Introduction to Missing Data

James Carpenter
james.carpenter@lshtm.ac.uk

Mike Kenward
mike.kenward@lshtm.ac.uk

London School of Hygiene & Tropical Medicine
www.missingdata.org.uk

Support from MRC, ESRC

November 20, 2012

Acknowledgements

Harvey Goldstein (LSHTM)
Rachael Hughes (Bristol)
Mike Kenward (LSHTM)
Geert Molenberghs (Limburgs University, Belgium)
Michael Spratt (Bristol)
Jonathan Sterne (Bristol)
Stijn Vansteelandt (Ghent University, Belgium)
Ian White (MRC Biostatistics Unit, Cambridge)
Many colleagues at GSK, especially Carly Donovan (Neuro. MDC, GSK)

Much of this material comes from '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.

Outline

1. Missing data — the need for a principled approach
2. Jargon unpacked
3. De facto (ITT) & De jure (Per-protocol) analyses
4. Discussion

Why is this necessary?

Missing data are common.

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

For example, [4] reviewed 71 ‘recently’ 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 guidelines state that the number of patients with missing data should be reported by treatment arm.

But [1] estimate that 65% of studies in PubMed journals do not report the handling of missing data.

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

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

The ICH E9 guideline, 1999

- Missing data are a potential source of bias
- Avoid if possible (!)
- With missing data, a trial 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

Guidelines downloadable from www.ich.org

The same principles apply to epidemiological research.

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

The EMA guideline, 2010

Guideline on Missing Data in Confirmatory Trials

Effective from 1 January 2011

Importance of

- considering missing data at the trial design stage
- specification in the protocol
- sensitivity analyses

Guideline downloadable from www.ema.europa.eu

The NRC / NAS document

The Prevention and Treatment of Missing Data in Clinical Trials

- Requested by the FDA
- Issued July 2010
- Importance of the *Estimand*
- Collecting data after withdrawal from treatment

Study validity and sensible analysis

Missing data are observations we intended to make but did not.

The sampling process involves both the selection of the units, and the process by which observations 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:
  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

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

		Stent	Angioplasty
Restenosis	No	54	43
	Yes	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, and population, under investigation) NRC call this the *estimand*.
- the information in the observed data
- the reason for missing data

Stent trial revisited

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

1. The reason for missing data does not depend on treatment or outcome; it is completely random
2. In the stent arm, outcomes are missing because they are good; specifically the chance of a good outcome 30% higher than amongst those observed.

   Conversely, in the angioplasty group, outcomes are missing completely randomly.

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 may conclude that trials 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 the following:

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 trial.

This approach broadly agrees with the E9 guideline.

Missing data mechanisms

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 about the treatment effect, missing observations are *Missing Completely at Random* (MCAR).

Eg: missing observations because of equipment failure at a clinic; patient could not attend because child was unwell.

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

Of course, results are less precise than when full data is 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 92 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

Plausibility of MCAR

For well designed studies, the proportion of MCAR data is likely to be small.

The chance of missing data will not depend on any observed covariates.

Although these points are *necessary* for MCAR data, they are not *sufficient* to guarantee it.

In an extreme example, patient withdrawal may depend on a sudden, unpredicted change in response. From the observed data, patients may appear MCAR. But in fact they are systematically different.

In practice, missing data in trials are rarely MCAR.

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 patient withdrawal may depend on baseline. After accounting for baseline, it is independent of follow-up response.

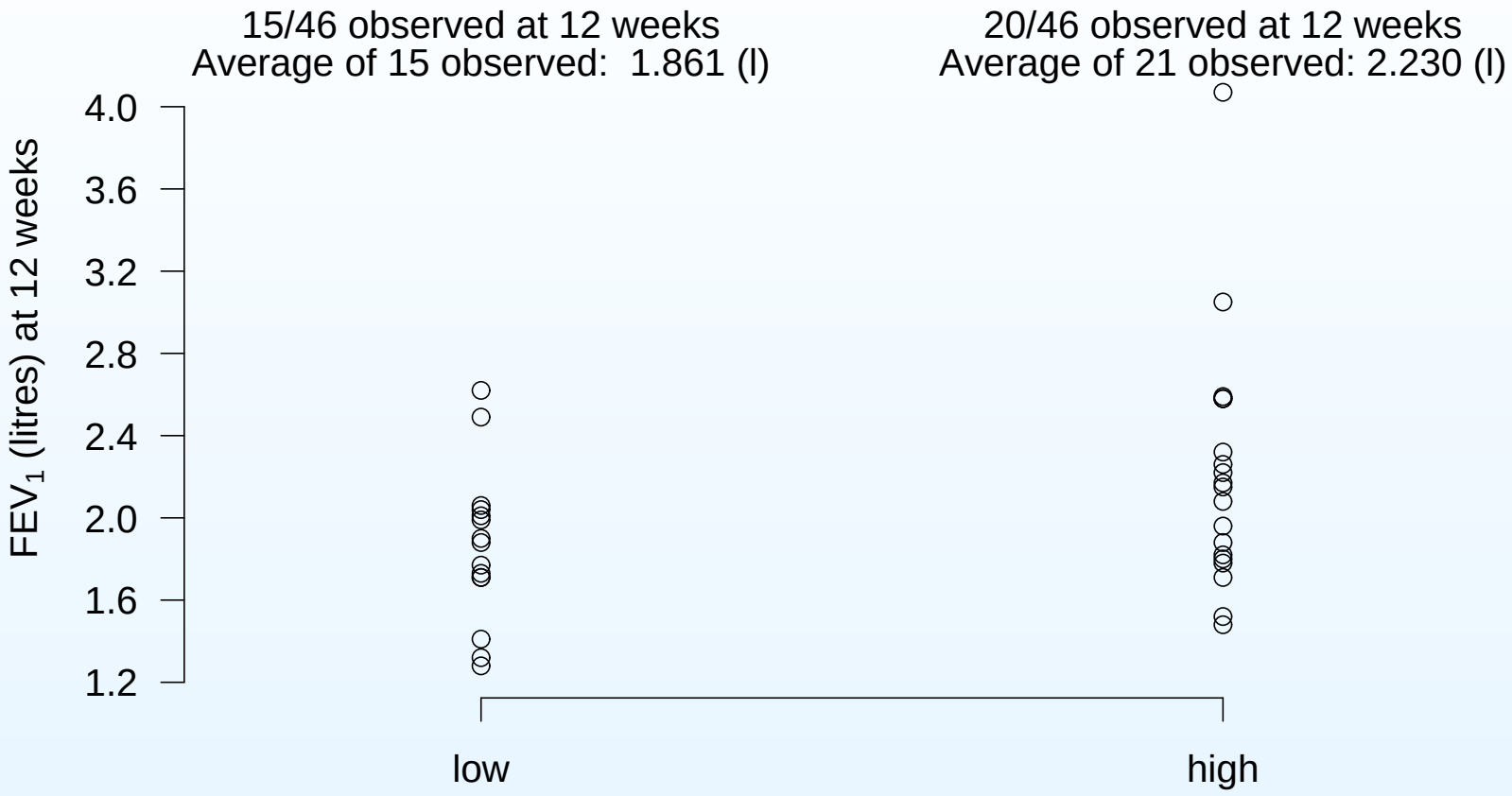
In this case simply analysing the observed data is invalid.

Thus simple summary statistics are invalid.

To obtain valid estimates, we have to include in the analysis the variables predictive of dropout. Often, we condition on them, eg. as covariates in a regression.

Data are Conditionally Randomly Missing

Example: asthma study



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

More on MAR

MAR is confusing jargon to the uninitiated — hence I prefer 'conditionally random dropout'

If we say 'Y is MAR', we mean we have fully observed variables, and conditional on these, Y is missing completely randomly.

We stress that the reason for missingness may depend on the unobserved variable, but *conditional on variables 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.

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 variable, *even after taking into account all the information in the observed variables*.

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.

Example: asthma data

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:

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

Common confusion over jargon

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

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

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

Jargon revisited

Another look at MAR

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).

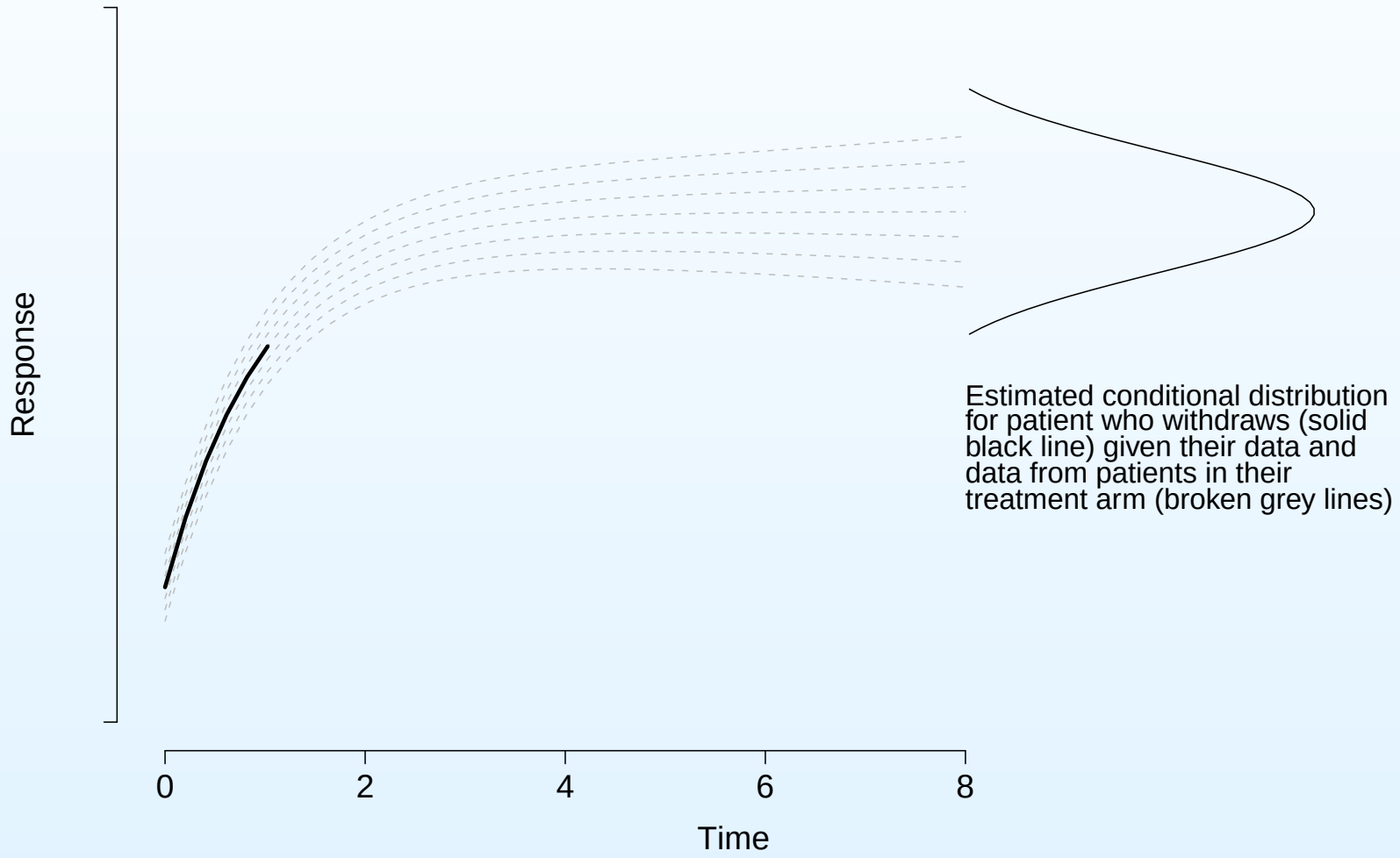
Still more on MAR

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 a patient with missing data the same conditional distribution of 'missing $|$ observed' as patient(s) who share the same observed data.

Schematic illustration



‘De-jure’ [Per-protocol or On-Treatment analysis]

Simply speaking, a per-protocol, or on-treatment, analysis seeks to estimate the outcome had patients adhered to the protocol.

This is also referred to as an *efficacy* question in contrast to an *effectiveness* question.

We define this *de-jure* question as ‘Does the treatment work under the best case scenario’.

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 directly addresses this *de-jure* question.

‘De-facto’ [ITT analysis]

With an ITT analysis, we seek to estimate the holistic effect of intending to treat a group of patients.
This is also referred to as an *effectiveness* question.

We define this *de-facto* question as ‘What would be the effect seen in practice if this treatment was applied to the population defined by the trial inclusion criteria.’
Thus if patients withdraw from treatment, but are still followed up, we have the data we need. [Retrieved Dropout]
However, if they are lost to follow-up when they discontinue treatment, a MAR likelihood analysis is less plausible.

Often some patients will be lost to follow-up when they discontinue treatment, and others will not.

In this case, we can use this information, most easily via multiple imputation, to estimate an ITT treatment effect. Stick with the course, and see [2].

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 ICH E9 and EMA guidance on Missing Data, it is sensible to consider carefully possible missingness mechanisms, and formulate appropriate analyses, before breaking the code.
- Ideally, such analyses should include assessing the sensitivity of MAR analyses to plausible MNAR mechanisms.
- This approach is preferable to using ad-hoc methods.

Summary II

- MAR analyses can often readily be done by ML or MI; 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
- MAR/likelihood analyses are thus particularly well suited to *de-jure* analyses
- *de-facto* analyses generally need a more subtle approach

References

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

[2] R J A Little and L Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52:471–483, 1996.

[3] 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.

[4] 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.