Inducing missing data in simulated datasets:–

the challenge of MAR and the need for replicability

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

https://missingdata.lshtm.ac.uk

Support from UK MRC,
grants MC_UU_00004/07, MC_UU_12023/29

Synthetic data workshop, Freiburg, 3rd June 25

Acknowledgements

Discussions with:

Tim Morris (UCL)

Overview

- Introduction: goals for synthetic data
- Two challenges:
  1. Simulating data under MAR / MNAR
  2. The need for replicability
- Discussion

Introduction : Goals

Ideally, we wish to use our synthetic data to both

1. test specific properties of methods on missing data issues *in isolation*, and
2. establish the performance of a method in a ‘recognised’ setting, that is an acceptable representation of actual data: in particular in posing a number of analytical challenges simultaneously.

This poses two particular challenges:

- How to simulate plausible missing/coarsening at random mechanisms, and
- The need for replicability.

Challenge of simulating MAR mechanisms

Suppose we wish to understand predictors/causes of not having any educational qualifications by age 23 from the 1956 UK Birth cohort study.

We might consider the following variables:

`fammove`	Number of family moves since child's birth (from 0 to 9) (from 0 to 35, high is good)
`bsag`	Behavioural score (from 0 to 70, high indicates more behavioural problems at 7 years)
`sex`	Child's sex (0 — boy; 1 — girl)
`care`	In care before 7 years old (0 — no; 1 — yes)
`soch7`	In social housing before 7 years old (0 — no; 1 — yes)
`bwt`	birthweight (ounces)
`mo_age`	Mother's age at child's birth (centred at 28 years)
`noqual2`	Binary variable, Child has no qualifications at 23 years of age (0 — at least 1 qualification; 1 — no qualifications)

With the following missingness (=0) patterns

```
17,631 observations
   % |  sex  mo_age  bwt  care  soch7 bsag  noqual2  fammove
  ---+-------------------------------------------------------
  40 |  1    1       1    1     1     1     1        1
  12 |  1    1       1    1     1     1     1        0
   8 |  1    1       1    1     1     1     0        1
   8 |  1    1       1    0     0     0     0        0
   6 |  1    1       1    1     1     1     0        0
   4 |  1    1       1    1     1     0     1        1
   3 |  1    1       1    0     0     1     1        1
   3 |  1    1       1    1     1     0     0        0
   2 |  1    1       1    1     1     0     1        0
   1 |  1    1       1    0     0     1     1        0
   1 |  1    1       1    0     0     1     0        0
   1 |  1    1       1    1     1     0     0        1
   1 |  1    1       0    1     1     1     1        1
   1 |  1    1       1    0     0     0     1        1
  <1 |  1    1       1    0     0     0     0        1
  <1 |  1    0       0    1     1     1     1        1
--- Over 30 more patterns with < 1% missing -----------
```

What patterns might be operating?

Missingness is non-monotone, so we can't assume a common MAR mechanism.

For fammove (pattern 2) we see predictors are:

```
------------------------------------------------------
miss_fammove |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
      mo_age |  -.0103527   .0042639    -2.43   0.015
        care |  -.4361036   .1646932    -2.65   0.008
       soch7 |   .0946555   .0506923     1.87   0.062
     noqual2 |  -.2831488   .0540953    -5.23   0.000
       _cons |    1.43391   .0337158    42.53   0.000
------------------------------------------------------
```

- future variables (noqual2) 'predict' past missing values.
    - there is either a common unmeasured cause *or* fammove is MNAR *or* both.

How to simulate MAR?

- In a frequentist way, so frequentist properties of methods can be evaluated. Rubin's original conception conditioned on the observed data [1, 2].
- Focus on the key patterns.
- Consistent with how the data were collected (e.g. survey waves, questionnaire structure, merged data).
- For non-monotone MAR, there can't be a common MAR mechanism. Consider:
    - Use observed data to split the data into patterns;
    - Make data MAR within the pattern (use common cause if wish to include 'future' predictors).
- While this might sound contrived, it is often reasonable.

Example: NCDS - first three patterns

pattern 1: missing fammove;
pattern 2: missing noqual2;
pattern 3 only sex, bwt, mo_age observed.

Birth data predictors at 5% level:

```
-------------------------------------------------------
    mpattern |      Coef.   Std. Err.      z    P>|z|
-------------+-----------------------------------------
0            |  (all observed)
-------------+-----------------------------------------
1            |  (fammove missing)
      mo_age |   .0137294   .0041246     3.33   0.001
-------------+-----------------------------------------
2            |  (noqual2 missing)
         sex |  -.2505346   .0559124    -4.48   0.000
-------------+-----------------------------------------
3            |  (only hospital data observed)
         sex |   -.261636   .0580699    -4.51   0.000
         bwt |  -.0327923   .0012812   -25.60   0.000
-------------------------------------------------------
```

How to simulate MNAR?

Two broad approaches (again, within patterns):

- simulate a cause of missing values (possibly also a cause of future missing values), then do not disclose this variable.
- take missing at random mechanism, and introduce additional dependence on the variable which will be partially observed.

It's important to respect the patterns when you do this, because analyses looking for the 'effect' of a variable are unlikely to be substantially biased if the MNAR mechanism is common across the levels of the variable.

Replicability is important

Some authors (e.g. [3]) have suggested that, to simulate missing data in order to evaluate methods (such as MI), only a single dataset is needed, to which the missingness mechanism can be applied multiple times.

Although this is attractive, it is not a reliable approach for evaluating the frequentest properties of estimators.

For example, Morris *et al* [4] argue:

- it is not possible to fix the dataset unless a selection model is used to generate missing values;
- this is unable to estimate the consequences of discrepancies between the models assumed by the imputer and the analyst.
    - e.g. if imputer assumes (correctly) simpler model than analyst, superefficiency property of MI will be missed.

- exception if interest only in bias

Discussion

- Simulating a realistic missing data mechanism for cohort data is not straightforward: for MAR consider contextually plausible patterns.
- Replicating both the data generating mechanism and the missing data mechanism is required to establish frequentist properties.
- if the full synthetic/simulated data represents a population, then a single realisation of the missingness mechanism might be acceptable, provided 'observed' data (including any missing values) are always re-drawn from that population.
- Another option, that does not require simulating a master dataset, is that users submit their model of interest, which is then fitted to the original data and used to generate imputed dataest(s) which are returned to the user.
- These points apply to all forms of coarsened data [5]

References

[1] S Seaman, S Galati, D Jackson, and J Carlin.
What Is Meant by ‘Missing at Random’?
Statistical Science, 28:257–268, 2013.

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

[3] H Oberman and G Vink.
Towards a standardized evaluation of imputation methodology.
Biometrical Journal, 6:xxx–yyy, 2024.

[4] T Morris, I R White, S Cro, J R Carpenter, and T M Pham.
Comment on oberman and vink: ‘Should we fix or simulate the complete data in simulation studies evaluating missing data methods?’.
Biometrical Journal, 66:xxx–yyy, 2024.

[5] D F Heitjan and D B Rubin.
Ignorability and coarse data.
The Annals of Statistics, pages 2244–2253, 1991.