Journal of Clinical Epidemiology 151 (2022) 185–188

KEY CONCEPTS IN CLINICAL EPIDEMIOLOGY

Handling missing data in clinical research

Martijn W. Heymans, Jos W.R. Twisk*
Department of Epidemiology and Data Science, Amsterdam UMC, Amsterdam, the Netherlands

Accepted 31 August 2022; Published online 21 September 2022

Journal Club

Handling missing data in clinical research

Monchai Suntipap, MD
M.Sc. Clinical Epidemiology Student
Academic Year 2022

Missing data

- Missing data reduced precision because there are fewer observed data.

- Missing outcomes can seriously impair the ability to make correct inferences from studies

- The missing data mechanism should be considered because it is important to handle missing data properly

- The topic of missing data imputation is still evolving and controversial in many aspects

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Goals of missing data replacement

- Minimize bias

- Maximize the use of available information

- Get good estimates of uncertainty

Missing data mechanism

01 Missing completely at random (MCAR)

02 Missing at random (MAR)

03 Missing not at random (MNAR)

Missing data mechanism

1. Missing completely at random (MCAR)

- Missing values are **randomly distributed** over the data sample

- *Ideal condition*, no systematic differences between the observed and missing data

- Missing values due to incidental circumstances

2. Missing at random (MAR)

- The probability of missing data is "**related to other variables"**

- Can be explained by ***associations*** with the observed data

- Systematic differences between the observed and missing data

3. Missing not at random (MNAR)

- The probability of missing data is **dependent on the values of the variable itself**

- Count not be explained by observed variables in the dataset

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Missing data mechanism Scenario

The study investigates which covariates are related to blood pressure in population

→ Blood pressure is missing

1. MCAR = Some people **were not able** to visit the research center = a strike in public transport

2. MAR = More data on blood pressure are **missing** for people with **high BMI**

3. MNAR = Case with the **highest values** for blood pressure **do not visit**

Exploring missing data

- There are a few methods proposed to explore the mechanism but the practical value is dubious

- The tabulation of the missingness pattern can be used to identify possible missing in the data

1.MCAR

- T-tests and logistic regression
  - Investigate if there is a relationship between variables *with and without* missing data
  - Variable with missing data can be coded 0 for the observed and 1 for the missing data

2.MAR and MNAR

- Missing data are related to unobserved data (*impossible to evaluate*)

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Table II. Number of subjects with observed and missing mean probing depth by age

Age, y	*Mean probing depth observed*	*Mean probing depth missing*	*Total*
≥25	51	13	64
<25	35	30	65

- Check whether the proportion of individuals whose mean probing depth is missing differs between the two age groups

1. The Pearson $X^2 = 9.69$, $P = 0.002$ (the evidence against the null hypothesis of no association between mean probing depth missing and age <25)

2. A logistic regression model with mean probing depth missing as the dependent variable and age<25 as the independent variable, and test for an association

(ORs = 3.36; 95% CI 1.54- 7.34, Wald P = 0.002)

Table II. Number of subjects with observed and missing mean probing depth by age

Age, y	*Mean probing depth observed*	*Mean probing depth missing*	*Total*
≥25	51	13	64
<25	35	30	65

P =0.002

- In this example, mean probing depth is more likely to be missing in younger individuals

-From an analysis, this means that the data is not MCAR

-To distinguish between MAR and MNAR, we need to know if the mean probing depth missing depends on the mean probing depth within each age group.

-We cannot determine if MAR holds or not from the observed data alone

Method to deal with missing data

1. Deletion (listwise deletion)

    - Complete-case analysis (CCA)

2. Single imputation (Replacement)

3. Multiple imputation (MI)

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Method to deal with missing data

1. Complete-case analysis (CCA)

- All missing values on one or more variables are excluded from the analysis

- CCA is commonly used but reduces sample size (reduced statistical efficiency of estimates)

- Less precise and increased potential for bias

Method to deal with missing data

2. Imputation(replacement) → Preserve sample size

2.1 Single imputation

- Average imputation
- Regression substitution (replacement with regression predicted value)

(a logistic model for dichotomous outcomes and a linear model for continuous outcomes)

Limitation

- Biased variances
- Under estimated SE
- Ignore natural random values

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

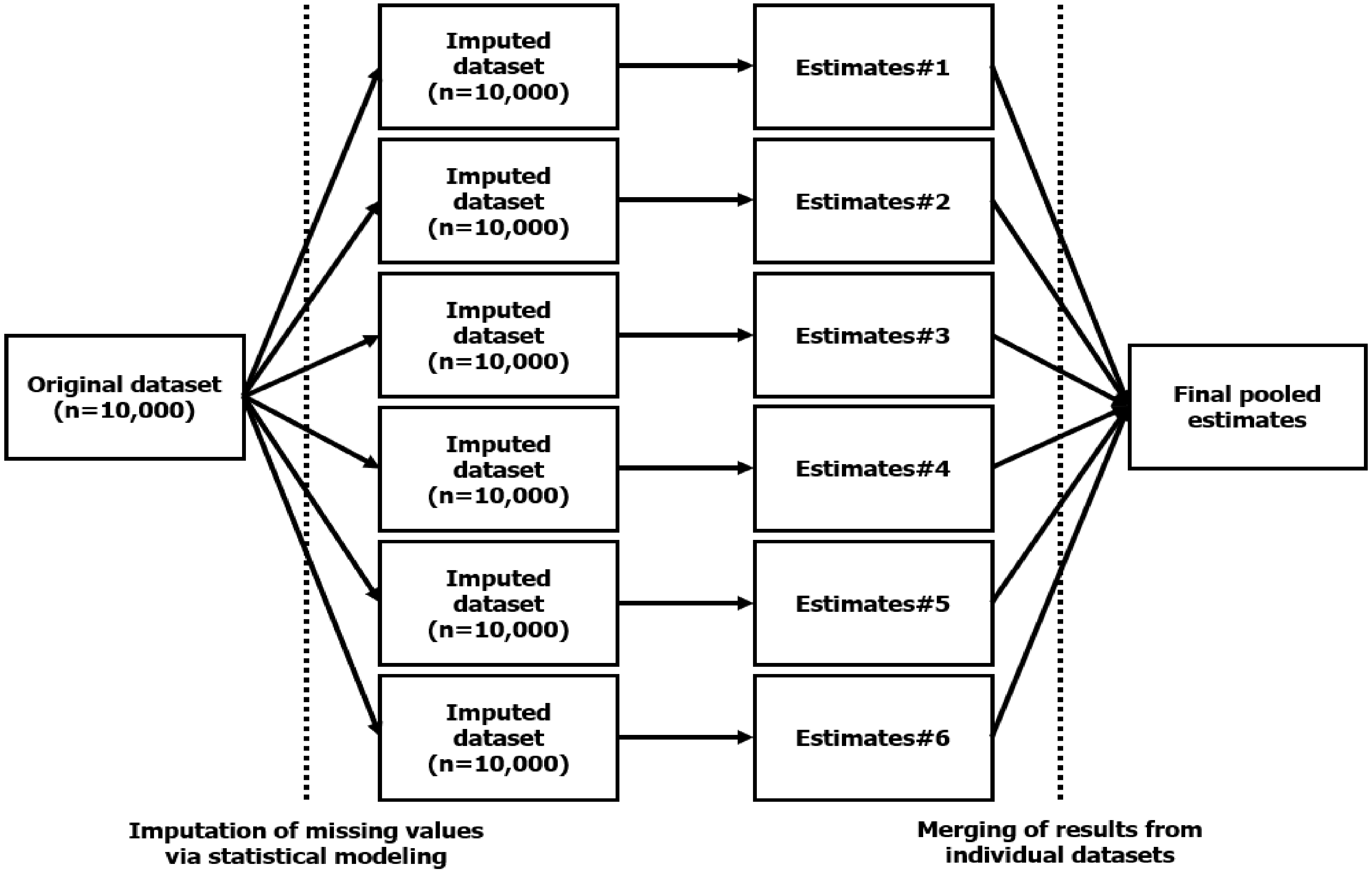
Method to deal with missing data

2.2 Multiple imputation

Three basic steps

- Imputation
  - Introduce random variation, generate several datasets , each with different imputed values
  - The Multivariate Imputation by Chained Equations (MICE) procedure is mostly used
- Analysis
  - Do analysis on each dataset
  - Multiple imputation (MI), inverse probability weighting (IPW), doubly robust inverse probability weighting (DR-IPW), and maximum likelihood estimation (MLE)
- Pooling
  - The results are summarized into one final estimate
  - The uncertainty about the missing data is reflected in the standard error of the pooled effect estimate

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022


Original dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Imputed
dataset
(n=10,000)
Estimates#1
Estimates#2
Estimates#3
Estimates#4
Estimates#5
Estimates#6
Final pooled
estimates
Imputation of missing values
via statistical modeling
Merging of results from
individual datasets

To impute or not to impute

Table 1. Handling missing data: an overview

Missing data mechanism	Analysis	Imputation
MCAR	Complete case analysis	No imputation necessary
MAR	No complete case analysis	Single imputation methods not valid
		Multiple imputation needed
MNAR	No complete case analysis	All imputation methods not valid

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Basic guidance

- MCAR
  - Ignorable; however, lower precision/power
  - Complete-case analysis (CCA)
  - Multiple imputation (MI) gives unbiased results
- MAR
  - Multiple imputation (MI) gives unbiased results
- MNAR
  - Complete-case analysis (CCA), while acknowledging limitations
  - MI should gives biased results

Summary

Table 1: Summary of missing data mechanisms

Missing data mechanism	Related to	Not related to	Probability to be missing	Valid analysis
MCAR		Observed or missing data	Equal for every data point	Complete case analysis, single and multiple imputation
MAR	Observed data	Missing data	Equal for data points within groups	Multiple imputation
MNAR	Missing data		Unequal and unknown	Sensitivity analysis

MAR: missing at random; MCAR: missing completely at random; MNAR: missing not at random.

Conclusion

- CCA may be valid when missing data are MCAR
- MI is only valid when missing data are MAR
- Single imputation leads to the underestimated SE of the effect estimates
- Regarding missing data, prevention is always better than treatment

Heymans MW, et al. Handling missing data in clinical research. J Clin Epidemiol 2022

Thank you