EP16: Missing Values in Clinical Research: Multiple Imputation

14. Strategies for using MICE

Nicole S. Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

In **early publications** on multiple imputation:

- 3 – 5 imputations are sufficient
- still is a common assumption in practice (Rubin 2004)

Reasoning behind using a small number of imputations:

- **storage of imputed data was "expensive"** (no longer the case)
- larger number of imputations would only have little advantage (Schafer 1997)

Number of Imputations

More **recent work** from various authors considers

- the efficiency of the pooled estimates
- reproducibility of the results,
- statistical power of tests or
- the width of the resulting confidence intervals compared to the width of the true confidence intervals

(White, Royston, and Wood 2011; Van Buuren 2012; Graham, Olchowski, and Gilreath 2007)

Number of Imputations

A **suggested rule of thumb** (White, Royston, and Wood 2011):

number of imputed datasets should be similar to the **percentage of incomplete cases**

Since this percentage depends on the size of the dataset, the **average percentage of missing values** per variable could be used as an alternative. (Van Buuren 2012)

Generally:

- using **more imputed datasets should be preferred**,
- especially in settings where the computational burden allows for it.

Even though results are unlikely to change with a larger number of imputations, it can increase the efficiency and reproducibility of the results.

What to do with large datasets?

In imputation, generally the **advice is to include as much information as possible** in the imputation models.

Using a large number of predictor variables

- makes the **MAR assumption more plausible** (and, hence, reduces bias due to MNAR missingness)
- can **reduce uncertainty** about the missing values

What to do with large datasets?

In imputation, generally the **advice is to include as much information as possible** in the imputation models.

Using a large number of predictor variables

- makes the **MAR assumption more plausible** (and, hence, reduces bias due to MNAR missingness)
- can **reduce uncertainty** about the missing values

This can **work well in small or medium sized datasets** (20 – 30 separate variables, i.e. without interactions, variables derived from others, …)

However, **in large datasets** (contain hundreds or thousands of variables) this is **not feasible**. (Van Buuren 2012)

What to do with large datasets?

For large datasets a possible strategy is to

- Include all **variables used in the analysis model(s)** (including the outcome!)
- Include auxiliary variables if
  - if they are **strong predictors of missingness**, or
  - if they have **strong associations with the incomplete variables**
  - **only if they do not have too many missing values** themselves (and are observed for most of the incomplete cases of the variable of interest)
  - only in those imputation models for which they are **relevant**
- Calculate **summary scores** from multiple items referring to the same concept and use the summary score as predictor variable.

How much missing is too much?

There is **no clear cut-off** for the proportion of missing values that can be handled adequately using MICE (or any other imputation method).

The amount of missingness that can be handled **depends on the information that is available** to impute it.

- Are there **strong predictor variabels** available & observed?
- Are there **sufficient observed cases** to get reliable estimates for the predictive distribution?

How much missing is too much?

There is **no clear cut-off** for the proportion of missing values that can be handled adequately using MICE (or any other imputation method).

The amount of missingness that can be handled **depends on the information that is available** to impute it.

- Are there **strong predictor variabels** available & observed?
- Are there **sufficient observed cases** to get reliable estimates for the predictive distribution?

Example:

- In a set of *N* = 50 cases, 50% missing values leaves 25 cases to estimate the parameters of the predictive distribution.
- In a large set of *N* = 5000 subjects, 50% missing cases leaves 2500 observed cases to estimate parameters.

Imputation of Outcomes

Usually, **missing outcome values are not imputed**.

Why?
When there are no auxiliary variables, imputation and analysis model are equal.

- Parameters of the imputation model are estimated on observed cases of the outcome.
- Imputed values will fit the assumed model perfectly.
- Including imputed cases in the analysis does not add any information.

Imputation of Outcomes

Exception:

- When very strong auxiliary variables are available.
- Outcomes may be imputed when one imputation is performed for several analysis models, because not imputing the outcome(s) would mean
  - excluding cases with missing outcome(s) from the imputation, or
  - excluding the outcome variable(s) as predictor(s).

Notes of caution & things to keep in mind

Multiple imputation is **not a quick and easy solution for missing data**.

It requires **care and knowledge** about

- the **data** to be imputed (and the context of the data),
- the statistical **method** used for imputation, and
- the **software** implementation used.

Moreover

- **Never accept default settings of software blindly.**
- **Question the plausibility of the MAR assumption.** If it is doubtful, use sensitivity analysis.

Notes of caution & things to keep in mind

- **Use as much information as possible**
  - include all covariates and the outcome
  - use auxiliary information
  - use the most detailed version of variables if possible
- **Avoid feedback** from derived variables to their originals.
- **Imputation models must fit the data** (correct assumption of error distribution and functional forms and possible interactions of predictor variables).
- Think carefully how to handle variables that are derived from other variables.
- Consider the impact the **visit sequence** may have.
- Choose an appropriate **number of imputations**.
- Make sure the imputation algorithm has **converged**.
- Use **common sense** when evaluating if the imputed values are plausible.

Imputation Methods

We have focussed on a few imputation methods that cover the most common types of data, but there are many more methods implemented.

Imputation methods implemented in the **mice** package:

mice.impute.2l.bin	mice.impute.logreg	mice.impute.panImpute
mice.impute.2l.lmer	mice.impute.logreg.boot	mice.impute.passive
mice.impute.2l.norm	mice.impute.mean	mice.impute.pmm
mice.impute.2l.pan	mice.impute.midastouch	mice.impute.polr
mice.impute.2lonly.mean	mice.impute.mnar.logreg	mice.impute.polyreg
mice.impute.2lonly.norm	mice.impute.mnar.norm	mice.impute.quadratic
mice.impute.2lonly.pmm	mice.impute.norm	mice.impute.rf
mice.impute.cart	mice.impute.norm.boot	mice.impute.ri
mice.impute.jomoImpute	mice.impute.norm.nob	mice.impute.sample
mice.impute.lda	mice.impute.norm.predict	NA

Note: That a method is implemented does not mean you need to / should use it.

Imputation methods

Imputation methods implemented in the **miceadds** package:

mice.impute.2l.binary	mice.impute.hotDeck
mice.impute.2l.contextual.norm	mice.impute.lm
mice.impute.2l.contextual.pmm	mice.impute.lm_fun
mice.impute.2l.continuous	mice.impute.lqs
mice.impute.2l.groupmean	mice.impute.ml.lmer
mice.impute.2l.groupmean.elim	mice.impute.plausible.values
mice.impute.2l.latentgroupmean.mcmc	mice.impute.pls
mice.impute.2l.latentgroupmean.ml	mice.impute.pmm3
mice.impute.2l.plausible.values	mice.impute.pmm4
mice.impute.2l.pls	mice.impute.pmm5
mice.impute.2l.pls2	mice.impute.pmm6
mice.impute.2l.pmm	mice.impute.rlm
mice.impute.2lonly.function	mice.impute.smcfcs
mice.impute.2lonly.norm2	mice.impute.tricube.pmm
mice.impute.2lonly.pmm2	mice.impute.tricube.pmm2
mice.impute.bygroup	mice.impute.weighted.norm
mice.impute.grouped	mice.impute.weighted.pmm

Imputation methods

Imputation methods implemented in the **micemd** package:

mice.impute.2l.2stage.bin	mice.impute.2l.glm.bin
mice.impute.2l.2stage.norm	mice.impute.2l.glm.norm
mice.impute.2l.2stage.pmm	mice.impute.2l.glm.pois
mice.impute.2l.2stage.pois	mice.impute.2l.jomo

Tips & Tricks

In complex settings, variables may need to be **re-calculated** or **re-coded** after imputation:

- Use `complete` to convert the imputed data from a `mids` object to a `data.frame`.
- Perform the necessary calculations.
- Convert the changed `data.frame` back to a `mids` object using the functions such as `as.mids()`, `miceadds::datalist2mids()`, `mitools::imputationList()`, ...

Tips & Tricks

In complex settings, variables may need to be **re-calculated** or **re-coded** after imputation:

- Use `complete` to convert the imputed data from a `mids` object to a `data.frame`.
- Perform the necessary calculations.
- Convert the changed `data.frame` back to a `mids` object using the functions such as `as.mids()`, `miceadds::datalist2mids()`, `mitools::imputationList()`, ...

Not just in imputation: Set a **seed value** to create reproducible results.

- in R: `set.seed()`
- in `mice()`: argument `seed`

MICE and MI in the bigger picture

Other R packages that do imputation

In Part III of this course we have seen some R packages that perform imputation or provide functionality for missing data other than **mice**.

Currently, there are **297 packages** available on CRAN that use the word **“missing”** in either the title or description of the package, **158** that use either **“impute” or “imputation”** and **74** that use the word **“incomplete”**.

Not all of these packages perform imputation or are useful for our purposes, but even if we excluded those packages, the number of useful packages for dealing with missing data would still be to large to mention them all.

➡ The mice package is often a good option, but certainly not the only option to perform imputation!

Other R Packages that do Imputation

CRAN Task View on Missing Data:

https://cran.r-project.org/web/views/MissingData.html

- gives an overview on the available R packages for missing data / imputation
- good point to start when searching for a package with a particular functionality

Imputation in other Software

In this second half of the course, we have focused on (multiple) imputation using R.

Naturally, R is not the only statistical software that can perform multiple imputation.

- **Stata, SAS and MPLUS** provide packages/functions to perform multiple imputation and pool the results.
- There are macros and additional packages available, e.g., **smcfcs** is implemented for **Stata** as well
- **SPSS** provides some functionality to perform MI

Other Approaches to Handle Missing Values

Finally, we should not forget that **MICE is not the only method to handle missing values**.

Besides MICE, **multiple imputation** can be performed in a **joint model approach** (as for instance implemented in the R package **jomo**).

Furthermore,

- **direct likelihood methods**,
- **fully Bayesian methods** (as implemented in **JointAI**), or
- **weighted estimating equations**

are valid alternative approaches when data are incomplete.

References

Graham, John W, Allison E Olchowski, and Tamika D Gilreath. 2007. “How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory.” *Prevention Science* 8 (3): 206–13.

Rubin, Donald B. 2004. “The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys.” *The American Statistician* 58 (4): 298–302. https://doi.org/10.1198/000313004X6355.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. CRC press.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/Crc Interdisciplinary Statistics. Taylor & Francis. https://stefvanbuuren.name/fimd/.

White, Ian R, Patrick Royston, and Angela M Wood. 2011. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.” *Statistics in Medicine* 30 (4): 377–99.