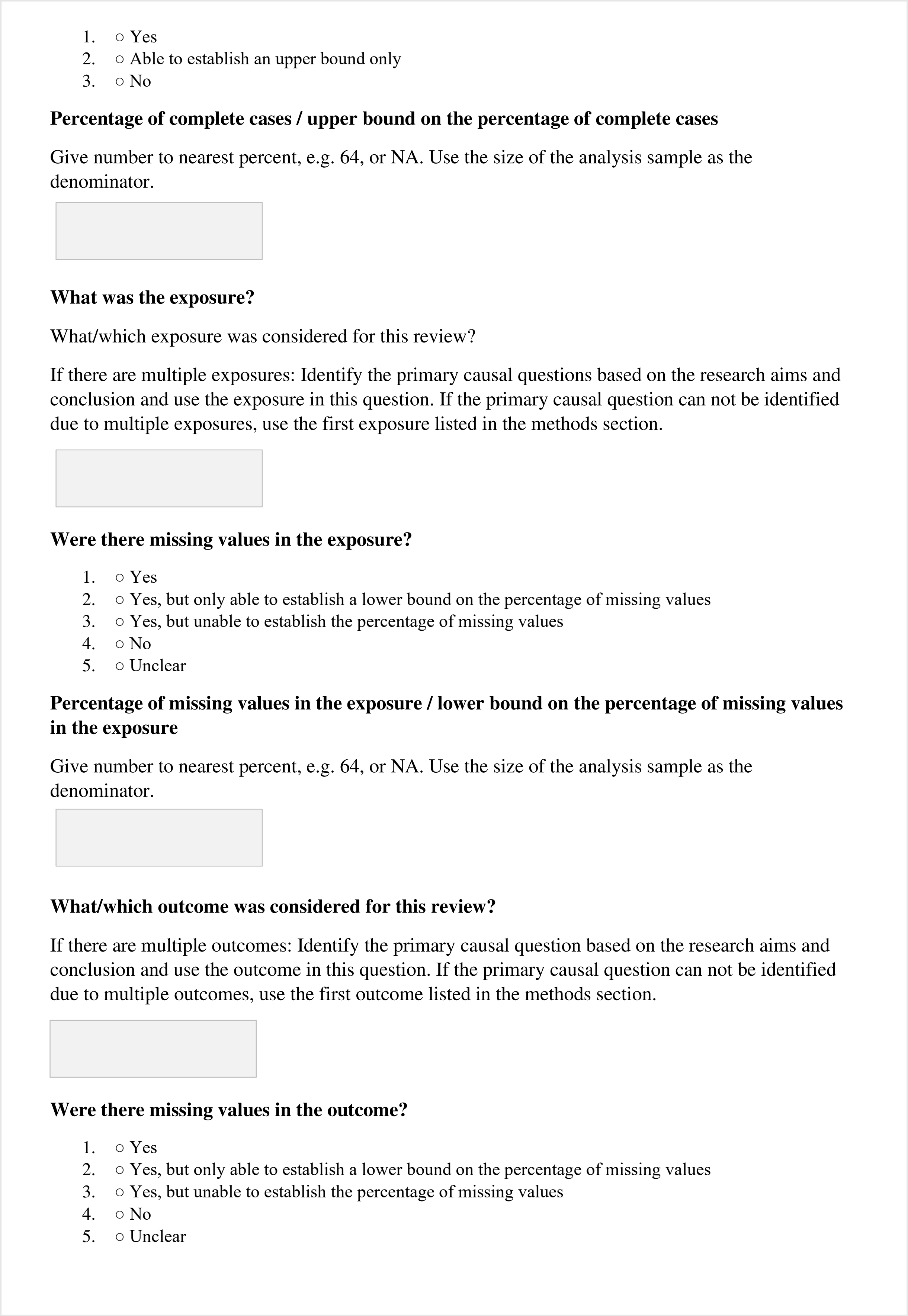
1. ○ Yes
2. ○ No
3. ○ NA
4. ○ Other

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.

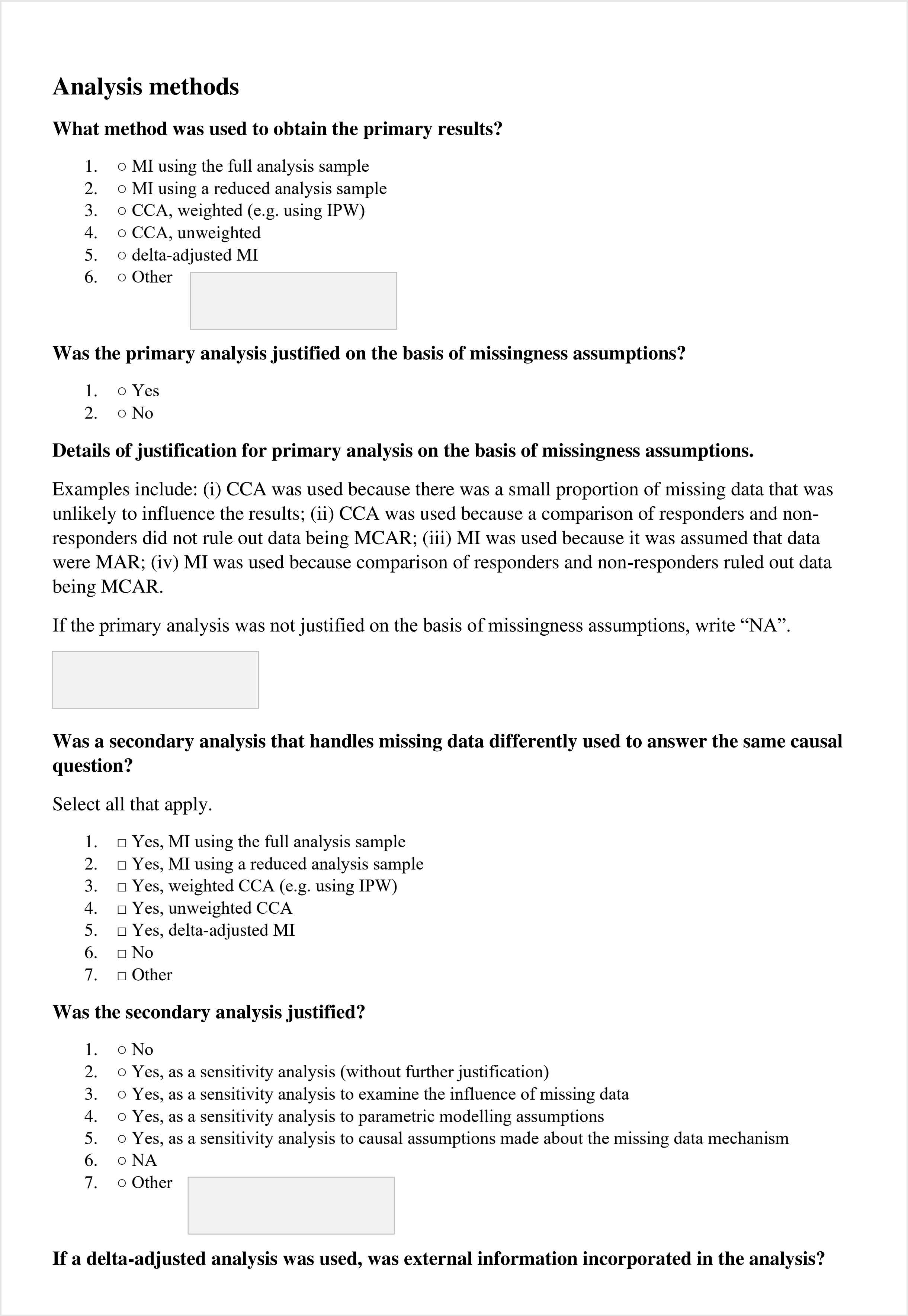
1. ○ Yes
2. ○ No

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?