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

November 17, 2009

Overview

Overview 2 / 37

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 course 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.

Course aims

- Develop an intuitive understanding of key concepts in the missing data literature;
- Understand the basis of multiple imputation (MI), and its pros and cons relative to other approaches;
- Discuss how to perform MI in practice, and avoid common pitfalls;
- Learn how to frame and to carry out simple sensitivity analyses, and
- Develop an awareness of current research questions.

Course programme

Four lectures:

1. Introduction; towards a principled approach to the issues raised by missing data
2. Shortcomings of ad-hoc methods and introduction to multiple imputation
3. Lecture: More on multiple imputation
4. Sensitivity analysis after multiple imputation

We will conclude with a Q & A session.

Lecture 1 5 / 37

Outline

1. Missing data — towards a principled approach
2. Common jargon
3. More on the Missing At Random assumption
4. Discussion

Why is this necessary?

Missing data are common.

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

For example, [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 [2] estimate that 65% of studies in PubMed journals do not report the handling of missing data.

Both [4] and [7] identified serious weaknesses in the description of missing data and the methodology adopted.

It is unlikely that the situation in observational research is much better.

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

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

The same principles apply to observational research.

The question is, how do we apply them in practice?

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). Then the posterior distribution is unbiased, correctly reflects loss of information due to 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

[6] 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).

Key points for analysis

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

Stent trial

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

1. Within each arm, the chance of a good response for the missing outcomes is 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 the same as that among the observed outcomes.

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)

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:

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.

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.

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

Example: asthma study

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

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

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 many 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.

How to proceed

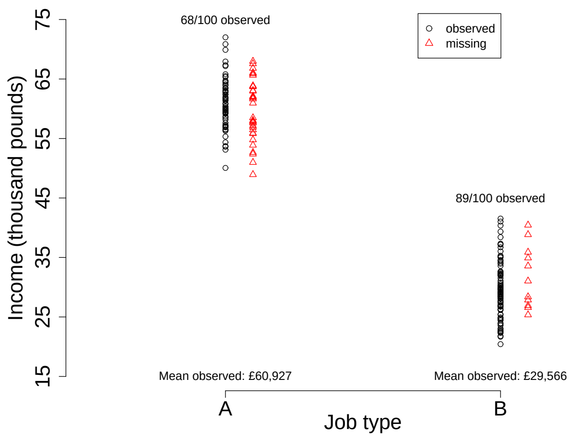
To obtain valid estimates, we have to include in the analysis the variables predictive of non-response.

For example, we may condition on them, eg. as covariates in a regression.

Of course, with several partially observed variables the issues are more complex.

‘Missing At Random’ means Data are Conditionally Randomly Missing

Example: true mean income $\pounds$45,000



Observed income: $\pounds 43,149.$

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

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 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.

A bit more on MAR

In general for any unit (typically individual) missing data are MAR if, given the observed data, the missingness mechanism does not depend on the missing values.

It turns out MAR holds even when each unit has a different MAR mechanism; this appears rather contrived in practice.

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*.

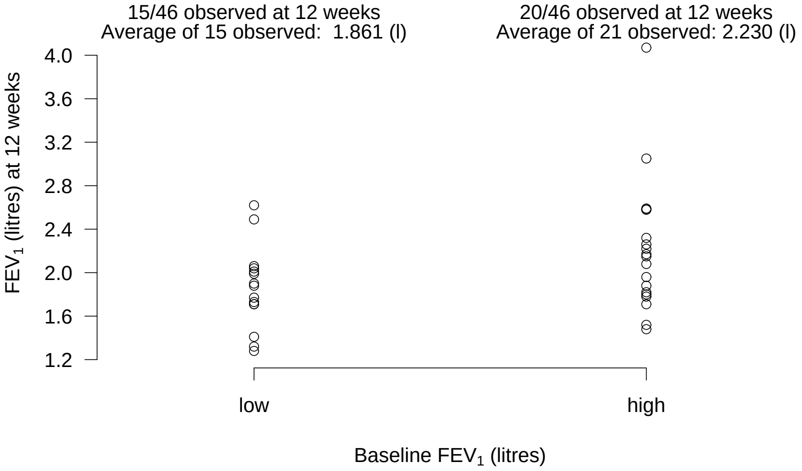
Under MNAR, we have to model both:

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 below, and Session 4.

Example: asthma data

First, a MAR analysis:



Overall average: $$\frac{46 \times 1.86 + 46 \times 2.23}{46 + 46} = 2.05(l)$$

Now an MNAR 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. We make them in the ‘pattern mixture’ setting.

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:

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

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*.

Jargon revisited

Another look at MAR

MAR is not an inherent property of the dataset; it is an assumption we make for a particular analysis, or set of analyses, of the dataset.

Let X be baseline (complete). Let Y be the (scalar) response, and R the indicator for seeing Y.

Assuming MAR, $[R|Y, X] = [R|X]$.

Rearranging, this implies $[Y|X, R] = [Y|X]$.

In other words the conditional distribution of the data is the same in the two groups of patients (those whose Y is observed, and those whose Y is missing).

Thus we can get a valid estimate of this distribution *from the observed data*. This is a key observation, which multiple imputation builds on.

With missing covariates in a regression, things are more complicated. We will address these issues as we go on.

32 / 37

More on MAR

Recall MAR means $[Y|X]$ is the same whether Y is observed or not.

This means that 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 are MNAR — see Session 4.

33 / 37

Implication

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

This has an interesting 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.

34 / 37

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, in advance.
- Ideally, such analyses should include assessing the sensitivity of MAR analyses to plausible MNAR mechanisms.
- The above approach is preferable to using ad-hoc methods.

Summary II

- MAR analyses can often readily be done by Multiple Imputation (MI) or (sometimes directly) by Maximum Likelihood (ML); MNAR are more tricky.
- With missing responses, MAR/likelihood analyses assume a patient who drops out has the same conditional distribution of missing given observed as a patient sharing the same observed values who does not drop out.
- In trials, MAR/likelihood analyses are thus particularly well suited to per-protocol analyses (on treatment analyses), while ITT analyses generally need a more subtle approach.

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. Free from http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml, 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] R J A Little and L Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52:471–483, 1996.

[6] 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.

[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.