

Q&A

Goals of imputation of missing data

- maintain relationship in the data
-
- keep uncertainty (noise) that's equal to the real uncertainty (outcome and imputation model)
--> coefficients are generated according to the imputations
- able to impute missing values when missing data on multiple variables
impute random value or mean and then in the second step iterate to improve the imputed value (always uncertainty around the coefficient for each imputation)

not the goal to find the best imputed values but to find the values that can be plotted best that maintain the variation and uncertainty (standard errors) of the given data

book recommended $m=50$ (imputations) - no clear answer to this

when doing infinite number of imputations we could catch all variation but because we don't do that (can't see all coefficients for imputed data) we have to take the uncertainty into account -> variation between datasets and correction factor

-> more imputations means less corrections later

in some cases it's not possible to do many imputations (large dataset 1M cases or complex) then only 5 imputations and relying to use formula to account for uncertainty and lack of imputations - so used instead of imputing many times

no reason on why one should use single imputation! just to understand multiple imputations

-> standard error is only smaller because it is BIASED

reading material more to understand the lectures better - not to hung up on the statistics and formulas - understand concept

goal of multiple imputations -

3 sources of variation (2.3.1 and 2.3.3) we don't have to understand the equation

2.20 also in the practicals (we don't need to know the difficult math/equations!!!!) won't be tested

Question 4. What is the adjusted odds ratio for vacc, and the corresponding 95% confidence interval? Is annual influenza vaccination effective in reducing the risk of hospitalization? calculate for one category and then for the other
divide the being vacc group by the non vacc group -> going over it on Friday (27. Nov)

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

single imputation

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

each participant only has one value (observed or imputed)

However, imputations are uncertain

Remember:

which is why we added some noise to add natural variation to the data

#1 Uncertainty due to natural variation

#2 Uncertainty of the estimated imputation model is

ignored treated them as though they were known but this uncertainty can be added in multiple imputation

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 instead of stopping after one go we draw a random sample from estimated imputation model
- 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

parameter uncertainty of imputation model + noise



Multiple imputation by regression

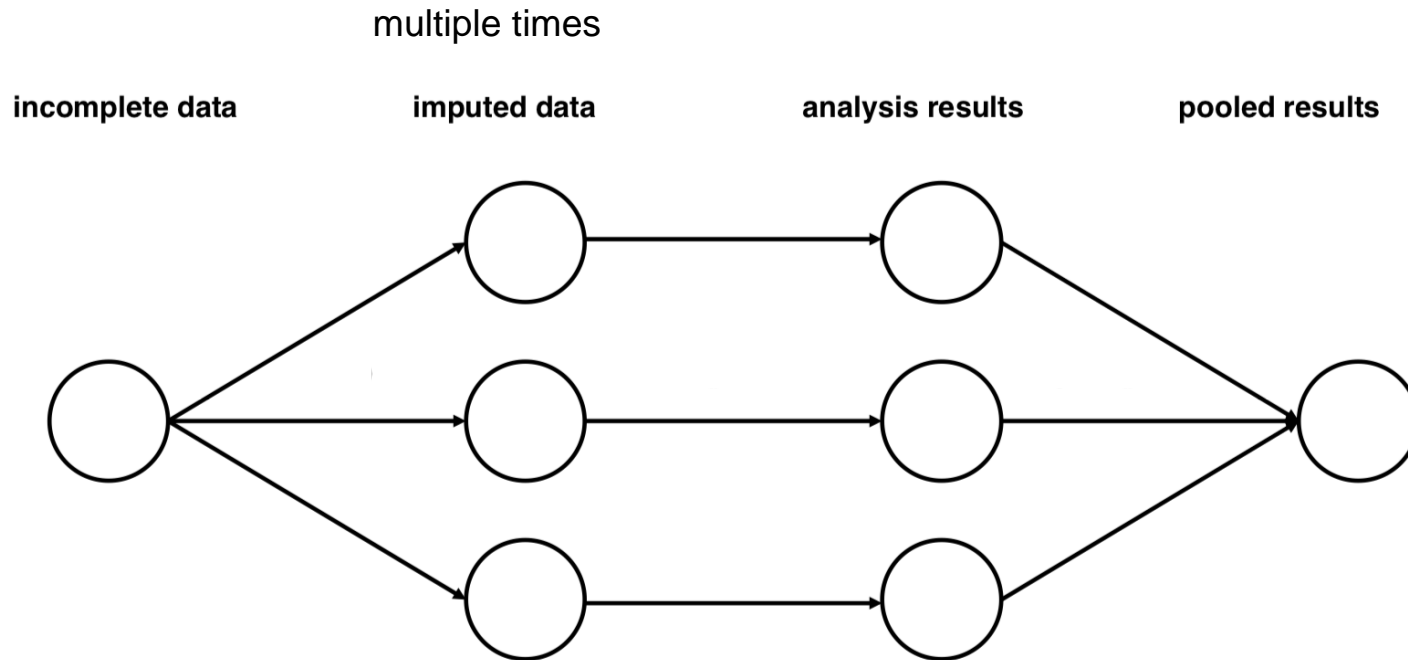
This results in m imputed data sets

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

imputed values differ over each set whereas the observed values are all the same



Multiple imputation by regression



can be treated as usual
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 \rightarrow 1 beta per X
- m SE's per beta are “*pooled*” (via Rubin's rules) \rightarrow accounting for between imputation variation \rightarrow 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
unbiased SE unlike single regression imputation -> unbiased p-values!!
- 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

so if there's more than one column missing per case as we use the existing values to estimate the missing ones - if there's multiple missing columns they are automatically biased

- 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 maybe more (to the extent 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**
final analysis model is model of outcome predicted by several predictors that are exactly there because of the outcome! which is why the outcome needs to be included
- **Dealing with non-normally distributed variables**
all imputations follow normal distribution - if not discrepancy
implement in mice as default option to not use normal dit.
but predicted mean matching instead of tests for normality that are usually not reliable
- **Plausibility of missing at random assumption**
- **Data that are missing not at random**
systematic difference between missing and observed data can be observed
hard to tell if you have all the info you need to explain the missing data
cannot tell/calculate if its missing at random or 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 advise to INCLUDE 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)

= the fact that all relations in analysis model should be represented in the imputation models



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: usually only when MCAR but here possible
 - 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!**

the important point is that usually NA is all over dataset but it can happen that missing data is ONLY in the outcome
- when outcomes are missing conditional only on data that we have collected e.g. MAR then it's not an issue as we are already accounting for these cases within the model therefore no imputation needed

might still be good to do imputations to reduce uncertainty



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)

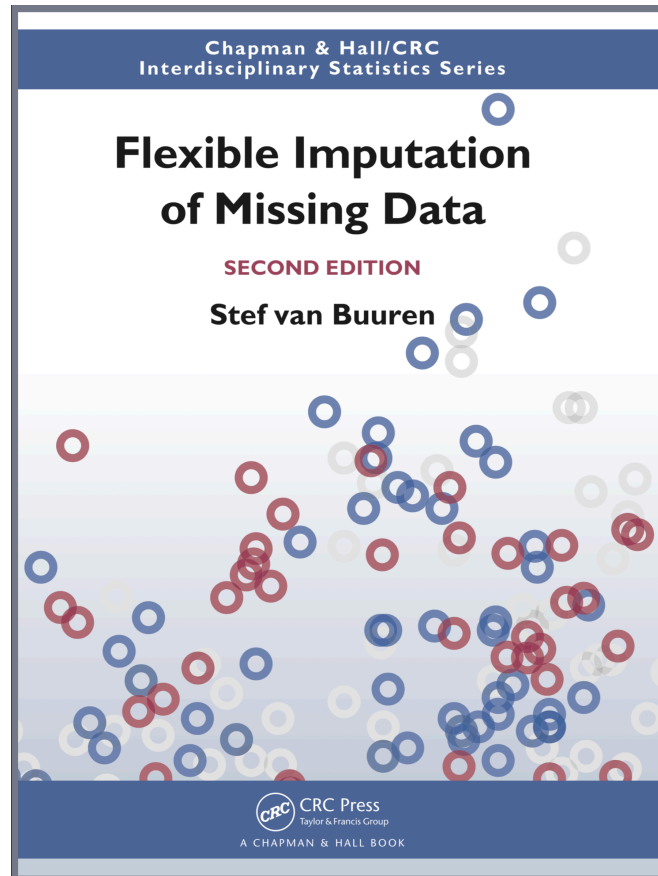
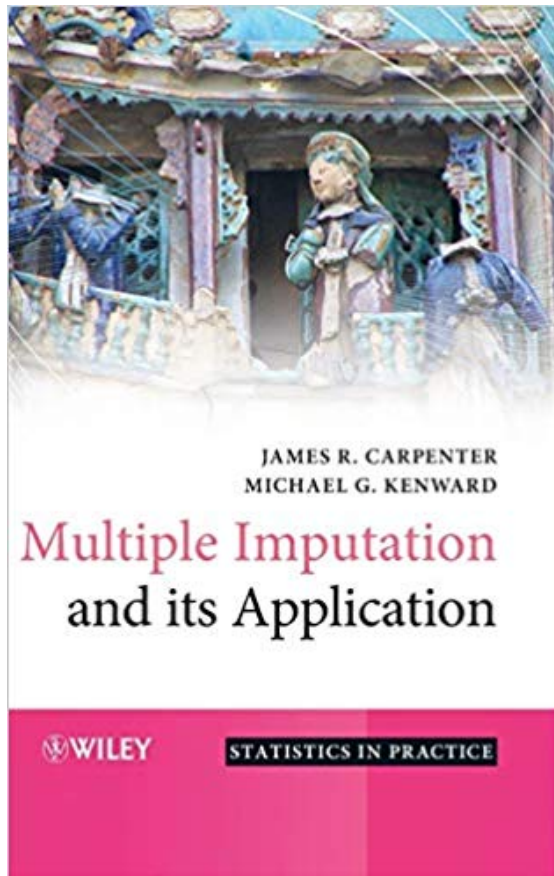


R packages for MICE

- [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