EP16: Missing Values in Clinical Research: Multiple Imputation

Summary I

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

Summary I

1. What is Multiple Imputation?

- Rubin's **idea**:
  - Missing values need to be represented by **multiple imputed values**.
  - A **model is necessary** to obtain good imputations.
- Imputed values are obtained from the **predictive distribution** of the missing data, given the observed data.
- Multiple completed datasets are created from the multiple imputed values.
- Multiple imputation has **three steps: imputation, analysis, pooling**

Summary II

2. Imputation step

- Two **sources of variation**:
  - **parameter uncertainty**
  - **random variation**
- **Two approaches** to MI for imputation of non-monotone multivariate missing data
  - **MICE/FCS**
  - **Joint model imputation**
- The MICE algorithm re-uses univariate imputation models by iterating through all incomplete variables, multiple times (**iterations**)
- **Multiple runs** through the algorithm are necessary to create multiple imputed datasets
- The **convergence of the chains** needs to be checked.

Summary III

3. Analysis step

- Analyse each imputed dataset the way you would analyse a complete dataset

4. Pooling

- Results from analyses of multiple imputed datasets can be summarized by taking the **average of the regression coefficients**
- For the total variance, **three sources of variation** need to be considered:
  - **within imputation variance**
  - **between imputation variance**
  - uncertainty due to finite number of imputations

Summary IV

5. A closer look at the imputation step

- Two **parametric approaches** for imputation:
  - **Bayesian** (sample from posterior distribution of parameters)
  - **Bootstrap** (uses bootstrap samples of the data to estimate parameters)
- **Predictive mean matching** is a semi-parametric alternative (it matches observed and missing cases based on their predicted values).
- In PMM we need to consider
  - **donor selection**
  - **matching type** (how parameters are sampled/estimated),
  - the **set of data** used to calculate/estimate the parameters.
- Bayesian and bootstrap imputation take into account the variation, while many **choices in PMM lead to underestimation of the variation**.

6. Know your data

Check the

- missing data pattern
- distribution of observed values
- associations & patterns in the observed values

- Think about why values are missing. Is MAR reasonable?
- Is additional information available (auxiliary variables)?

7. Imputation with mice

Specification of

- imputation method
- predictor matrix
- visit sequence

Further tailoring of the imputation using

- passive imputation
- post processing

8. Convergence & diagnostics

- Logged events
- convergence: traceplots
- distribution of observed and imputed values (conditional on other variables)

9. Analysis & pooling

- **mice** functions ‘with()‘.R and ‘pool()‘.R
- alternative pooling using **mitools**
- additional functions in **mice**: ‘pool.r.squared()‘.R, ‘pool.compare()‘.R
- additional functions in other packages: ‘miceadds::micombine.chisquare()‘.R, ‘miceadds::micombine.F()‘.R

10. Additional functions in mice

- ‘complete()‘.R, ‘mids2spss()‘.R
- ‘ibind()‘.R, ‘cbind.mids()‘.R

- MICE requires **congenial & compatible imputation models** to work well.
- When this is not the case, (naive) use of MICE can lead to **biased results**.
- Common settings that require special attention are
  - **non-linear functional forms & interaction terms**
  - **longitudinal data**
  - **survival data**

- When using the package **mice**, there are choices that can **reduce bias**
  -
  - JAV approach reduces bias in settings with interactions or non-linear associations
  - **special 2-level imputation methods** are available for longitudinal data
  - The **Nelson-Aalen estimator** can be used instead of the time variable for imputing survival data when effects are not too large.
- Generally, **problems** are more severe when
  - **proportions of missing values are large**,
  - effect sizes are large,
  - little other **covariate information** is available.

  (Note that in the examples we had all of the above.)

- In settings where MICE may not provide valid imputations, **alternative approaches** are available and should be considered.
- R packages that provide such alternative approaches are for example:
  - **JointAI** (non-linear, longitudinal & survival)
  - **smcfcs** (non-linear & survival)
  - **jomo** (non-linear, longitudinal & survival)
- These packages are very young.
  - Hence, they may still have some problems.
    ➡ **Use them carefully!** (and email the maintainer about problems)
  - They are under **active development**, so resolutions of bugs and features are frequently added.