Comparison of multiple imputation methods for systematically and sporadically missing multilevel data

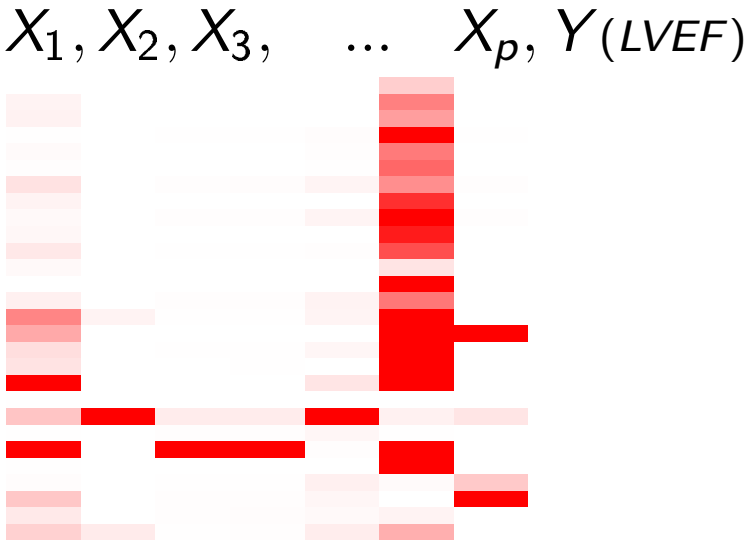
V. Audigier, I. White , S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. van Buuren, M. Resche-Rigon

INSERM, UMR 1153, ECSTRA team, Saint-Louis Hospital, Paris

MODAL Seminar, November 22th, Lille

Motivation: GREAT data (Great Network, 2013)

- Risk factors associated with short-term mortality in acute heart failure
- 28 observational cohorts, 11685 patients, 2 **binary** and 8 **continuous** variables (patient characteristics and potential risk factors)
- sporadically and systematically **missing data**



Aim: explain the relationship between biomarkers (BNP, AFIB,...) and the left ventricular ejection fraction (LVEF)

$$y_{ik} = \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_{ik} \sim \mathcal{N}\left(0, \sigma^2\right)$$

$\hat{\beta}$ **and associated variability** $var\left(\widehat{\beta}\right)$

Methods to handle missing values

Missing data are often assumed to be missing at random (MAR)

Ad-hoc methods

- **Complete-case analysis**
  - → generally leads to biased estimates
  - → increases standard errors

- **Single imputation**
  - → leads to unbiased estimates
  - → standard errors are downwardly biased

Methods to handle missing values

Relevant methods

- **Likelihood approaches**
  - Frequentist framework: EM algorithm
  - Bayesian framework: Data Augmentation

  → leads to unbiased estimates
  → specific to the analysis model
  → not always feasible

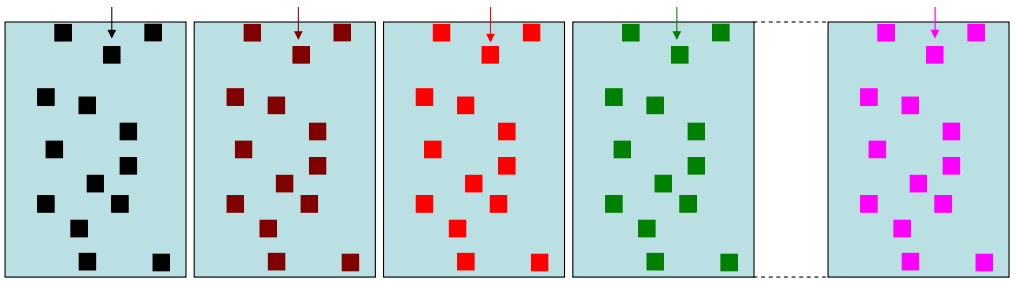
- **Multiple imputation**

  → leads to unbiased estimates
  → can be used for several analysis models

Multiple imputation (Rubin, 1987)

1. Generate a set of M parameters θ_m of an **imputation model** to generate M plausible imputed data sets

$$P\left(X^{miss}|X^{obs},\theta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(X^{miss}|X^{obs},\theta_M\right)$$



2. Fit the **analysis model** on each imputed data set: $\hat{\beta}_m, \widehat{\text{Var}}\left(\hat{\beta}_m\right)$

3. Combine the results: $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$

$$T = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{Var}}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

$\Rightarrow$ **Provide estimation of the parameters and of their variability**

MI for multilevel data

Two standard ways to perform MI

- **Fully conditional specification** (FCS, MICE): a conditional imputation model for each variable
- **Joint modelling** (JM): a joint imputation model for all variables

The imputation model (joint or conditional) needs to be in line with the data

- need to account for the **heterogeneity** between clusters
- need to account for the **types** of data (continuous and binary)

MI for multilevel data

Type and name	Handles missing data:			Coded in R
	Sporadic?	Systematic?	in binary variable?	
JM-Pan	yes	yes	no	yes, (Pan)
JM-REALCOM	yes	yes	yes	no
JM-jomo	**yes**	**yes**	**yes**	**yes, (jomo)**
JM-Mplus	yes	yes	yes	no
FCS-2lnorm	yes	no	no	yes, (mice)
FCS-1stage	**yes using variant**	**yes**	**yes**	**yes**
FCS-2stage	**yes**	**yes**	**yes using variant**	**yes**

Outline

Continuous variables

Heteroscedastic random effect model as **imputation model**

$$\begin{aligned} \mathbf{y}_{ik} &= \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \\ b_k &\sim \mathcal{N}(0, \Psi) \\ \varepsilon_{ik} &\sim \mathcal{N}(0, \Sigma_k) \end{aligned}$$

Multiple imputation under this model

1. generating M sets of parameters $\theta_m = \left(\beta^m, \Psi^m, (\Sigma_k^m)_{1\leq k\leq K}\right)$
   - $\rightarrow$ **Bayesian** formulation: draw θ_m from its posterior distribution
   - $\rightarrow$ **asymptotic** method: estimate θ_m, draw θ_m from the asymptotic distribution of the estimator
2. imputing the data according each set θ_m
   - draw $b_k^m | \mathbf{y}_k^{obs}, \theta_m$
   - draw $\mathbf{y}_{ik}^{miss} | \theta_m, b_k^m$

Continuous variables

Heteroscedastic random effect model as **imputation model**

$$\begin{aligned} \mathbf{y}_{ik} &= \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \\ b_k &\sim \mathcal{N}(0, \Psi) \\ \varepsilon_{ik} &\sim \mathcal{N}(0, \Sigma_k) \end{aligned}$$

Multiple imputation under this model

1. generating M sets of parameters $\theta_m = \left(\beta^m, \Psi^m, (\Sigma_k^m)_{1\leq k\leq K}\right)$
2. imputing the data according each set θ_m
   - draw $b_k^m | \mathbf{y}_k^{obs}, \theta_m$
   - draw $\mathbf{y}_{ik}^{miss} | \theta_m, b_k^m$

Specific issues

1. how to generate Σ_k without $\mathbf{y}_{ik}$? (systematic)
2. how to draw b_k^m without $\mathbf{y}_{ik}$ (systematic) or given $\mathbf{y}_{ik}$ (sporadic)?

FCS-1stage (Jolani et al., 2015)

Conditional imputation models

$$y_{ik} = \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

1. generate $\theta_m = (\beta_m, \Psi_m, \sigma^2{}_m) \qquad 1 \leq m \leq M$
   - prior: non-informative (Jeffreys)
   - posterior distribution

$$\beta_m \sim \mathcal{N}\left(\widehat{\beta}, var\left(\widehat{\beta}\right)\right) \quad \Psi_m^{-1} \sim \mathcal{W}\left(K, \widehat{bb}^{\top}\right) \quad \sigma_m^2 \sim \text{Inv-}\Gamma\left(\tfrac{n-p}{2}, \tfrac{(n-p)\widehat{\sigma^2}}{2}\right)$$

$\rightarrow$ **requires REML estimate**

2. impute in each cluster k with **systematically missing data**
   - draw $b_k \sim \mathcal{N}(0, \Psi_m)$
   - impute data according to the imputation model

FCS-1stage (Jolani et al., 2015)

Conditional imputation models

$$y_{ik} = \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

1. generate $\theta_m = (\beta_m, \Psi_m, {\sigma^2}_m) \qquad 1 \leq m \leq M$
   - prior: non-informative (Jeffreys)
   - posterior distribution

$$\beta_m \sim \mathcal{N}\left(\widehat{\beta}, var\left(\widehat{\beta}\right)\right) \quad \Psi_m^{-1} \sim \mathcal{W}\left(K, \widehat{bb}^\top\right) \quad \sigma_m^2 \sim \text{Inv-}\Gamma\left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma^2}}{2}\right)$$

$\rightarrow$ **requires REML estimate**

2. impute in each cluster k with **sporadically missing data**
   - draw $b_k \sim \mathcal{N}\left(\mu_{b_k|y_k}, \Psi_{b_k|y_k}\right)$
   - impute data according to the imputation model

FCS-2stage (Resche-Rigon and White, 2016)

Conditional imputation models

$$y_{ik} = \mathbf{x}_{ik}\left(\beta + b_k\right) + \varepsilon_{ik} \qquad b_k \sim \mathcal{N}\left(0, \Psi\right) \qquad \varepsilon_{ik} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

$\rightarrow$ **the same imputation model, with heteroscedastic assumption**

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$
   - estimate θ and var $\left(\widehat{\theta}\right)$ with **a two-stage estimator**
   - draw θ_m from the **asymptotic** distribution of the estimator with expectation $\widehat{\theta}$ and variance $\widehat{\text{var}}\left(\widehat{\theta}\right)$
2. impute in each cluster k

FCS-2stage (Resche-Rigon and White, 2016)

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$

   stage 1 fit a linear model to each observed cluster

   $$\widehat{\beta}_k = \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} \mathbf{X}_k^\top \mathbf{y}_k$$

   stage 2 combine the estimates

$$\widehat{\beta}_k = (\beta + b_k) + \varepsilon'_k \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon'_k \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

Two estimators available: **REML** or **method of moments**

$\rightarrow \widehat{\Psi}$, $\widehat{\beta}$ and their associated (asymptotic) variances

FCS-2stage (Resche-Rigon and White, 2016)

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$

   stage 1 fit a linear model to each observed cluster

$$\widehat{\beta}_k = \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} \mathbf{X}_k^\top \mathbf{y}_k \quad \widehat{\sigma}_k = \frac{\|\mathbf{y}_k - \mathbf{X}_k \widehat{\beta}_k\|^2}{n_k - p - 1}$$

   stage 2 combine the estimates

$$\log \widehat{\sigma}_k = (log\ \sigma + s_k) + \varepsilon_k'' \quad s_k \sim \mathcal{N}(0, \Psi_s) \quad \varepsilon_k'' \sim \mathcal{N}(0, var(log\widehat{\sigma}_k))$$

$$\widehat{\beta}_k = (\beta + b_k) + \varepsilon_k' \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_k' \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

Two estimators available: **REML** or **method of moments**

$\rightarrow$ $log\widehat{\sigma}, \widehat{\Psi}_s$ and their associated (asymptotic) variances

$\rightarrow$ $\widehat{\Psi}$, $\widehat{\beta}$ and their associated (asymptotic) variances

FCS-2stage (Resche-Rigon and White, 2016)

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$

   stage 1 fit a linear model to each observed cluster

$$\widehat{\beta}_k = \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} \mathbf{X}_k^\top \mathbf{y}_k \quad \widehat{\sigma}_k = \frac{\|\mathbf{y}_k - \mathbf{X}_k \widehat{\beta}_k\|^2}{n_k - p - 1}$$

   stage 2 combine the estimates

$$\log \widehat{\sigma}_k = (log\ \sigma + s_k) + \varepsilon_k'' \quad s_k \sim \mathcal{N}(0, \Psi_s) \quad \varepsilon_k'' \sim \mathcal{N}(0, var(log\widehat{\sigma}_k))$$

$$\widehat{\beta}_k = (\beta + b_k) + \varepsilon_k' \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_k' \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

   Two estimators available: **REML** or **method of moments**

   → $log\widehat{\sigma}, \widehat{\Psi}_s$ and their associated (asymptotic) variances

   → $\widehat{\Psi}$, $\widehat{\beta}$ and their associated (asymptotic) variances

2. impute in each cluster k with **systematically missing data**
   - draw b_k **from their marginal distribution**
   - impute data according to the imputation model

FCS-2stage (Resche-Rigon and White, 2016)

1. generate $\theta_m = \left(\beta_m, \Psi_m, \left(\sigma_1^2, \ldots, \sigma_K^2\right)_m\right)$

   stage 1 fit a linear model to each observed cluster

   $$\widehat{\beta}_k = \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} \mathbf{X}_k^\top \mathbf{y}_k \quad \widehat{\sigma}_k = \frac{\|\mathbf{y}_k - \mathbf{X}_k \widehat{\beta}_k\|^2}{n_k - p - 1}$$

   stage 2 combine the estimates

   $$\log \widehat{\sigma}_k = (log\ \sigma + s_k) + \varepsilon_k'' \quad s_k \sim \mathcal{N}(0, \Psi_s) \quad \varepsilon_k'' \sim \mathcal{N}(0, var(log\widehat{\sigma}_k))$$
   $$\widehat{\beta}_k = (\beta + b_k) + \varepsilon_k' \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_k' \sim \mathcal{N}\left(0, var\left(\widehat{\beta}_k\right)\right)$$

   Two estimators available: **REML** or **method of moments**

   $\rightarrow$ $log\widehat{\sigma}, \widehat{\Psi}_s$ and their associated (asymptotic) variances

   $\rightarrow$ $\widehat{\Psi}$, $\widehat{\beta}$ and their associated (asymptotic) variances

2. impute in each cluster k with **sporadically missing data**
   - draw b_k **conditionally to** $\widehat{\beta_k}$
   - impute data according to the imputation model

JM-jomo (Quartagno and Carpenter, 2016)

$$y_{ik} = \mathbf{x}_{ik}\beta + \mathbf{z}_{ik}b_k + \varepsilon_{ik} \qquad b_k \sim \mathcal{N}(0, \Psi) \qquad \varepsilon_{ik} \sim \mathcal{N}(0, \Sigma_k)$$

1. Bayesian formulation to generate $\theta_m = (\beta_m, \Psi_m, \Sigma_m)_{1 \leq m \leq M}$
   - (informative) prior:
     $\beta \propto 1, \quad \Psi^{-1} \sim W(\nu_1, \Lambda_1), \quad \Sigma_k^{-1} | \nu_2, \Lambda_2 \sim W(\nu_2, \Lambda_2)$
   - posterior: unknown explicitly **but...**
     - most of conditional posterior distributions are known

       $\rightarrow$ **Gibbs sampler**
     - do not require REML estimate
     - unknown conditional distributions can be simulated by MCMC

2. Imputation (given by step 1)

Binary variables

- **FCS-1stage** (Jolani et al., 2015)
  - fit a logistic model with mixed effect to all clusters

  → sporadically missing values not handled

- **FCS-2stage** (Resche-Rigon and White, 2016)
  - fit a logistic model with fixed effect to each cluster
  - combine estimates using a meta-analysis

  → large clusters are required

- **JM-jomo** (Quartagno and Carpenter, 2016)
  - probit link: outcomes are latent normal variables, variance for errors are fixed to 1
  - draw latent normal variables
  - derive categories

  → more time consuming

Summary

method	Bayesian / asymptotic	prior for covariance matrices	heteroscedasticity assumption for errors	binary variables
FCS-1stage	Bayesian	Jeffrey	no	probit link
FCS-2tage	asymptotic		yes	logistic link
JM-jomo	Bayesian	Wishart	yes	logistic link

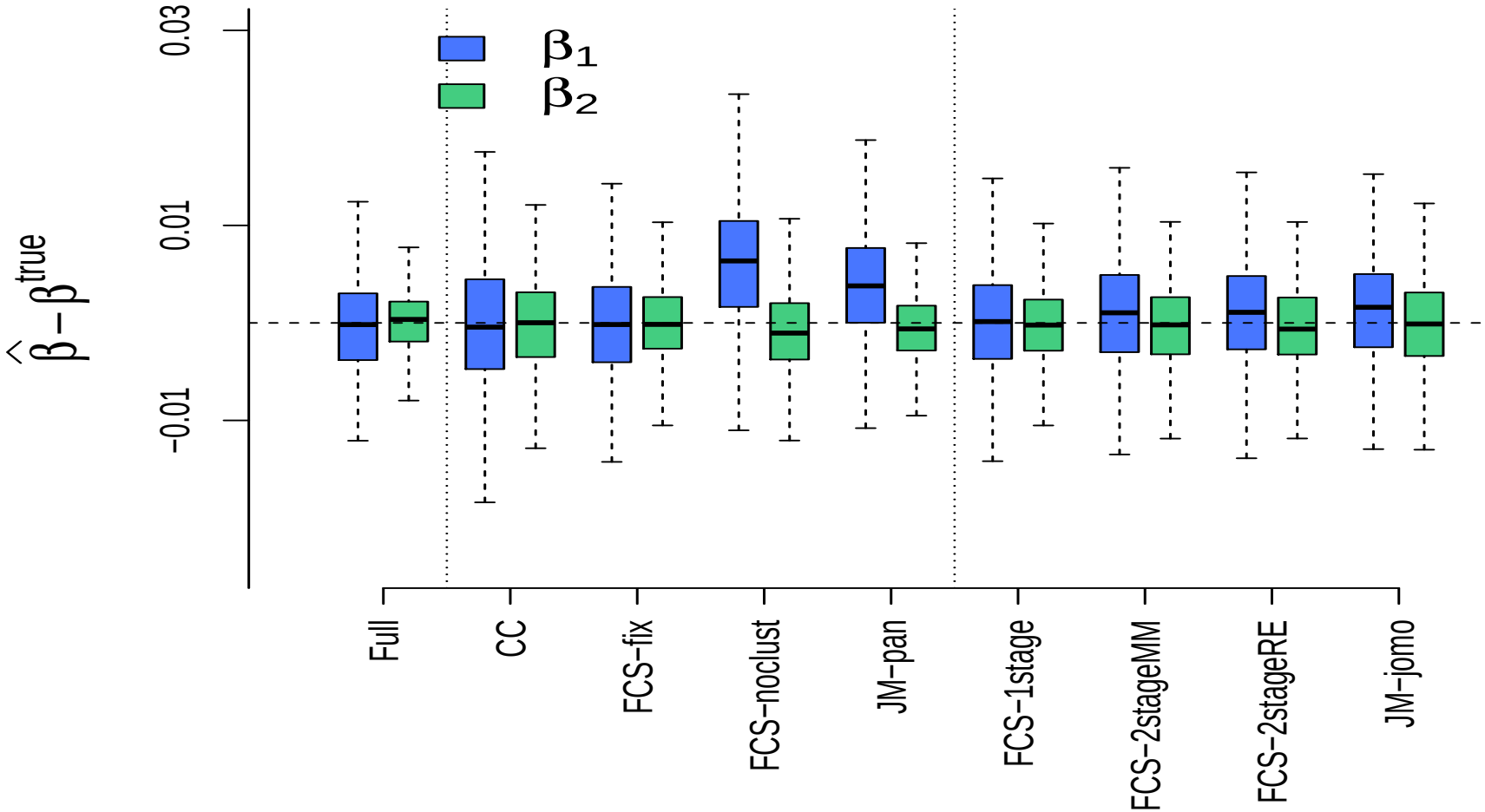
Simulation design

- **Data generation**: 500 incomplete data sets are independently simulated ($n = 11685, K = 28, 18 \leq n_k \leq 1834$)
  - $y_{ik} = \beta^{0} + \beta^{1}\mathbf{x}_{ik}^{(1)} + \beta^{2}\mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1\mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$ with $\beta = (.72, -.11, .03)$, $\Psi = \begin{bmatrix} .0077 & .0015 \\ .0015 & .0004 \end{bmatrix}$, $\sigma = .15$
  - $(\mu_k, \nu_k, \xi_k) \sim \mathcal{N}\left(0, \begin{bmatrix} .12 & .001 & .001 \\ .001 & .12 & .001 \\ .001 & .001 & .12 \end{bmatrix}\right)$
  - $\mathbf{x}_{ik}^{(1)} : \mathcal{N}(2.9 + \mu_k, .36)$
  - $\mathbf{x}_{ik}^{(2)} : logit\left(P\left(\mathbf{x}_{ik}^{(2)} = 1\right)\right) = 4.2 + \nu_k$
  - $\mathbf{x}_{ik}^{(3)}$: $\mathcal{N}(2.9 + \xi_k, .36)$
  - add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ with $\pi_{syst} = .25$ and $\pi_{spor} = .25$

Simulation design

- **Methods**
  - JM-jomo, FCS-1stage, FCS-2stage
  - Full, CC, FCS-fix, FCS-noclust, JM-pan
  - $M = 5$ imputed arrays
- **Estimands**: β and $var\left(\widehat{\beta}\right)$
- **Criteria**: bias, rmse, variance estimate, coverage

Base-case configuration



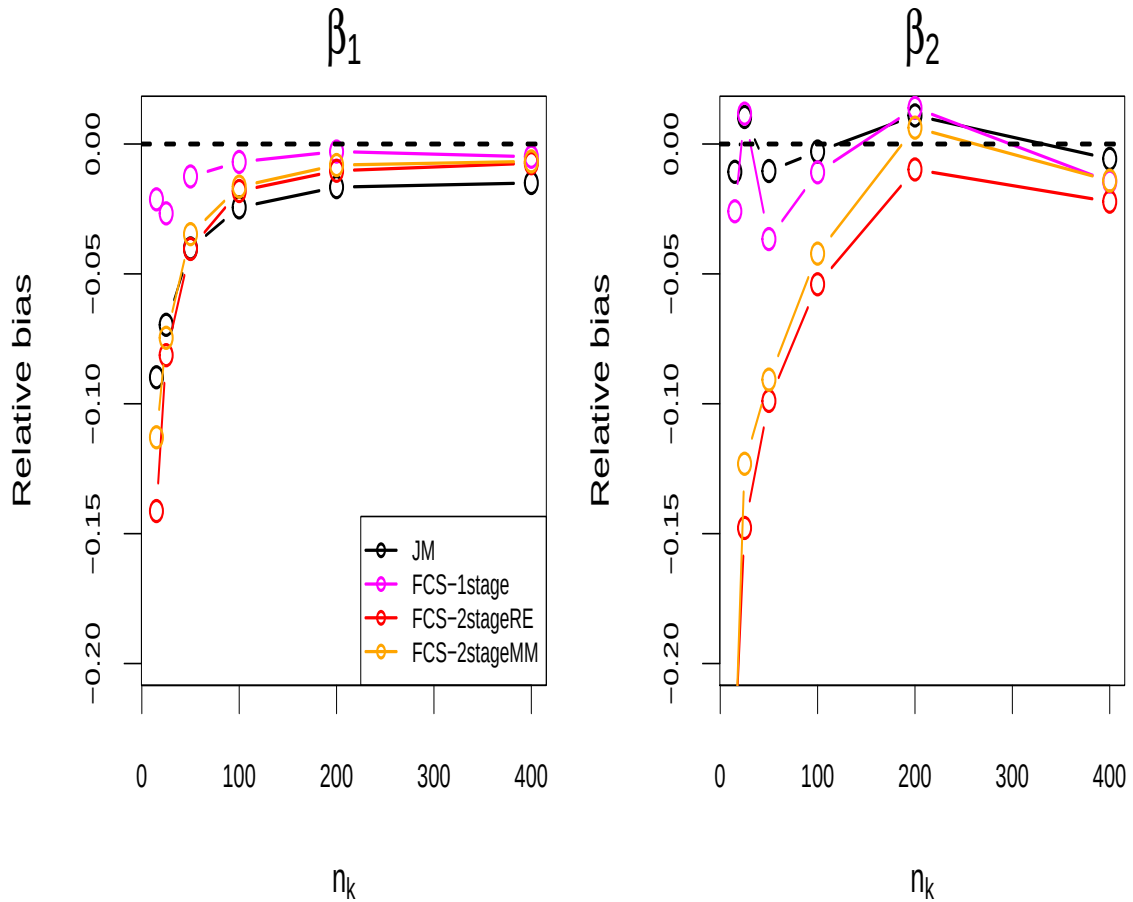
Base-case configuration

Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0070	0.0053	0.0071	0.0053	92.2	94.4	
FCS-fix	0.0043	0.0043	0.0058	0.0042	**85.6**	94.6	0.732
FCS-noclust	0.0041	0.0043	0.0067	0.0046	**59.6**	92.4	0.601
JM-pan	0.0048	0.0042	0.0058	0.0039	**83.0**	95.2	0.006
FCS-1stage	0.0049	0.0046	0.0056	0.0043	**90.4**	96.8	63.43
FCS-2stagemm	0.0059	0.0054	0.0058	0.0044	95.0	97.0	0.538
FCS-2stagere	0.0059	0.0049	0.0058	0.0044	95.0	96.2	1.304
JM-jomo	0.0066	0.0069	0.0057	0.0050	96.8	**97.6**	6.739

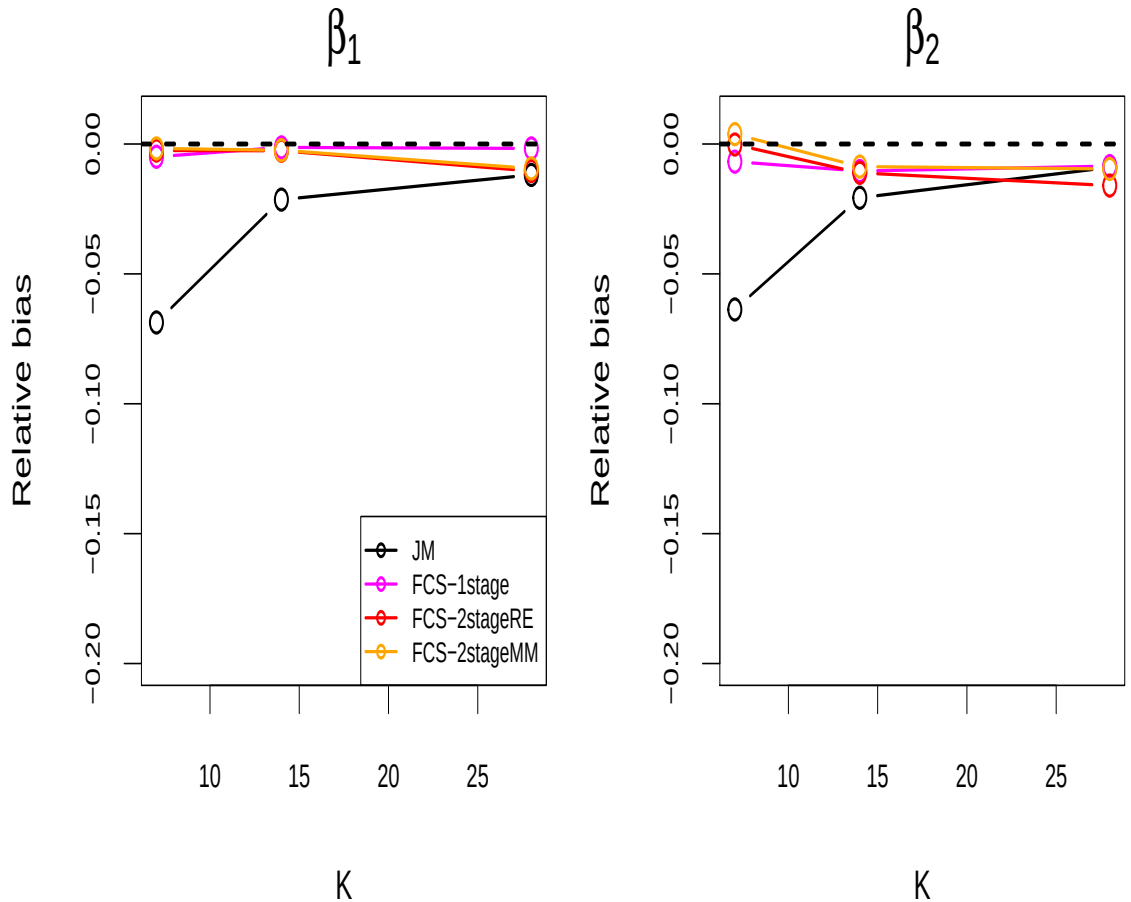
Base-case configuration

Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0070	0.0053	0.0071	0.0053	92.2	94.4	
FCS-fix	0.0043	0.0043	0.0058	0.0042	85.6	94.6	0.732
FCS-noclust	0.0041	0.0043	0.0067	0.0046	59.6	92.4	0.601
JM-pan	0.0048	0.0042	0.0058	0.0039	83.0	95.2	0.006
FCS-1stage	0.0049	0.0046	0.0056	0.0043	90.4	96.8	**63.43**
FCS-2stagemm	0.0059	0.0054	0.0058	0.0044	95.0	97.0	0.538
FCS-2stagere	0.0059	0.0049	0.0058	0.0044	95.0	96.2	1.304
JM-jomo	0.0066	0.0069	0.0057	0.0050	96.8	97.6	6.739

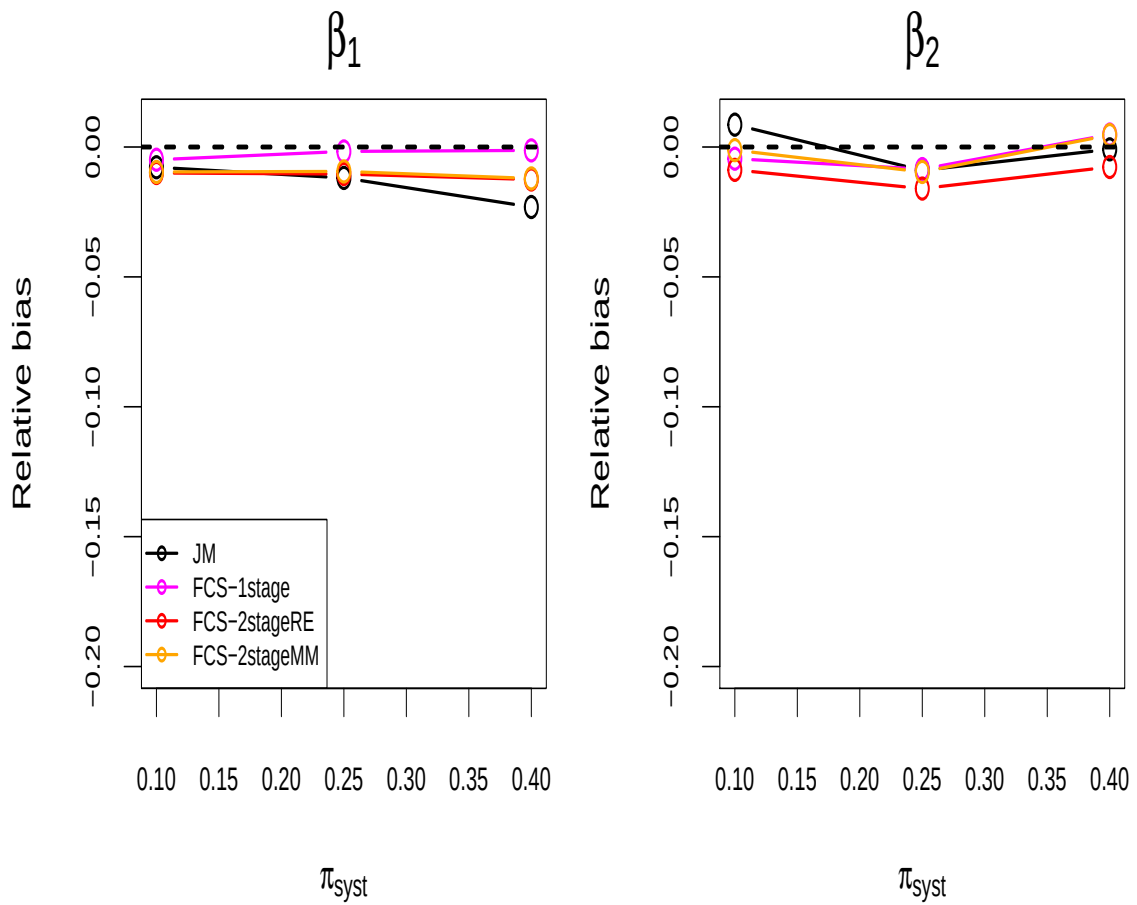
Robustness to the cluster size: point estimate



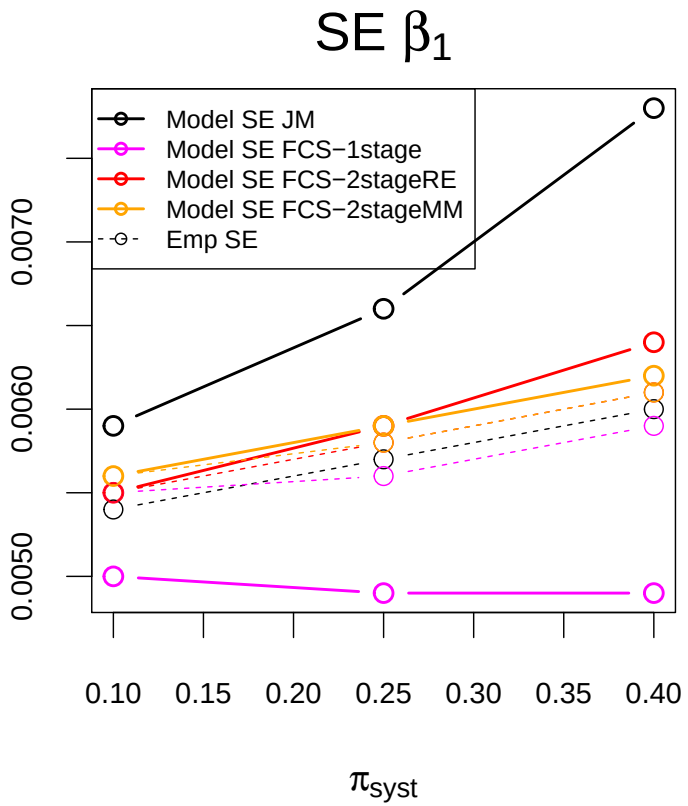
Robustness to the number of clusters: point estimate

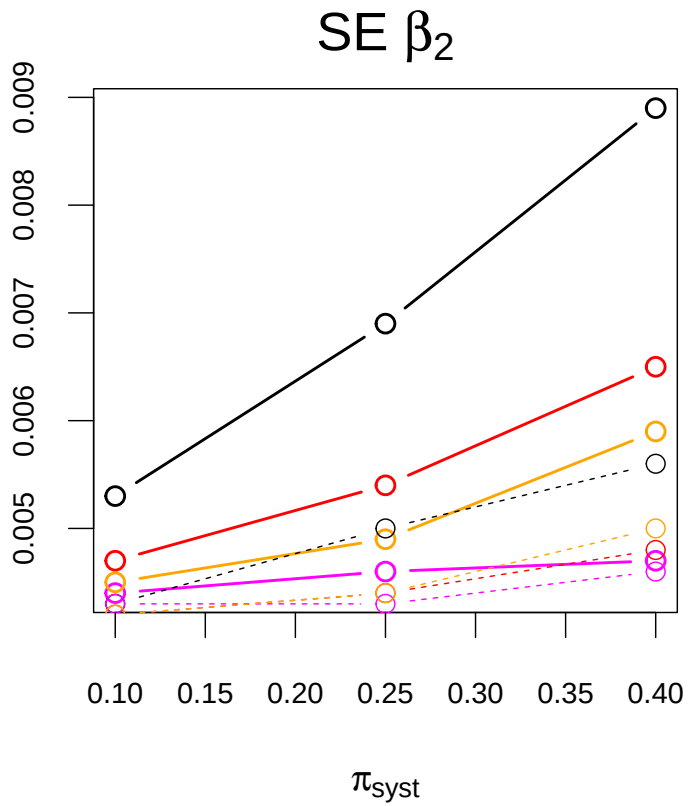


Robustness to π_{syst}: point estimate

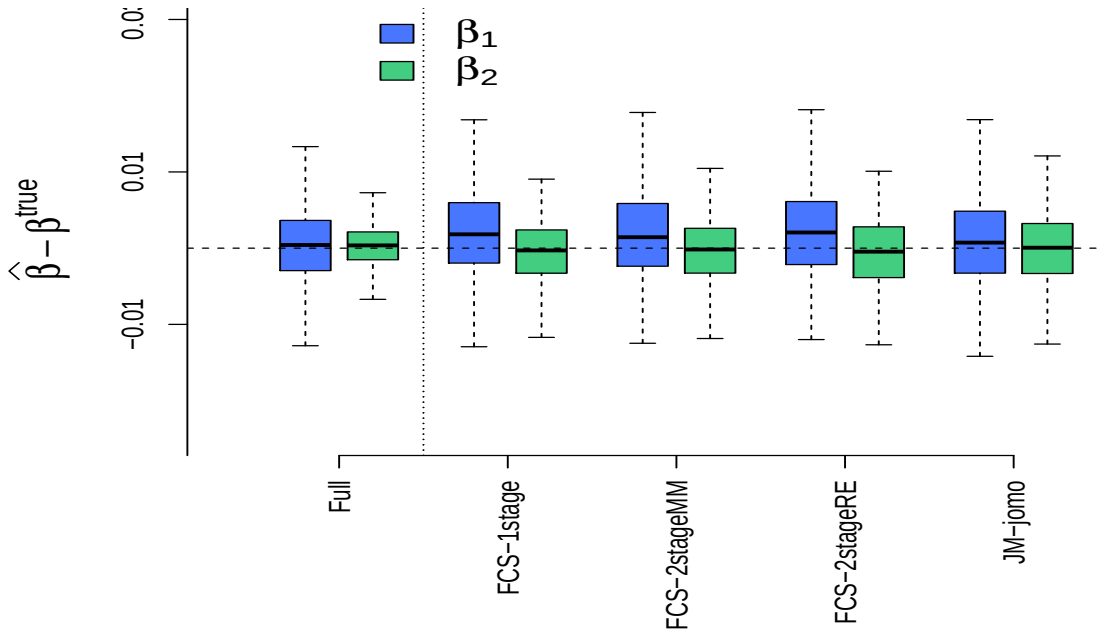


Robustness to π_{syst}: variance estimate





Robustness to the type of imputed variables



Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0050	0.0029	0.0049	0.0028	95.0	95.0	
FCS-1stage	0.0057	0.0044	0.0059	0.0043	92.0	95.2	**103.665**
FCS-2stagemm	0.0063	0.0051	0.0060	0.0044	94.0	96.2	0.652
FCS-2stagere	0.0056	0.0045	0.0061	0.0044	**90.4**	95.0	1.572
JM-jomo	0.0074	0.0072	0.0064	0.0047	97.0	98.6	5.612

Other configurations

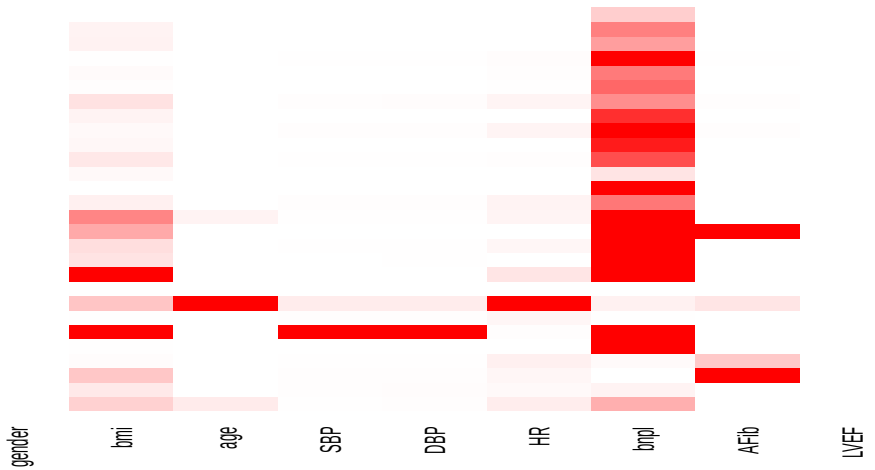
Methods have similar performances when

- the missing data mechanism is MAR
- the outcome of the analysis model is binary
- the variance of random effects is higher or smaller
- binary variables are generated using a probit link

Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays

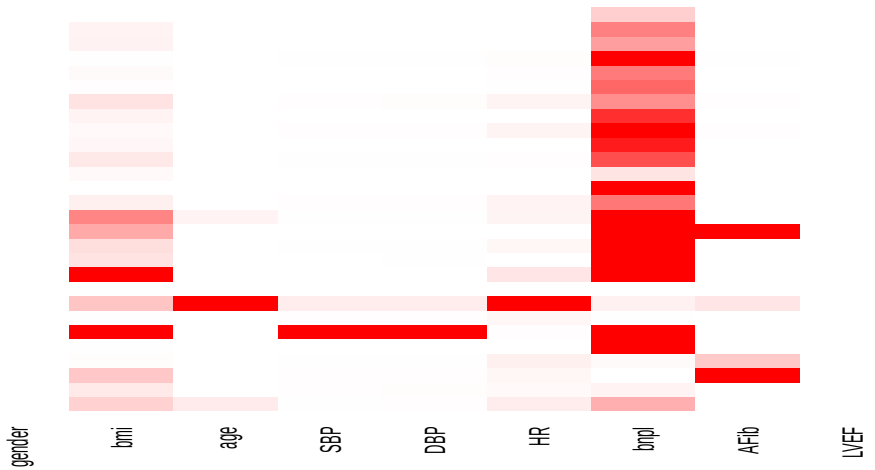


		CC	JM	FCS-1stage	FCS-2stagere	FCS-2stagemm
β_{BNP}	Est	**-0.1132**	**-0.0891**	**-0.0902**	**-0.0854**	**-0.1009**
	ModelSE	0.0108	0.0078	0.0153	0.0099	0.0112
β_{AFIB}	Est	0.0268	0.0216	0.0251	0.0215	0.0273
	ModelSE	0.0071	0.0046	0.0047	0.0040	0.0045
Time			94.0	18609.3	361.3	31.8

Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays

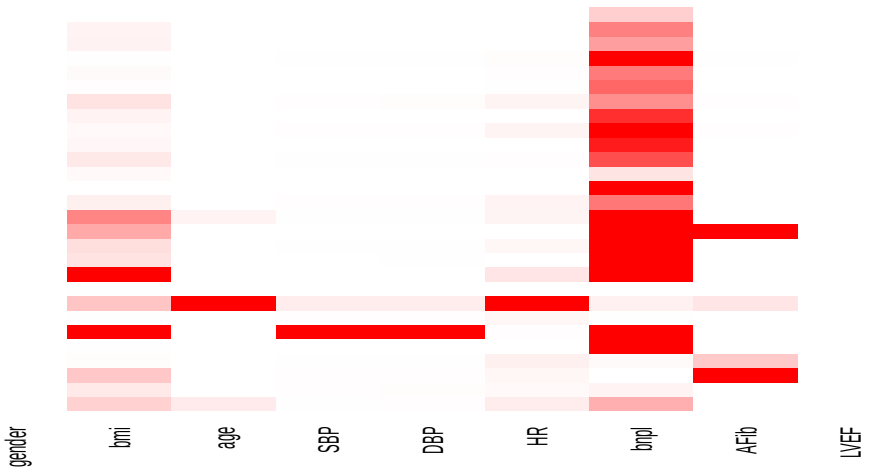


		CC	JM	FCS-1stage	FCS-2stagere	FCS-2stagemm
β_{BNP}	Est	-0.1132	-0.0891	-0.0902	-0.0854	-0.1009
	ModelSE	**0.0108**	0.0078	0.0153	0.0099	0.0112
β_{AFIB}	Est	0.0268	0.0216	0.0251	0.0215	0.0273
	ModelSE	**0.0071**	0.0046	0.0047	0.0040	0.0045
Time			94.0	18609.3	361.3	31.8

Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays

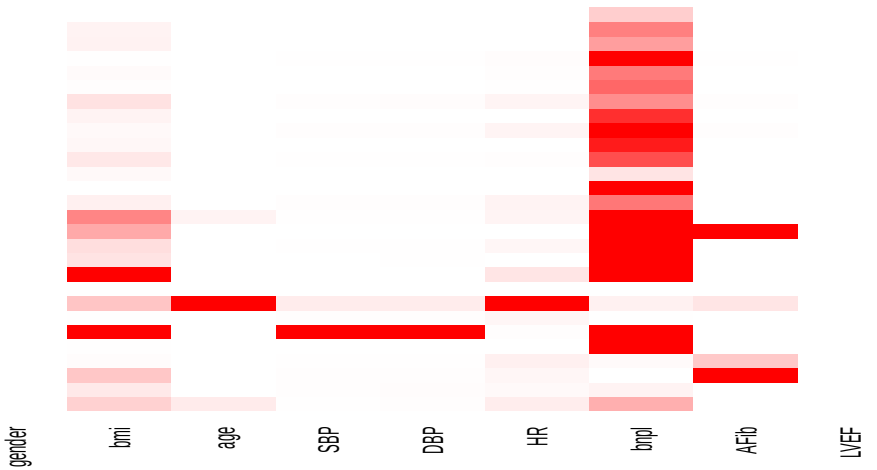


		CC	JM	FCS-1stage	FCS-2stagere	FCS-2stagemm
β_{BNP}	Est	-0.1132	-0.0891	-0.0902	-0.0854	-0.1009
	ModelSE	0.0108	0.0078	**0.0153**	0.0099	**0.0112**
β_{AFIB}	Est	0.0268	0.0216	0.0251	0.0215	0.0273
	ModelSE	0.0071	0.0046	0.0047	0.0040	0.0045
Time			94.0	18609.3	361.3	31.8

Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays

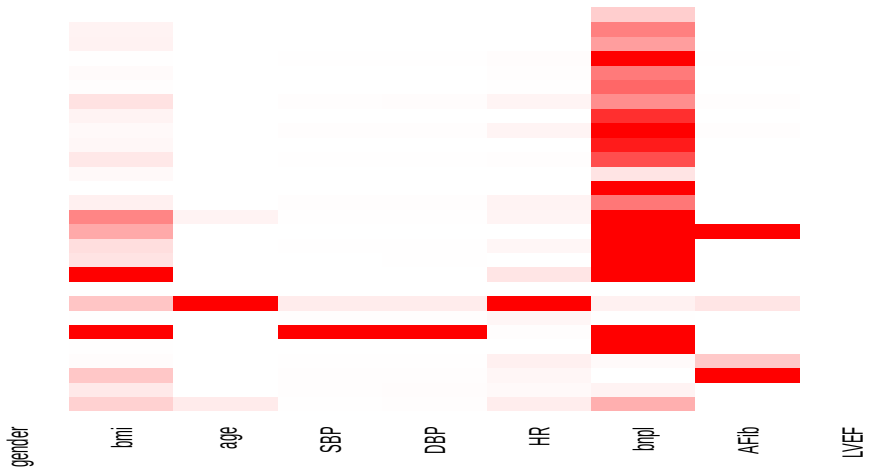


		CC	JM	FCS-1stage	FCS-2stagere	FCS-2stagemm
β_{BNP}	Est	-0.1132	-0.0891	-0.0902	-0.0854	-0.1009
	ModelSE	0.0108	**0.0078**	0.0153	0.0099	0.0112
β_{AFIB}	Est	0.0268	0.0216	0.0251	0.0215	0.0273
	ModelSE	0.0071	0.0046	0.0047	0.0040	0.0045
Time			94.0	18609.3	361.3	31.8

Appplication to GREAT data

Explain the relationship between biomarkers easily measurable (BNP, AFIB) and the left ventricular ejection fraction

- **y**=LVEF,
  X = BNP, AFib
- MI using $M = 20$ imputed arrays



		CC	JM	FCS-1stage	FCS-2stagere	FCS-2stagemm
β_{BNP}	Est	-0.1132	-0.0891	-0.0902	-0.0854	-0.1009
	ModelSE	0.0108	0.0078	0.0153	0.0099	0.0112
β_{AFIB}	Est	0.0268	0.0216	0.0251	0.0215	0.0273
	ModelSE	0.0071	0.0046	0.0047	0.0040	0.0045
Time			94.0	**18609.3**	361.3	31.8

Conclusion

An overview of MI methods for multilevel data

- FCS-1stage, FSC-2stage and JM-jomo all appear to perform well
- Outperform had-hoc methods

FCS-2stage	• MM version provides a quick way to obtain first results • for large clusters
FCS-1stage	• relevant with few systematically missing values • time consuming with binary variables
JM-jomo	• advised with a lot of incomplete categorical variables • be careful with few clusters

Methods are implemented in R

- mice package for FCS methods
- jomo package for JM-jomo

Limits and perspectives

Limits

- congeniality

$$y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$$

- convergence of the FCS approaches
- computational time

Perspectives

- correction for FCS-2stage with small clusters?
- handle logistic link for FCS-1stage?

References I

Great Network. Managing acute heart failure in the ED - case studies from the acute heart failure academy, 2013. http://www.greatnetwork.org.

D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.

S. van Buuren. Multiple imputation of multilevel data. In *The Handbook of Advanced Multilevel Analysis*. Routledge, Milton Park, UK, 2010.

S. Jolani, T. P. A. Debray, H. Koffijberg, S. van Buuren, and K. G. M. Moons. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine*, 34(11):1841–1863, 2015.

M. Resche-Rigon, I. R. White, J. Bartlett, S.A.E. Peters, S.G. Thompson, and on behalf of the PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine*, 32(28):4890–4905, 2013. ISSN 1097-0258. doi: 10.1002/sim.5894.

M. Resche-Rigon and I. White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *smmr*, 2016.

J. L. Schafer. Imputation of missing covariates under a multivariate linear mixed model. Technical report, Dept. of Statistics, The Pennsylvania State University, 1997.

M. Quartagno and J. R. Carpenter. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17):2938–2954, 2016. ISSN 1097-0258.