# Missing data: best practice and beyond in flexible modelling and causal inference

James R. Carpenter

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

London School of Hygiene and Tropical Medicine &

MRC Clinical Trials Unit at UCL

www.missingdata.lshtm.ac.uk

Support from ESRC, MRC, DFG, EU

$29^{th}$ July 2020

# Acknowledgements

This work is joint with:

Sébastian Bailly (INSERM 1042, Grenoble)

Clemênce Leyrat (LSHTM)

Matteo Quartagno (MRC-CTU at UCL) — `jomo`

Elizabeth Williamson (LSHTM)

We are grateful to STRATOS for the invitation to contribute to this session.

JC co-chairs STRATOS Missing Data Topic Group and EW is a member of the STRATOS Design Topic Group.

# Overview

- Handling missing data: a perspective of the state of the art
- Using multiple imputation for missing data in non-linear & hierarchical models
- Missing data in marginal structural models
  - Illustrative example
  - Common methods used and their assumptions
  - Simulation study
- Discussion

# A perspective on the state of the art

- Rubin published his classification of missing data mechanisms in 1976 [1], and his classic book on multiple imputation for surveys in 1987 [2].
- There are two algorithms for multiple imputation of missing data: *joint modelling (JM)* (c.f. [3]) and *full conditional specification (FCS)* (c.f. [4, 5]). Joint modelling is more natural for multilevel/hierarchical structures, and FCS for cross-sectional data, involving a mix of variable types (e.g. interval censored variables) and questionnaire features such as skips.
- Either FCS or JM (and often both) are now implemented in all standard software packages.

# Challenges for practitioners

- The key challenge for practitioners is choosing an appropriate imputation model.
- This needs to be consistent with the scientific model. Analysts also need to choose which auxiliary variables, not in the scientific model, to additionally include in the imputation model.
- The STRATOS missing data topic group has developed guidance (https://arxiv.org/abs/2004.14066); see also [6], and STRATOS workshop at August 2020 ISCB.

# Further challenge: handling non-linear relationships

A further challenge is how to handle non-linear relationships in the multiple imputation, particularly if combined with multilevel structure, e.g. for observations $i$ on units $j$ :

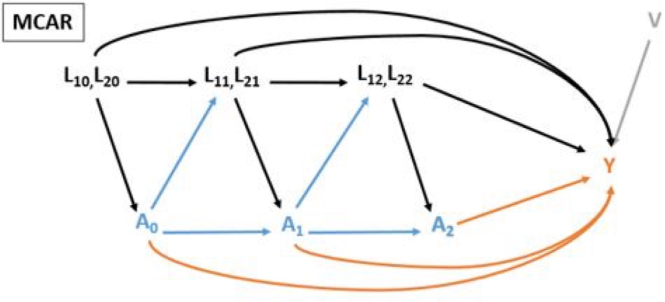
$$
\begin{aligned}
Y_{i,j} &= (\beta_0 + u_{0,j}) + (\beta_1 u_{1,j}) x_{1,i,j} + \beta_2 x_{2,i,j} + \beta_3 x^2_{2,i,j} + \epsilon_{i,j} \\
\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} &\overset{iid}{\sim} N(\mathbf{0}, \Sigma) \\
\epsilon_{i,j} &\overset{iid}{\sim} N(0, \sigma^2_e)
\end{aligned}
$$

# Solutions

- In single-level (cross sectional) models, one approach with FCS is to include all interactions in the imputation models [7]; however, this can get cumbersome, and is not always appropriate.
- A theoretically preferable approach is to construct imputations consistent with the non-linear structure [8]; this is available as `smcfcs` in R and Stata.
- Building on the approach proposed by [9], this has now been implemented in the R package for multilevel modelling, `jomo` [10], as `smcjomo`.

# Missing data in Marginal Structural Models (MSMs)

MSMs were developed by Robins and co-workers, to estimate intervention effects from observational data affected by time varying confounding, for example:




where $Y$ is the continuous outcome, $A_0, A_1, A_2$ time-varying binary treatments, $L$'s are time varying confounders (one binary, one continuous) and $V$ is an additional variable predictive of outcome.

# The challenge

- Estimation follows a two-stage process:
  1. weights — based on the inverse of the probability of a patient receiving the treatment they actually received — are estimated to create a pseudo-population in which treatment and confounders are independent.
  2. a weighted regression (using the weights derived in the first stage) including only the treatment history can be used to obtain estimate the causal effect of the treatment regimens of interest.
- In practice, the weights can be estimated using pooled logistic regression, in which each person-time interval is considered as an observation. This pooled logistic regression model must include the confounders and their relevant interactions to ensure the distributions of confounders are balanced between treatment groups
- However, there is no consensus on the appropriate method to use when the confounder data have a non-monotone missingness pattern.

# Five commonly used missing data methods for partially observed confounders

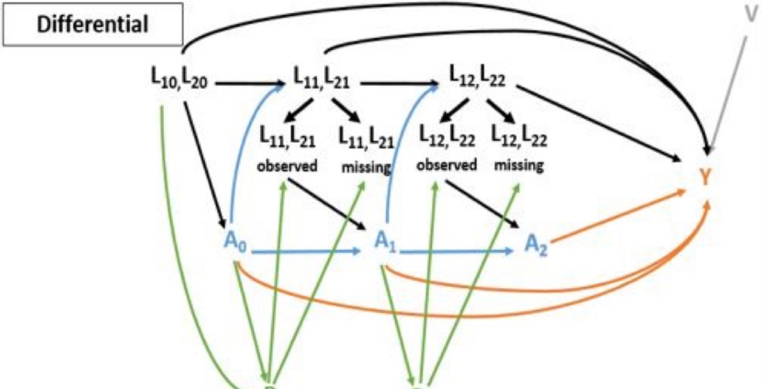
| Method | Missing data on … | Assumptions | Unbiased in MSMs when… | Advantages | Limitations |
|---|---|---|---|---|---|
| Complete case (CC) | Covariates<br>Treatment<br>Outcome | Missing data are MCAR | MCAR | Straightforward | May be inefficient because of the loss in sample size |
| Last Observation Carried Forward (LOCF) | Covariates<br>Treatment<br>Outcome except at baseline | The true, but missing, value is the same as the last available measurement<br><br>OR<br><br>The treatment decision depends on the previous available measurement rather than the true (unobserved) one | Constant | Straightforward<br><br>Discards fewer patients from the analysis than CC | It can lead to too narrow confidence intervals<br><br>Patients are discarded if baseline measurements are missing |

# ...continued

| | | | | | |
|---|---|---|---|---|---|
| Multiple imputation (MI) | Covariates Treatment Outcome | Missing data are MAR[a]<br><br>The imputation model is correctly specified | MCAR<br>MAR\|A,L<br>MAR\|A,L,Y<br>MAR\|A,L,V | Maintains the original sample size | May be computationally intensive<br><br>Challenging for a large number of time points |
| Inverse-probability-of-missingness weighting (IPMW) | Covariates Treatment Outcome | Missing data are MAR given the treatment and the covariates, but not the outcome<br><br>The weight model is correctly specified | MCAR<br>MAR\|A,L<br>MAR\|A,L,V<br>Constant | Faster than MI for large datasets<br><br>Weights simultaneously address confounding and missing data | May be inefficient for small and moderate sample size |
| Missingness Pattern Approach (MPA) | Covariates | The partially observed covariate is no longer a confounder once missing<br><br>e.g. the treatment decision depends on the confounder value only when a measurement is available | Differential | Relatively simple to implement<br><br>Assumptions do not relate to Rubin's taxonomy so may work when standard methods do not | Does not handle missing data on the exposure or outcome<br><br>Challenging when the number of missingness patterns is large |

# Details of simulation study

- We simulated data from $n = 10,000$ individuals with about 40% missing data in the confounders, and used 5000 replications.
- Values were informed by a motivating study of sleep apnoea. Full details in the supplementary materials of the forthcoming paper [11].



# Results

True MSM is:

$$Y_i = \beta_{int} + 1.163a_{0,i} + 1.677a_{1,i} + 2a_{2,i} + \epsilon_i;\ \epsilon_1 \overset{iid}{\sim} N(0, \sigma^2)$$

Absolute bias and coverage rate (%) for the 5 methods to handle missing data in each scenario considered at time 2:

| Scenario | CC | | LOCF | | MPA | | MI | | IPMW | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Bias* | *Coverage* | *Bias* | *Coverage* | *Bias* | *Coverage* | *Bias* | *Coverage* | *Bias* | *Coverage* |
| MCAR | 0.000 | 98.1 | 0.100 | 80.2 | 0.094 | 82.8 | 0.002 | 97.9 | 0.000 | 97.0 |
| MAR\|AL | 0.002 | 98.1 | 0.122 | 71.3 | 0.115 | 75.6 | 0.004 | 97.7 | 0.002 | 97.2 |
| MAR\|ALY | -0.547 | 0.0 | 0.000 | 98.7 | -0.437 | 0.0 | 0.002 | 98.3 | -0.663 | 0.0 |
| MAR\|ALV | -0.002 | 97.9 | 0.104 | 79.3 | 0.103 | 79.7 | 0.001 | 97.8 | -0.003 | 96.8 |
| Constant | 0.001 | 98.1 | 0.001 | 97.8 | 0.165 | 49.9 | 0.095 | 83.8 | 0.001 | 97.6 |
| Differential | -0.004 | 97.9 | 0.034 | 96.2 | 0.001 | 96.8 | -0.048 | 94.9 | -0.003 | 96.9 |

Full details in [11]

# Summary & Discussion

- Multiple imputation provides a very general, applicable, method for handling missing data. It is particularly useful with missing covariates.
- The most common challenges are making imputation models consistent with the substantive model, and choosing appropriate auxiliary variables. The former can be addressed using `smcfcs` (Stata and R) and `jomo` and `smcjomo` in R.
- In MSMs, a variety of proposals for handling missing data have been made; we summarised them and reviewed their assumptions.
- Our results show that:
  - It is important to reflect carefully on the likely missing data mechanisms. If they assumptions of one of the similar methods really hold, this is preferable
  - MI had the best across-the-board performance, and is always worth doing, at least as a secondary analysis.
  - Better coverage could be obtained by accounting for weight estimation (e.g., with MI, [12]).

# References I


[1] D B Rubin.
Inference and missing data.
*Biometrika*, 63:581–592, 1976.

[2] D B Rubin.
*Multiple imputation for nonresponse in surveys*.
New York: Wiley, 1987.

[3] J L Schafer.
*Analysis of incomplete multivariate data*.
London: Chapman and Hall, 1997.

[4] S van Buuren, J P L Brand, C G M Groothuis-Oudshoorn, and D B Rubin.
Fully conditional specification in multivariate imputation.
*Journal of Statistical Computation and Simulation*, 76:1049–1064, 2006.

[5] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger.
A multivariate technique for multiply imputing missing values using a sequence of regression models.
*Survey Methodology*, 27:85–95, 2001.

[6] J R Carpenter.
Missing data: a framework for practice.
*Biometrical journal (revision under review)*, xx:yyyy–zzzz, 2020.

[7] K Tilling, E Williamson, M Spratt, J A C Sterne, and J R Carpenter.
Appropriate inclusion of interactions was needed to avoid bias in multiple imputation.
*Journal of Clinical Epidemiology*, 80:107–115, 2016.

[8] J W Bartlett, S Seaman, I R White, and J R Carpenter.
Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model.
*Statistical Methods in Medical Research*, 24:462–487, 2015.

# References II


[9] H Goldstein, J R Carpenter, and W Browne.
Fitting multilevel multivariate models with missing data in responses and covariates, which may include interactions and non-linear terms.
*Journal of the Royal Statistical Society, Series A*, 177:553–564, 2014.

[10] M Quartagno, S Grund, and J R Carpenter.
`jomo`: a flexible package for two-level level joint modelling multiple imputation.
*The R Journal*, XX:YYY–ZZZ, 2020.

[11] C Leyrat, J R Carpenter, S Bailly, and E J Williamson.
Common methods for missing data in marginal structural models: what works and why.
*Revision submitted to American Journal of Epidemiology*, XXX:YYY–ZZZ, 2020.

[12] M Quartagno, J R Carpenter, and H Goldstein.
Multiple Imputation with Survey Weights: A Multilevel Approach.
*Journal of Survey Statistics and Methodology*, 09 2019.
smz036.