Statistical modelling with missing data using multiple imputation

James R. Carpenter
London School of Hygiene & Tropical Medicine
james.carpenter@lshtm.ac.uk

www.missingdata.org.uk

Support for JRC from ESRC, MRC, & German Research Foundation

March 23, 2010

Overview 2 / 40

Acknowledgements

John Carlin, Lyle Gurrin, Helena Romaniuk, Kate Lee (Melbourne)
Mike Kenward, Harvey Goldstein (LSHTM)
Geert Molenberghs (Limburgs University, Belgium)
James Roger (GlaxoSmithKline Research)
Sara Schroter (BMJ, London)
Jonathan Sterne, Michael Spratt, Rachael Hughes (Bristol)
Stijn Vansteelandt (Ghent University, Belgium)
Ian White (MRC Biostatistics Unit, Cambridge)

Parts of this session relevant to clinical trials are based on a peer-reviewed book, 'Missing data in clinical trials — a practical guide' (joint with Mike Kenward), commissioned by the UK National Health Service, available free on-line at www.missingdata.org.uk. Also available from the website are expanded notes and practicals.

2 / 40

Outline

Session 1:

- Issues raised by non-trivial proportions of missing data; the central role of assumptions
- A principled approach to missing data
- Missing data mechanisms: unpacking common jargon
- When is complete case analysis OK?
- Summary

Session 2:

- Multiple imputation: an intuitive introduction
- Example: cancer epidemiology
- Pitfalls
- Sensitivity analysis
- Publishing analyses that use multiple imputation
- Discussion

3 / 40

Issues raised by missing data 5 / 40

Why is this necessary?

Missing data are common. However, they are usually inadequately handled in both observational and experimental research.

For example, Wood *et al* (2004)[7] reviewed 71 published BMJ, JAMA, Lancet and NEJM papers.

- 89% had partly missing outcome data.
- In 37 trials with repeated outcome measures, 46% performed complete case analysis.
- Only 21% reported sensitivity analysis.

Further, CONSORT[a] guidelines state that the number of patients with missing data should be reported by treatment arm [exposure group].

But Chan *et al* (2005)[2] estimate that 65% of studies in PubMed journals do not report the handling of missing data.

5 / 40

[a] Consolidated Standards of Reporting Trials, an international guideline for reporting trials. See http://www.consort-statement.org

In addition...

Sterne *et al* (2009)[6] searched four major medical journals (NEJM, Lancet, BMJ, JAMA) from 2002–7 for articles involving original research in which multiple imputation was used. They reported

> 'Although multiple imputation is increasingly regarded as a standard method, many aspects of its implementation can vary and very few of the papers that were identified provided full details of the approach that was used, the underlying assumptions that were made, or the extent to which the results could confidently be regarded as more reliable than any other approach to handling the missing data (such as, in particular, restricting analyses to complete cases).'

See also Klebanoff and Cole (2008)[4].

6 / 40

A starting point: the E9 guideline on conducting RCTs, 1999

The International Conference on Harmonisation (ICH) issued the E9 guideline on statistical aspects of carrying out and reporting trials in 1999[3]; see also www.ich.org.

With regard to missing data, in summary it says:

- Missing data are a potential source of bias.
- Avoid if possible (!)
- With missing data, a trial[study] may still be regarded as valid if the methods are *sensible*, and preferably *predefined*.
- There can be no universally applicable method of handling missing data.
- The sensitivity of conclusions to methods should thus be investigated, particularly if there are a large number of missing observations.

These principles are generally applicable.

The question is, how do we use them in analyses?

Study validity and sensible analysis

Data are sometimes missing by design, but our focus is on observations we intended to make but did not.

The sampling process involves both the selection of the units, and the process by which observations on those units [i.e. the *items*] become missing — the *missingness mechanism*.

Thus for sensible inference, we need to take account of the missingness mechanism

By *sensible* we mean:

- Frequentist: nominal properties hold. Eg:
  Estimators consistent; confidence intervals attain nominal coverage.
- Bayesian:
  We have used the appropriate likelihood (usually the same as for the frequentist analysis). Thus the posterior distribution is unbiased, and correctly reflects the loss of information due to the missingness mechanism.

Why there can be no universal method

In contrast with the sampling process, which is usually known, the missingness mechanism is usually unknown.

The data alone cannot usually definitively tell us the sampling process.

Likewise, the missingness pattern, and its relationship to the observations, cannot identify the missingness mechanism.

With missing data, extra assumptions are thus required for analysis to proceed.

The validity of these assumptions cannot be determined from the data at hand.

Assessing the sensitivity of the conclusions to the assumptions should therefore play a central role.

Example: Stent vs Angioplasty trial

[5] report the following data (restenosis is a poor outcome):

		Stent	Angioplasty
Restenosis	No (Good)	54	43
	Yes (Poor)	32	37
	Unknown	24	30
Total randomised		110	110

Observed outcomes: OR in favour of stent:
1.45 (95% CI 0.78–2.70).

10 / 40

Key points for analysis

- the question (i.e. the hypothesis under investigation)
- the information in the observed data
- the reason for missing data

11 / 40

Stent trial

Consider the impact of two possible assumptions about the reason for missing data:

1. Within each arm, the odds of a good response for the missing outcomes is *exactly* the same as that among the observed outcomes.
2. In the stent arm, outcomes are missing because they are good; specifically the chance of a good outcome is 30% higher than that among the observed outcomes.

   On the other hand in the angioplasty group, the chance of a good response for the missing outcomes is *exactly* the same as that among the observed outcomes.

12 / 40

Implications

		Assumption 1		Assumption 2	
		Stent	Angio.	Stent	Angio.
Outcome	Good	69	59	74	59
	Poor	41	51	36	51
	Total	110	110	110	110
		OR: 1.45; (95% CI 0.85–2.48)		OR: 1.78; (95% CI 1.03–3.08)	

13 / 40

A principled approach

14 / 40

Towards a systematic approach

Given this example we might conclude that studies with non-trivial missing data must be discarded.

However, although some information is irretrievably lost, we can often salvage a lot.

The success of the salvage operation depends on:

1. whether we can identify plausible reasons for the data being missing (called *missingness mechanisms*), and
2. the sensitivity of the conclusions to different missingness mechanisms.

A possible systematic approach is as follows:

14 / 40

A systematic approach

Investigators discuss possible missingness mechanisms, say A–E, possibly informed by a (blind) review of the data, and consider their plausibility. Then

1. Under most plausible mechanism A, perform valid analysis, draw conclusions
2. Under similar mechanisms, B–C, perform valid analysis, draw conclusions
3. Under least plausible mechanisms, D–E, perform valid analysis, draw conclusions

Investigators discuss the implications, and arrive at a valid interpretation of the study in the light of the possible mechanisms causing the missing data.

For trialists, this approach broadly agrees with the E9 guideline.

15 / 40

Common jargon

Missing data mechanisms (see [1], ch. 1)

It follows from this that the missing data mechanism plays a central role in informing the analysis.

Fortunately, it turns out that there are three broad classes of mechanism, each with distinct implications for the analysis.

In practice, to obtain sensible answers, we therefore have to:

1. postulate a missingness mechanism;
2. identify its class, and
3. perform a valid analysis for that class of missingness mechanism.

We now consider these three classes.

16 / 40

I: Missing completely at random

If the missingness mechanism is unrelated to any inference we wish to draw, missing observations (items) are *Missing Completely at Random* (MCAR).

Eg: missing observations because a page of the questionnaire was missing; missing data because of a data processing error; missing data because of a change in data collection procedure.

In this case analysing only those with observed data gives sensible results.

Of course, results are less precise than when full data are observed.

Data are randomly missing

17 / 40

Example: asthma study

Response is FEV_1 as % of that predicted for a healthy patient of the same age, height etc.

	Full data 91 observations	10 obs MCAR case 1	10 obs MCAR case 2	Missing 10 largest obs
mean:	69.7	70.6	69.2	66.3
SE:	1.96	2.05	2.16	1.88

18 / 40

Plausibility of MCAR

We have said that if data are MCAR, the mechanism causing the missing data will not depend on covariates relevant to the analysis.

For well designed trials, and in some observational settings the proportion of MCAR data is likely to be small.

Although the above is *necessary* for MCAR, it is not *sufficient* to guarantee it.

In an extreme example, items may be missing from longitudinal follow-up because of a sudden, unpredicted change in response. From the observed data, items may appear MCAR. But in fact they are systematically different.

II: Missing at random

If, given the observed data, the missingness mechanism does not depend on the unseen data, then we say the missing observations are *Missing at Random* (MAR).

For example, the probability of a missing observation may depend on an earlier observation. After accounting for the earlier observation, the chance of seeing the missing observation is independent of its value.

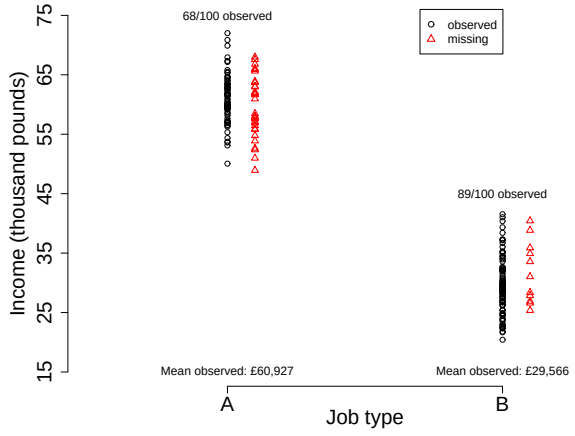
In this case simply analysing the observed data is invalid. We have two threats:

- bias — the fully observed subset of data is not representative, and
- loss of information — we have thrown away information on cases with even 1 missing observation.

Thus simple summary statistics are invalid as estimates of population parameters.

'Missing At Random' means Data are Conditionally Randomly Missing

Example: true mean income £45,000



Observed income: $\pounds 43,149$.

MAR estimate: $\dfrac{100 \times 60,927 + 100 \times 29,566}{200} = \pounds 45,246$

More on MAR

MAR is confusing jargon — it is a conditional independence statement.

Suppose we have two variables, Y (partially observed) and X (fully observed).

If we say 'Y is MAR', we mean that given, or conditional on, X, observations on Y are missing completely randomly.

- We can then get a valid estimate of $[Y|X]$ directly from those with no missing data.
- We *cannot directly* get a valid estimate of $[X|Y]$ from those with no missing data.

22 / 40

To recap...

We stress that the reason for missingness may depend on the unobserved values, but *conditional on data we observe* they are independent.

As we cannot assess any residual dependence between missingness mechanism and Y, we can never know if MAR holds.

Nevertheless, it is often a useful starting point; particularly as it makes the analysis much simpler.

23 / 40

III: Missing Not At Random

If data are neither MCAR nor MAR, we say they are Missing Not at Random (MNAR).

The missingness mechanism depends on the unobserved data, *even after taking into account all the information in the observed data*.

Under MNAR, we have to model both:

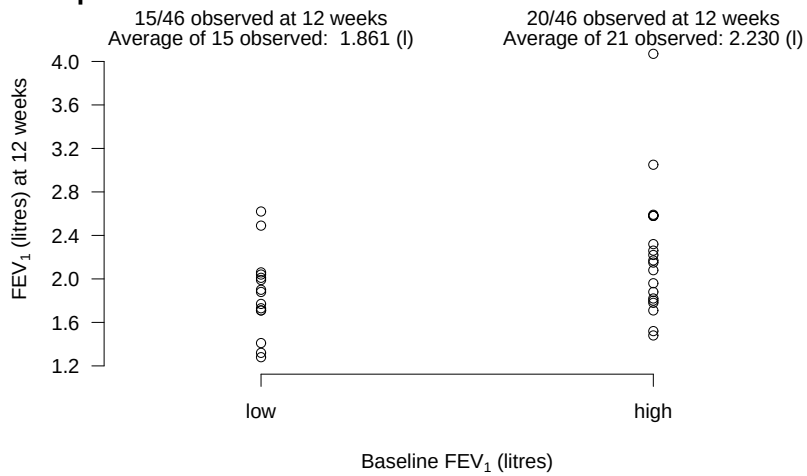
1. the response of interest, and
2. the missingness mechanism.

This is considerably harder! Often there is little to choose between various models for (2), but they may give quite different conclusions.

The 'pattern mixture' approach is sometimes a convenient way to proceed — see the example below, and the relevant section in the next session.

24 / 40

Example: asthma data



Estimated average assuming MAR: $\dfrac{46 \times 1.86 + 46 \times 2.23}{46 + 46} = 2.05(l)$

MNAR 'pattern mixture' analysis

First, notice that MAR means conditional on baseline, the distribution of observed and missing response is the same.

Suppose the asthma data are MNAR.

To estimate the average % predicted FEV, we have to make additional assumptions.

For example: suppose we say that patients who withdraw have response 10% below that predicted assuming MAR.

Then our new estimate of the average response at the end of the trial is:

$$\frac{1}{92}(1.861 \times 15 + 1.675 \times 31 + 2.230 \times 20 + 2.007 \times 26) = 1.920(l).$$

Jargon revisited

27 / 40

Implications for complete case analysis

28 / 40

Complete case analysis

Suppose we are interested in the regression of Y on multivariate X.

For each individual (unit) let $R_i = 1$ be the indicator for a complete case.

If, for all i, $[R_i|Y_i, X_i] = [R_i|X_i]$, that is the probability of a complete case, given (possibly partially observed) X_i does not depend on Y_i, then the complete case analysis is valid — though it may be quite inefficient.

To see this, consider

$$\begin{aligned}[Y_i|X_i, R_i = 1] &= \frac{[Y_i, X_i, R_i = 1]}{[X_i, R_i = 1]} \\ &= \frac{[R_i = 1|X_i][Y_i, X_i]}{[R_i = 1|X_i][X_i]} \text{ (under above assumption)} \\ &= [Y_i|X_i].\end{aligned}$$

Note we do not need the same mechanism for each individual, i, but taking this to the limit becomes very contrived!

28 / 40

Why are we interested in MCAR, MAR, MNAR?

These are the assumptions the statistical methods for the analysis of partially observed datasets — in particular multiple imputation — rest on.

Therefore, when we go beyond complete case analysis, we have to consider the missingness mechanism.

In particular, we have to consider whether including additional variables, not in the model of interest, can make the MAR assumption (and the analyses that use it) more plausible.

Such additional variables can also help us recover information.

29 / 40

Beyond a complete case analysis I: assuming MCAR

If data are MCAR[a], then a Complete Case (CC) analysis loses information through two routes:

1. An individual only has to have missing data on one variable in the model to be excluded. In practice this can lead to a lot of individuals being omitted from a CC analysis.
2. The information about the missing values in variables that are not in the model of interest (e.g. they may be on the causal path) is also excluded.

Given the cost of collecting data this is undesirable.

We can recover this information using an appropriate statistical method — e.g. multiple imputation.

30 / 40

[a] We do not require the same MCAR mechanism for each individual

Beyond a complete case analysis II: assuming MAR

It will often be plausible that the data are MAR with a mechanism depending on Y.

1. In this case the complete case analysis is invalid.
2. However, valid inference can be drawn using multiple imputation under the MAR assumption.

Further, suppose we have additional covariates Z :

1. If these have information about the missing values, we can include this using multiple imputation under MAR.
2. If, *in addition*, they are important for the missingness mechanism to be plausibly MAR, including them will reduce bias in the parameter estimates.

31 / 40

Missing data mechanisms for different individuals

For a valid analysis under MAR — i.e. without specifying the missingness mechanism explicitly — we do not need to assume the same missingness mechanism for each individual (or unit).

However, for each individual we must assume that, *given their observed data*, the probability mechanism for their unseen data is conditionally independent of their unseen data values.

32 / 40

More detail

To see this, for individual i let (Y_{Mi}, Y_{Oi}, Z_i) denote their missing and observed data (including covariates) and auxiliary variables. Let R_i be a vector of missing data indicators. Provided

$$[R_i, Y_{Mi}, Y_{Oi}, Z_i] = [R_i|Y_{Oi}, Z_i][Y_{Oi}, Y_{Mi}, Z_i]$$

then for that individual

$$[Y_{Oi}, Z_i] = \int [Y_{Mi}, Y_{Oi}, Z_i]\, dY_{Mi}$$

is their appropriate contribution to the likelihood.

Although mathematically each person can have a different missingness mechanism (i.e. form of $[R_i|Y_{Oi}, Z_i]$) in reality this is contrived.

Nevertheless if different groups of the data have different MAR mechanisms, provided we include in the analysis the key variables in those mechanisms, a missing at random analysis is plausible.

33 / 40

Beyond a complete case analysis III: assuming MNAR

Recall we are considering the regression of Y on X. We have two broad cases

1. If Y is MNAR, then a complete case analysis is invalid.
2. if Y is observed, but X is MNAR conditionally independent of Y, complete case analysis is valid, but may be inefficient.

In both cases, if we can find auxiliary variables Z which make MAR more plausible, we may prefer an MI analysis that includes this information.

However, case (2) cautions us a CC analysis may be valid when an ill-thought out analysis using a more sophisticated method is not.

Using multiple imputation with pattern mixture models provides a natural way of exploring the effect of MNAR mechanisms.

34 / 40

Notes

35 / 40

Common confusion over jargon

The term *ignorable* is sometimes wrongly identified with *MCAR*

In the literature, *ignorable* essentially means MAR. Analyses valid under MAR are also valid under MCAR.

Thus ignorable is an adjective for the missing data mechanism: if it is MAR, then — provided you are doing a likelihood analysis — you can *ignore* specifying a detailed model for it.

By contrast, analyses based on the observed data (marginal summary statistics, most generalised estimating equations) are *only valid under MCAR*.

35 / 40

MAR and randomised clinical trials

Recall that if we say Y is MAR given X this means $[Y|X]$ is the same whether Y is observed or not.

Thus if we have two patients, the first with data $[y, x]$, and the second missing Y but with the same x value, they have the *same conditional distribution* $[Y|X = x]$.

In other words, a MAR analysis gives units with missing data the same conditional distribution of 'missing $|$ observed' as unit(s) who share the same observed data.

Making these conditional distributions different gives the 'pattern mixture' model for data that under a MNAR mechanism.

36 / 40

Implication

If we think conditional distributions are different for patients with Y missing/observed, data are MNAR.

This has an important implication for clinical trials, with longitudinal response.

Simply speaking an on-treatment analysis seeks to estimate the outcome had patients adhered to the protocol.

If we can assume data are MAR, and patients withdraw when they violate the protocol (stop treatment), then given the previous slide, a likelihood based analysis (or an equivalent multiple imputation analysis) directly addresses this question.

Analyses addressing other questions need a more subtle approach.

37 / 40

Summary

38 / 40

Summary I

- Missing data introduce ambiguity into the analysis, beyond the familiar sampling imprecision.
- Extra assumptions about the missingness mechanism are needed; these assumptions can rarely be verified from the data at hand.
- Sensitivity analysis is therefore important.
- The assumptions fall into three broad classes, MCAR, MAR and MNAR, with different implications for the analysis.
- In line with E9, it is sensible to consider carefully possible missingness mechanisms, and formulate appropriate analyses, *a priori*.
- Ideally, such analyses should include assessing the sensitivity of MAR analyses to plausible MNAR mechanisms.
- This approach is preferable to using ad-hoc methods.

38 / 40

Summary II

- Complete Case (CC) analysis is only valid if the missing data mechanism does not involve the response variable; even then it will often be inefficient.
- A carefully constructed Multiple Imputation (MI) analysis is thus usually preferable to a complete case analysis, as
  1. the assumptions it rests on are likely to be more plausible;
  2. it will be more efficient, and
  3. it will reduce bias.
- However,
  1. All analyses of partially observed data rest on inherently untestable assumptions.
  2. Thus, if the MI and CC analyses differ, there is a responsibility to explain to the reader which assumptions/mechanisms are causing this.

39 / 40

References

[1] James R Carpenter and Michael G Kenward. *Missing data in clinical trials — a practical guide*. Birmingham: National Health Service Co-ordinating Centre for Research Methodology. Freely downloadable from `www.missingdata.org.uk`, accessed 15 December 2009, 2008.

[2] A-W Chan and Douglas G Altman. Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, 365:1159–1162, 2005.

[3] ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline. *Statistics in Medicine*, 18:1905–1942, 1999.

[4] M A Klebanoff and S R Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168:355–357, 2008.

[5] M P Savage, Douglas J. S. Jr, D L Fischman, C J Pepine, King S. B. 3rd, J A Werner, S R Bailey, P A Overlie, S H Fenton, J A Brinker, M B Leon, and S Goldberg. Stent placement compared with balloon angioplasty for obstructed coronary bypass grafts. Saphenous Vein De Novo Trial Investigators. *New England Journal of Medicine*, 337:740–747, 1997.

[6] J A C Sterne, I R White, J B Carlin, M Spratt, P Royston, M G Kenward, A M Wood, and J R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339:157–160, 2009.

[7] Angela M Wood, Ian R White, and Simon G Thompson. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376, 2004.

40 / 40