Imputation of missing covariates: when standard methods may fail

Nicole S. Erler[1,2], Dimitris Rizopoulos[1], Oscar H. Franco[2], Emmanuel M.E.H. Lesaffre[1,3]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, the Netherlands
[2] Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands
[3] L-Biostat, KU Leuven, Leuven, Belgium

Motivation (1)

Vitamin D concentration during fetal life and bone health at age 6

- bone mineral content (BMC)
- serum vitamin D concentration (⊛)
- sun exposure (⊛), season at measurement (⊛)
- gender, age at measurement
- . . . (⊛)

(⊛) incomplete

Analysis model:

$$BMD = (age + VitD + VitD^2) \times gender + season + sun\_exposure + \ldots$$

Motivation (2)

Maternal sugar-sweetened bevarage consumption and child's body composition

- child BMI at up to 13 time points
- maternal sugar-sweetened bevarage consumption (SBC)
- child's physical activity, TV watching (⊛)
- gender, age at measurement
- . . . (⊛)

(⊛) incomplete

Analysis model:

$$BMI_{ij} = SBC_i + age_{ij} + \ldots + u_{0i} + u_{1i} \times age_{ij}$$

Standard for imputation: Multiple Imputation (MI)

impute ➡ analyze ➡ pool

Standard for imputation: Multiple Imputation (MI)

impute ➡ analyze ➡ pool

fully conditional specification (**FCS**)
chained equations (**MICE**)

joint model imputation

Standard for imputation: Multiple Imputation (MI)

impute ➡ analyze ➡ pool

fully conditional specification (**FCS**)
chained equations (**MICE**)

joint model imputation

⬇

In iteration $k = 1, \ldots, K$:

for variable $j = 1, \ldots, p$:

- Draw **parameter** $\hat{\boldsymbol{\theta}}_j^k \sim p(\boldsymbol{\theta}_j^k \mid \boldsymbol{x}_j^{obs}, \hat{\boldsymbol{X}}_{-j}^k)$
- Draw **imputation** $\hat{\boldsymbol{x}}_j^k \sim p(\boldsymbol{x}_j^{mis} \mid \boldsymbol{x}_j^{obs}, \boldsymbol{X}_{-j}^k, \hat{\boldsymbol{\theta}}_j^k)$

} e.g. regression with **all other variables** in the lin. predictor

Standard for imputation: Multiple Imputation (MI)

impute ➡ analyze ➡ pool

fully conditional specification (**FCS**)
chained equations (**MICE**)

joint model imputation

⬇

In iteration $k = 1, \ldots, K$:

for variable $j = 1, \ldots, p$:

- Draw **parameter** $\hat{\boldsymbol{\theta}}_j^k \sim p(\boldsymbol{\theta}_j^k \mid \boldsymbol{x}_j^{obs}, \hat{\boldsymbol{X}}_{-j}^k)$
- Draw **imputation** $\hat{\boldsymbol{x}}_j^k \sim p(\boldsymbol{x}_j^{mis} \mid \boldsymbol{x}_j^{obs}, \boldsymbol{X}_{-j}^k, \hat{\boldsymbol{\theta}}_j^k)$

} e.g. regression with **all other variables** in the lin. predictor

➡ keep last iteration ➡ 1 imputed data set ➡ repeat **m** times

Requirements for MICE

- **all** relevant **variables** must be included
  - covariates (from all analyses)
  - the **outcome**
- **compatibility:** a joint model exists that has the imputation models as its conditional distributions
- **congeniality:** compatibility between analysis model and imputation model
- imputation models should **fit the data**
- **M(C)AR** (in most implementations)

When MICE might fail

Imputation model not congenial with analysis:

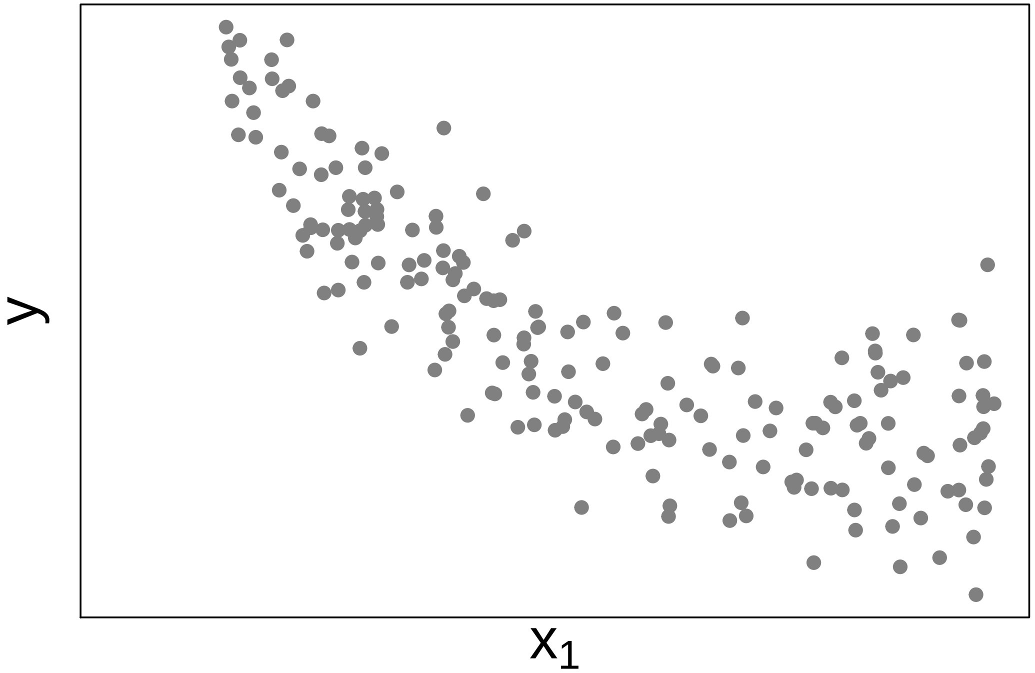
- quadratic, logarithmic, . . . effects
- interactions between covariates

Complex (non univariate) outcomes:

- survival
- longitudinal

Uncongeniality

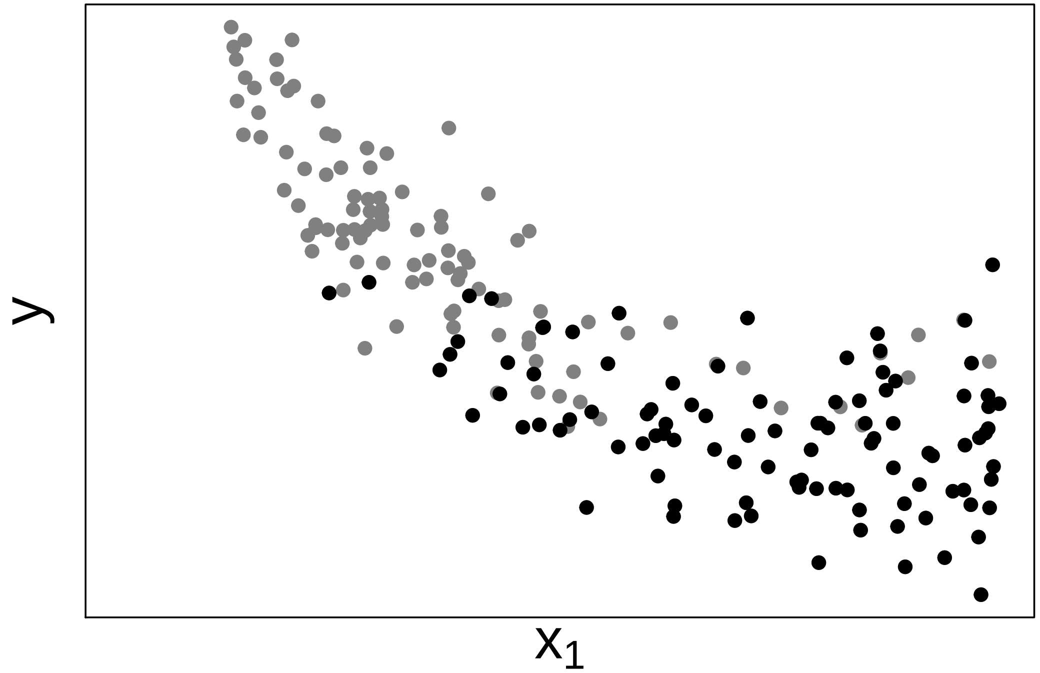
True model:	$y = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2 x_1^2} + \ldots$	(quadratic association)
Imputation model:	$x_1 = \theta_{10} + \theta_{11} y + \ldots$	(linear association)



Uncongeniality

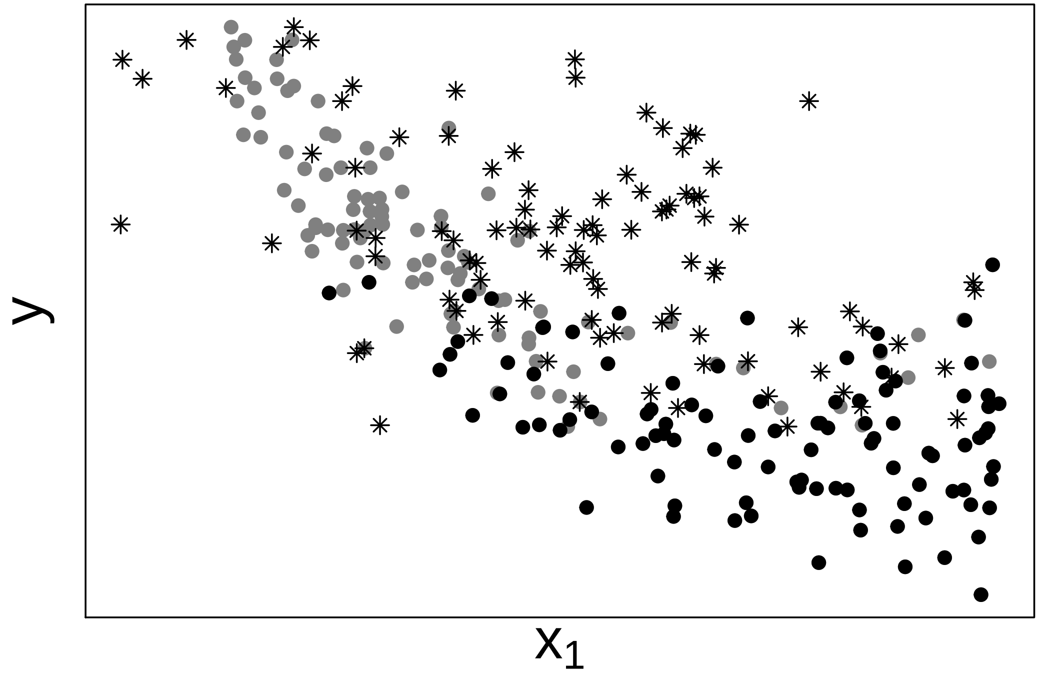
True model: $y = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2 x_1^2} + \ldots$ (quadratic association)

Imputation model: $x_1 = \theta_{10} + \theta_{11} y + \ldots$ (linear association)



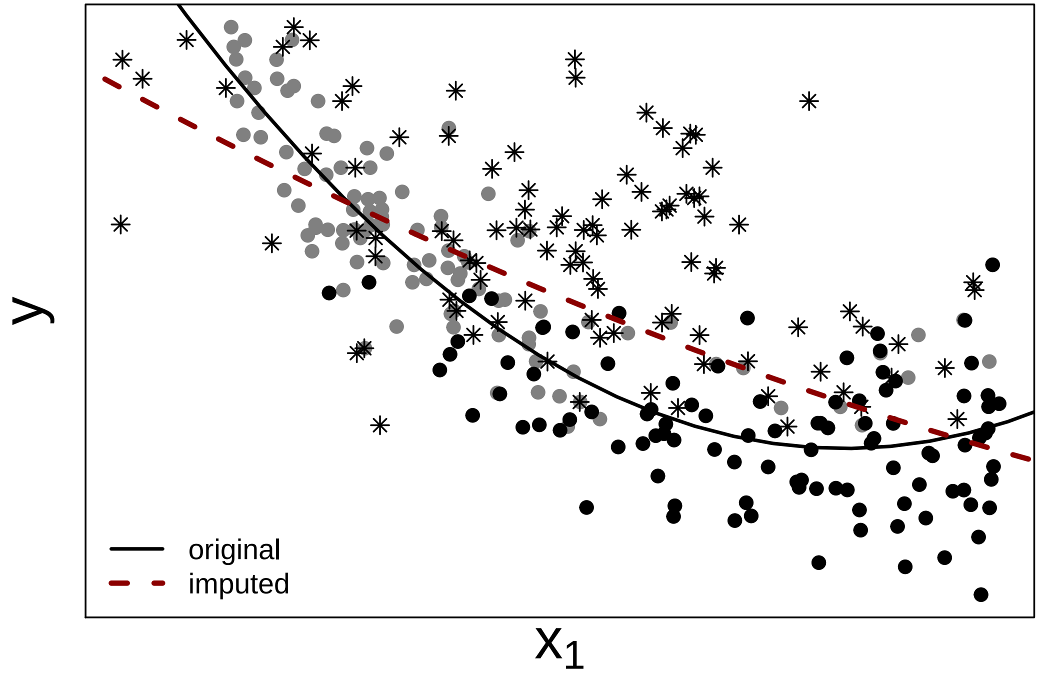
Uncongeniality

True model:	$y = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2 x_1^2} + \ldots$	(quadratic association)
Imputation model:	$x_1 = \theta_{10} + \theta_{11} y + \ldots$	(linear association)



Uncongeniality

True model:	$y = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta_2 x_1^2} + \ldots$	(quadratic association)
Imputation model:	$x_1 = \theta_{10} + \theta_{11} y + \ldots$	(linear association)



Simple approaches

- **passive normal imputation:**
  standard MICE ➡ calculate interactions & non-lin. terms afterwards

(Can be done in SPSS)

Simple approaches

- **passive normal imputation:**
  standard MICE ➡ calculate interactions & non-lin. terms afterwards

- **predictive mean matching (pmm)** (also passive)
  use pmm instead of linear regression for imputation

(Can be done in SPSS)

Simple approaches

- **passive normal imputation:**
  standard MICE ➡ calculate interactions & non-lin. terms afterwards

- **predictive mean matching (pmm)** (also passive)
  use pmm instead of linear regression for imputation

- **just another variable**
  - calculate interactions & non-lin. terms before imputation
  - add as columns to data set

(Can be done in SPSS)

Some advanced approaches

- **smcfcs: S**ubstantive **M**odel **C**ompatible **FCS**
  ➡ MICE type approach

Some advanced approaches

- **smcfcs: S**ubstantive **M**odel **C**ompatible **FCS**
  ➡ MICE type approach

- **jomo:** joint modeling MI using multivariate normal distribution
  ➡ joint model MI

Some advanced approaches

- **smcfcs:** **S**ubstantive **M**odel **C**ompatible **FCS**
  ➡ MICE type approach

- **jomo:** joint modeling MI using multivariate normal distribution
  ➡ joint model MI

- **JointAI:** joint analysis and imputation
  ➡ not MI, but simultaneous analysis & imputation

Some advanced approaches

- **smcfcs: S**ubstantive **M**odel **C**ompatible **FCS**
  ➡ MICE type approach
- **jomo:** joint modeling MI using multivariate normal distribution
  ➡ joint model MI
- **JointAI:** joint analysis and imputation
  ➡ not MI, but simultaneous analysis & imputation

Explicitly take into account the **analysis model** in the sampling distribution for $\hat{\boldsymbol{x}}_j$

Simulation study (I): Data setup

Models: linear regression with

- interaction
- logarithmic or quadratic effect
- combinations

Simulation study (I): Data setup

Models: linear regression with

- interaction
- logarithmic or quadratic effect
- combinations

Missing values:

- in one or two covariates
- MAR, depending on outcome (and other covariate)
- 20%, 40%, 60%

Simulation study (I): Methods

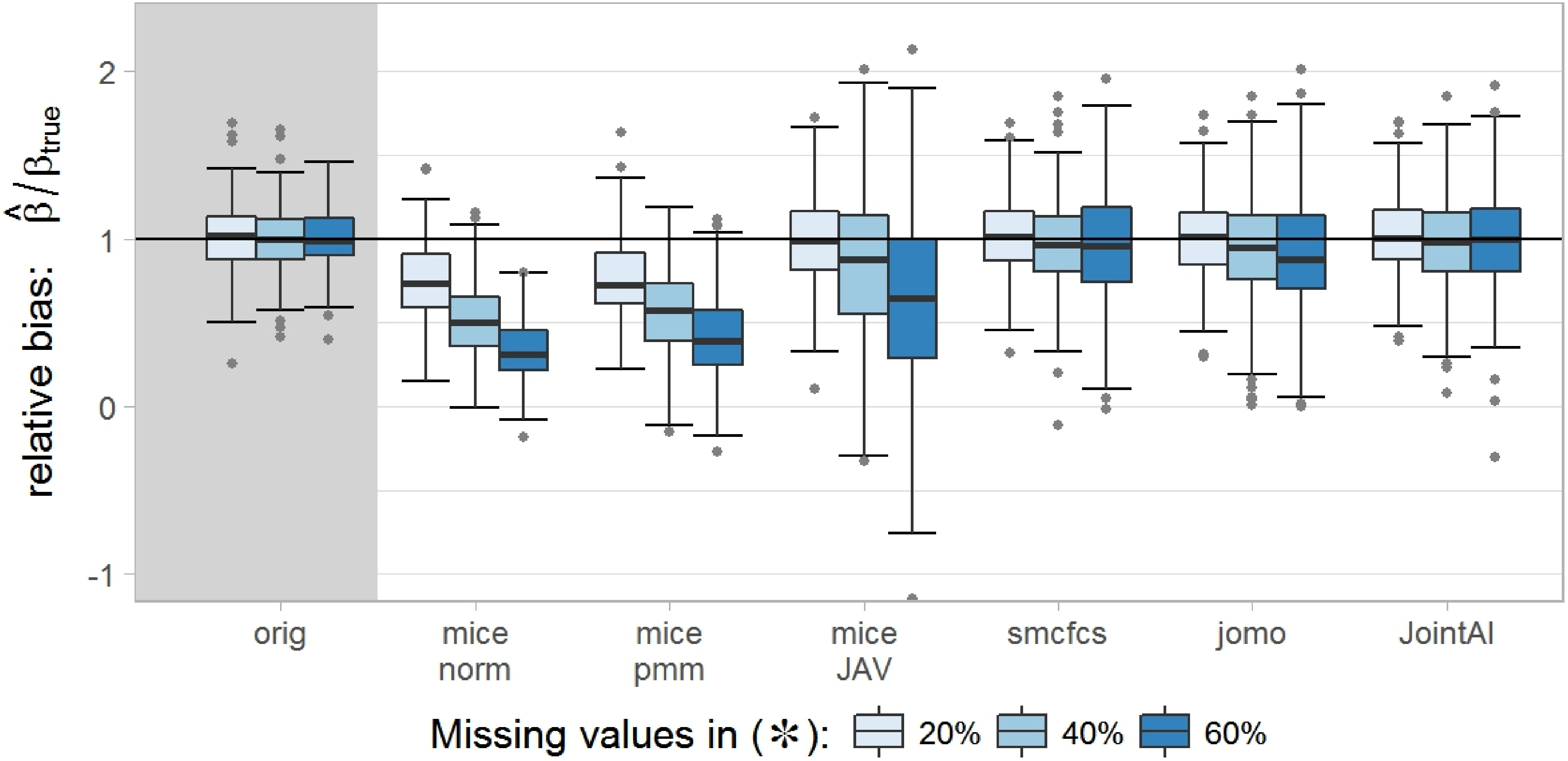
Approaches using the **mice** package:

- **norm**
- **pmm**
- **JAV** (using `pmm`)

other packages:

- **smcfcs:** `smcfcs()`
- **jomo:** `jomo.lm()`
- **JointAI:** `lm_imp()`

qdr. with interaction: $y \sim c_1 + (c_2^{(*)} + c_2^{2(*)}) \times b^{(*)}$ (effect of $c_2^2 \times b$)



Summary of Simulation Study (I)

	interaction	log	quadratic	interact & qdr
norm	😐		☹	☹
pmm	😐	🙂	😐	☹
JAV	😐	🙂	🙂	☹
smcfcs	🙂	⚡	🙂	🙂
jomo	😐	⚡	🙂	😐
JointAI	🙂	🙂	🙂	🙂

When MICE might fail

Imputation model not congenial with analysis:

- quadratic, logistic, . . . , effects
- interactions between covariates

✓

Complex (non univariate) outcomes:

- survival
- longitudinal

Imputation for survival data (Cox PH model)

Outcome: **event time** (T) and **event indicator** (D)

MICE strategies: represent outcome by including

- D
- T and/or $f(T)$
- Nelson-Aalen estimator of $H_0(T)$

White & Royston (2009). Imputing missing covariate values for the Cox model. *Stat Med* 28(15), 1982–1998.

Bartlett et al.(2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*, 24(4), 462 - 487.

Imputation for survival data (Cox PH model)

Outcome: **event time** (T) and **event indicator** (D)

MICE strategies: represent outcome by including

- D
- T and/or $f(T)$
- Nelson-Aalen estimator of $H_0(T)$

➡ use **D** + **Nelson-Aalen**

small bias towards zero when large covariate effect

White & Royston (2009). Imputing missing covariate values for the Cox model. *Stat Med* 28(15), 1982–1998.

Bartlett et al.(2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*, 24(4), 462 - 487.

Imputation for survival data (Cox PH model)

Outcome: **event time** (T) and **event indicator** (D)

MICE strategies: represent outcome by including

- D
- T and/or $f(T)$
- Nelson-Aalen estimator of $H_0(T)$

➡ use **D** + **Nelson-Aalen**

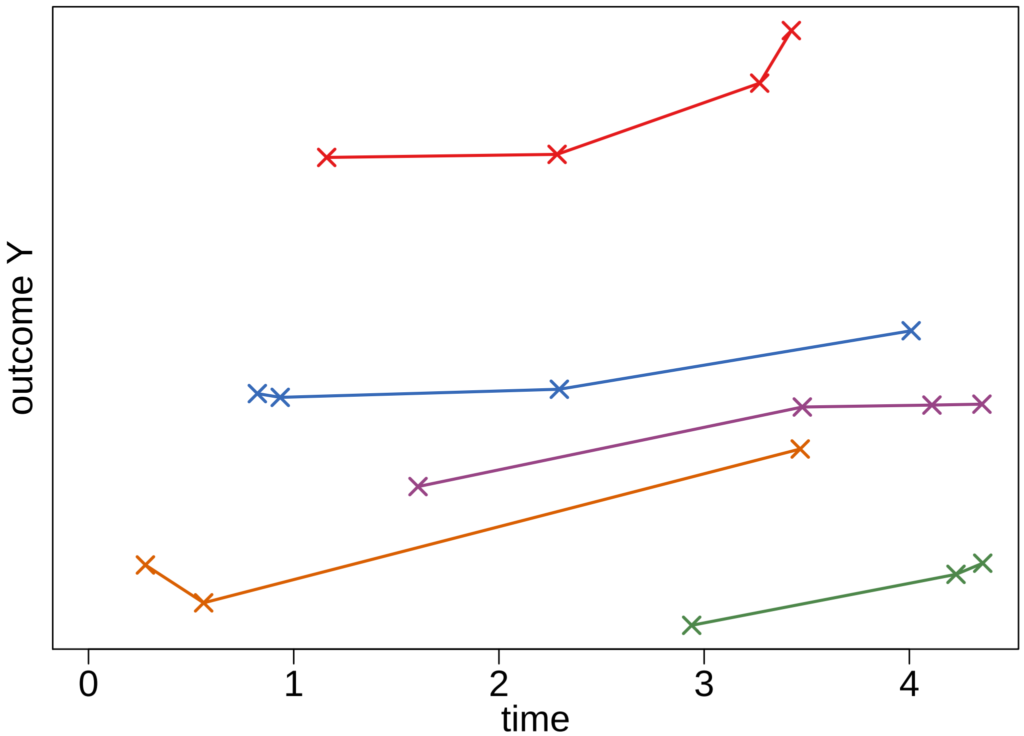
small bias towards zero when large covariate effect

smcfcs:
unbiased in simulation study ➡ improvement over MICE

White & Royston (2009). Imputing missing covariate values for the Cox model. *Stat Med* 28(15), 1982–1998.

Bartlett et al.(2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*, 24(4), 462 - 487.

Multi-level imputation



id	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	1.16
1	✓	✓	NA	✓	✓	2.28
1	✓	✓	NA	✓	✓	3.27
1	✓	✓	NA	✓	✓	3.42
2	✓	NA	✓	✓	✓	0.82
2	✓	NA	✓	✓	✓	0.93
2	✓	NA	✓	✓	✓	2.29
2	✓	NA	✓	✓	✓	4.01
3	✓	✓	NA	✓	NA	2.94
3	✓	✓	NA	✓	NA	4.23
3	✓	✓	NA	✓	NA	4.36
⋮	✓	✓	✓	NA	✓	⋮

Multi-level imputation: strategies

Imputation in long format:

- **clustering** needs to be taken into account
- **consistency** of incomplete baseline covariates

id	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	1.16
1	✓	✓	NA	✓	✓	2.28
1	✓	✓	NA	✓	✓	3.27
1	✓	✓	NA	✓	✓	3.42
2	✓	NA	✓	✓	✓	0.82
2	✓	NA	✓	✓	✓	0.93
2	✓	NA	✓	✓	✓	2.29
2	✓	NA	✓	✓	✓	4.01
3	✓	✓	NA	✓	NA	2.94
3	✓	✓	NA	✓	NA	4.23
3	✓	✓	NA	✓	NA	4.36
⋮	✓	✓	✓	NA	✓	⋮

Multi-level imputation: strategies

Imputation in long format:

- **clustering** needs to be taken into account
- **consistency** of incomplete baseline covariates

Imputation in wide format:

difficult with **unbalanced data**, ideas:

- create **intervals** to balance data
- use **summary** of the outcome:
  - only baseline observation
  - random effects from preliminary model

id	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	1.16
1	✓	✓	NA	✓	✓	2.28
1	✓	✓	NA	✓	✓	3.27
1	✓	✓	NA	✓	✓	3.42
2	✓	NA	✓	✓	✓	0.82
2	✓	NA	✓	✓	✓	0.93
2	✓	NA	✓	✓	✓	2.29
2	✓	NA	✓	✓	✓	4.01
3	✓	✓	NA	✓	NA	2.94
3	✓	✓	NA	✓	NA	4.23
3	✓	✓	NA	✓	NA	4.36
⋮	✓	✓	✓	NA	✓	⋮

Simulation study (II): Data setup

Models: linear mixed model with random intercept & slope

- interaction
- quadratic effect
- interaction & quadratic effect

Missing values: (as before)

- in one or two covariates
- MAR, depending on outcome (and other covariate)
- 20%, 40%, 60%

Simulation study (II): Methods

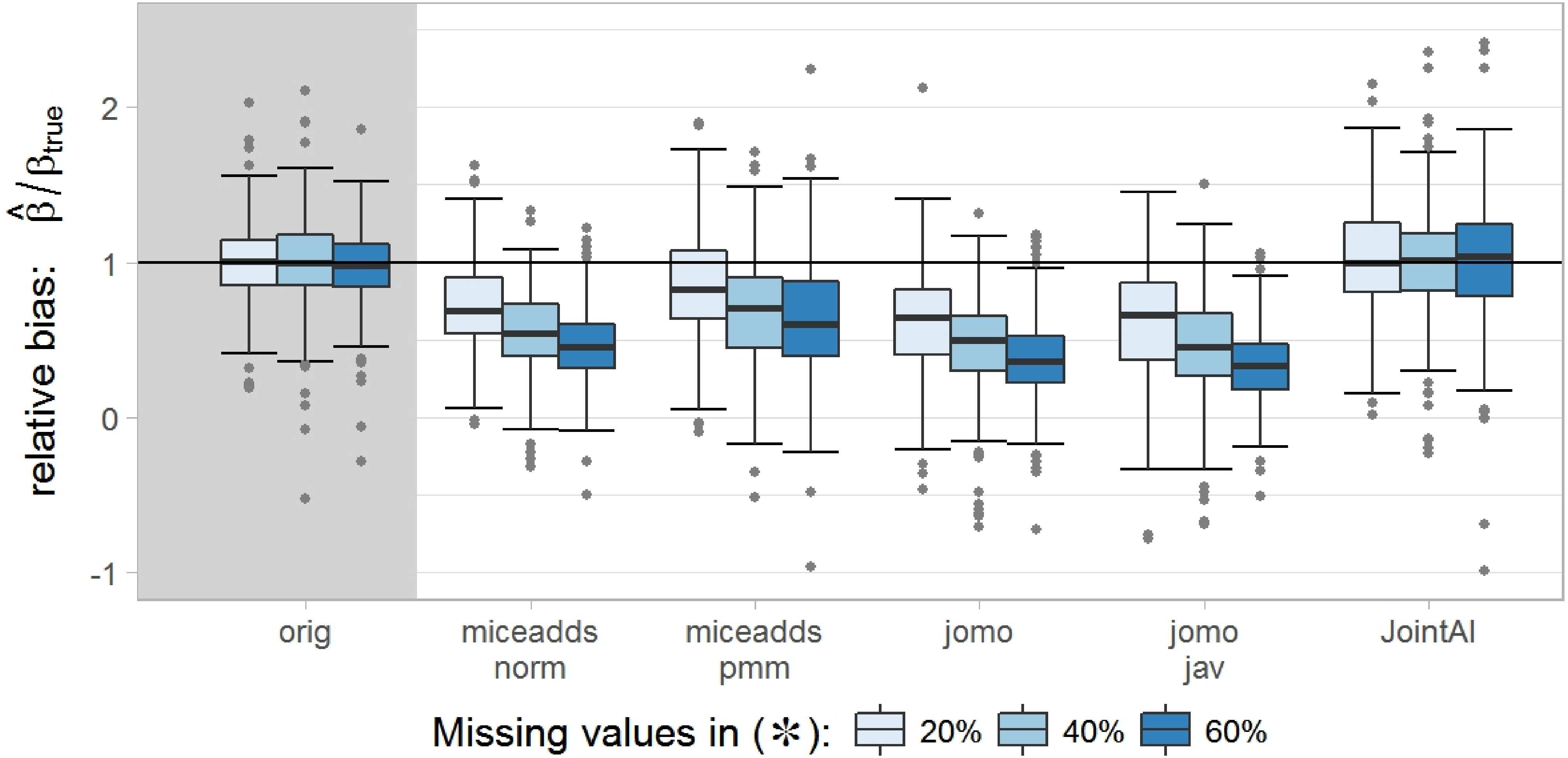
Approaches using **MICE**:

	mice	**miceadds**
norm	`2lonly.norm`	`2lonly.function` (+ `norm` & `logreg`)
pmm	`2lonly.pmm`	`2lonly.function` (+ `pmm3` & `logreg`)

other packages:

- **jomo:**
  - (`jomo.lmer()`: problems with missing baseline covariates)
  - `jomo2()`: no functionality for non-linear terms ➡ JAV
- **JointAI:** `lme_imp()`

interaction & qdr.: $y \sim c_1 \times b^{(*)} + c_2^{(*)} + c_2^{2(*)} + t + (t \mid id)$ (effect of c_2^2)



Summary of Simulation Study (II)

	longitudinal	interaction	quadratic & interaction
norm			
pmm			
jomo			
jomo JAV			
JointAI			

Discussion

- **Missing data is common** challenge
- standard implementations may be **biased**
- but more and more software is available
  - **extensions of mice** package
  - stand-alone packages: **smcfcs, jomo, JointAI, . . .**
- easy to use:

```
library(JointAI)

lme_imp(fixed = y ~ c1 * b + c2 + I(c2^2) + time,
        random = ~ time|id,
        data = DF, n.iter = 1000)
```

(https://github.com/NErler/JointAI)

Thank you for your attention.

n.erler@erasmusmc.nl
N_Erler
NErler

Dep. Biostatistics: www.erasmusmc.nl/biostatistiek
ErasmusAGE: www.erasmusage.com