UMC Utrecht

Imputation of Missing Data

Thomas Debray, PhD
Assistant Professor
T.Debray@umcutrecht.nl

Jeroen Hoogland, GradStat
Junior Researcher
J.Hoogland-2@umcutrecht.nl

Methods to handle missing data
Simple methods (ad hoc)

Complete case analysis (CC)

Available case analysis (AC)

Missing indicator method

Overall mean/median imputation

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

Multiple imputation by regression

Key problem in SI: **all values are treated as observed**

However, imputations are uncertain

Remember:

#1 Uncertainty due to natural variation
#2 Uncertainty of the estimated imputation model is ignored

SI cannot convey all of this uncertainty

Multiple imputation by regression

Key change → **multiple** imputation

Univariate missingness

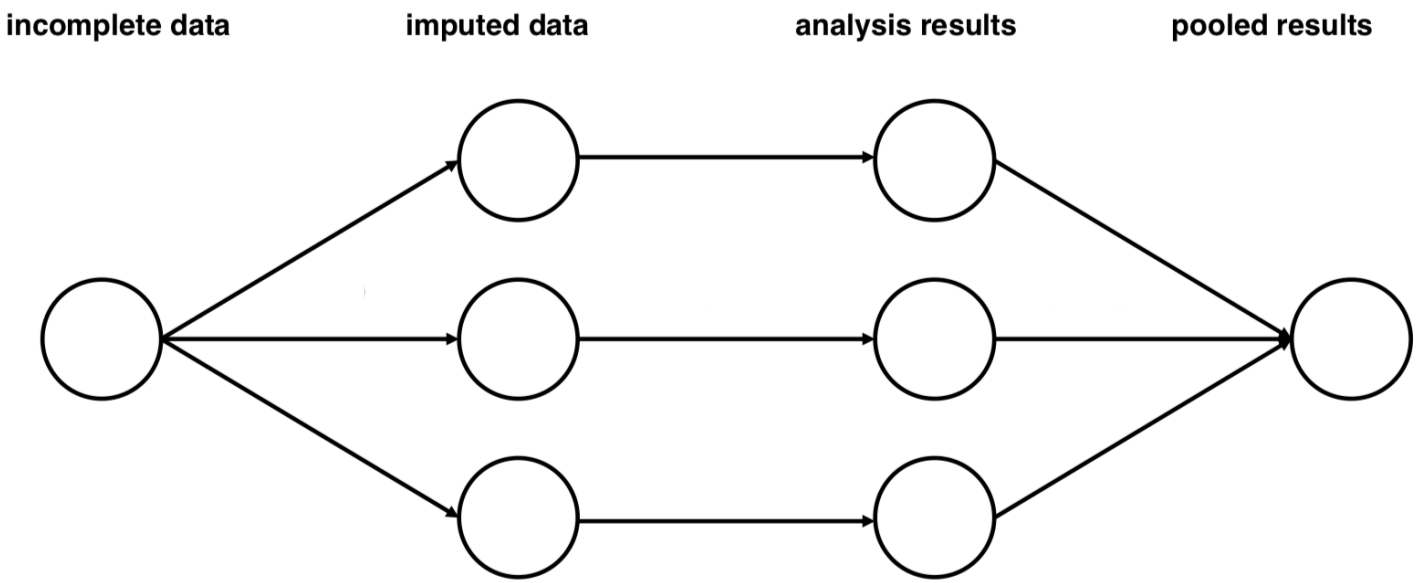
- Estimate imputation model for the missing variable in the complete data
- For each patient with a missing value
  - Draw a random sample from estimated imputation model parameters (regression coefficients, residual error)
  - Use sampled parameters to generate prediction (solution to #2)
  - Add noise to the prediction (solution to #1)
- Repeat many times to generate many imputed data sets

Multiple imputation by regression

This results in *m* imputed data sets

Note that the imputed values change over imputed data sets and observed value don't → this conveys imputation uncertainty

Multiple imputation by regression



Analysis as usual on complete data!

Multiple imputation by regression

Analysis and pool of the results

- Study the association (X's on Y) in each of the *m* imputed (completed) data sets
- *m* beta's of X's on Y are averaged → 1 beta per X
- *m* SE's per beta are "*pooled*" (via Rubin's rules) → accounting for between imputation variation → yielding 1 SE per X

Multiple imputation by regression

- Under M(C)AR:
  - MI provides unbiased beta estimates (as SI)
  - Provides unbiased SEs (p-values) --> since properly accounting for the uncertainty surrounding the imputations

- Under MNAR: even when missing X's or Y are partly MNAR, applying imputation methods that assume MAR can reduce bias based on the MAR part of the missing data mechanism

Multiple imputation by regression

Multivariate missingness

Aforementioned approach is ineffective when we have patients with >1 missing value

- We can only use patients without missing values to estimate the imputation models
- Patients with missing values for X_1 may still inform imputation of X_2

Iterative procedures provide a solution
(**discussed tomorrow**)

Multiple imputation by regression

General recommendations

- **Imputations should properly reflect all uncertainty**
  - Both the imputation model error component and the uncertainty on imputation model parameters
- **Imputation models should be as flexible as possible**
  - At least include the level of complexity necessary for the analysis model and mabye more (to the extend that other variables may also carry information on missing data)
  - Flexibility also relates to interactions and functional form in the analysis model (see Seaman et al. doi: 10.1186/1471-2288-12-46)

Imputation models should be as flexible as possible

More complex imputation models are possible

- Generalized additive models
- Neural networks
- Random forests
- …

Common pitfalls of multiple imputation

Clearly described by Sterne et al. (doi: 10.1136/bmj.b2393)

- Omitting the outcome variable from the imputation procedure
- Dealing with non-normally distributed variables
- Plausibility of missing at random assumption
- Data that are missing not at random
- Computational problems

Imputation (single or multiple) with or without outcome?

- Missingness on determinants commonly relates to other patient characteristics, including (directly or indirectly) the outcome
- Advice (SI + MI) = use *all* observed patient data, i.e. all other determinants (X's) *plus* the outcome
- If the outcome is ignored during imputation, the association between the imputed predictor and outcome will evaporate
- This also relates to the congeniality problem (more on this tomorrow)

Special case: handling missing outcomes (in RCTs or observational studies)

When i) only outcome data is missing and ii) the analysis are based on maximum likelihood (which is very common):

- CCA with covariate adjustment yields unbiased estimates, of both betas and SEs, when:
  - missing outcome data are MAR, and
  - All predictors of missingness (all covariates/confounders) of the outcome are included as covariates in the adjustment model → fully adjusted model!
  - Irrespective of the relations between the covariates and treatment (i.e., holds for RCTs and observational studies)
  - No imputation needed!

Special case: handling missing outcomes (in RCTs or observational studies)

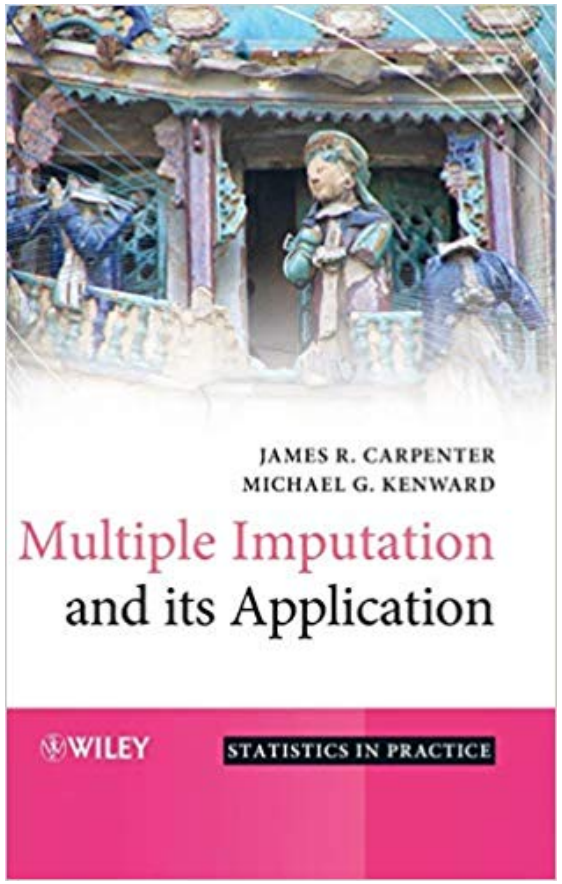
When i) only outcome data is missing and ii) the analysis are based on maximum likelihood (which is very common):

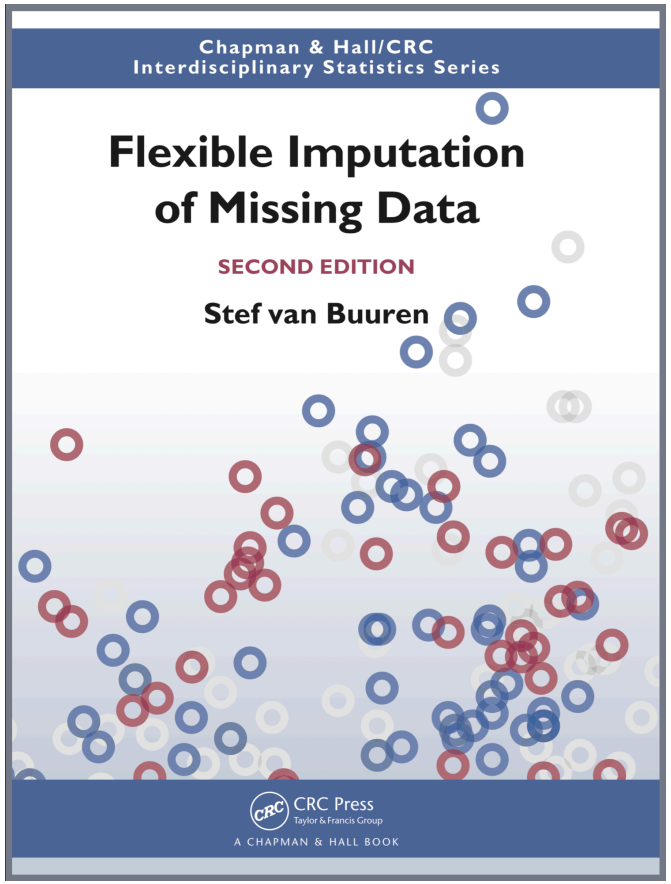
- But
  - MI allows also for incorporation of post-randomization variables (e.g. secondary endpoints, never included in adjusted model)
  - MI can handle simultaneously both missing outcome and predictors (which is common)

R packages for MICE

- mice - the MICE algorithm as described by Van Buuren and Groothuis-Oudshoorn (2011)
- mi – an implementation of MICE by Gelman et al. Additional methods for assessing convergence
- micemd – addons for MICE to impute multilevel data, by Audigier et al.
- rms – an implementation of MICE by Harrell

Recommended reading





Key references

(next to those mentioned in the slides)

- A gentle introduction to imputation of missing values (Donders JCE 2006)
- Handling missing data in multivariable diagnostic research: a clinical example (van der Heijden JCE 2006).
- Using the outcome variable to impute missing values of predictor variables: a self fulfilling prophecy? (Moons JCE 2006)
- To Impute is better than to ignore (Janssen JCE 2010)
- Dealing with missing values when validating a prediction model (Janssen Clin Chem 2009)
- Imputation of missing outcomes in observational and randomised studies (Groenwold AJE 2012)
- Little et al; New Engl J Med 2012
- Groenwold RH, Moons KG, Vandenbroucke JP. Randomized trials with missing outcomes: what to report and how to analyze. *CMAJ 2014*