EP16: Missing Values in Clinical Research: Multiple Imputation

1. What is Multiple Imputation?

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

✉ n.erler@erasmusmc.nl

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and
- increased **availability of computers**.

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and
- increased **availability of computers**.

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and
- increased **availability of computers**.

Goal: data should be usable by (Rubin, 1996)

- a **large number of analysts**, who commonly have to rely on

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and
- increased **availability of computers**.

Goal: data should be usable by (Rubin, 1996)

- a **large number of analysts**, who commonly have to rely on
- standard **software that can only handle complete data**, and usually

History & Ideas

- Developed by **Donald B. Rubin** in the 1970s
- to handle missing values in **public use databases** (e.g., census data provided by the government),
- motivated by the **increase in missing values**, and
- increased **availability of computers**.

Goal: data should be usable by (Rubin, 1996)

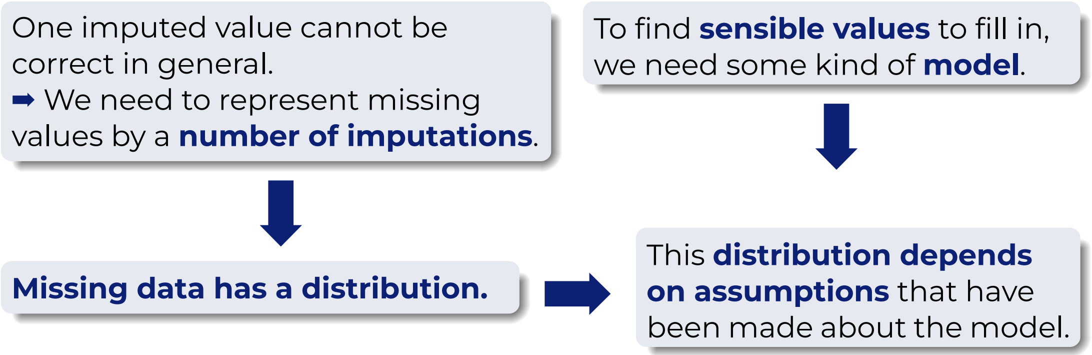
- a **large number of analysts**, who commonly have to rely on
- standard **software that can only handle complete data**, and usually
- are **not experts in handling incomplete data**.

History & Ideas (Rubin, 2004)

One imputed value cannot be correct in general.
➡ We need to represent missing values by a **number of imputations**.

To find **sensible values** to fill in, we need some kind of **model**.

History & Ideas (Rubin, 2004)



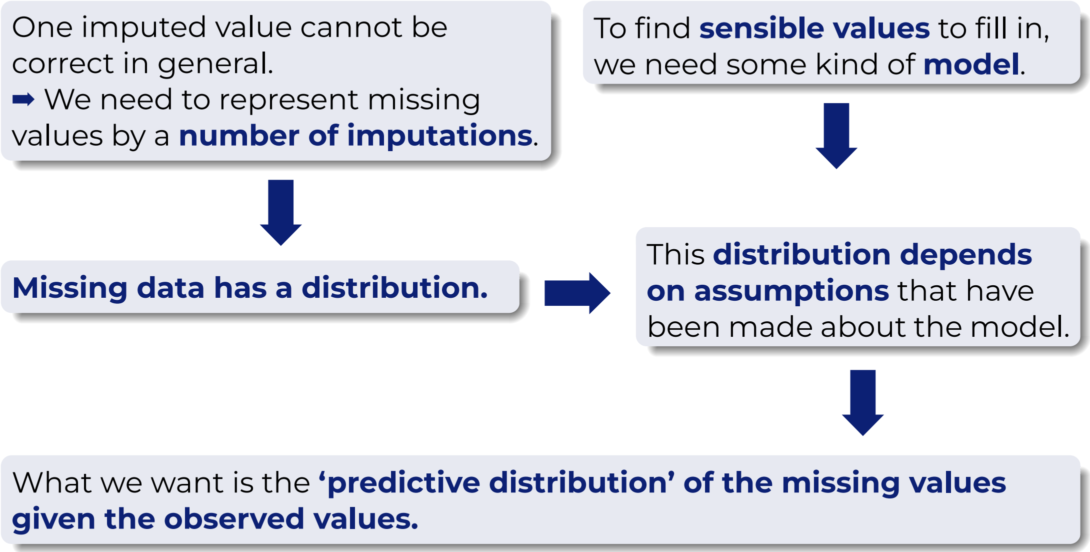
One imputed value cannot be correct in general.
➡ We need to represent missing values by a **number of imputations**.

To find **sensible values** to fill in, we need some kind of **model**.

Missing data has a distribution.

This **distribution depends on assumptions** that have been made about the model.

History & Ideas (Rubin, 2004)



One imputed value cannot be correct in general.
➡ We need to represent missing values by a **number of imputations**.

Missing data has a distribution.

To find **sensible values** to fill in, we need some kind of **model**.

This **distribution depends on assumptions** that have been made about the model.

What we want is the **‘predictive distribution’ of the missing values given the observed values.**

History & Ideas

How to obtain that predictive distribution?

History & Ideas

How to obtain that predictive distribution?

- fit a model to the observed data (“respondents”), and to
- obtain for each “nonrespondent” the conditional distribution of the missing data (given the observed data) as if he/she was a respondent.

➡ We assume that **nonrespondents are just like respondents**, and obtain the predictive distribution from the model of the respondents’ data.

History & Ideas

How to obtain that predictive distribution?

- fit a model to the observed data ("respondents"), and to
- obtain for each "nonrespondent" the conditional distribution of the missing data (given the observed data) as if he/she was a respondent.

➡ We assume that **nonrespondents are just like respondents**, and obtain the predictive distribution from the model of the respondents' data.

Example: survey including age, gender and height

10 – 12 year old boys answered (on average) that they are 1.45m tall.

➡ We assume that boys aged 10 to 12 who did not report their height are also around 1.45m tall.

History & Ideas

How to represent the multiple imputed values?
For each missing value, we now have multiple imputed values.

History & Ideas

How to represent the multiple imputed values?
For each missing value, we now have multiple imputed values.

- For each set of imputed values, create a dataset
  (datasets agree in the observed values but imputed values differ).
- Analyse each dataset.
- Combine the results from all analyses.

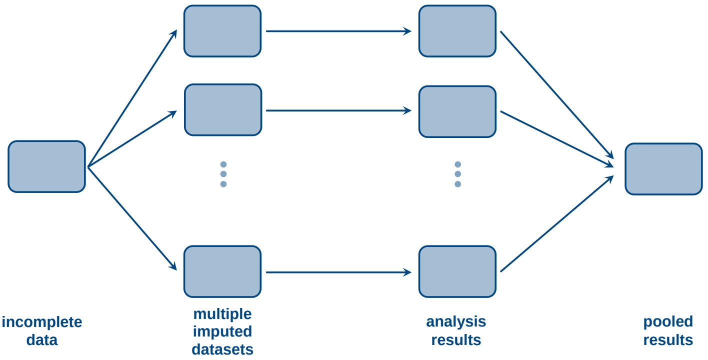
History & Ideas

How to represent the multiple imputed values?
For each missing value, we now have multiple imputed values.

- For each set of imputed values, create a dataset (datasets agree in the observed values but imputed values differ).
- Analyse each dataset.
- Combine the results from all analyses.

➡ We can describe how (much) the **results vary between the imputed datasets**, and calculate summary measures.

Three Steps



In summary:

1. **Imputation:** impute multiple times ➡ multiple completed datasets
2. **Analysis:** analyse each of the datasets
3. **Pooling:** combine results, taking into account additional uncertainty

References

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.2307/2291635

Rubin, D. 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