Imputation model misspecification: How robust are Bayesian methods?

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

[1]Department of Biostatistics, Erasmus MC, Rotterdam, the Netherlands, [2]K.U.Leuven, L-Biostat, Leuven, Belgium

Erasmus MC University Medical Center Rotterdam

Introduction

Nowadays: Availability of imputation methods in standard software facilitates automated imputation of incomplete data. For example in R:

mice	jomo	JointAI
Multiple Imputation (MI) using **chained equations**	Joint Model MI using a **multivariate normal model** (MVN)	**Bayesian joint model sequential factorization** imputation

Impute missing values by **draws from the (posterior) predictive distribution** of an incomplete variable, conditional on (all) other variables.

➡ The **predictive distributions need to fit the data well!**

However:

- imputation models are specified automatically by the software
- in practice **often no effort is made to check the validity of the postulated models**

Robustness of MI - Some Findings from Literature

Normal imputation model for non-normal data

- MICE & MVN robust for inference about the mean
- **more flexible distributions necessary** when interest is **in quantiles**
- MICE: non-/semi-parametric methods often better

Bounded variables

- imputation outside range **acceptable for inference on mean**
- problematic for variance, quantiles, shape, ...

Comparison between approaches

- MVN & MICE similarly robust
- misspecified MICE better than compl. case analysis
- doubly robust IPW may be even better than MICE

Structure of the linear predictor

- flexible models can outperform normal imputation & pred. mean matching
- e.g.: GAMLSS, penalized regression

Sequential Factorization Imputation

Fully Bayesian approach allowing **simultaneous analysis and imputation**:

- factorize **joint distribution** as **sequence of conditional distributions**,
- one of which one is the analysis model of interest:

$$p(y, \boldsymbol X, \boldsymbol\theta) \propto \underbrace{p(y \mid \boldsymbol X_c, x_1, \dots, x_p, \boldsymbol\theta_y)}_{\text{analysis model}} \underbrace{p(x_1 \mid \boldsymbol X_c, \boldsymbol\theta_{x_1}) \dots p(x_p \mid \boldsymbol X_c, x_1, \dots, x_{p-1}, \boldsymbol\theta_{x_p})}_{\text{conditional distributions}}$$
$$\underbrace{\pi(\boldsymbol\theta_y)\pi(\boldsymbol\theta_{x_1})\dots\pi(\boldsymbol\theta_{x_p})}_{\text{priors}},$$

Notation:

$\boldsymbol X = (\boldsymbol X_c, \boldsymbol X_{mis})$ design matrix of completely observed and incomplete covariates

$\boldsymbol X_{mis} = (x_1, \dots, x_p)$, $\quad \boldsymbol\theta = (\boldsymbol\theta_y^T, \boldsymbol\theta_{x_1}^T, \dots, \boldsymbol\theta_{x_p}^T)^T$,

$\boldsymbol X_{<\ell} = (x_1, \dots, x_{\ell-1})^T$

- Draw imputations from the **Posterior Predictive Distribution** (PPD) (e.g., for a covariate x_ℓ):

$$p(x_\ell \mid y, \boldsymbol X_c, \boldsymbol X_{-\ell}, \boldsymbol\theta) \propto p(y \mid \boldsymbol X_c, \boldsymbol X_{mis}, \boldsymbol\theta_y) \underbrace{p(x_\ell \mid \boldsymbol X_c, \boldsymbol X_{<\ell}, \boldsymbol\theta_{x_\ell})}_{\text{cond. distr. of } x_\ell}$$
$$\underbrace{\left\{\prod_{k=\ell+1}^{p} p(x_k \mid \boldsymbol X_c, \boldsymbol X_{<k}, \boldsymbol\theta_{x_k})\right\}}_{\text{cond. distr. of } x_{\ell+1}, \dots, x_p} \pi(\boldsymbol\theta_y)\pi(\boldsymbol\theta_{x_\ell}) \prod_{k=\ell+1}^{p} \pi(\boldsymbol\theta_{x_k}),$$

- **PPD** specified indirectly ➡ **direct evaluation of its fit not possible**

Our Research Question:

How robust is sequential factorization imputation to misspecification of conditional distributions?

Investigating Robustness by Simulation

- Analysis model: linear regression with 4 covariates

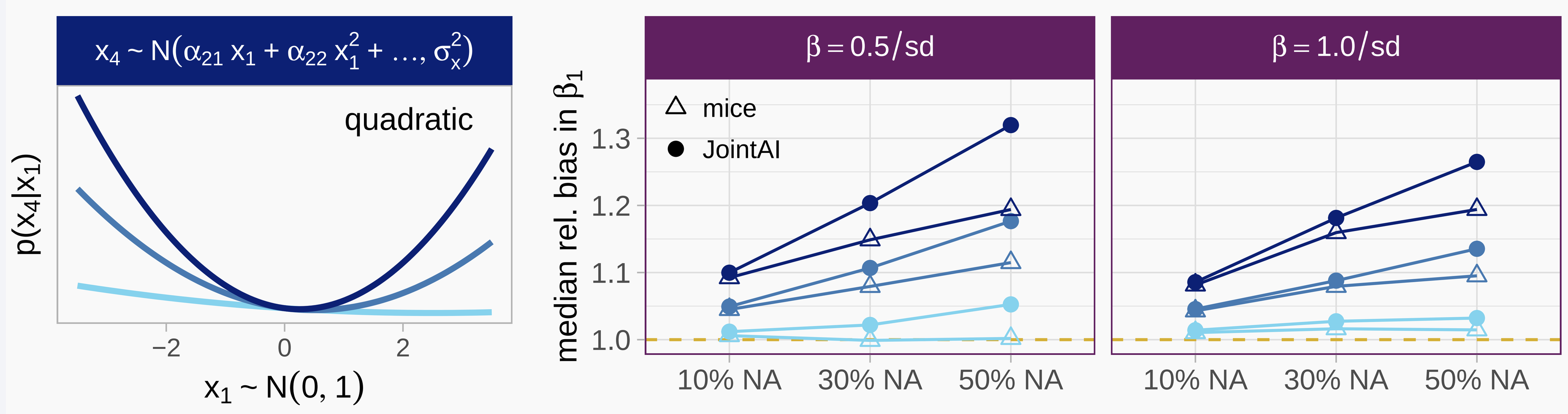
$$\begin{aligned}
y &\sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \sigma_y^2) && \text{standardized } \beta = 0.1,\ 0.5 \text{ or } 1\\
x_1 &\sim N(0, 1) \quad \text{or} \quad x_1 \sim Gamma(5, 10) && (complete)\\
x_2 &\sim Bin(0.5) && (complete)\\
x_3 &\sim Bin\left(\text{expit}\{\alpha_{10} + \alpha_{11} x_1 + \alpha_{12} x_2\}\right) && \Big\}\ (10\%,\ 30\% \text{ or } 50\% \text{ MAR})\\
x_4 &| x_1, x_2, x_3, \alpha_{2\cdot} \text{ depending on scenario} &&
\end{aligned}$$

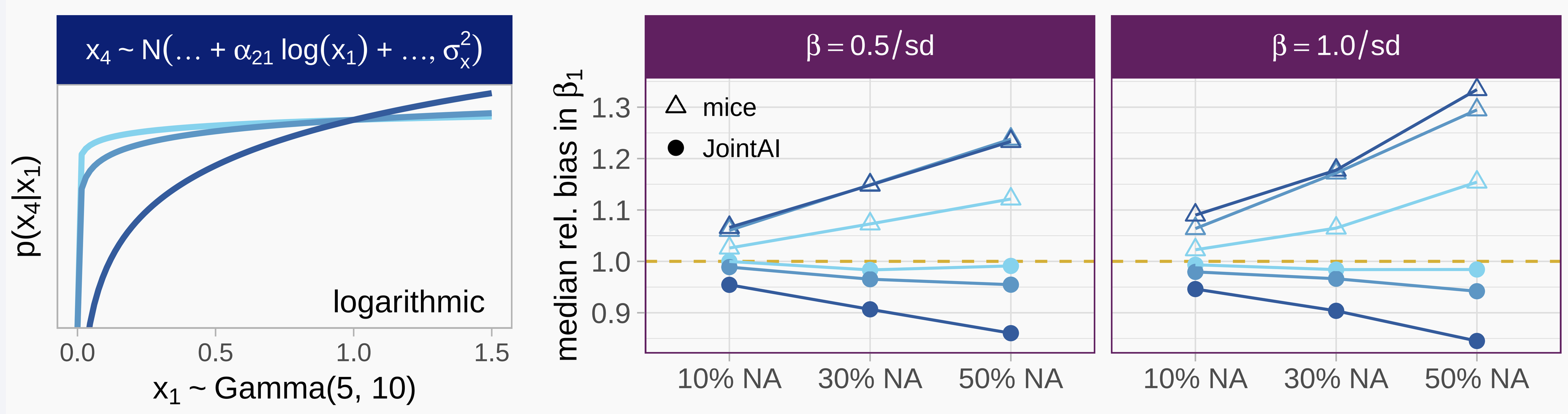
- **Misspecification** of the conditional distribution of x_4:
  - wrongly assuming **linear association** with other covariates,
  - omission of an important **interaction effect**,
  - disregard **skewness or multimodality** by mis-specification of the **residual distribution** or **sequence of cond. distributions**
- Imputation under a **naive model assuming normality & lin. associations**, using
  - **sequential factorization imputation** (R package **JointAI**)
  - as comparison: **MICE** (R package **mice**, with pred. mean matching)
- **Performance evaluation:**
  - **relative bias** ($\hat\beta_{imp}/\hat\beta_{complete}$)
  - **coverage of** true parameter by the **95% confidence/credible intervals** (CI)

How Robust is Sequential Factorization Imputation?

- Overall: methods performed worse with more missing values and larger β.
- Missingness proportion had stronger impact on performance than size of β.
- Settings with small standardized $\beta = 0.1$ had negligible bias for most scenarios.

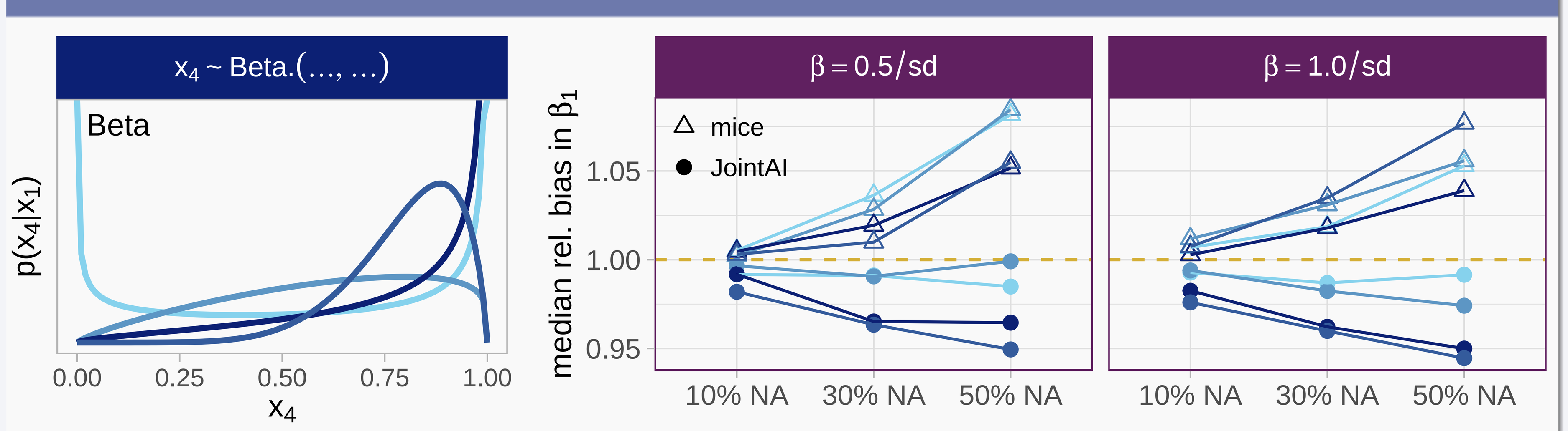
Non-linear association between x_4 and x_1:

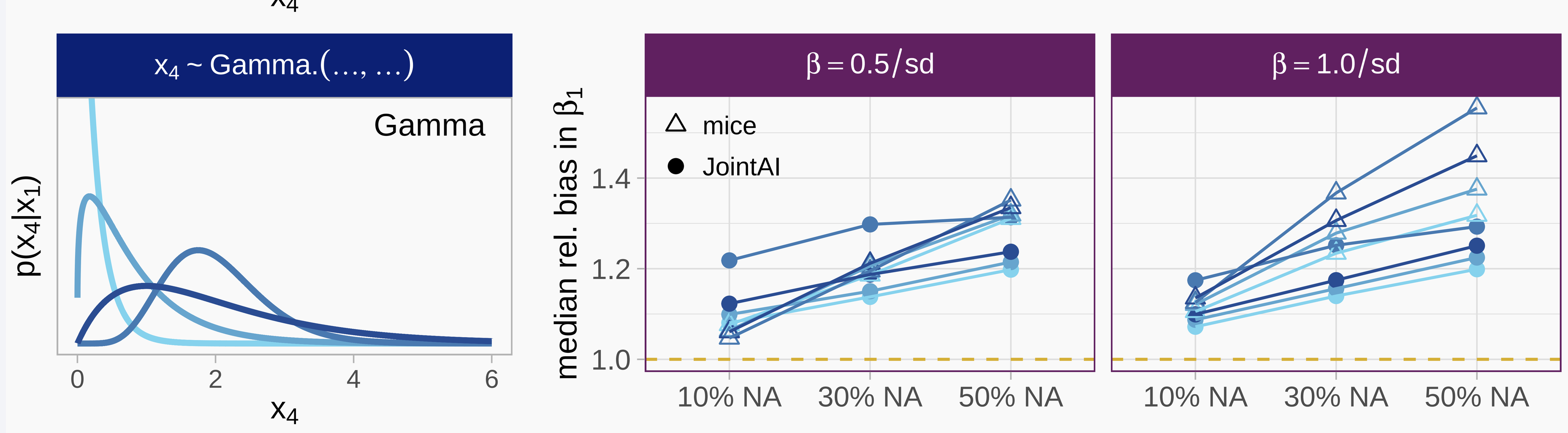




quadratic: coverage in JointAI ≥ 0.4, MICE ≥ 0.6; bias in $\hat\beta_4$: for MICE larger than for JointAI
logarithmic: coverage in MICE ≥ 0.6 (JointAI ≥ 0.9); MICE also biased in all other $\hat\beta$

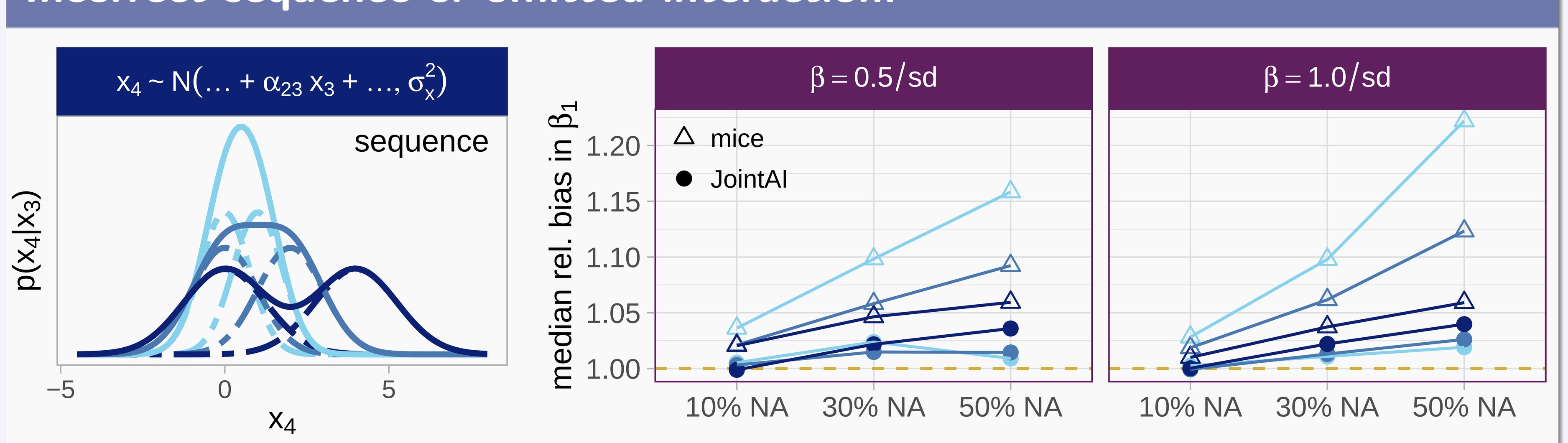
Non-normal conditional distribution of x_4:

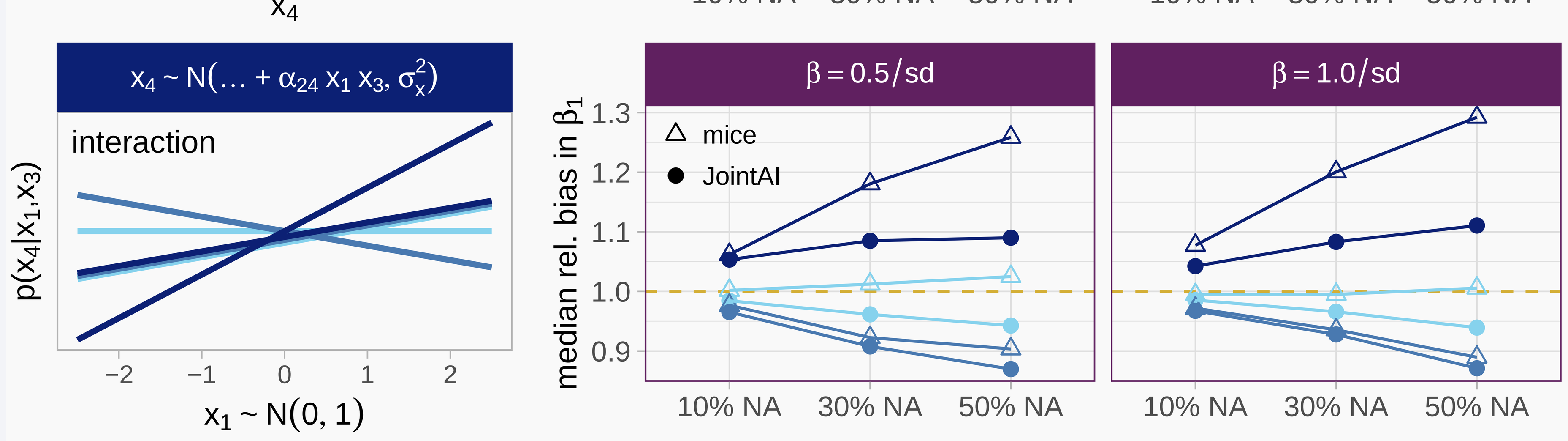




Beta: MICE also biased for $\hat\beta_4$ (JointAI not), coverage for both ≈ 0.95
Gamma: bias in all $\hat\beta$, worse for MICE; coverage JointAI ≥ 0.7, MICE ≥ 0.25

Incorrect sequence or omitted interaction:





sequence: bias in all $\hat\beta$, but worse for MICE; coverage MICE ≥ 0.65, JointAI ≥ 0.85
interaction: MICE more severely biased in all $\hat\beta$, coverage MICE ≥ 0.5, JointAI ≥ 0.75

Conclusions

- **Misspecification** of the cond. distributions **translates to** misspecified **imputation models**.
- In most of our scenarios: **JointAI** performed **(slightly) better than MICE**.
- **Fit of the cond. distributions needs to be validated** to obtain unbiased results.
- **More flexible models are needed** to assure appropriate performance **in practice**, where imputation is often used in a "black-box" manner.

n.erler@erasmusmc.nl N Erler NErler