MRC | Biostatistics Unit

Strategies for handling missing data in randomised trials

NIHR statistical meeting
London, 13th February 2012

Ian White
MRC Biostatistics Unit, Cambridge, UK

Plan

1. Why do missing data matter?
2. Popular analysis methods and their assumptions
3. Which methods are best in a RCT?
4. Intention-to-treat analysis strategy for randomised trials with missing outcomes
5. Sensitivity analysis

- Work with: James Carpenter & Stuart Pocock (LSHTM), Nick Horton (USA), Simon Thompson (BSU)

Why do missing data matter?

1. Loss of power (cf. power with no missing data)
   - can't regain lost power
2. Any analysis must make an untestable assumption about the missing data
   - wrong assumption $\Rightarrow$ biased estimates
3. Some popular analyses with missing data get biased standard errors
   - resulting in wrong p-values and confidence intervals
4. Some popular analyses with missing data are inefficient
   - confidence intervals wider than they need be

What to do: loss of power

Can't solve by analysis (but can exacerbate it!)

Approach by design:

- Minimise amount of missing data
  - good communications with participants
  - aim to follow everyone up
  - make repeated attempts using different methods
- Reduce the impact of missing data
  - collect reasons for missing data
  - collect information predictive of missing values

What to do: analysis

A suitable method of analysis would:

- Make the correct assumption about the missing data
- Give an unbiased estimate (under that assumption)
- Give an unbiased standard error (so that P-values and confidence intervals are correct)
- Be efficient (make best use of the available data)

BUT we can never be sure what is the correct assumption
→ sensitivity analyses are essential

US report: “The Prevention and Treatment of Missing Data in Clinical Trials”

- Commissioned by Food & Drug Administration
- Written by a panel of top statisticians
- National Research Council (2010)

1. Introduction and Background
2. Trial Designs to Reduce the Frequency of Missing Data
   - focus on estimands (pre-trial)
3. Trial Strategies to Reduce the Frequency of Missing Data
4. Drawing Inferences from Incomplete Data
   - covers it all
5. Principles and Methods of Sensitivity Analyses
   - lots of suggestions
6. Conclusions and Recommendations

Plan

1. Why do missing data matter?
2. **Popular analysis methods and their assumptions**
3. Which methods are best in a RCT?
4. Intention-to-treat analysis strategy for randomised trials with missing outcomes
5. Sensitivity analysis

Note: missing data are most commonly in the outcome, but may also occur in baseline covariates

How to approach the analysis

- Start by knowing:
  - extent of missing data
  - pattern of missing data (e.g. how many people with time 1 missing have time 2 observed?)
  - predictors of missing data and of outcome
- Principled approach to missing data:
  - identify a plausible assumption (needs discussion between statisticians and clinicians)
  - choose an analysis method that's valid under that assumption
- Some analysis methods are simple to describe but have complex and/or implausible assumptions

The analysis toolkit

Simple methods

- Last observation carried forward (LOCF)
- Complete-case analysis
- Mean imputation
- Missing indicator method
- Regression imputation

More complex methods

- Multiple imputation
- Likelihood-based methods
- Inverse probability weighting (IPW)

Properties of analysis methods

Method	For missing covariate	For missing outcome
LOCF	Not applicable	OK under LOCF ass^{n}
Complete cases	Inefficient	Single Y: OK under MAR Repeated Y: inefficient
Mean imputation	OK in RCT Fails to control confounding in epi	SE ↓↓↓
Missing indicator		Not applicable
Regression imputation	OK under MAR (no Y in imp. model)	SE ↓↓
Multiple imputation	OK under MAR	OK under MAR
Maximum likelihood		
IPW	Inefficient or complex	OK under MAR Simple patterns only

OK means valid & efficient

Missing at random (MAR)

- The probability that data are missing
  - may depend on the values of the observed data
  - does not depend on the values of the missing data (conditional on the values of the observed data)
- Example: blood pressure (BP) data are MAR if
  - older individuals are more likely to have their BP recorded (and age is observed and included in the analysis)
  - but at any age, individuals with low and high BP are equally likely to have their BP recorded

A comment on MAR

- A lot of statistical literature seems to regard MAR as the correct starting point for analyses with missing data
- I think the correct assumption depends on the clinical context
- A general argument in favour of MAR is that it tends to become more plausible as more variables are included in the model

A comment on LOCF

- Assumes last observation is representative of the missing value
  - i.e. mean change after drop-out is zero
- Can't verify this assumption from the data
  - not implied by mean change in observed data is zero
- Analysts rarely give a good justification, and instead justify LOCF (wrongly) on the grounds that
  - it is conservative: *not true in general*
  - it respects ITT by analysing all individuals

- Recall principled approach to missing data:
  - identify a plausible assumption
  - choose analysis that's valid under that assumption

Plan

1. Why do missing data matter?
2. Popular analysis methods and their assumptions
3. **Which methods are best in a RCT?**
4. Intention-to-treat analysis strategy for randomised trials with missing outcomes
5. Sensitivity analysis

In this section I'm going to assume we are working on a trial where have decided that MAR is a reasonably plausible assumption, or at least a good starting point

Missing outcomes in a RCT under MAR: 1. Single outcome

- Under MAR, cases with missing Y contribute no information
  - complete-cases analysis is correct!
- Regress outcome (Y) on randomised group (Z), adjusting for baseline covariates (X)
  - analysis of covariance, ANCOVA
  - this is the likelihood-based method
- Which X?
  - to make MAR valid, adjust for X that predict both outcome and missingness
  - to gain power, adjust for X that predict outcome
- Can improve on complete-cases analysis with composite outcomes or auxiliary information – see later

Missing outcomes in a RCT under MAR:
2. Repeated outcome

Repeated quantitative outcome:

- Use a mixed model (likelihood-based)
- Include all observed outcome data
- Exclude any individuals with no post-baseline observations
- Include X's as before
- Software: Stata xtmixed, SAS proc mixed, R lme()
- There are some pitfalls
  - Don't allow a treatment effect at baseline
  - Allow a different treatment effect at each follow-up time
  - If possible, use unstructured variance-covariance matrix

Repeated binary outcome:

- May be worth using multiple imputation

What about multiple imputation?

- Idea of multiple imputation (tutorial: White et al, 2011)
  - Impute missing data *m* times from observed data
  - Analyse the *m* completed data sets
  - Combine estimates by Rubin's rules
- If imputation model = analysis model, MI is the same as fitting a [mixed] model to the observed data
  - but MI has additional random error
  - so why do MI?
- MI may be of value in a RCT
  - if auxiliary information (e.g. compliance or other trial outcomes) can be included in the imputation model
  - as a way to do sensitivity analyses
  - with composite outcomes
  - with repeated binary outcome

Missing baselines

Missing baselines in RCTs are a completely different problem from missing outcomes

- Not a source of bias: baseline adjustment is used to gain precision
- Complete cases analysis is a very bad idea
- Almost anything else is OK (White & Thompson, 2005)
  - in particular, mean imputation or missing indicator method are OK
  - provided randomisation is respected

- The above is only true when estimating the effect of a randomised intervention on outcome

Plan

1. Why do missing data matter?
2. Popular analysis methods and their assumptions
3. Which methods are best in a RCT?
4. **Intention-to-treat analysis strategy for randomised trials with missing outcomes**
5. Sensitivity analysis

Intention-to-treat (ITT) principle

- Include everyone randomised …
- … in the group to which they were assigned (whether or not they completed the intervention)

What does ITT mean with missing outcome data?

- "The statistical analysis of a clinical trial generally requires the imputation of values to those data that have not been recorded" (CPMP, 2001)
- "Although those participants [who drop out] cannot be included in the analysis, it is customary still to refer to analysis of all available participants as an intention-to-treat analysis" (Altman et al, 2001)
- "Full set analysis generally requires the imputation of values or modelling for the unrecorded data" (Eur. Medicines Agency, 2010)
- "We replaced mention of 'intention to treat' analysis, a widely misused term, by a more explicit request for information about retaining participants in their original assigned groups" (CONSORT, 2010)

Difficulties with ITT

- Including all randomised individuals in the analysis isn't enough to make an analysis valid
- The desire to include all randomised individuals in the analysis
  - reduces emphasis on the appropriate assumptions
  - leads to uncritical use of simple imputation methods, esp. Last Observation Carried Forward (LOCF)
  - leads to unnecessary use of complex methods, esp. multiple imputation
  - biases against MAR-based analyses

Strategy for intention to treat analysis with incomplete observations

(White et al, BMJ, 2011)

1. Attempt to follow up all randomised participants, even if they withdraw from allocated treatment
2. Perform a main analysis of all observed data that is valid under a plausible assumption about the missing data
3. Perform sensitivity analyses to explore the effect of departures from the assumption made in the main analysis
4. Account for all randomised participants, at least in the sensitivity analyses

Example: QUATRO trial

- European multicentre RCT to evaluate the effectiveness of adherence therapy in improving quality of life for people with schizophrenia (Gray *et al*, 2006)
- Primary outcome: quality of life measured by the SF-36 MCS scale at baseline and 52-week follow up

	Intervention	Control
Total n	204	205
Missing outcome	14%	6%
Mean of observed outcomes	40.2	41.3
SD of observed outcomes	12.0	11.5

QUATRO trial: ITT analysis strategy

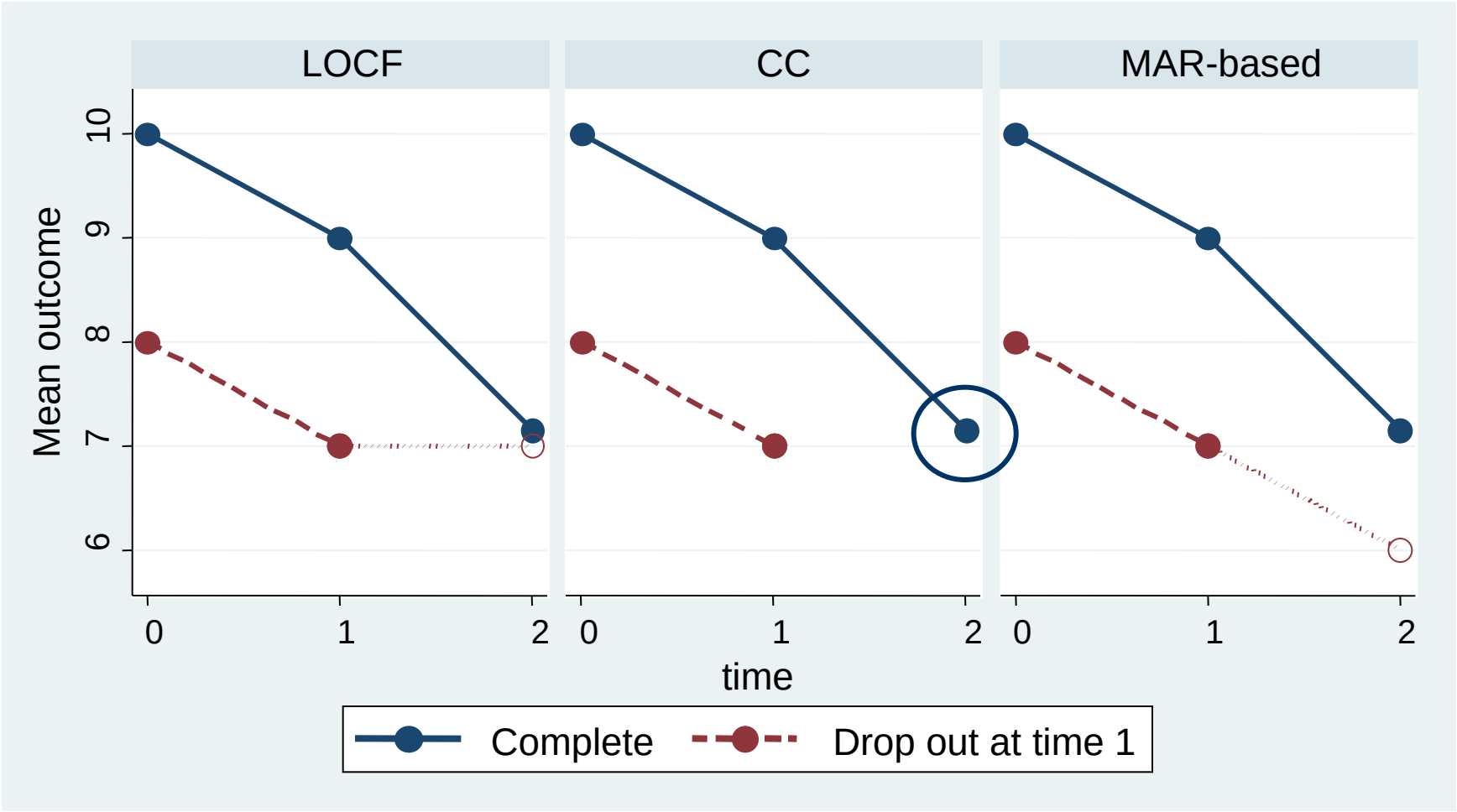
1. We did attempt to follow up all randomised individuals
2. Main assumption: no difference between missing and observed values, once adjusted for baseline variables (MAR)
   Main analysis: analysis of covariance on complete cases
   - intervention effect = -0.33 (s.e. 1.11)
3. Sensitivity analysis: consider possible differences between missing and observed values, allowed to be different in each arm
   - coming next
4. All randomised individuals were included in the sensitivity analyses

Plan

1. Why do missing data matter?
2. Popular analysis methods and their assumptions
3. Which methods are best in a RCT?
4. Intention-to-treat analysis strategy for randomised trials with missing outcomes
5. **Sensitivity analysis**

How to do sensitivity analyses?

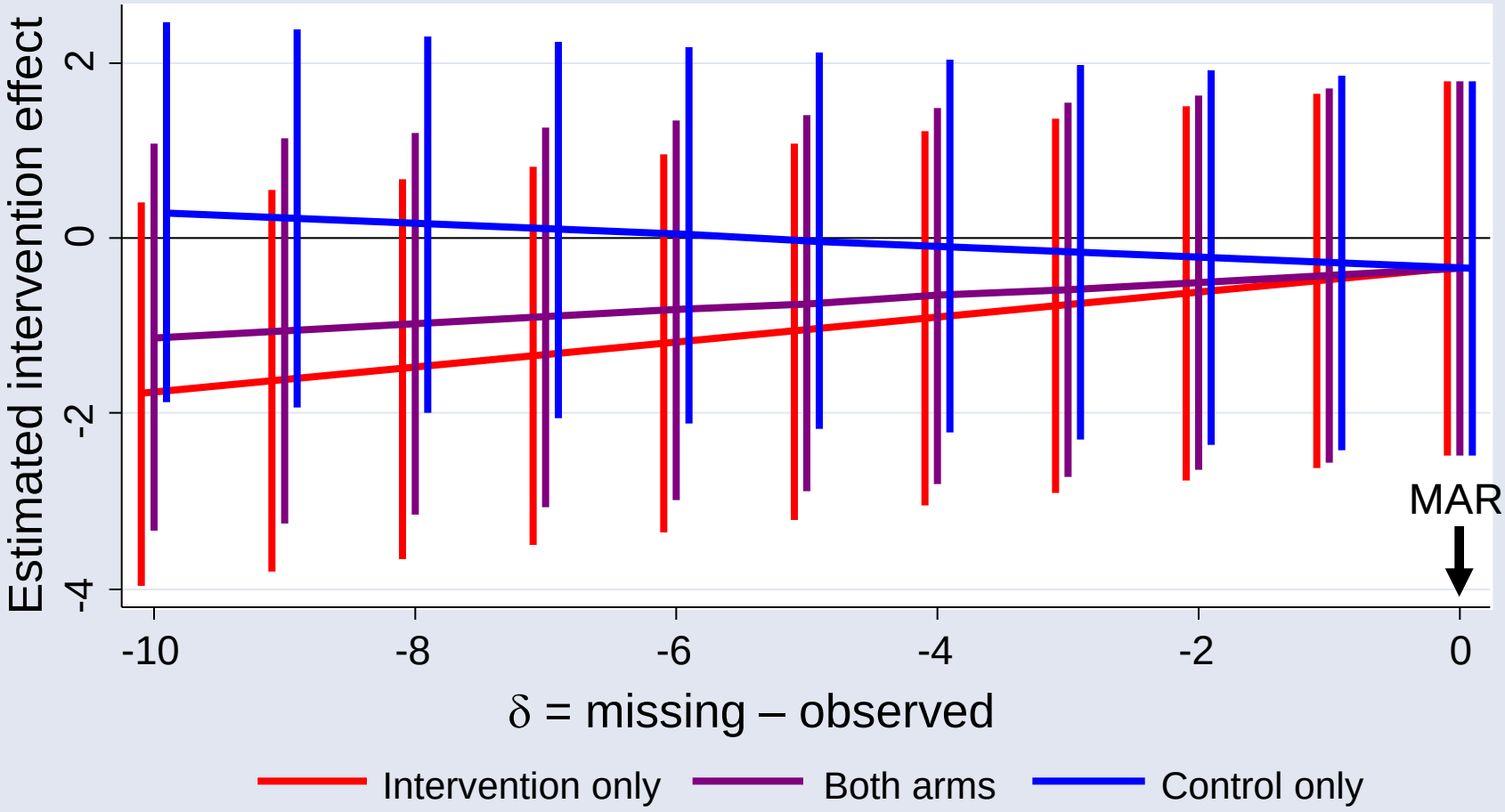
- *Not* LOCF for main analysis, CC for sensitivity analysis



How to do sensitivity analyses?

- *Not* LOCF for main analysis, CC for sensitivity analysis
- Instead, specify the numerical value of "sensitivity parameter(s)" governing the degree of departure from the main assumption (Kenward *et al*, 2001)
  - e.g. the degree of departure from MAR
  - "Principled sensitivity analysis"
- My approach:
  - let δ = mean of missing data – mean of observed data
  - so $\delta = 0$ is MAR
  - get plausible range of δ from subject matter
  - vary δ in both arms
  - vary δ in one arm (δ=0 in other arm)
- Methods: White et al (2007) or **`rctmiss`** software

Example: QUATRO data



Outcome SD ≈ 12

Conclusions & discussion

- Missing baselines: use simple methods that respect randomisation
- Missing outcomes: focus on assumptions, not methods
- ANCOVA and mixed models are often the best strategy for missing outcomes in RCTs
  - use MI with auxiliary data (e.g. compliance) or possibly as a way to do sensitivity analyses
- An intention-to-treat analysis strategy should include all individuals in sensitivity analyses
  - but not necessarily in main analyses
- Sensitivity analyses can be done in various ways
  - install my software **`rctmiss`** in Stata using **`net from http://www.mrc-bsu.cam.ac.uk/IW_Stata/missing`**

References

- Altman D *et al* (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine* 134: 663–694.
- Committee for Proprietary Medicinal Products (2001). Points to consider on missing data. http://www.emea.europa.eu/pdfs/human/ewp/177699EN.pdf
- European Medicines Agency (2010). Guideline on missing data in confirmatory clinical trials. http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500096793
- Gray R *et al* (2006). Adherence therapy for people with schizophrenia: European multicentre randomised controlled trial. *British Journal of Psychiatry* 189: 508–514.
- Kenward MG, Goetghebeur EJT, Molenberghs G (2001). Sensitivity analysis for incomplete categorical tables. *Statistical Modelling* 1: 31–48.
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. http://www.nap.edu/catalog.php?record_id=12955.
- Schulz KF, Altman DG, Moher D (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 340: c332.
- White IR, Horton N, Carpenter J, Pocock SJ (2011). An intention-to-treat analysis strategy for randomised trials with missing outcome data. *BMJ* 342:d40.
- White IR, Thompson SG (2005). Adjusting for partially missing baseline measurements in randomised trials. Statistics in Medicine 24:993–1007.
- White IR, Wood A, Royston P (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*; **30**: 377–399.