Dealing with Missing Data

Challenges and Solutions

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

N_Erler www.nerler.com NErler

13 January 2020

Erasmus MC
University Medical Center Rotterdam

Handling Missing Values is Easy!

Functions automatically exclude missing values:

```
## [...]
## Residual standard error:  2.305 on 69 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.09255, Adjusted R-squared:  0.02679
## F-statistic:  1.407 on 5 and 69 DF, p-value:  0.2325
```

Handling Missing Values is Easy!

Functions automatically exclude missing values:

```
## [...]
## Residual standard error:  2.305 on 69 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.09255, Adjusted R-squared:  0.02679
## F-statistic:  1.407 on 5 and 69 DF, p-value:  0.2325
```

Imputation is super easy:

```r
library("mice")
imp <- mice(mydata)
```

However ...

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

Handling Missing Values Correctly is Not So Easy!

Complete case analysis is usually **biased**

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

violation ➡ bias

Imputation ???

Remind me,
how did that imputation
thing work again???

Imputation

Imputation

filling in missing values with (good) "guesses"

Imputation

Imputation

filling in missing values with (good) "guesses"

Important:

Missing values ➡ **uncertainty**
This needs to be taken into account!!!

Imputation

Imputation

filling in missing values with (good) "guesses"

Important:

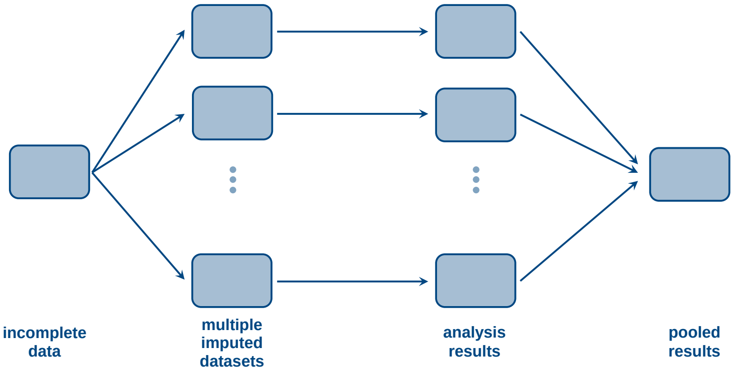
Missing values ➡ **uncertainty**
This needs to be taken into account!!!

Donald Rubin (in the 1970s):
Represent each missing value with **multiple imputed values**

Multiple Imputation

Note:
Imputation is not the only approach to handle missing values.
(Also: maximum likelihood, inverse probability weighting, ...)

Multiple Imputation



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

Imputation Step

Two main approaches

Joint Model Multiple Imputation

- the "original" approach
- often using a multivariate normal distribution

Imputation Step

Two main approaches

Joint Model Multiple Imputation

- the "original" approach
- often using a multivariate normal distribution

Multiple Imputation with Chained Equations (MICE)

- also: **Fully Conditional Specification** (**FCS**)
- now often considered the gold standard

Multiple Imputation with Chained Equations (MICE)

For each incomplete variable, specify a $\underbrace{\text{model using } \textbf{all other variables}}_{\text{full conditionals}}$:

x_1	x_2	x_3	x_4	$\dots$
✓	✓	NA	NA	$\dots$
NA	✓	✓	NA	$\dots$
✓	NA	NA	✓	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

For each incomplete variable, specify a $\underbrace{\text{model using } \textbf{all other variables}}_{\text{full conditionals}}$:

x_1	x_2	x_3	x_4	$\dots$
✓	✓	NA	NA	$\dots$
NA	✓	✓	NA	$\dots$
✓	NA	NA	✓	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

$$
\begin{aligned}
x_1 &\sim x_2 + x_3 + x_4 + \dots \\
x_2 &\sim x_1 + x_3 + x_4 + \dots \\
x_3 &\sim x_1 + x_2 + x_4 + \dots \\
x_4 &\sim x_1 + x_2 + x_3 + \dots \\
&\vdots
\end{aligned}
$$

Multiple Imputation with Chained Equations (MICE)

For each incomplete variable, specify a $\underbrace{\text{model using } \textbf{all other variables}}_{\text{full conditionals}}$:

x_1	x_2	x_3	x_4	$\dots$
✓	✓	NA	NA	$\dots$
NA	✓	✓	NA	$\dots$
✓	NA	NA	✓	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

$$x_1 \sim x_2 + x_3 + x_4 + \dots$$
$$x_2 \sim x_1 + x_3 + x_4 + \dots$$
$$x_3 \sim x_1 + x_2 + x_4 + \dots$$
$$x_4 \sim x_1 + x_2 + x_3 + \dots$$
$$\vdots$$

For example:

- linear regression
- logistic regression
- ...

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess

x_1	x_2	x_3	x_4	$\dots$
✓	✓	NA	NA	$\dots$
NA	✓	✓	NA	$\dots$
✓	NA	NA	✓	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess
- update x_1 based on initial values of $x_2, x_3, x_4, \ldots$

x_1	x_2	x_3	x_4	$\ldots$
✓	✓	NA	NA	$\ldots$
NA	✓	✓	NA	$\ldots$
✓	NA	NA	✓	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess
- update x_1 based on initial values of $x_2, x_3, x_4, \ldots$
- update x_2 based on new x_1 and initial values of $x_3, x_4, \ldots$
- ...

x_1	x_2	x_3	x_4	$\ldots$
✓	✓	NA	NA	$\ldots$
NA	✓	✓	NA	$\ldots$
✓	NA	NA	✓	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess
- update x_1 based on initial values of $x_2, x_3, x_4, \ldots$
- update x_2 based on new x_1 and initial values of $x_3, x_4, \ldots$
- $\ldots$
- update x_1 again, based on updated $x_2, x_3, x_4, \ldots$
- $\ldots$

x_1	x_2	x_3	x_4	$\ldots$
✓	✓	NA	NA	$\ldots$
NA	✓	✓	NA	$\ldots$
✓	NA	NA	✓	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess
- update x_1 based on initial values of $x_2, x_3, x_4, \ldots$
- update x_2 based on new x_1 and initial values of $x_3, x_4, \ldots$
- $\ldots$
- update x_1 again, based on updated $x_2, x_3, x_4, \ldots$
- $\ldots$
- until **convergence**

x_1	x_2	x_3	x_4	$\ldots$
✓	✓	NA	NA	$\ldots$
NA	✓	✓	NA	$\ldots$
✓	NA	NA	✓	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Multiple Imputation with Chained Equations (MICE)

MICE is an iterative algorithm:

- start with initial guess
- update x_1 based on initial values of $x_2, x_3, x_4, \ldots$
- update x_2 based on new x_1 and initial values of $x_3, x_4, \ldots$
- $\ldots$
- update x_1 again, based on updated $x_2, x_3, x_4, \ldots$
- $\ldots$
- until **convergence**

x_1	x_2	x_3	x_4	$\ldots$
✓	✓	NA	NA	$\ldots$
NA	✓	✓	NA	$\ldots$
✓	NA	NA	✓	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Values from **last iteration ➡ one imputed dataset**

MICE Makes Assumptions

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

Missing Data Mechanisms

Missing Completely At Random (MCAR)

Missing At Random (MAR)

Missing Not At Random (MNAR)

Missing Data Mechanisms

Missing Completely At Random (MCAR)

$$p(R \mid X_{obs}, X_{mis}) = p(R)$$

Missingness is independent of all data.

questionnaire got lost in mail

Missing At Random (MAR)

Missing Not At Random (MNAR)

Missing Data Mechanisms

Missing Completely At Random (MCAR)

$$p(R \mid X_{obs}, X_{mis}) = p(R)$$

Missingness is independent of all data.

questionnaire got lost in mail

Missing At Random (MAR)

$$p(R \mid X_{obs}, X_{mis}) = p(R \mid X_{obs})$$

Missingness depends only on observed data.

overweight participants are less likely to report their chocolate consumption (and we know their weight)

Missing Not At Random (MNAR)

Missing Data Mechanisms

Missing Completely At Random (MCAR)

$$p(R \mid X_{obs}, X_{mis}) = p(R)$$

Missingness is independent of all data.

questionnaire got lost in mail

Missing At Random (MAR)

$$p(R \mid X_{obs}, X_{mis}) = p(R \mid X_{obs})$$

Missingness depends only on observed data.

overweight participants are less likely to report their chocolate consumption (and we know their weight)

Missing Not At Random (MNAR)

$$p(R \mid X_{obs}, X_{mis}) \neq p(R \mid X_{obs})$$

Missingness depends (also) on unobserved data.

overweight participants are less likely to report their weight

MICE Makes Assumptions

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

In case of MNAR:
MICE ➡ bias

MICE Makes Assumptions

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

Imputation Model Misspecification

$$
\begin{aligned}
x_1 &\sim x_2 + x_3 + x_4 + \dots \\
x_2 &\sim x_1 + x_3 + x_4 + \dots \\
x_3 &\sim x_1 + x_2 + x_4 + \dots \\
x_4 &\sim x_1 + x_2 + x_3 + \dots \\
&\vdots
\end{aligned}
$$

For example:

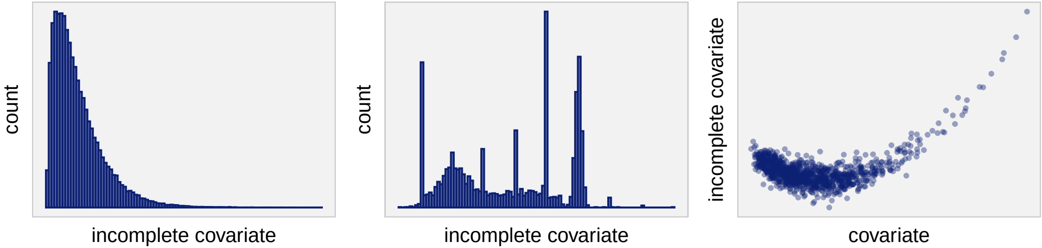
- linear regression
- logistic regression
- ...

Imputation Model Misspecification

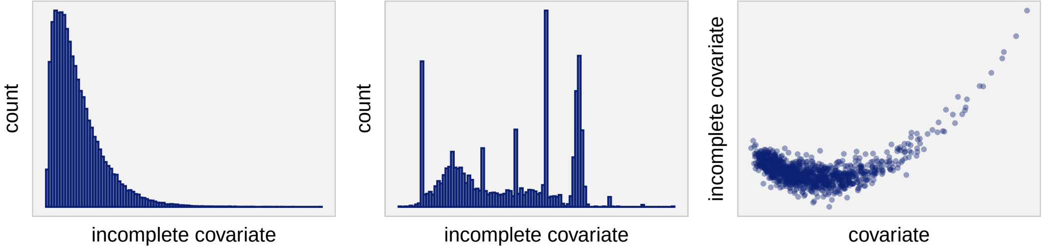
$$x_1 \sim x_2 + x_3 + x_4 + \ldots$$
$$x_2 \sim x_1 + x_3 + x_4 + \ldots$$
$$x_3 \sim x_1 + x_2 + x_4 + \ldots$$
$$x_4 \sim x_1 + x_2 + x_3 + \ldots$$
$$\vdots$$

For example:

- linear regression
- logistic regression
- ...

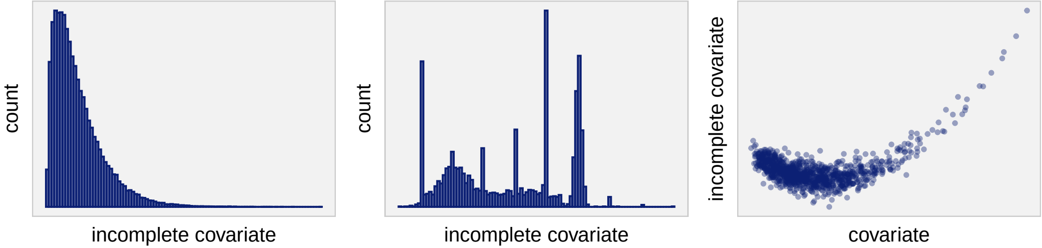


Imputation Model Misspecification



- misspecification of the **residual distribution**
- misspecification of the **association structure**

Imputation Model Misspecification

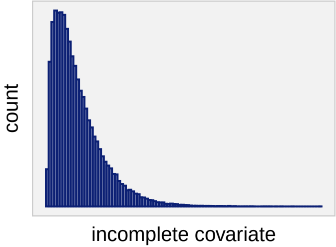


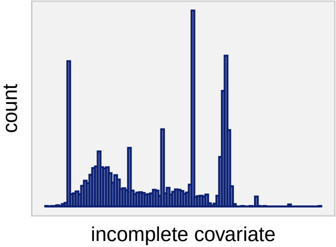
- misspecification of the **residual distribution**
- misspecification of the **association structure**

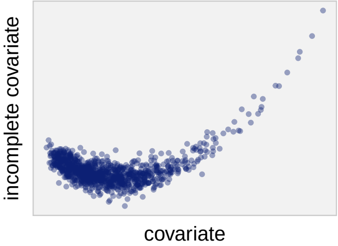
Partial solutions:

- Predictive Mean Matching
- Passive imputation

Imputation Model Misspecification







- misspecification of the **residual distribution**
- misspecification of the **association structure**

Partial solutions:

- Predictive Mean Matching
- Passive imputation

But...

- can get tedious
- requires knowledge (about data & methods)
- users often inexperienced and/or lazy

MICE Makes Assumptions

(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

Model misspecification ➡ bias

MICE Makes Assumptions

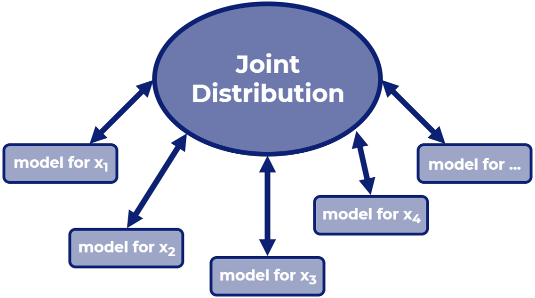
(**Imputation**) methods make certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
- **compatibility** and **congeniality**

Compatibility & Congeniality

Compatibility

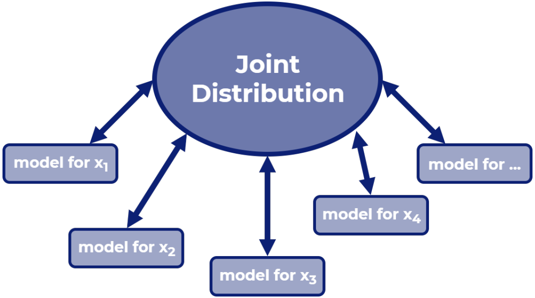
A **joint distribution** exists, that has the full conditionals (imputation models) as its conditional distributions.



Compatibility & Congeniality

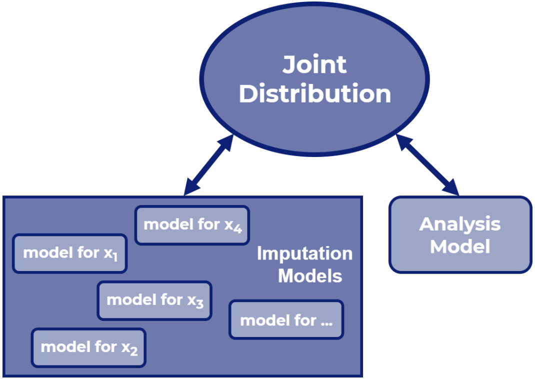
Compatibility

A **joint distribution** exists, that has the full conditionals (imputation models) as its conditional distributions.



Congeniality

The imputation model is compatible with the **analysis model**.

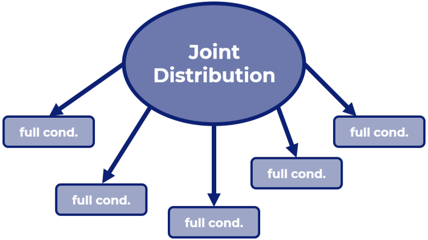


Compatibility & Congeniality in MICE

MICE is based on the idea of

Gibbs sampling

Exploits the fact that a joint distribution is fully determined by its full conditional distributions.

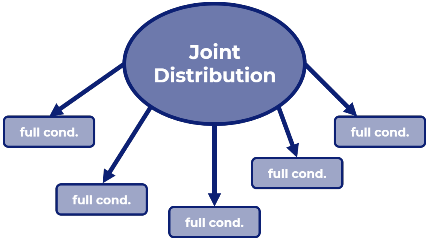


Compatibility & Congeniality in MICE

MICE is based on the idea of

Gibbs sampling

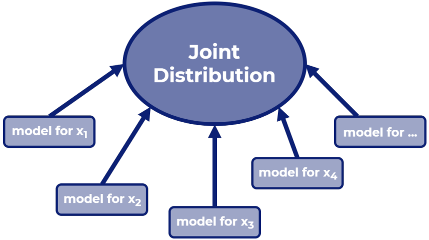
Exploits the fact that a joint distribution is fully determined by its full conditional distributions.



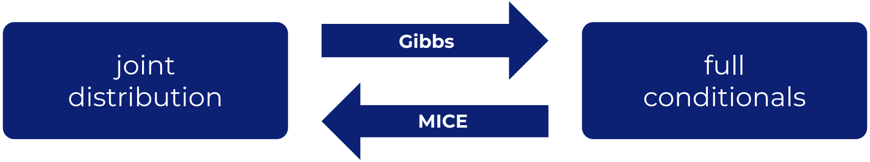
But:

In MICE

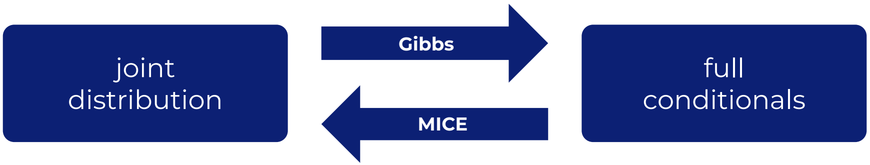
Imputation models are specified directly
➡ **no guarantee** that a corresponding joint distribution exists



Compatibility & Congeniality in MICE



Compatibility & Congeniality in MICE



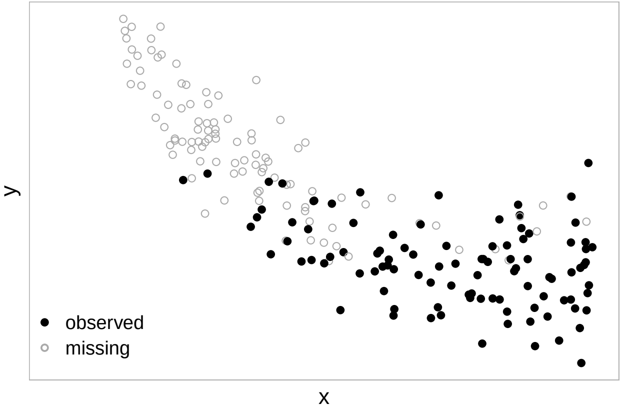
Is this a problem?

- often not
- but it can be when
  - imputation/analysis **models contradict each other**
  - **different assumptions** are made during analysis and imputation
  - the **outcome cannot easily be included** in the imputation models

Example 1: Contradicting Models

Analysis model with a **quadratic association**:

$$y = \beta_0 + \beta_1 x + \boldsymbol{\beta_2 \mathbf{x}^2} + \ldots$$

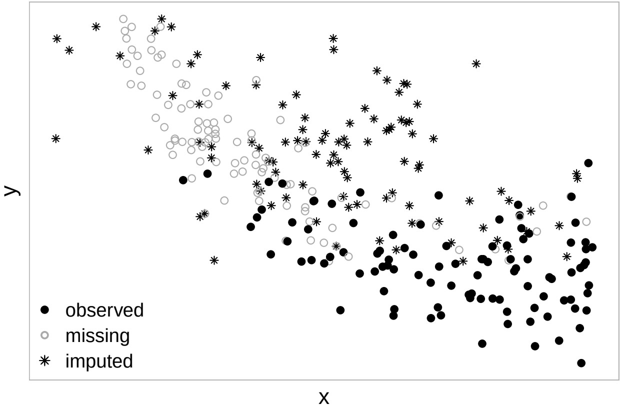


Example 1: Contradicting Models

Imputation model for x (when using MICE naively):

$$x = \theta_{10} + \theta_{11} y + \ldots,$$

i.e., a **linear relation** between x and y is assumed.

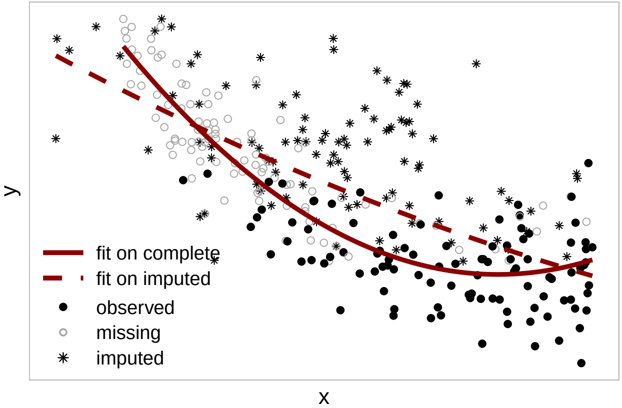


Example 1: Contradicting Models

Imputation model for x (when using MICE naively):

$$x = \theta_{10} + \theta_{11} y + \dots,$$

i.e., a **linear relation** between x and y is assumed.

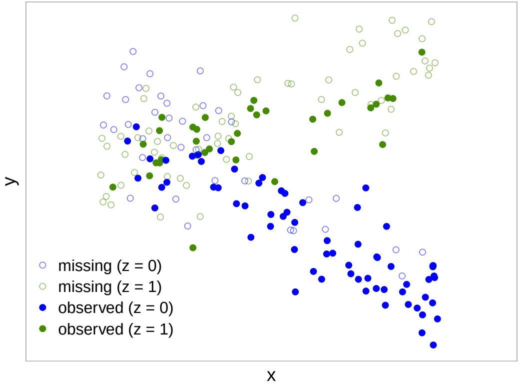


Example 2: Contradicting Models

Analysis model with **interaction term**:

$$y = \beta_0 + \beta_x x + \beta_z z + \color{red}{\beta_{xz} xz} + \ldots,$$

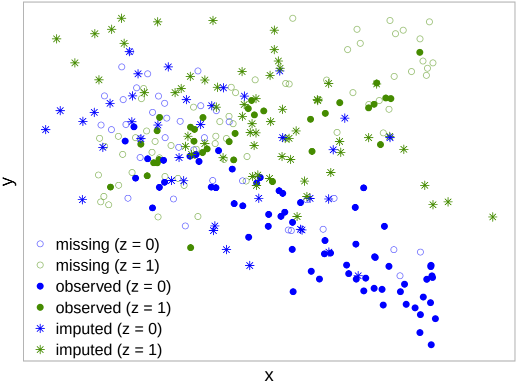
i.e., y again has a **non-linear relationship** with x



Example 2: Contradicting Models

Imputation model for x (when using MICE naively):

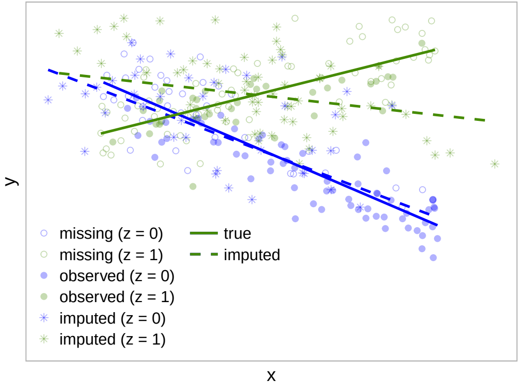
$$x = \theta_{10} + \theta_{11} y + \theta_{12} z + \ldots,$$



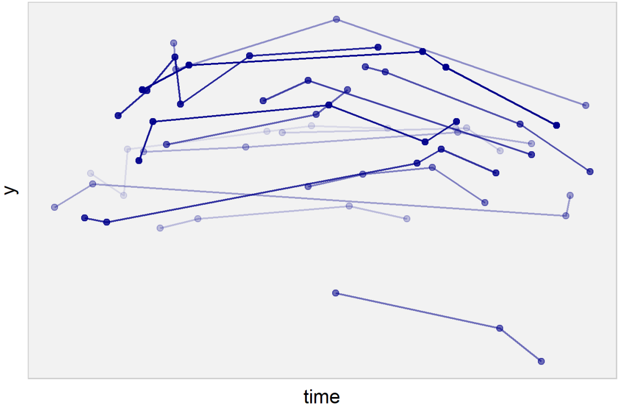
Example 2: Contradicting Models

Imputation model for x (when using MICE naively):

$$x = \theta_{10} + \theta_{11} y + \theta_{12} z + \ldots,$$



Example 3: Longitudinal / Multi-level Data



id	time	y	x
1	0.34	0.12	✓
1	0.65	-0.04	✓
1	0.68	0.30	✓
1	1.97	0.44	✓
1	2.38	0.48	✓
1	3.09	0.46	✓
2	2.11	0.43	NA
2	3.72	0.46	NA
2	3.82	0.46	NA
2	4.13	0.29	NA
⋮	⋮	⋮	⋮

Example 3: Longitudinal / Multi-level Data

Imputation in **long format**

- rows are treated as independent
- imputations in baseline covariates will vary over time

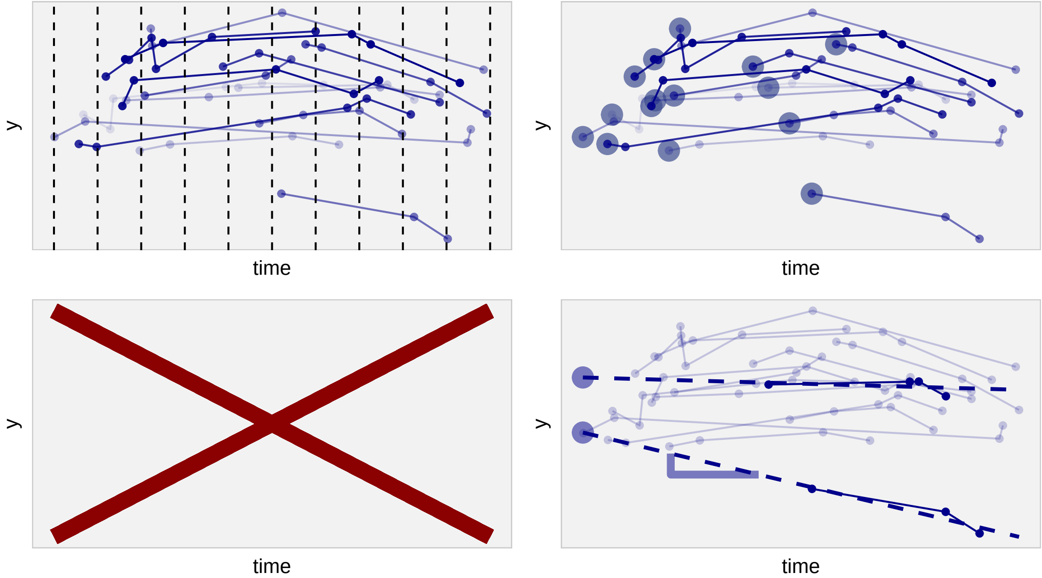
➡ **bias**

Can we use data in **wide format** (one row per subject)?

- can be very inefficient
- not always possible

id	time	y	x
1	0.34	0.12	✓
1	0.65	-0.04	✓
1	0.68	0.30	✓
1	1.97	0.44	✓
1	2.38	0.48	✓
1	3.09	0.46	✓
2	2.11	0.43	NA
2	3.72	0.46	NA
2	3.82	0.46	NA
2	4.13	0.29	NA
⋮	⋮	⋮	⋮

Example 3: Longitudinal / Multi-level Data



Compatibility & Congeniality in MICE

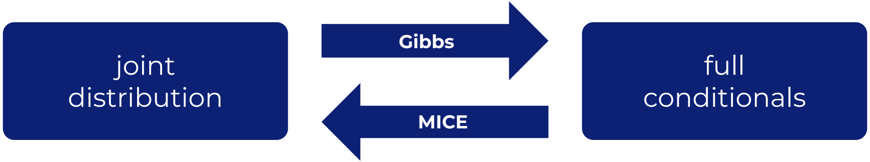
Lack of compatibility / congeniality can become a **problem for MICE** in settings with

- Non-linear associations
  - non-linear effects
  - interaction terms
  - ...
- complex outcomes
  - multi-level settings
  - time-to-event outcomes
  - ...

What can we do in these settings?

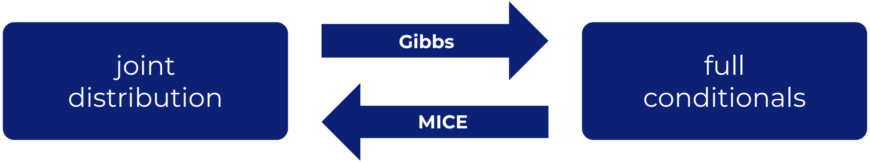
Imputation in Complex Settings

Remember, the **problem** is



Imputation in Complex Settings

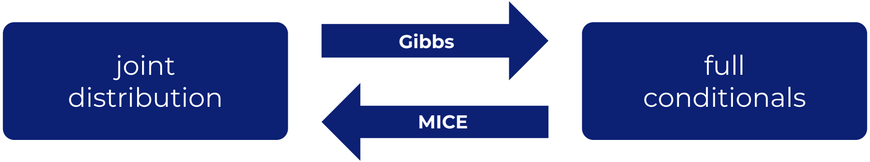
Remember, the **problem** is



➡ **Solution:** Start with the joint distribution!

Imputation in Complex Settings

Remember, the **problem** is

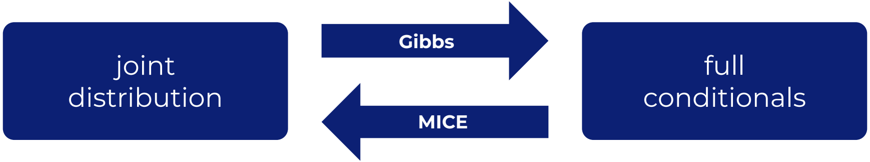


➡ **Solution:** Start with the joint distribution!

New problem:
What is the multivariate distribution of **multiple variables** of **different types**?

Imputation in Complex Settings

Remember, the **problem** is



➡ **Solution:** Start with the joint distribution!

New problem:
What is the multivariate distribution of **multiple variables** of **different types**?

☹ Usually, the joint distribution is not of any known form.

Joint Model Imputation

Multivariate Normal Model

Approximate the joint distribution by a known multivariate (usually normal) distribution

- this is **Joint Model** Multiple Imputation

☺ assures compatibility & congeniality

☹ can't handle non-linear associations

Joint Model Imputation

Multivariate Normal Model

Approximate the joint distribution by a known multivariate (usually normal) distribution

- ▶ this is **Joint Model** Multiple Imputation
- ☺ assures compatibility & congeniality
- ☹ can't handle non-linear associations

Sequential Factorization

Factorize the joint distribution into (a sequence of) conditional distributions.

- ☺ assures compatibility & congeniality
- ☺ can handle non-linear associations

Sequential Factorization

A **joint distribution** $p(y, x)$ can be written as the product of conditional distributions:

$$p(y, x) = p(y \mid x)\, p(x)$$

(or alternatively $p(y, x) = p(x \mid y)\, p(y)$)

Sequential Factorization

A **joint distribution** $p(y, x)$ can be written as the product of conditional distributions:

$$p(y, x) = p(y \mid x)\, p(x)$$

(or alternatively $p(y, x) = p(x \mid y)\, p(y)$)

This can be **extended for more variables**:

$$p(y, x_1, \ldots, x_p) = p(y \mid x_1, \ldots, x_p)\, p(x_1 \mid x_2, \ldots, x_p)\, p(x_2 \mid x_3, \ldots, x_p)\, \ldots\, p(x_p)$$

Sequential Factorization in the Bayesian Framework

Joint Distribution

$$p(y, X, \theta) = \underbrace{p(y \mid X, \theta)}_{\substack{\text{analysis}\\ \text{model}}} \; \underbrace{p(X \mid \theta)}_{\substack{\text{imputation}\\ \text{part}}} \; \underbrace{p(\theta)}_{\text{priors}}$$

θ contains regr. coefficients, variance parameters, …

Sequential Factorization in the Bayesian Framework

Joint Distribution

$$p(y, X, \theta) = \underbrace{p(y \mid X, \theta)}_{\substack{\text{analysis}\\\text{model}}} \quad \underbrace{p(X \mid \theta)}_{\substack{\text{imputation}\\\text{part}}} \quad \underbrace{p(\theta)}_{\text{priors}}$$

θ contains regr. coefficients, variance parameters, ...

Imputation part

$$\begin{aligned} p(\overbrace{x_1, \ldots, x_p, X_{compl.}}^{X} \mid \theta) \quad = \quad & p(x_1 \mid X_{compl.}, \theta) \\ & p(x_2 \mid X_{compl.}, x_1, \theta) \\ & p(x_3 \mid X_{compl.}, x_1, x_2, \theta) \\ & \ldots \end{aligned}$$

Sequential Factorization in the Bayesian Framework

Extension for a Multi-level Setting

$$\underbrace{\mathbf{p}(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \theta)}_{\substack{\text{analysis}\\\text{model}}} \; \underbrace{p(X \mid \theta)}_{\substack{\text{imputation}\\\text{part}}} \; \underbrace{\mathbf{p}(\mathbf{b} \mid \theta)}_{\substack{\text{random}\\\text{effects}}} \; \underbrace{p(\theta)}_{\text{priors}}$$

Sequential Factorization in the Bayesian Framework

Extension for a Multi-level Setting

$$\underbrace{\mathbf{p}(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\substack{\text{analysis}\\\text{model}}} \; \underbrace{p(X \mid \theta)}_{\substack{\text{imputation}\\\text{part}}} \; \underbrace{\mathbf{p}(\mathbf{b} \mid \boldsymbol{\theta})}_{\substack{\text{random}\\\text{effects}}} \; \underbrace{p(\theta)}_{\text{priors}}$$

Extension for a Time-to-Event Outcome

$$\underbrace{\mathbf{p}(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \boldsymbol{\theta})}_{\substack{\text{analysis}\\\text{model}}} \; \underbrace{p(X \mid \theta)}_{\substack{\text{imputation}\\\text{part}}} \; \underbrace{p(\theta)}_{\text{priors}}$$

Sequential Factorization in the Bayesian Framework

Extension for a Multi-level Setting

$$\underbrace{\mathbf{p}(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\text{analysis model}} \; \underbrace{p(X \mid \theta)}_{\text{imputation part}} \; \underbrace{\mathbf{p}(\mathbf{b} \mid \boldsymbol{\theta})}_{\text{random effects}} \; \underbrace{p(\theta)}_{\text{priors}}$$

Extension for a Time-to-Event Outcome

$$\underbrace{\mathbf{p}(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \boldsymbol{\theta})}_{\text{analysis model}} \; \underbrace{p(X \mid \theta)}_{\text{imputation part}} \; \underbrace{p(\theta)}_{\text{priors}}$$

Extension for a Multivariate Outcome

$$\underbrace{\mathbf{p}(\mathbf{y}_1, \mathbf{y}_2 \mid \mathbf{X}, \boldsymbol{\theta})}_{\text{analysis model}} \; \underbrace{p(X \mid \theta)}_{\text{imputation part}} \; \underbrace{p(\theta)}_{\text{priors}}$$

MICE vs Sequential Factorization

Imputation in MICE

$$p(x_1 \mid y, X_{compl.}, x_2, x_3, x_4, \ldots, \theta)$$
$$p(x_2 \mid y, X_{compl.}, x_1, x_3, x_4, \ldots, \theta)$$
$$p(x_3 \mid y, X_{compl.}, x_1, x_2, x_4, \ldots, \theta)$$
$$\ldots$$

Sequential Factorization

$$p(y \mid X_{compl.}, x_1, x_2, x_3, \ldots, \theta)$$
$$p(x_1 \mid X_{compl.}, \theta)$$
$$p(x_2 \mid X_{compl.}, x_1, \theta)$$
$$p(x_3 \mid X_{compl.}, x_1, x_2, \theta)$$
$$\ldots$$

MICE vs Sequential Factorization

Imputation in MICE

$$p(x_1 \mid \mathbf{y}, X_{compl.}, x_2, x_3, x_4, \ldots, \theta)$$
$$p(x_2 \mid \mathbf{y}, X_{compl.}, x_1, x_3, x_4, \ldots, \theta)$$
$$p(x_3 \mid \mathbf{y}, X_{compl.}, x_1, x_2, x_4, \ldots, \theta)$$
$$\ldots$$

Sequential Factorization

$$\mathbf{p}(\mathbf{y} \mid \mathbf{X}_{\mathbf{compl.}}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \theta)$$
$$p(x_1 \mid X_{compl.}, \theta)$$
$$p(x_2 \mid X_{compl.}, x_1, \theta)$$
$$p(x_3 \mid X_{compl.}, x_1, x_2, \theta)$$
$$\ldots$$

No issues with

- complex outcomes, e.g.:
  - multi-level
  - survival
- non-linear effects
- congeniality
- compatibility

MICE vs Sequential Factorization

Imputation in MICE

$$p(x_1 \mid y, X_{compl.}, x_2, x_3, x_4, \ldots, \theta)$$
$$p(x_2 \mid y, X_{compl.}, x_1, x_3, x_4, \ldots, \theta)$$
$$p(x_3 \mid y, X_{compl.}, x_1, x_2, x_4, \ldots, \theta)$$
$$\ldots$$

Sequential Factorization

$$p(y \mid X_{compl.}, x_1, x_2, x_3, \ldots, \boldsymbol{\theta})$$
$$p(x_1 \mid X_{compl.}, \theta)$$
$$p(x_2 \mid X_{compl.}, x_1, \theta)$$
$$p(x_3 \mid X_{compl.}, x_1, x_2, \theta)$$
$$\ldots$$

Analysis model part of specification

- ➡ parameters of interest directly available
- ➡ no need for pooling
- ➡ simultaneous analysis and imputation

Joint Analysis and Imputation in R

Sequential Factorization is implemented in the R package **JointAI**

Joint Analysis and Imputation in R

Sequential Factorization is implemented in the R package **JointAI**

Bayesian analysis of **incomplete data** using

- (generalized) linear regression
- (generalized) linear mixed models
- ordinal (mixed) models
- parametric (Weibull) time-to-event models
- Cox proportional hazards models

Joint Analysis and Imputation in R

Sequential Factorization is implemented in the R package **JointAI**

Bayesian analysis of **incomplete data** using

- (generalized) linear regression
- (generalized) linear mixed models
- ordinal (mixed) models
- parametric (Weibull) time-to-event models
- Cox proportional hazards models

- **on CRAN:** https://CRAN.R-project.org/package=JointAI
- **GitHub:** https://github.com/NErler/JointAI
- **website:** https://nerler.github.io/JointAI/

Joint Analysis and Imputation in R

type	standard regression		mixed model	
	outcome	covariate	outcome	covariate
normal	✓	✓	✓	✓
lognormal	(soon)	✓	(soon)	✓
Gamma	✓	✓	✓	✓
beta	(soon)	✓	(soon)	(soon)
binomial	✓	✓	✓	✓
poisson	✓	(soon)	✓	✓
ordinal	✓	✓	✓	✓
multinomial	(soon)	✓	(soon)	(soon)

Available soon:

- Joint models (of longitudinal & time-to-event data)
- Multivariate models

JointAI: How does it work?

Requirements:

- R (https://cran.r-project.org/)
- **JAGS** (Just Another Gibbs Sampler; https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/)

JointAI: How does it work?

Requirements:

- R (https://cran.r-project.org/)
- **JAGS** (Just Another Gibbs Sampler; https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/)

Installation:

```
install.packages("JointAI")
```

JointAI: How does it work?

Requirements:

- R (https://cran.r-project.org/)
- **JAGS** (Just Another Gibbs Sampler; https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/)

Installation:

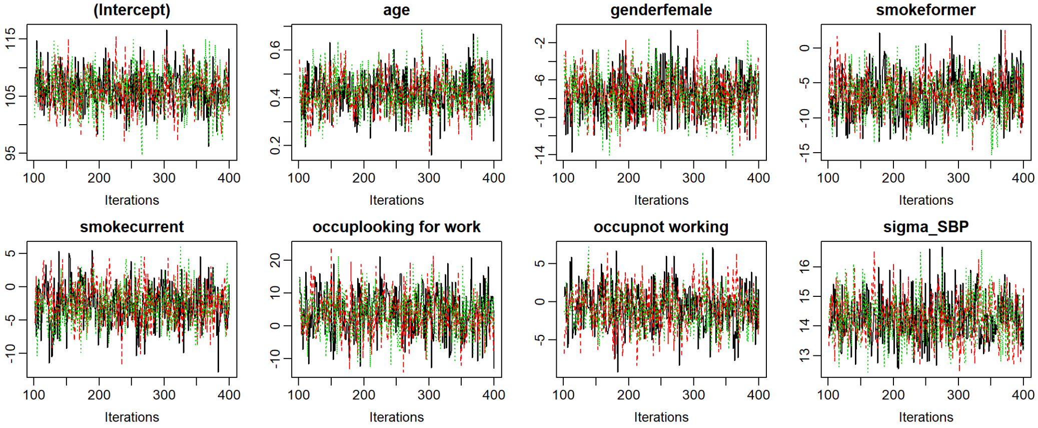
```
install.packages("JointAI")
```

Usage:

```
library("JointAI")
res <- lm_imp(SBP ~ age + gender + smoke + occup, data = NHANES,
              n.iter = 300)
```

JointAI: How does it work?

```
traceplot(res)
```



JointAI: How does it work?

```r
summary(res)
```

```
##
##  Linear model fitted with JointAI
##
## Call:
## lm_imp(formula = SBP ~ age + gender + smoke + occup, data = NHANES,
##     n.iter = 300)
##
## Posterior summary:
##                          Mean     SD    2.5%   97.5% tail-prob. GR-crit
## (Intercept)           106.222 3.3979  99.461 112.961     0.0000    1.00
## age                     0.427 0.0798   0.278   0.583     0.0000    1.00
## genderfemale           -7.450 2.2718 -11.755  -3.072     0.0000    1.00
## smokeformer            -6.692 3.0297 -12.342  -0.885     0.0267    1.03
## smokecurrent           -2.658 3.0229  -8.450   3.313     0.3711    1.01
## occuplooking for work   3.817 6.4037  -9.487  16.087     0.5044    1.01
## occupnot working       -0.869 2.6858  -6.110   4.256     0.7511    1.02
##
## Posterior summary of residual std. deviation:
##           Mean    SD 2.5% 97.5% GR-crit
## sigma_SBP 14.3 0.753 12.8  15.8   0.999
##
##
## MCMC settings
## [...]
```

What is left to do?

(**Imputation**) methods make certain **assumptions**, e.g.:

- ? **missingness** is M(C)AR
- ? the incomplete variable has a certain conditional **distribution**
- ? all associations are **linear**
- ☑ **compatibility** and **congeniality**

What is left to do?

(**Imputation**) methods make certain **assumptions**, e.g.:

- ? **missingness** is M(C)AR
- ? the incomplete variable has a certain conditional **distribution**
- ? all associations are **linear**
- ☑ **compatibility** and **congeniality**

➡ extension to MNAR using pattern mixture model

What is left to do?

(**Imputation**) methods make certain **assumptions**, e.g.:

- ? **missingness** is M(C)AR
- ? the incomplete variable has a certain conditional **distribution**
- ? all associations are **linear**
- ☑ **compatibility** and **congeniality**

➡ extension to MNAR using pattern mixture model

➡ non-parametric Bayesian methods

What is left to do?

(**Imputation**) methods make certain **assumptions**, e.g.:

- ? **missingness** is M(C)AR ➡ extension to MNAR using pattern mixture model
- ? the incomplete variable has a certain conditional **distribution** ➡ non-parametric Bayesian methods
- ? all associations are **linear** ➡ semi-parametric methods
- ☑ **compatibility** and **congeniality**

Take-Home Message

- handling missing values **correctly: not that easy**
- all methods have **assumptions**
  violation ➡ bias

Take-Home Message

- handling missing values **correctly: not that easy**
- all methods have **assumptions**
  violation ➡ bias
- good use of (imputation) methods requires
  - **knowledge** of the **data**
  - **knowledge** of the **methods**
  - **knowledge** of the **software**
  - **time & patience!**

Take-Home Message

- handling missing values **correctly: not that easy**
- all methods have **assumptions**
  violation ➡ bias
- good use of (imputation) methods requires
  - **knowledge** of the **data**
  - **knowledge** of the **methods**
  - **knowledge** of the **software**
  - **time & patience!**
- **JointAI** aims to **facilitate correct handling of missing values** by
  - assuring compatibility & congeniality
  - simultaneous analysis & imputation
  - especially for complex settings

Take-Home Message

- handling missing values **correctly: not that easy**
- all methods have **assumptions**
  violation ➡ bias
- good use of (imputation) methods requires
  - **knowledge** of the **data**
  - **knowledge** of the **methods**
  - **knowledge** of the **software**
  - **time & patience!**
- **JointAI** aims to **facilitate correct handling of missing values** by
  - assuring compatibility & congeniality
  - simultaneous analysis & imputation
  - especially for complex settings
- **There is no magical solution** that will always work in all settings.

Thank you for your attention.

n.erler@erasmusmc.nl
N_Erler
NErler
www.nerler.com

Erasmus MC
University Medical Center Rotterdam