Handling missing data in the analysis:

practical guidance for structuring the analysis, choosing the tools, and reporting the results

James R. Carpenter (on behalf of Topic Group 1)

James.Carpenter@lshtm.ac.uk · J.Carpenter@ucl.ac.uk

STRATOS INITIATIVE

Hamburg, Thu 31st March 2022

Acknowledgements

Katherine J Lee (Melbourne)

Kate Tilling (Bristol)

Rosie Cornish (Bristol)

Roderick J A Little (Michigan)

Melanie L Bell (Arizona)

Els Goetghebeur (Ghent)

Joseph W Hogan (Brown)

The STRATOS initiative

Overview

- Motivation
- Case study
- Structuring the analysis: TARMOS framework
- Choosing the tools: CC and/or IPW and/or MI?
- Application
- Towards structured reporting
- Discussion

Motivation: level 2 (experienced statistical knowledge)

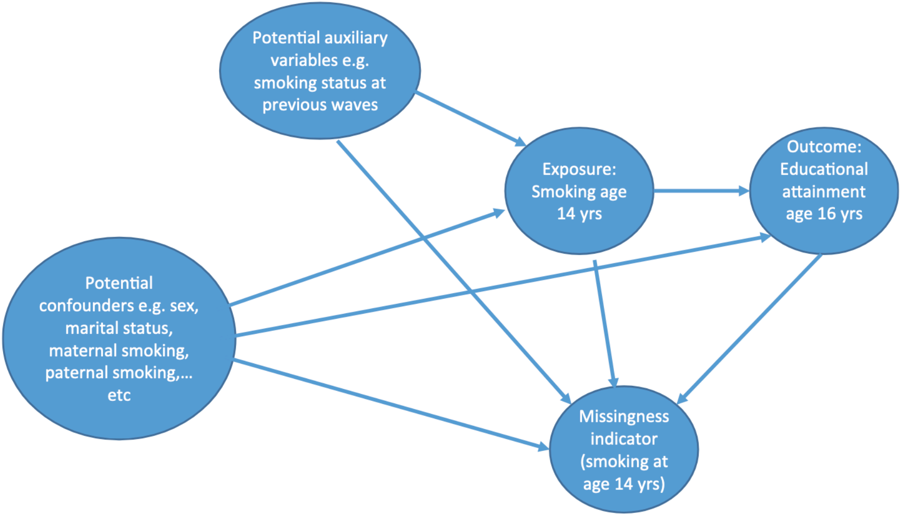
> *A lot of people arrive at doing MI the way I did, i.e. borrow a [Stata] do-file from someone who has done MI on a similar dataset, tinker with the variables in the MI command, run it, see that the imputed estimates aren't so different, write-up and publish.*
>
> *This means that incorrect approaches are likely to propagate virally...Therefore, to the hands-on MSc student, it boils down to "what should I put into my imputation?" "what should I leave out of my imputation".*

— Missing data course participant

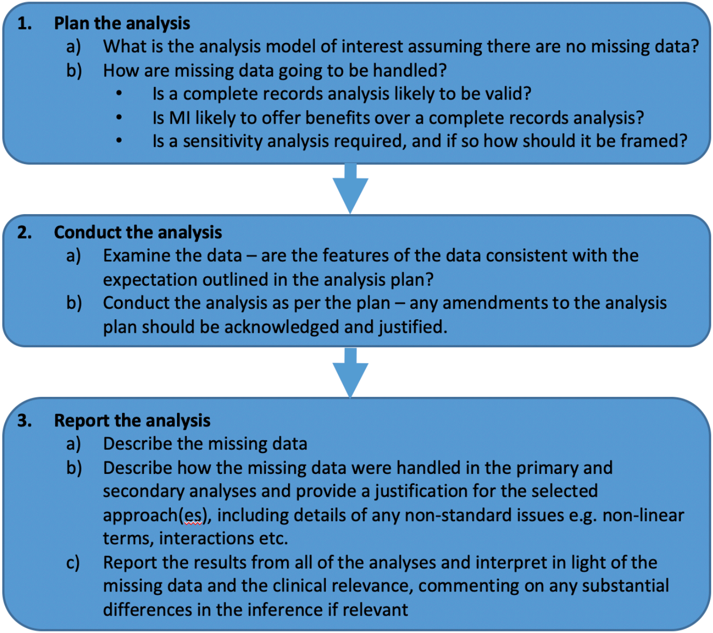
Case study

- ALSPAC (Avon Longitudinal Study of Parents and Children) initially recruited 14,591 pregnant women living around Bristol (UK) in the early 1990s. More women were recruited subsequently [1].
- ALSPAC suffers from attrition (highest in infancy and late adolescence) and sporadic missingness.
- We investigate whether there is association between smoking at 14 years and educational achievement (GCSE score) at 16 years.
- Data from 14,684 adolescents are available, but there are missing data in all variables except sex.

Causal diagram


Potential auxiliary variables e.g. smoking status at previous waves
Exposure: Smoking age 14 yrs
Outcome: Educational attainment age 16 yrs
Potential confounders e.g. sex, marital status, maternal smoking, paternal smoking,… etc
Missingness indicator (smoking at age 14 yrs)

Structuring the analysis: TARMOS framework [2]



1. **Plan the analysis**
   a) What is the analysis model of interest assuming there are no missing data?
   b) How are missing data going to be handled?
      - Is a complete records analysis likely to be valid?
      - Is MI likely to offer benefits over a complete records analysis?
      - Is a sensitivity analysis required, and if so how should it be framed?

2. **Conduct the analysis**
   a) Examine the data – are the features of the data consistent with the expectation outlined in the analysis plan?
   b) Conduct the analysis as per the plan – any amendments to the analysis plan should be acknowledged and justified.

3. **Report the analysis**
   a) Describe the missing data
   b) Describe how the missing data were handled in the primary and secondary analyses and provide a justification for the selected approach(es), including details of any non-standard issues e.g. non-linear terms, interactions etc.
   c) Report the results from all of the analyses and interpret in light of the missing data and the clinical relevance, commenting on any substantial differences in the inference if relevant

Choosing the tools[3]

Meth.	When to use	When to avoid
CC	• when the probability that a case is complete depends on the covariates but, given these, not the outcome — unbiased (though not fully efficient).	• when estimating the mean of an incomplete outcome if data are not MCAR • when there are good auxiliary variables (MI or IPW more efficient)
IPW	• when there are useful auxiliary variables (more efficient than CC); • Under MAR	• when MI is also valid, because IPW is generally less efficient as (i) only reweights complete cases and (ii) cannot use incomplete auxiliary variables.
MI	• when useful auxiliary variables & MAR holds (more efficient than a CC analysis); • when MAR holds	• when not confident the imputation model is (i) consistent with the scientific model and (ii) correctly specified (risk of bias)

Note on auxiliary variables

- Useful auxiliary variables need to be good predictors of the missing values.
- If they are additionally good predictors of the propensity of data to be complete, they correct for bias (if data are MAR).
- If they *only* predict the propensity of data to complete, they add noise, and may induce bias.

See Spratt *et al* [4].

Application

In ALSPAC:

- dropout is associated with many of the covariates in the analysis model (so not MCAR), and in particular the outcome (educational attainment). So CC likely biased.
- 51% have missing data on smoking status at 14 years, and we expect missingness to be associated with the outcome.
  - There are a number of strong auxiliary variables, such as smoking status at previous and later waves, which are observed when the exposure of interest is missing in some observations (but which also have missing values).
  - Given this, MI has the potential to reduce bias and improve precision over a complete records analysis.
- It's reasonable that missingness in smoking at 14 is associated with smoking itself (even given other covariates) — hence we should conduct sensitivity analysis.

Sensitivity analysis using MI

- A simple way to allow different relationships in the complete and incomplete records is using a pattern- mixture approach.

- We assume that the value of the variable (or log odds, conditional on the other variables in the imputation model) is different in those observed and unobserved by a value, δ.

Sensitivity analysis using MI

- A simple way to allow different relationships in the complete and incomplete records is using a pattern- mixture approach.

- We assume that the value of the variable (or log odds, conditional on the other variables in the imputation model) is different in those observed and unobserved by a value, δ.

- In this example, we proceed as follows [7]:
  1. Perform MI under MAR;
  2. In each imputed dataset, regress `smoke` on the other variables in the imputation model;
  3. Add δ to the constant term from Step 2.
  4. Re-impute the missing values of `smoke`
  5. Refit the scientific model to each imputed dataset and combine the results using Rubin's rules.

Results

Analysis	N	Estimated smoking		% miss. smoke. values
		effect (95% CI)		imputed as ‘smokers’
Primary: MI	14,684	−10.8	(−12.2, −9.4)	13.3
Supp.: CC	3,153	−7.9	(−9.1, −6.7)	N/A
Sens, $\delta = 0.1$	14,684	−10.9	(−12.4, −9.4)	14.2
Sens, $\delta = 0.5$	14,684	−11.0	(−12.3, −9.6)	18.1
Sens, $\delta = 10$	14,684	−4.3	(−4.7, −3.8)	99.8

Choosing the number of imputations

- Much has been written about how to choose the number of imputations (e.g., [5], ch. 2).
- Ideally, we should choose the number of imputations so that our results are reproducible at the presented precision.

Choosing the number of imputations

- Much has been written about how to choose the number of imputations (e.g., [5], ch. 2).
- Ideally, we should choose the number of imputations so that our results are reproducible at the presented precision.
- Stata calculates the Monte-Carlo error on all MI results.
- We can use this to simply estimate how many more imputations are needed so that the results are reproducible at the presented precision.

Example

- Suppose that after k_1 (say 15) imputations the Monte-Carlo estimate of our p-value has error $e_1 = 0.003$.
- Then, approximately,

  $$\sqrt{\frac{\nu^2}{k_1}} = e_1,$$

  so that $\nu = e_1\sqrt{k_1} = 0.003 \times \sqrt{15} \approx 0.012$.
- If we want an error of, say, $e_2 = 0.001$, then in total we need

  $$k_2 = \left(\frac{\nu}{e_2}\right)^2 = \left(\frac{0.012}{0.001}\right)^2 = 144$$

  imputations, in other words at least $k = 144 - 15 = 129$ additional imputations.

Suggestions for structured reporting[1]

- There should be a pre-specified plan detailing how missing data will be handled in the analysis
- Any deviations from the pre-specified plan should be acknowledged and justified

The *Methods* section should:

- Describe the pattern of missing values
- For each of primary, supplementary and sensitivity analyses, state the assumptions and provide enough detail for reproducibility
- Provide a justification for the selected approaches, including details of any non-standard issues e.g. non-linear terms, interactions

[1] using supplementary material as necessary

...ctd

The *Results* section should:

- report the extent of missing data using appropriate summaries, and the reasons for missing values where possible.
- report the results from all of the analyses

The *Discussion* section should

- discuss of the plausibility of missing data assumptions, and what this means for the clinical interpretation, especially if there are substantial differences in the inferences from different analyses.

Summary

- Analysts should be clear about the scientific model(s) (i.e. how to do the analysis if no data are missing). Causal graphs are helpful here.
- Exploring the pattern of missing data, and the predictors of a complete record (both from the data and in discusion with collaborators) is crucial to establish whether a complete record analysis is sufficient.
  - causal graphs can be useful here too [6].
- Little *et al* [3] give practical guidance for choosing between CC, IPW, MI.
- Sensitivity analysis will often be needed. This sounds scary, but is actually quite simple with MI [7], [8].
- Reporting remains a challenge — often the analysis is not reproducible — the above suggestions are intended as a first step towards an agreed framework.

References I

[1] Andy Boyd, Jean Golding, John Macleod, Debbie A Lawlor, Abigail Fraser, John Henderson, Lynn Molloy, Andy Ness, Susan Ring, and George Davey Smith.
Cohort Profile: The 'Children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children.
International Journal of Epidemiology, 42(1):111–127, 04 2012.

[2] Katherine J. Lee, Kate M. Tilling, Rosie P. Cornish, Roderick J.A. Little, Melanie L. Bell, Els Goetghebeur, Joseph W. Hogan, and James R. Carpenter.
Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework.
Journal of Clinical Epidemiology, 134:79–88, 2021.

[3] R J A Little, J R Carpenter, and K Lee.
A comparison of three popular methods for handling missing data: complete case analysis, wieghting and multiple imputation *sociological methods and research (accepted)*, 2022.

[4] M Spratt, J R Carpenter, J A C Sterne, B Carlin, J Heron, J Henderson, and K Tilling.
Strategies for Multiple Imputation in Longitudinal Studies.
American Journal of Epidemiology, 172:478–487, 2010.

[5] J R Carpenter and M G Kenward.
Multiple Imputation and its Application.
Chichester, Wiley, 2013.

[6] Rhian M Daniel, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola.
Using causal diagrams to guide analysis in missing data problems.
Statistical Methods in Medical Research, ?:?, 2011.

[7] James R. Carpenter and Melanie Smuk.
Missing data: A statistical framework for practice.
Biometrical Journal, 63(5):915–947, 2021.

References II

[8] J R Carpenter.
Multiple Imputation-Based Sensitivity Analysis, in W iley *Statistics Reference Online*, ISBN: 9781118445112; doi:10.1002/9781118445112.stat07852, 2019.