Modern Approaches to Handling Missing Data in Clinical Trials

Thomas R. Sullivan
South Australian Health & Medical Research Institute
thomas.sullivan@sahmri.com

Outline

- Missing data in trials – the problem
- Ad hoc approaches
- Framework for analysis
- Multiple imputation
- Sensitivity analysis
- Summary

Missing data

- Values that are not available but would have been meaningful for analysis had they been observed. Reasons could include:
  - participant withdrawal
  - loss to follow-up
  - skipped appointments/tests/questions
  - data lost accidentally
- Whether data are meaningful for analysis depends on what's being estimated – e.g. treatment effect in all randomised or compliers only
- Vast majority of trials have at least some missing data

Missing data – the problem

- Untestable assumptions must be made about the missing data – getting this wrong can lead to bias
- Reduced power compared to an analysis with complete data
- Reduced generalisability
- Ultimately can lead to inappropriate conclusions

⇒ **Try to prevent missing data in the first place!**

Another problem: common use of inappropriate methods of analysis

Complete case analysis

ID	Weight t1	Weight t2
1	76	.
2	79	86
3	60	60
4	54	.
5	91	94

- Assumes participants with complete data representative of those with missing data – can lead to bias if assumption wrong
- Not recommended for longitudinal data as inefficient
- Can be OK for outcomes measured at a single time point if analyses adjusted for baseline predictors of the outcome and missing data

⇒ *Think about associations with missing data when specifying covariates for adjustment*

Last observation carried forward

ID	Weight t1	Weight t2
1	76	.
2	79	86
3	60	60
4	54	.
5	91	94

Last observation carried forward

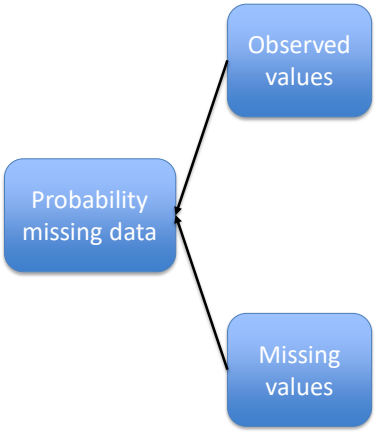
ID	Weight t1	Weight t2
1	76	**76**
2	79	86
3	60	60
4	54	**54**
5	91	94

- Assumes outcomes do not change after dropout – unrealistic in many settings
- Replaced values analysed as if they were observed, which fails to account for uncertainty due to missing data
- Widely condemned in the trials literature
- Many journals recommend avoiding the method, e.g. JAMA

Better approach

- Start with a sensible assumption about the cause(s) of missing data – good data collection and clinical input important
- Choose a statistical method that produces valid results under this assumption
  - unbiased estimates of the treatment effect and its standard error
  - efficient, i.e. makes the best use of the observed data
- Investigate the sensitivity of conclusions to the assumption made about the missing data
- Report on all of this – state and justify the assumptions made

Missing data assumptions



Missing completely at random
Probability missing data does not depend on observed or missing values, *e.g. data forms accidentally lost*

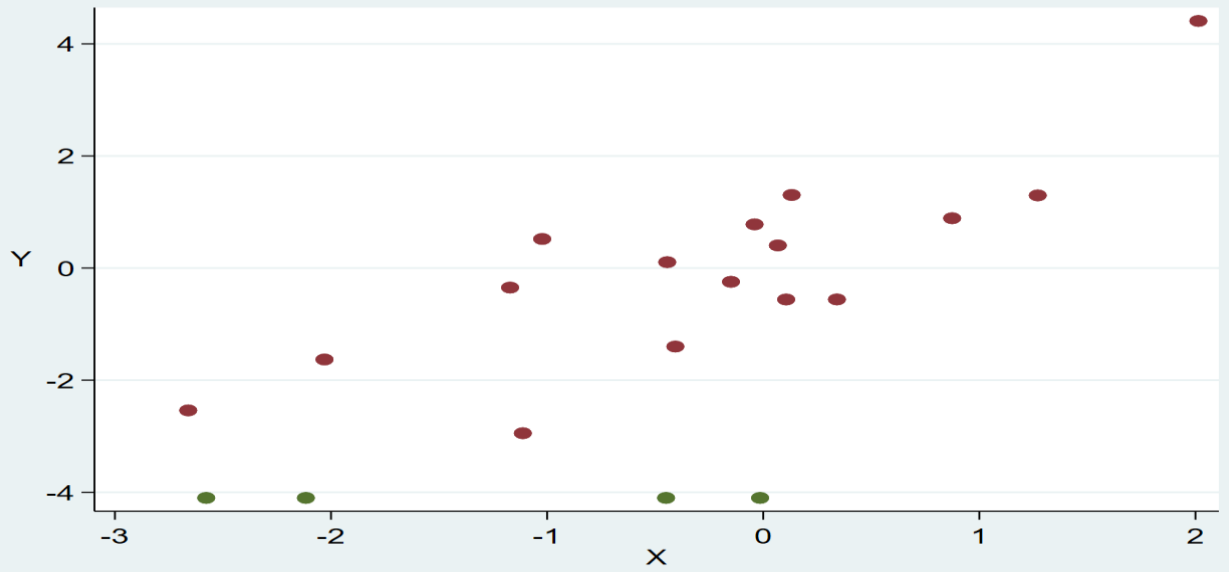
Missing at random
Probability missing data depends on observed but not on missing values, *e.g. those with a higher recorded baseline weight more likely to dropout*

Missing not at random
Probability missing data also depends on the values of the missing data, *e.g. subject skips the assessment because they've gained too much weight*

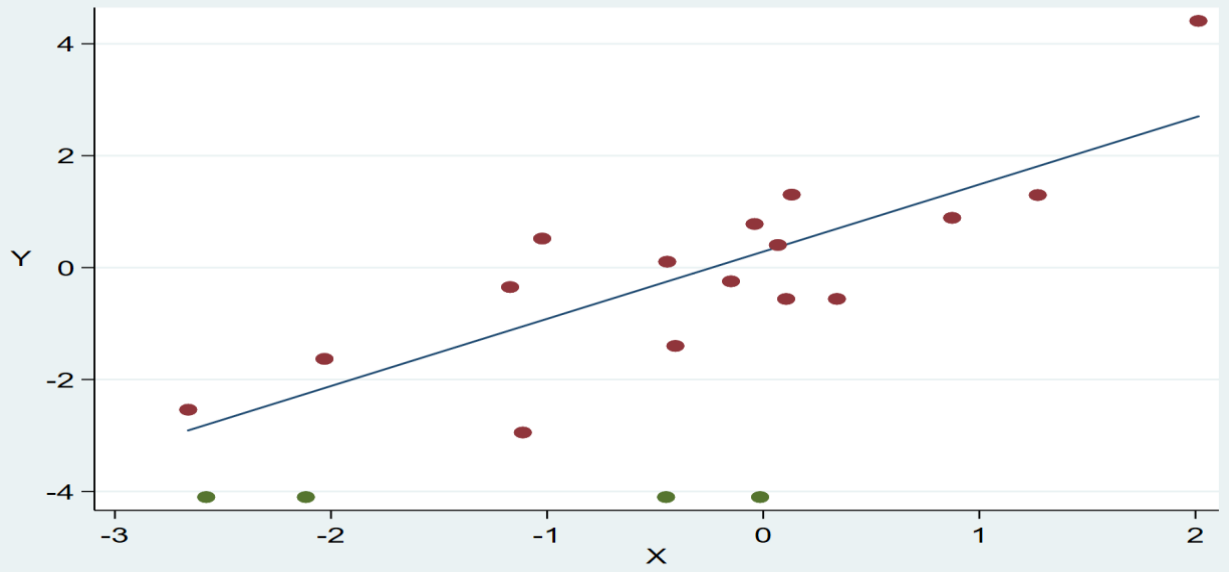
Missing at random

- Sensible starting assumption for the analysis of many clinical trials
- More realistic than the stricter missing completely at random assumption
- Statistical methods that produce appropriate results under a missing at random assumption include:
  - maximum likelihood methods, e.g. mixed models
  - inverse probability weighting
  - multiple imputation

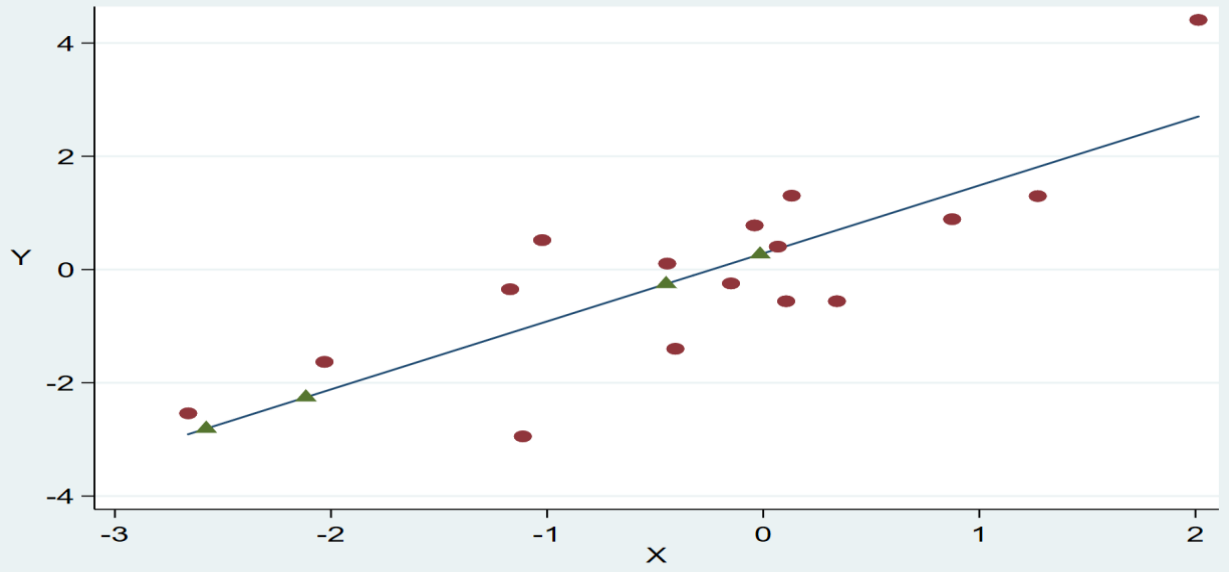
Regression example



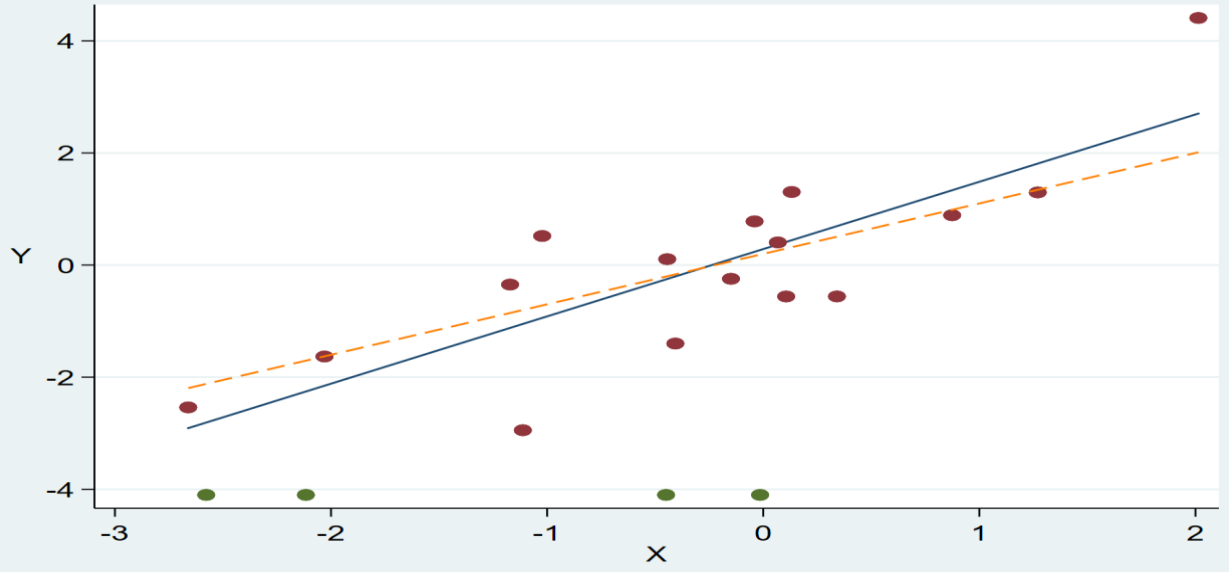
Regression example



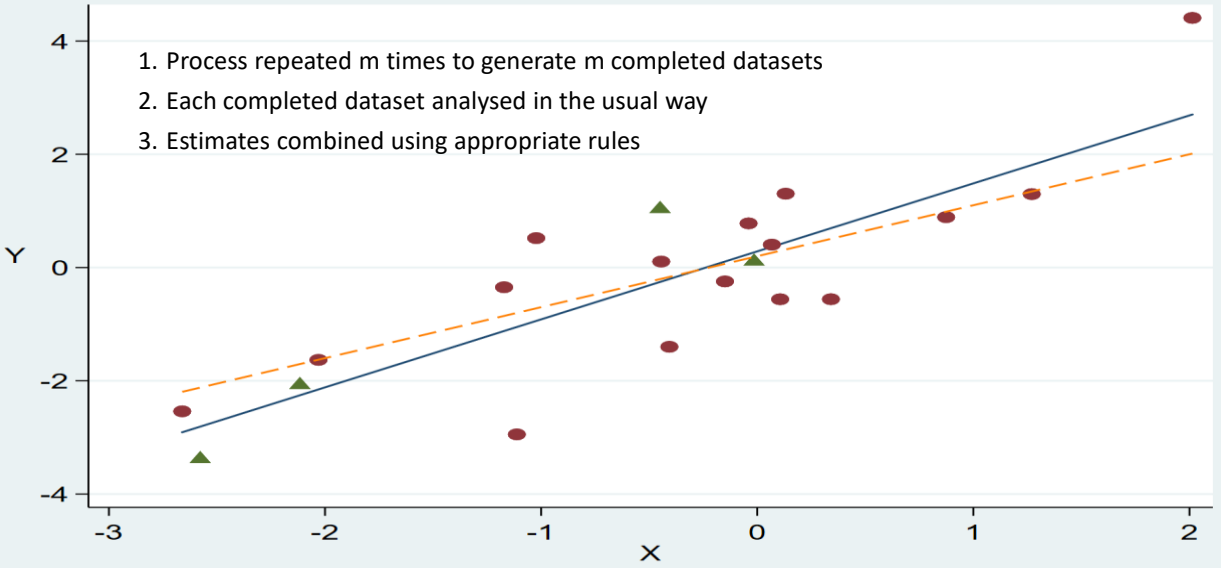
Regression example



Multiple imputation



Multiple imputation



Multiple imputation

- Increasingly popular in trials due to its flexibility – can be used with any intended analysis model
- Great for sensitivity analyses
- Can benefit from the inclusion of auxiliary variables:
  - not a part of the analysis model but included in the imputation model to improve the prediction of missing values
  - want variables associated with the outcome (precision gains) or both the outcome and the probability of missing data on the outcome (bias reduction)
  - baseline or post-randomisation

⇒ *Think about potential auxiliary variables when designing a trial*

MI by treatment group

- If possible, better to fit separate & identical imputation models to each treatment group:
  - easier to specify
  - facilitates subgroup analyses for any baseline characteristic included in the imputation model
  - produces unbiased average treatment effects when subgroup analyses not of interest
  - allows for unequal variances between treatment groups
  - not much lost precision when there is no effect modification

Sensitivity analysis

- See how results change under alternative plausible assumptions about the missing data
- Alternative assumptions should contradict the main assumption
  - wouldn't use complete case analysis as a sensitivity analysis to multiple imputation
- For primary analysis under a missing at random assumption, consider missing not at random assumptions

Pattern mixture models

- Let δ = mean difference in weight at final time point between missing and observed values
  - missing at random: conditional on observed data, δ = 0
  - missing not at random: conditional on observed data, δ ≠ 0
- Specify plausible values for δ and vary between treatment groups

Analysis	Assumption	δ intervention group	δ control group
Primary	MAR	0	0
Sensitivity 1	MNAR	+5kg	0
Sensitivity 2	MNAR	0	+5kg
Sensitivity 3	MNAR	+5kg	+5kg

- Multiple imputation: just add δ to the imputed values!

Summary

- Prevent missing data in the first place
- Start with a sensible assumption about the missing data and choose an analysis method valid under this assumption
- Missing at random often reasonable – maximum likelihood methods, inverse probability weighting and multiple imputation
- Collect information to support the analysis – e.g. reasons for missing data and auxiliary variables
- Conduct sensitivity analyses

References

- Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Medical Research Methodology. 2014;14(1):118.
- Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. Statistical Methods in Medical Research. 2014;23(5):440-59.
- National Research Council Panel on Handling Missing Data in Clinical Trials. The prevention and treatment of missing data in clinical trials. Washington (DC): National Academies Press (US); 2010.
- Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations. Statistics in Medicine. 2016;35(17):2865-75.
- Rubin D. Inference and missing data. Biometrika. 1976;63(3):581-92.
- Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Statistical Methods in Medical Research. 2018;27(9):2610-26.
- White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. Clinical Trials. 2012;9(4):396-407.
- Yamaguchi Y, Ueno M, Maruo K, Gosho M. Multiple imputation for longitudinal data in the presence of heteroscedasticity between treatment groups. Journal of Biopharmaceutical Statistics. 2019:1-19.