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

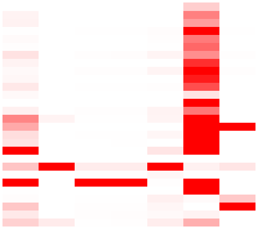
V. Audigier, I. White , S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. van Buuren, M. Resche-Rigon

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

ISCB 2016, August 21-25th, Birmingham

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** $var\left(\widehat{\beta}\right)$

MI methods 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

Some MI methods to impute **multilevel data**

- FCS-2lnorm (van Buuren, 2010): continuous / sporadic
- **FCS-1step** (Jolani et al., 2015; Resche-Rigon et al., 2013): mixed / systematic
- **FCS-2step** (Resche-Rigon and White, 2016): mixed / systematic and sporadic
- JM-Pan (Schafer, 1997): continuous / systematic and sporadic
- **JM-jomo** (Quartagno and Carpenter, 2016): mixed / systematic and sporadic

However, only **FCS-1step**, **FCS-2step** and **JM-jomo** handle **systematically missing** values and **mixed data**

Continuous variables

Heteroscedastic mixed-effects 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 $\theta_m = (\beta^m, \Psi^m, \Sigma_k^m)$
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-1step (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 $\theta_m = (\beta_m, \Psi_m, {\sigma^2}_m) \qquad 1 \leq m \leq M$
    - estimate θ and $var\left(\widehat{\theta}\right)$ by REML
    - draw θ_m from an appropriate distribution with expectation $\widehat{\theta}$, and variance $\widehat{var}\left(\widehat{\theta}\right)$
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-1step (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 $\theta_m = (\beta_m, \Psi_m, {\sigma^2}_m) \qquad 1 \leq m \leq M$
   - estimate θ and $var\left(\widehat{\theta}\right)$ by REML
   - draw θ_m from an appropriate distribution with expectation $\widehat{\theta}$, and variance $\widehat{var}\left(\widehat{\theta}\right)$
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-2step (Resche-Rigon and White, 2016)

Conditional imputation models

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

→ **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 $var\left(\widehat{\theta}\right)$ by **a two-step estimator**:

     step a fit $y_{ik} = \mathbf{x}_{ik}\beta_k + \varepsilon_{ik}$ to each cluster

     step b combine the $\widehat{\beta}_k$ and $\widehat{\sigma}_k^2$ by multivariate meta-analysis (by REML or MM)
   - draw θ_m from an appropriate distribution with expectation $\widehat{\theta}$, and variance $\widehat{var}\left(\widehat{\theta}\right)$
2. impute in each cluster k
   - 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

JM-jomo (Quartagno and Carpenter, 2016)

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

1. Bayesian formulation to generate $\theta_m = (\beta_m, \Psi_m, \Sigma_m)_{1 \leq m \leq M}$
   - prior: $\Sigma_k^{-1} \sim W(\nu_1, \Lambda_1)$, $\Psi^{-1} \sim W(\nu_2, \Lambda_2)$, $\beta \propto 1$
   - posterior: unknown $\rightarrow$ **Gibbs sampler**

$$\begin{aligned}
\beta^{(\ell+1)} &\sim P\left(\beta | X^{obs}, X^{miss\,(\ell)}, \Sigma^{(\ell)}, b^{(\ell)}\right)\\
b_k^{(\ell+1)} &\sim P\left(b_k | X^{obs}, X^{miss\,(\ell)}, \beta^{(\ell+1)}, \Psi^{(\ell)}, \Sigma_k^{(\ell)}\right)\\
\Psi^{-1\,(\ell+1)} &\sim P\left(\Psi^{-1} | X^{obs}, X^{miss\,(\ell)}, b^{(\ell+1)}\right)\\
\Sigma_k^{-1\,(\ell+1)} &\sim P\left(\Sigma_k^{-1} | X^{obs}, X^{miss\,(\ell)}, b_k^{(\ell+1)}\right)\\
X_k^{miss\,(\ell+1)} &\sim P\left(X_k^{miss} | X^{obs}, \beta^{(\ell+1)}, \Psi^{(\ell+1)}, \Sigma^{(\ell+1)}, b_k^{(\ell+1)}\right)
\end{aligned}$$

2. Imputation (given by step 1)

Binary variables

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

  $\rightarrow$ sporadically missing values not handled

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

  $\rightarrow$ large clusters are required

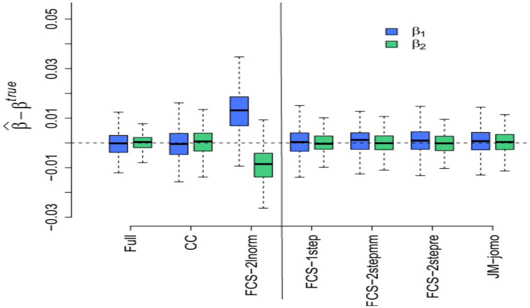
- **JM-jomo** (Quartagno and Carpenter, 2016)
    - draw latent normal variables
    - derive categories

  $\rightarrow$ more time consuming

Simulation design

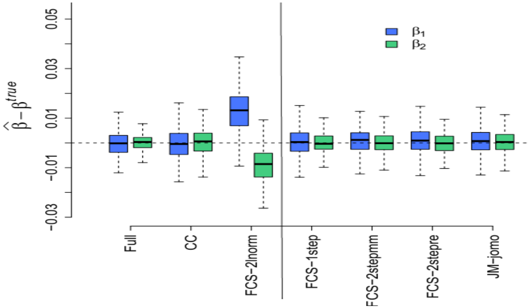
- **Data generation**: 500 incomplete data sets are independently simulated ($n = 11685$, $K = 28$, $18 \leq n_k \leq 1834$)
  - $y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$
    with $\beta = (.72, .11, .03)$, $\Psi = \begin{bmatrix} .0077 & .0015 \\ .0015 & .0004 \end{bmatrix}$, $\sigma = .15$
  - $\mathbf{x}_{ik}^{(1)} : \mathcal{N}\left(\mu + \mu_k, .36\right)$
  - $\mathbf{x}_{ik}^{(2)} : logit\left(P\left(\mathbf{x}_{ik}^{(2)} = 1\right)\right) = \nu + \nu_k$
  - add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ varying π_{syst} and π_{spor}
- **Analysis**: β and $var\left(\widehat{\beta}\right)$ estimated by applying MI methods using $M = 5$ imputed arrays
- **Criteria**: bias, rmse, variance estimate, coverage

Results: $\pi_{syst} = .1$, $\pi_{spor} = .375$



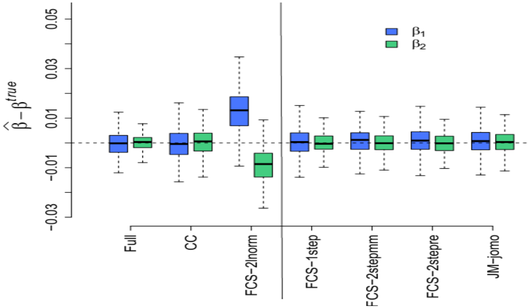
Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0063	0.0052	0.0066	0.0053	90.4	93.4	
FCS-2lnorm	0.0086	0.0044	0.0086	0.0068	71.8	57.0	2.078
FCS-1step	0.0050	0.0044	0.0055	0.0043	91.6	95.4	82.722
FCS-2step-mm	0.0055	0.0047	0.0055	0.0042	94.2	95.5	0.77
FCS-2step-re	0.0056	0.0045	0.0056	0.0042	93.2	95.9	1.835
JM-jomo	0.0059	0.0053	0.0054	0.0043	95.8	96.0	6.3

Results: $\pi_{syst} = .1,\ \pi_{spor} = .375$



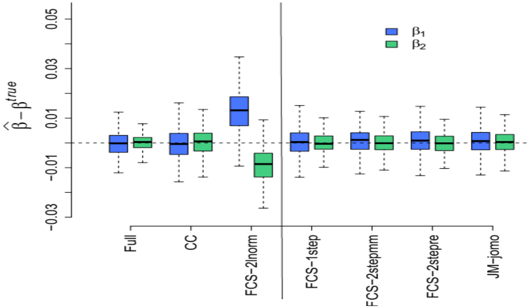
Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0063	0.0052	**0.0066**	**0.0053**	90.4	93.4	
FCS-2lnorm	0.0086	0.0044	0.0086	0.0068	71.8	57.0	2.078
FCS-1step	0.0050	0.0044	**0.0055**	**0.0043**	91.6	95.4	82.722
FCS-2step-mm	0.0055	0.0047	**0.0055**	**0.0042**	94.2	95.5	0.77
FCS-2step-re	0.0056	0.0045	**0.0056**	**0.0042**	93.2	95.9	1.835
JM-jomo	0.0059	0.0053	**0.0054**	**0.0043**	95.8	96.0	6.3

Results: $\pi_{syst} = .1,\ \pi_{spor} = .375$



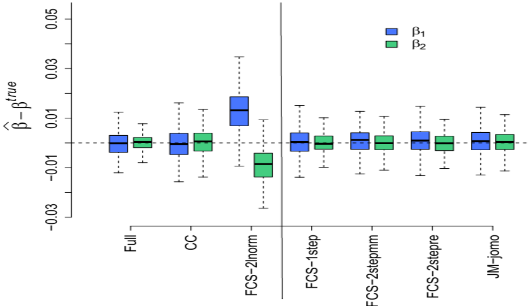
Method	$\sqrt{\widehat{var}\left(\widehat{\beta}\right)}$		$\sqrt{var\left(\widehat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0063	0.0052	0.0066	0.0053	90.4	93.4	
FCS-2lnorm	0.0086	0.0044	0.0086	0.0068	71.8	57.0	2.078
FCS-1step	**0.0050**	0.0044	**0.0055**	0.0043	91.6	95.4	82.722
FCS-2step-mm	0.0055	0.0047	0.0055	0.0042	94.2	95.5	0.77
FCS-2step-re	0.0056	0.0045	0.0056	0.0042	93.2	95.9	1.835
JM-jomo	0.0059	0.0053	0.0054	0.0043	95.8	96.0	6.3

Results: $\pi_{syst} = .1$, $\pi_{spor} = .375$



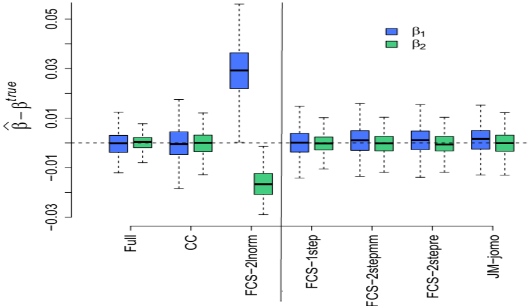
Method	$\sqrt{\widehat{var}\left(\hat{\beta}\right)}$		$\sqrt{var\left(\hat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0063	0.0052	0.0066	0.0053	90.4	93.4	
FCS-2lnorm	0.0086	0.0044	0.0086	0.0068	71.8	57.0	2.078
FCS-1step	0.0050	0.0044	0.0055	0.0043	91.6	95.4	82.722
FCS-2step-mm	0.0055	0.0047	0.0055	0.0042	94.2	95.5	0.77
FCS-2step-re	0.0056	0.0045	0.0056	0.0042	93.2	95.9	1.835
JM-jomo	**0.0059**	**0.0053**	**0.0054**	**0.0043**	95.8	96.0	6.3

Results: $\pi_{syst} = .1$, $\pi_{spor} = .375$



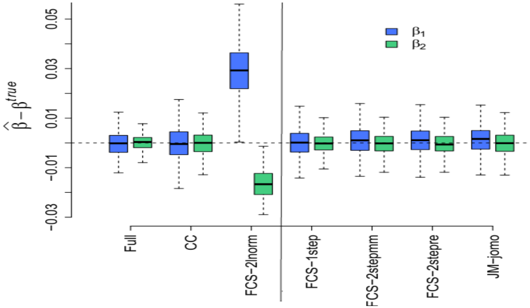
Method	$\sqrt{\widehat{var}\left(\hat{\beta}\right)}$		$\sqrt{var\left(\hat{\beta}\right)}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0063	0.0052	0.0066	0.0053	90.4	93.4	
FCS-2lnorm	0.0086	0.0044	0.0086	0.0068	71.8	57.0	2.078
FCS-1step	0.0050	0.0044	0.0055	0.0043	**91.6**	95.4	82.722
FCS-2step-mm	0.0055	0.0047	0.0055	0.0042	94.2	95.5	0.77
FCS-2step-re	0.0056	0.0045	0.0056	0.0042	93.2	95.9	1.835
JM-jomo	0.0059	0.0053	0.0054	0.0043	**95.8**	**96.0**	6.3

Results: $\pi_{syst} = .25, \pi_{spor} = .25$



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-2lnorm	0.0105	0.0038	0.0106	0.0058	19.8	11.8	1.967
FCS-1step	0.0049	0.0046	0.0056	0.0043	**90.4**	96.8	63.43
FCS-2step-mm	0.0059	0.0054	0.0058	0.0044	95.0	97.0	0.538
FCS-2step-re	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

Results: $\pi_{syst} = .25, \pi_{spor} = .25$



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-2lnorm	0.0105	0.0038	0.0106	0.0058	19.8	11.8	1.967
FCS-1step	0.0049	0.0046	0.0056	0.0043	90.4	96.8	**63.43**
FCS-2step-mm	0.0059	0.0054	0.0058	0.0044	95.0	97.0	**0.538**
FCS-2step-re	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**

Conclusion

An overview of MI methods for multilevel mixed data

- standard methods are irrelevant
- FCS-1step, FSC-2step and JM-jomo all appear to perform well
- inference performances are quite similar
- FCS-2step is quicker to perform

Perspectives

- a larger simulation study (MAR mechanism, number of clusters, size of clusters,...)
- a precise guidance

References I

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

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.