Statistical Science
2018, Vol. 33, No. 2, 160–183
https://doi.org/10.1214/18-STS646

Multiple Imputation for Multilevel Data with Continuous and Binary Variables

Vincent Audigier, Ian R. White, Shahab Jolani, Thomas P. A. Debray, Matteo Quartagno, James Carpenter, Stef van Buuren and Matthieu Resche-Rigon

Abstract. We present and compare multiple imputation methods for multilevel continuous and binary data where variables are systematically and sporadically missing. The methods are compared from a theoretical point of view and through an extensive simulation study motivated by a real dataset comprising multiple studies. The comparisons show that these multiple imputation methods are the most appropriate to handle missing values in a multilevel setting and why their relative performances can vary according to the missing data pattern, the multilevel structure and the type of missing variables. This study shows that valid inferences can only be obtained if the dataset includes a large number of clusters. In addition, it highlights that heteroscedastic multiple imputation methods provide more accurate inferences than homoscedastic methods, which should be reserved for data with few individuals per cluster. Finally, guidelines are given to choose the most suitable multiple imputation method according to the structure of the data.

Key words and phrases: Missing data, systematically missing values, multilevel data, mixed data, multiple imputation, joint modelling, fully conditional specification.

1. INTRODUCTION

When individual observations are nested in clusters, statistical analyses generally need to reflect this structure; this is usually referred to as a multilevel structure, where individuals constitute the lower level, and clusters the higher level. This situation often arises in fields including survey research, educational science, sociology, geography, psychology and clinical studies.

Missing data affect most datasets in these fields, and multilevel data present specific patterns of miss-

Vincent Audigier is Associate Professor, CNAM, Cedric MSDMA, Paris, France (e-mail: vincent.audigier@cnam.fr). Ian R. White is Professor, MRC Biostatistics Unit, Cambridge Institute of Public Health, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: ian.white@ucl.ac.uk). Shahab Jolani is Assistant Professor, Department of Methodology and Statistics, School CAPHRI, Care and Public Health Research Institute, Maastricht University, The Netherlands (e-mail: s.jolani@maastrichtuniversity.nl). Thomas P. A. Debray is Assistant Professor, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands (e-mail: T.Debray@umcutrecht.nl). Matteo Quartagno is Research Fellow, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: Matteo.Quartagno@lshtm.ac.uk). James Carpenter is Professor, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: james.carpenter@lshtm.ac.uk). Stef van Buuren is Professor, Department of Methodology & Statistics, FSS, University of Utrecht, The Netherlands; Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands (e-mail: S.vanBuuren@uu.nl). Matthieu Resche-Rigon is Professor, Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, Paris, France; Université Paris Diderot—Paris 7, Sorbonne Paris Cité, UMR-S 1153, Paris, France; INSERM, UMR 1153, Equipe ECSTRA, Hôpital Saint-Louis, Paris, France (e-mail: matthieu.resche-rigon@univ-paris-diderot.fr).

ing values. Some variables may be fully unobserved for some clusters because they were not measured, or because they were not defined consistently across clusters. Resche-Rigon et al. (2013) named such missing data patterns *systematically missing values*. Nowadays, systematically missing values are becoming increasingly common because of a greater availability of data coming from several sources, including different sets of variables (Riley et al., 2016). Examples of data with systematically missing values are numerous, even in the absence of multilevel structure, and include Bos et al. (2003), Mullis et al. (2003) and Blossfeld, Günther Roŝbach and von Maurice (2011) in educational sciences, Kunkel and Kaizar (2017), Global Research on Acute conditions Team (GREAT) Network (2013) in medicine, and Carrig et al. (2015) in sociology. As opposed to systematically missing values, *sporadically missing values* are missing data specific to each individual observation. Often both types of missing data occur in a multilevel dataset. For instance, in educational research, questionnaires may be too long to be administered to all students and, therefore, only a subset of items may be asked of each class, leading to systematically missing values. In addition, some students in each class may not answer some questions, leading to sporadically missing values.

Multiple imputation (MI) is a common strategy to deal with missing values in statistical analysis (Schafer, 1997, Rubin, 1987, Little and Rubin, 2002). It involves first specifying a distribution in accordance with the data, the *imputation model*, under which M imputations are drawn from their posterior (or approximated posterior) predictive distribution given the observed values. Thus, M complete datasets are generated. Second, a standard statistical analysis is performed on each imputed dataset, leading to M estimates of the *analysis model*'s parameters. Finally, the estimates are pooled according to Rubin's rules (Rubin, 1987). The standard assumption when using MI is ignorability (Schafer, 1997, page 11), implying that missing values occur *at random* (Rubin, 1976), that is, the probability of missingness depends solely on observed data. Several MI methods have been proposed, differing mainly in the assumed form of the imputation model. Among methods assuming a parametric imputation model, two strategies are the most common: joint modelling (JM) imputation when a multivariate joint distribution is specified for all variables, and fully conditional specification (FCS) when a conditional distribution is defined for each incomplete variable (van Buuren et al., 2006, Raghunathan et al., 2001).

In the standard statistical framework where data are complete, multilevel data induce dependence between observations and require dedicated analysis models accounting for this dependency. The linear mixed effects model is one such model. In the same way, with missing values, imputation models need to take into account dependency between observations, since otherwise the prediction variance of the missing values cannot be properly reflected. Indeed, when an incomplete variable is part of a linear mixed effect analysis model, but it is imputed ignoring the multilevel structure, the imputed values can be unsuitable (Reiter, Raghunathan and Kinney, 2006). Thus, biases can occur even when applying appropriate statistical methods on inappropriately imputed data.

To account for the multilevel structure with sporadically missing values, imputation models are generally based on regression models including a fixed or a random intercept for cluster (Drechsler, 2015). Methods using a fixed intercept treat the identifier of each cluster as a dummy variable. They are generally parametric, using normal regression for instance, but imputation according to semi-parametric method can also be relevant for complex datasets (Vink, Lazendic and van Buuren, 2015, Little, 1988). However, using a random intercept is generally preferable because fixed intercept inflates the true variability between clusters (Andridge, 2011, Graham, 2012). MI methods using a random intercept, with the short names we use for them in this paper, include JM for multivariate panel data (Schafer and Yucel, 2002), *JM-pan*; multilevel JM multiple imputation (Quartagno and Carpenter, 2016a), *JM-jomo*; JM for realistically complex social science data (Goldstein, Bonnet and Rocher, 2007, Goldstein et al., 2009, Carpenter and Kenward, 2013), *JM-REALCOM*; JM based on latent variable model (Asparouhov and Muthén, 2010), *JM-Mplus*; JM using random-covariances and mixed-effects models (Yucel, 2011), *JM-RCME*; FCS for multivariate panel data (Schafer and Yucel, 2002), *FCS-pan*; FCS by Bayesian multilevel imputation (Enders, Keller and Levy, 2017), *FCS-blimp*; FCS using two-level normal model (van Buuren, 2011), *FCS-2lnorm*; FCS using generalized linear mixed model (Jolani et al., 2015, Jolani, 2018), *FCS-GLM*; and FCS based on a two-stage estimator (Resche-Rigon and White, 2016), *FCS-2stage*. These methods differ in the form of the imputation model (joint or not), but also in their ability to account for different types of variables (continuous, binary, or others). In particular, JM-pan, JM-RCME, FCS-pan, FCS-2lnorm do not accommodate binary variables.

TABLE 1

Summary of MI methods' properties for multilevel data based on random intercept [*JM-pan* (*Schafer and Yucel*, 2002), *JM-REALCOM* (*Goldstein*, *Bonnet and Rocher*, 2007, *Goldstein et al.*, 2009, *Carpenter and Kenward*, 2013), *JM-jomo* (*Quartagno and Carpenter*, 2016a), *JM-Mplus* (*Asparouhov and Muthén*, 2010), *JM-RCME* (*Yucel*, 2011), *FCS-pan* (*Schafer and Yucel*, 2002), *FCS-blimp* (*Enders*, *Keller and Levy*, 2017) *FCS-2lnorm* (*van Buuren*, 2011), *FCS-GLM* (*Jolani et al.*, 2015), *FCS-2stage* (*Resche-Rigon and White*, 2016)]

Method	**Handles missing data:**				**Coded in R**
(form-name)	**Sporadic?**	**Systematic?**	**in continuous variable?**	**in binary variable?**	**(R Core Team, 2016)**
JM-pan	yes	yes	yes	no	yes, package *pan*
JM-REALCOM	yes	yes	yes	yes	no
JM-jomo	yes	yes	yes	yes	yes, package *jomo*
JM-Mplus	yes	yes	yes	yes	no
JM-RCME	yes	yes	yes	no	no
FCS-pan	yes	yes	yes	no	yes, package *mice*
FCS-blimp	yes	yes	yes	yes	no
FCS-2lnorm	yes	no	yes	no	yes, package *mice*
FCS-GLM	yes[1]	yes	yes	yes	yes, add-on for *mice*
FCS-2stage (REML or MM)	yes	yes	yes	yes[1]	yes, add-on for *mice*

[1] Using variant reported in this paper.

Systematically missing values imply identifiability issues for imputation models which include a fixed intercept for cluster. Thus, only methods using random effects are appealing. Most of the MI methods first proposed to impute multilevel data were based on random intercept models, but systematically missing values were not considered. Therefore, not all of these methods are tailored for the imputation of such missing data. Table 1 summarizes the MI methods available for multilevel data.

In this paper, we compare the most relevant MI methods for dealing with clustered datasets with systematically and sporadically missing variables, continuous and binary. Among JM methods for multilevel data, we focus on the JM-jomo method proposed in Quartagno and Carpenter (2016a), which can be seen as a generalisation of JM-REALCOM and JM-Mplus, while among FCS methods, we focus on the FCS-GLM method presented in Jolani et al. (2015) and on the FCS-2stage method proposed in Resche-Rigon and White (2016). We do not focus on FCS-blimp which can be seen as a univariate version of JM-jomo.

The paper is organized as follows. First, we present the three MI methods for handling multilevel data with systematically and sporadically missing values (Section 2). Both FCS methods have theoretical deficiencies in this general setting, so we propose improvements for them: accounting for binary variables in FCS-2stage, and accounting for continuous sporadically missing values in FCS-GLM. Second, these MI methods are compared through a simulation study (Section 3). Third, MI methods are applied to a real data analysis (Section 4). Finally, practical recommendations are provided (Section 5).

2. MULTIPLE IMPUTATION FOR MULTILEVEL CONTINUOUS AND BINARY DATA

2.1 Methods

2.1.1 *Univariate missing data pattern.* Random variables will be indicated in italics, while fixed values will be denoted in roman letters. Vectors will be in lower case, while matrices will be in upper case. Let $\mathbf{Y}_{n\times p} = (\mathbf{y}_1, \dots, \mathbf{y}_p)$ be an incomplete data matrix for n individuals in rows and p variables in columns. Let i be the index for the individuals $(1 \le i \le n)$ and j for the columns $(1 \le j \le p)$. $\mathbf{Y}$ is stratified into K clusters of size n_k where k denotes the index for the cluster $(1 \le k \le K)$. $\mathbf{y}_{jk}$ denotes the n_k-vector corresponding to the vector $\mathbf{y}_j$ restricted to individuals within cluster k. Let $(\mathbf{y}_j^{\text{obs}}, \mathbf{y}_j^{\text{miss}})$ be the missing and observed parts of $\mathbf{y}_j$ and let $\mathbf{Y}^{\text{obs}} = (\mathbf{y}_1^{\text{obs}}, \dots, \mathbf{y}_p^{\text{obs}})$ and $\mathbf{Y}^{\text{miss}} = (\mathbf{y}_1^{\text{miss}}, \dots, \mathbf{y}_p^{\text{miss}})$.

In order to propose a unified presentation of the three MI methods, we assume in this section that the variable y_p is the only incomplete variable and is continuous. Extension to several incomplete variables, continuous or binary, will be discussed in the next section.

The imputation step in MI aims to draw missing values from the predictive distribution $P(Y^{\text{miss}}|Y^{\text{obs}})$. To achieve this goal, an imputation model with parameter $\boldsymbol{\theta}$ is specified and realisations of the predictive distribution of missing values can be obtained by:

Step (1) drawing $\boldsymbol{\theta}$ from $P(\boldsymbol{\theta}|Y^{\text{obs}})$, its posterior distribution.

Step (2) drawing missing data according to $P(Y^{\text{miss}}|Y^{\text{obs}}, \boldsymbol{\theta})$, their predictive distribution for a given $\boldsymbol{\theta}$.

For a single continuous incomplete variable (y_p), the posterior distribution can be specified by letting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi}, (\boldsymbol{\Sigma}_k)_{1\leq k\leq K})$ be the parameters of a linear mixed effects model:

$$\begin{aligned} y_{pk} &= \mathbf{Z}_k\boldsymbol{\beta} + \mathbf{W}_k b_k + \varepsilon_k, \\ b_k &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \\ \varepsilon_k &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k), \end{aligned} \tag{1}$$

where y_{pk} denotes the incomplete variable restricted to the cluster k, $\mathbf{Z}_k$ $(n_k \times q)$ and $\mathbf{W}_k$ $(n_k \times q')$ are the known covariate matrices corresponding to two subsets of $(\mathbf{y}_{1k}, \dots, \mathbf{y}_{(p-1)k})$, $\boldsymbol{\beta}$ is the q-vector of regression coefficients of fixed effects, b_k is the q'-vector of random effects for cluster k, $\boldsymbol{\Psi}$ $(q' \times q')$ is the between cluster variance matrix, and $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbb{I}_{n_k}$ $(n_k \times n_k)$ is the variance matrix within cluster k. Model (1) is the imputation model used in FCS-GLM and FCS-2stage and potentially in JM-jomo for the case of a univariate missing data pattern.

Drawing parameters of the imputation model from their posterior distribution [Step (1)] can be achieved by several approaches (Little and Rubin, 2002, pages 200–222). A first approach uses explicit Bayesian modelling of (1), specifying a prior distribution for $\boldsymbol{\theta}$ and drawing from its posterior distribution. This approach is used in FCS-GLM and in JM-jomo: FCS-GLM uses a noninformative Jeffreys prior distribution, while JM-jomo uses a conjugate prior distribution.

A second approach uses the asymptotic distribution of a frequentist estimator of $\boldsymbol{\theta}$. More precisely, the parameters of this distribution are estimated from the data, and a value of $\boldsymbol{\theta}$ is drawn from this asymptotic distribution. FCS-2stage is based on this principle: the estimator used is called the *two-stage estimator* in IPD meta-analysis (Simmonds et al., 2005, Riley et al., 2008). It is also possible to use the Maximum Likelihood (ML) estimator (Resche-Rigon et al., 2013), but the two-stage estimator has the advantage of being easier and quicker to compute than the ML estimator for linear mixed effects models.

When the variable y_p is only sporadically missing, the posterior distribution of the parameters only involves individuals that are observed (Rubin, 1987, page 165), so that both approaches easily handle missing data. However, systematically missing values complicate Step (1) for both approaches. Using Bayesian modelling, simulating the posterior distribution of $\boldsymbol{\theta}$ generally requires a Gibbs sampler (Geman and Geman, 1984), but the posterior distribution of $\boldsymbol{\Sigma}_k$ cannot be updated from the data at each iteration for systematically missing clusters. Similarly, $\boldsymbol{\Sigma}_k$ cannot be estimated from observed data by using the asymptotic method. Thus, MI methods developed for sporadically missing data cannot be directly used to impute systematically missing data. The problem is overcome by assuming a distribution across $(\boldsymbol{\Sigma}_k)_{1\leq k\leq K}$, as proposed in JM-jomo and in FCS-2stage, or by assuming $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k, as proposed in FCS-GLM.

To draw missing values according to the parameters drawn at Step (1), missing values are predicted according to model (1) and Gaussian noise is added to the prediction [Step (2)]. However, the random coefficients are not strictly parameters of this model and, therefore, are not directly given by Step (1). Thus, to obtain realisations for $(b_k)_{1\leq k\leq K}$, each random coefficient is drawn from its distribution conditional on y_k^{obs} and the parameters generated from Step (1). Imputation can then be performed. If data are sporadically missing, these conditional distributions are derived by the classic calculation for Gaussian vectors; if data are systematically missing, random coefficients are drawn from their marginal distribution.

From this unified presentation, we now present the methods for several incomplete variables, which can also be binary.

2.1.2 *Multivariate missing data pattern.*

2.1.2.1 *JM-jomo* (Quartagno and Carpenter, 2016a). To multiply impute multilevel data with several incomplete continuous variables, JM approaches are based on the multivariate version of model (1) where covariate matrices are matrices $(n_k \times p)$ of ones:

$$\begin{aligned} Y_k &= \mathbf{1}\boldsymbol{\beta} + \mathbf{1}b_k + \varepsilon_k, \\ b_k &\sim \mathcal{N}(0, \boldsymbol{\Psi}), \\ \varepsilon_k^V &\sim \mathcal{N}(0, \boldsymbol{\Sigma}_k). \end{aligned} \tag{2}$$

$\boldsymbol{\beta}$ is the $1 \times p$ matrix of regression coefficients of fixed effects, and b_k is the $1 \times p$ matrix of random effects. The superscript V indicates the vectorisation of a matrix by stacking its columns. $\boldsymbol{\Psi}$ $(p \times p)$ is the between cluster variance matrix and $\boldsymbol{\Sigma}_k$ $(pn_k \times pn_k)$ is the block diagonal variance matrix within cluster k.

Note that model (2) includes all variables on the left-hand side of imputation model $[Y_k = (Y_k^{\text{miss}}, Y_k^{\text{obs}})]$.

Other modelling could be considered by including complete variables on the left-hand or right-hand side. The proposed model has the advantage of limiting overfitting when the number of variables is small compared to the number of individuals (Quartagno and Carpenter, 2016a).

To perform MI according to this imputation model, a Bayesian approach is used with the following independent prior distributions for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi}, (\boldsymbol{\Sigma}_k)_{1 \le k \le K})$:

$$\boldsymbol{\beta} \propto 1, \tag{3}$$

$$\boldsymbol{\Psi}^{-1} \sim \mathcal{W}(\nu_1, \boldsymbol{\Lambda}_1), \tag{4}$$

$$\boldsymbol{\Sigma}_k^{-1} | \nu_2, \boldsymbol{\Lambda}_2 \sim \mathcal{W}(\nu_2, \boldsymbol{\Lambda}_2) \quad \text{for all } 1 \le k \le K, \tag{5}$$

$$\nu_2 \sim \chi^2(\eta), \quad \boldsymbol{\Lambda}_2^{-1} \sim \mathcal{W}(\nu_3, \boldsymbol{\Lambda}_3), \tag{6}$$

where $\mathcal{W}(\nu, \boldsymbol{\Lambda})$ denotes the Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{\Lambda}$, and η denotes the degrees of freedom of the chi-squared distribution. The prior distributions for the covariance matrices are informative (Gelman, 2006); to make them as vague as possible, hyperparameters are set as $\nu_1 = p$, $\boldsymbol{\Lambda}_1 = \mathbb{I}_p$, $\nu_3 = pK$, $\boldsymbol{\Lambda}_3 = \mathbb{I}_{pK}$ and $\eta = pK$. From this modelling, the posterior distribution of $\boldsymbol{\theta}$ can be derived. The derived posterior distributions as well as the technical details to obtain realisations from them are available in Supplementary Material [Audigier et al., 2018, Equations (15)–(17)].

In summary, the parameters of the imputation model are drawn from their posterior distribution with a multivariate missing data pattern by using a Data-Augmentation (DA) algorithm (Tanner and Wong, 1987): given current values $\boldsymbol{\theta}^{(\ell)}$ for $\boldsymbol{\theta}$ and $\mathbf{Y}^{\text{miss}^{(\ell)}}$ for Y^{miss}, the components of $\boldsymbol{\theta}$ are successively updated according to their posterior distribution given $(Y^{\text{obs}}, Y^{\text{miss}^{(\ell)}})$, providing $\boldsymbol{\theta}^{(\ell+1)}$. Then, $\boldsymbol{\theta}^{(\ell+1)}$ can be used to draw $Y^{\text{miss}^{(\ell+1)}}$ according to model (2). To obtain M independent realisations from the posterior distribution, the algorithm is run through a burn-in period (to reach the convergence to the posterior distribution) and then realisations are drawn by spacing them with several iterations (to ensure independence). Note that the number of iterations for the burn-in period and the number of iterations between realisations need to be carefully checked (Schafer, 1997, pages 160–169). Moreover, since generating $\boldsymbol{\theta}$ in its predictive distribution using the DA algorithm also requires imputation of missing data, Step (1) and Step (2) of MI (see Section 2.1.1) are not distinguished here.

This method allows imputation of datasets with systematically and sporadically missing values. In particular, despite systematically missing values, the posterior distribution for $\boldsymbol{\Sigma}_k$ can be updated at each step of the DA algorithm by considering observed values from other clusters.

To deal with binary variables, a probit link and a latent variables framework have been proposed (Goldstein et al., 2009). Let L be the set of continuous variables joined with a set of latent variables corresponding to the binary variables, so that $L = (L^{\text{miss}}, L^{\text{obs}})$. At the end of each cycle of the DA algorithm, given current parameters $\boldsymbol{\theta}^{(\ell)}$ and random coefficients $b^{(\ell)}$, L^{miss} is drawn conditionally on L^{obs}.

Like $P(Y^{\text{miss}}, Y^{\text{obs}})$ for continuous incomplete variables, $P(L^{\text{miss}}, L^{\text{obs}})$ is a multivariate Gaussian distribution. Thus, drawing missing latent variables consists of drawing L from a Gaussian distribution under the positivity or negativity constraint imposed by observed binary values, which is straightforward (Carpenter and Kenward, 2013, pages 96–98). Next, binary data from Y^{miss} are derived from the previously drawn latent variables: the outcome 1 is drawn if the latent variable takes a positive value, and 0 otherwise.

The JM-jomo method is an extension of the JM-RCME (Yucel, 2011), JM-Mplus (Asparouhov and Muthén, 2010), JM-REALCOM (Goldstein, Bonnet and Rocher, 2007, Goldstein et al., 2009, Carpenter and Kenward, 2013) and the JM-pan method (Schafer and Yucel, 2002); JM-jomo additionally allows for heteroscedasticity of the imputation model and imputation of binary (and more generally categorical) variables, while JM-RCME only handles heteroscedasticity, and JM-REALCOM and JM-Mplus propose imputation of categorical variables, but allows only for homoscedasticity with continuous variables.

2.1.2.2 *FCS-GLM* (Jolani et al., 2015). Instead of using a JM approach, fully conditional specification can be used to multiply impute a dataset with several incomplete variables. The principle is to successively simulate from the predictive distributions of the missing values of each incomplete variable conditionally on the other variables. Thus, instead of specifying a joint imputation model as (2), only the conditional distribution of each incomplete variable is required. Compared to JM approaches, FCS approaches make it easier to model complex dependence structures.

Jolani et al. (2015) use a FCS approach to perform multiple imputation of systematically missing variables only. For continuous incomplete variables, the conditional imputation model is model (1) assuming homoscedastic error terms, that is, $\sigma_k = \sigma$ for all k.

To draw missing values of y from their predictive distribution, a Bayesian formulation of the univariate linear mixed effects model based on noninformative independent priors is used. Details on the posterior distributions are available in the Supplementary Material [Audigier et al., 2018, Equations (18)–(20)]; we underline that they depend on the maximum likelihood estimates of the imputation model's parameters.

Imputation of a systematically missing variable y is performed as follows:

Step $(1')$ $\boldsymbol{\theta}$ is drawn according to the posterior distribution,

Step $(2')$ $P(y^{\text{miss}}|Y^{\text{obs}}, \boldsymbol{\theta})$ is simulated by:

- drawing b_k from $\mathcal{N}(0, \boldsymbol{\Psi})$ for all clusters in $1 \leq k \leq K$ where y_k is systematically missing,
- drawing y_k^{miss} from $\mathcal{N}(\mathbf{Z}_k\boldsymbol{\beta} + \mathbf{W}_k b_k, \sigma^2 \mathbb{I}_{n_k})$ for all clusters in $1 \leq k \leq K$ where y_k is sporadically missing.

Binary variables are imputed in the same way by considering a generalized linear mixed model (GLMM) with a logit link.

FCS-GLM was originally developed to impute systematically missing variables only. We extend it to also impute sporadically missing continuous variable following the rationale of Resche-Rigon et al. (2013). To achieve this goal, Step $(1')$ is essentially the same, and the main difference lies in Step $(2')$: each b_k is drawn conditionally on y_k^{obs}, instead of being drawn from its marginal distribution. However, when y is a binary variable, this conditional distribution is analytically intractable because of the logit link. Therefore, binary sporadically missing variables are handled as binary systematically missing ones, which can potentially introduce bias and a lack of variability in the imputed values.

van Buuren (2011), Enders, Keller and Levy (2017) and Schafer and Yucel (2002) also proposed FCS approaches (FCS-2lnorm, FCS-blimp, FCS-pan, respectively) which use a conjugate prior to reflect the posterior distribution of the parameter of the imputation model. FCS-2lnorm is based on the model (1) as conditional imputation model and thus allows heteroscedasticity of errors for continuous variables. However, it cannot be directly applied to systematically missing clusters because of nonidentifiability of (σ_k^2) for $1 \leq k \leq K$. On the contrary, FCS-blimp and FCS-pan assume homoscedasticity only.

2.1.2.3 *FCS-2stage* (Resche-Rigon and White, 2016). FCS-2stage is another FCS method drawing the parameters of the imputation model by using an asymptotic strategy: an estimator is evaluated from the observed data and the posterior distribution is then approximated (cf. Section 2.1.1). This estimator is a two-stage estimator (Simmonds et al., 2005, Riley et al., 2008). Often used in IPD meta-analysis, it has the advantage of being quicker to compute than the usual one-stage estimator required for the previous method (through the expressions of posterior distributions).

More precisely, for a continuous incomplete variable y, the conditional imputation model (1) is rewritten as follows:

$$\begin{aligned} y_k &= \mathbf{Z}_k(\boldsymbol{\beta} + b_k) + \varepsilon_k, \\ b_k &\sim \mathcal{N}(0, \boldsymbol{\Psi}), \\ \varepsilon_k &\sim \mathcal{N}(0, \sigma_k^2 \mathbb{I}_{n_k}). \end{aligned} \tag{7}$$

Note that for clarity, the method is presented for the case $\mathbf{Z}_k = \mathbf{W}_k$. Extension to the more general imputation model is given in Resche-Rigon and White (2016). The parameter of this model is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi}, (\sigma_k)_{1 \leq k \leq K})$.

To fit the two-stage estimator, at stage one, the ML estimator of a linear model is computed on each available cluster:

$$\widehat{\boldsymbol{\beta}}_k = (\mathbf{Z}_k^\top \mathbf{Z}_k)^{-1} \mathbf{Z}_k^\top \mathbf{y}_k. \tag{8}$$

Then, at stage two, the following random effects model is used:

$$\widehat{\boldsymbol{\beta}}_k = \boldsymbol{\beta} + b_k + \varepsilon'_k \tag{9}$$

with $b_k \sim \mathcal{N}(0, \boldsymbol{\Psi})$ and $\varepsilon'_k \sim \mathcal{N}(0, \sigma_k^2 (\mathbf{Z}_k \mathbf{Z}_k^\top)^{-1})$. $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ may be estimated by REML; alternatively, Resche-Rigon and White (2016) suggest using the method of moments (MM), which is even faster, especially with high dimensional $\boldsymbol{\beta}$ (DerSimonian and Laird, 1986, Jackson, White and Riley, 2013).

We explain in the Supplementary Material how the asymptotic distribution of such an estimator can be derived with incomplete data, as well as how realisations from this distribution can be obtained [Audigier et al., 2018, Equations (23)–(26)]. Following such developments, imputation of variable y is performed as follows:

Step $(1'')$ $\boldsymbol{\theta}$ is drawn according to the asymptotic posterior.

Step $(2'')$ $y^{\text{miss}}|Y^{\text{obs}}, \boldsymbol{\theta}$ is generated by:

- drawing b_k from $\mathcal{N}(0, \boldsymbol{\Psi})$ for $1 \leq k \leq K$ if y_k is systematically missing or conditionally on $\widehat{\boldsymbol{\beta}}_k$ if y_k is sporadically missing,

TABLE 2

Synthesis of the modelling assumptions of the MI methods JM-jomo, FCS-GLM and FCS-2stage

	JM-jomo	**FCS-GLM**	**FCS-2stage**
Heteroscedasticity assumption	yes	no	yes
Link function for binary variables	probit	logit	logit
Strategy for proper MI	Bayesian modelling based on conjugate prior	Bayesian modelling based on Jeffrey prior	asymptotic method based on a two-stage estimator

- drawing y_{ki}^{miss} from $\mathcal{N}(\mathbf{z}_{ki}(\boldsymbol{\beta} + b_k), \sigma_k^2)$ for all k $(1 \leq k \leq K)$ and for all i $(1 \leq i \leq n_k)$ such that y_{ki}, the observation i in cluster k, is missing.

Originally, this method was proposed to handle incomplete continuous variables only. We extend it to handle binary variables with both sporadically and systematically missing values by applying a logit link in the imputation model (7). In this case, the two-stage estimator is based on logistic models at stage one. The missing values can be imputed according to a scheme similar to that for continuous variables.

Table 2 sums up the main modelling assumptions of each MI method. In the next section, the consequences of the differences for inference from incomplete data are highlighted.

2.2 Properties

2.2.1 *FCS or JM.* Comparisons between FCS and JM methods have been extensively studied (van Buuren, 2007, Lee and Carlin, 2010, Zhao and Yucel, 2009, Wagstaff and Harel, 2011, Kropko et al., 2014, Hughes et al., 2014, Resche-Rigon and White, 2016, Erler et al., 2016), particularly in settings without clustering. It is generally believed that FCS methods are less likely to yield biased imputations because they allow for more flexibility than JM methods. For multilevel data, the lack of flexibility for JM methods has been recently highlighted when the analysis model includes random slopes corresponding to incomplete variables (Enders, Mistler and Keller, 2016). However, FCS-methods raise other issues like selection of variables for conditional models. Furthermore, the theoretical background of FCS is not well understood and constitutes a current topic of research (Zhu and Raghunathan, 2015, Liu et al., 2014, Bartlett et al., 2015). Indeed, unlike in Gibbs samplers, convergence towards a joint posterior distribution cannot generally be proven (Kropko et al., 2014, Hughes et al., 2014, van Buuren, 2012, page 117). Nevertheless, simulation shows that this weakness might not affect the quality of imputation without clustering (van Buuren et al., 2006). In addition, estimation of conditional distributions is more computationally intensive than the estimation of a joint distribution.

2.2.2 *One-stage or two-stage estimator.* The two FCS approaches use different estimators of the imputation model: the FCS-GLM method uses the one-stage estimator of parameters of model (1), while the two-stage estimator uses the rewritten model (7). The one-stage estimator has the drawback of being computationally intensive and slow to converge (Schafer and Yucel, 2002), particularly with binary variables (Noh and Lee, 2007). The two-stage estimator solves this computational time issue, but tends to have a larger variance (Mathew and Nordström, 2010) and requires large clusters with binary outcome to avoid separability problems (Albert and Anderson, 1984) and to reduce the small-sample bias of the ML estimator (Firth, 1993). Furthermore, by using a limited number of observations at stage one, the FCS-2stage method is more prone to suffer overfitting if the number of covariates or the number of missing values is large.

2.2.3 *Heteroscedasticity.* JM-jomo and FCS-2stage allow for heteroscedasticity of the imputation model, whereas FCS-GLM assumes homoscedastic error variances. It has previously been demonstrated that data generated from a joint homoscedastic model [similar to model (2)] can yield heteroscedastic conditional distributions (Resche-Rigon and White, 2016). As a result, imputation models allowing for heteroscedasticity tend to yield more reliable imputations. Previous simulation studies seem to support this point (van Buuren, 2011, Resche-Rigon and White, 2016). However, homoscedasticity can be a useful assumption when studies are very small, since it overcomes overfitting issues by shrinking cluster-specific parameter estimates towards their weighted average.

2.2.4 *Bayesian modelling or asymptotic strategy for Step* (1). JM-jomo and FCS-GLM consider an explicit Bayesian specification of the imputation model, which implies that uncertainty of $\boldsymbol{\theta}$ is fully propagated. Conversely, FCS-2stage only propagates the asymptotic uncertainty, and may therefore be problematic in small samples. Regardless, in large samples, both approaches should yield similar results (Little and Rubin, 2002, page 216).

As a direct result of Bayesian modelling, JM-jomo and FCS-GLM require the specification of a prior distribution for $\boldsymbol{\theta}$. Various priors have been proposed for hierarchical models such as model (1) (Robert, 2007, pages 456–506). In general, it is recommended to use proper prior distributions when working with multivariate linear mixed effects models (Schafer and Yucel, 2002), as this helps to avoid convergence issues of the Gibbs sampler. To this purpose, JM-jomo considers conjugate prior distributions. A major advantage of using conjugate prior distributions is that drawing from the posterior distribution avoids systematic recourse to MCMC methods. In particular, when using a univariate linear mixed effects model with fully observed covariates, the posterior distribution becomes analytically tractable.

In contrast to JM-jomo, FCS-GLM uses the Jeffreys prior for drawing imputations. This prior is derived from the sampling distribution and can therefore be regarded as noninformative.

In conclusion, all methods are likely to yield different posterior distributions, particularly in the presence of small sample sizes.

2.2.5 *Binary variables*. JM-jomo uses a probit link to model binary variables, while both FCS approaches use a logit link. Although both link functions tend to yield similar predictions, the probit link is more convenient for imputation purposes in multilevel data. The underlying reason is that conditional distributions of random coefficients can be easily simulated with a probit link, because it is based on latent normal variables for which these conditional distributions are well known, but not for mixed models with a logistic link. As a result, imputation of sporadically missing values is achieved in the same way as systematically missing variables for FCS-GLM, that is, by ignoring the relationship between the random effects and the observed values on the imputed variables. It implies that this method is not relevant with binary sporadically missing variables including few missing values. Conversely, for FCS-2stage it is still possible to draw random coefficients from the conditional distribution, by considering the distribution of the random coefficients conditionally on the ML estimates given at stage one. Nevertheless, because of the asymptotic unbiasedness property of the ML estimator for logistic regression models (used at stage one), the performance of the FCS-2stage method for binary variables deteriorates when all clusters contain few observed individuals.

3. SIMULATIONS

3.1 Simulation Design

We consider a simulation study to assess the relative performance of the MI methods described in Section 2. As discussed there, we anticipate that FCS-GLM is problematic when imputing binary sporadically missing variables, that FCS-2stage is problematic in datasets with few participants and/or clusters, and that JM-jomo is sensitive to the proportion of missing values because of the influence of the prior distribution. For this reason, we vary the proportion of systematically and sporadically missing values, the data type of imputed variables, the size of included clusters and the size of the total dataset. Other settings are also investigated to cover a large range of practical cases. For all investigated configurations, we generate $T = 500$ complete datasets, after which we introduce missing values. Afterwards, we apply the MI methods on each incomplete dataset by considering $M = 5$ completed datasets, and obtain parameter estimates from the multiply imputed datasets.

3.1.1 *Data generation*. For each simulation, we generate a dataset with four variables (y, x_1, x_2, x_3); x_1 and x_3 are continuous variables, and x_2 is a binary variable. The outcome variable y (continuous or binary) is defined according to a GLMM with covariates x_1 and x_2. We use x_3 as an auxiliary variable explaining the missing data mechanism. More precisely, data are simulated as follows:

1. Draw K realisations of the triplet of variables (v_1, v_2, v_3) so that

$$(v_1, v_2, v_3) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}}), \tag{10}$$

   where $\boldsymbol{\Sigma}_{\mathbf{v}}$ is a (3×3) covariance matrix.
2. Draw two continuous variables x_1 and x_3 so that

$$(x_{1ki}, x_{3ki}) \sim \mathcal{N}\bigl((\alpha_1 + v_{1k}, \alpha_3 + v_{3k}), \boldsymbol{\Sigma}_{\mathbf{x}}\bigr), \tag{11}$$

   where α_1 and α_3 are the fixed intercepts, v_{1k} and v_{3k} are the random intercepts for cluster k drawn at the previous step and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is a (2×2) covariance matrix.

3. Draw a binary variable x_2 according to the model

$$\operatorname{logit}\big(P(x_{2ki} = 1)\big) = \alpha_2 + v_{2k}, \tag{12}$$

where α_2 is a fixed intercept and v_{2k} is the random intercept for cluster k, drawn from (10).

4. Draw a response variable y:

- for a continuous variable y,

$$\begin{aligned} y_{ki} &= \beta^{(0)} + \beta^{(1)} x_{1ki} + \beta^{(2)} x_{2ki} \\ &\quad + u_k^{(0)} + u_k^{(1)} x_{1ki} + \varepsilon_{ki}, \end{aligned} \tag{13}$$

where $\varepsilon_{ki} \sim \mathcal{N}(0, \sigma_y^2)$ and $(u_k^{(0)}, u_k^{(1)})$ are the random effects for cluster k so that $(u_k^{(0)}, u_k^{(1)}) \sim \mathcal{N}(0, \Psi)$ with $\Psi = \begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{01} & \psi_{11} \end{pmatrix}$;

- for a binary variable y,

$$\begin{aligned} \operatorname{logit}\big(P(y_{ki} = 1)\big) &= \beta^{(0)} + \beta^{(1)} x_{1ki} + \beta^{(2)} x_{2ki} \\ &\quad + u_k^{(0)} + u_k^{(1)} x_{1ki} \end{aligned} \tag{14}$$

with the same assumption for $(u_k^{(0)}, u_k^{(1)})$.

The parameters are chosen to mimic the structure of an individual patient data meta-analysis data set (named GREAT data). The dataset consists of 28 observational cohorts with characteristics, potential risk factors for acute heart failure and outcomes of 11,685 patients. One challenge consists in explaining the left ventricular ejection fraction (LVEF), which is observed by ultrasound, from biomarkers that are easier to measure, such as brain natriuretic peptide (BNP), a blood biomarker, and atrial fibrillation (AFIB). More detail on the data are given in Appendix A.

Data on the variable BNP is used to motivate the distribution of the continuous covariates x_1 and x_3, while the variable AFIB is used to motivate the distribution of the binary covariate x_2. We tune the covariate distribution according to the posterior distribution estimated by the fully Bayesian approach (Section 2.1.2.1) implemented in the R package *jomo* (Quartagno and Carpenter, 2016b). In addition, we used complete-case analysis to estimate parameters of the analysis model (13). Thus, unless otherwise specified, the parameters are $K = 28$, $18 \leq n_k \leq 1093$,

$$\boldsymbol{\Sigma}_{\mathbf{v}} = \begin{pmatrix} 0.12 & 0.001 & 0.001 \\ 0.001 & 0.12 & 0.001 \\ 0.001 & 0.001 & 0.12 \end{pmatrix},$$

$$(\alpha_1, \alpha_3) = (2.9, 2.9), \quad \boldsymbol{\Sigma}_{\mathbf{x}} = \begin{pmatrix} 0.36 & 0.108 \\ 0.108 & 0.36 \end{pmatrix},$$

$$\alpha_2 = 0.42, \quad (\beta^{(0)}, \beta^{(1)}, \beta^{(2)}) = (0.72, -0.11, 0.03),$$

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.0077 & -0.0015 \\ -0.0015 & 0.0004 \end{pmatrix},$$

and $\sigma_Y = 0.15$.

This defines the base-case configuration. Then, these parameters will be varied one by one. Details about the parameters used for each case are provided in Appendix B.1.1 in Table 9.

3.1.2 *Missing data mechanisms.* Variables are independently systematically missing on (x_1, x_2) with probability π_{sys}. In addition, for clusters where a covariate is not systematically missing, sporadically missing values are generated with probability π_{spor}. Unless otherwise specified, $\pi_{sys} = 0.25$ and $\pi_{\text{spor}} = 0.25$, so that the proportion of missing values on x_1 and x_2 is roughly 0.44. Two missing data mechanisms are considered: a MCAR mechanism, where sporadically missing data are generated independently of the data, and a MAR mechanism, where sporadically missing values occur according to the observed values of the auxiliary variable x_3. In both cases systematically missing values remain MCAR.

3.1.3 *Methods.* The simulation study evaluates a total of 10 methods. The reference methods are as follows:

- Full—Analysis of original dataset, before introduction of missing values,
- CC—Case-wise deletion of individuals with incomplete data.

We consider three methods that allow imputation of sporadically and systematically missing data in multilevel data by adopting random effects distributions. The performance of these methods is of primary interest in the current simulation study:

- JM-jomo (Quartagno and Carpenter, 2016a),
- FCS-GLM (Jolani et al., 2015),
- FCS-2stage (Resche-Rigon and White, 2016) (estimation using REML and MM).

We also consider five ad-hoc methods that were not designed to be used in multilevel data with a combination of sporadically and systematically missing values. Nevertheless, these methods are evaluated to highlight the relative merits of the dedicated methods:

- JM-pan (Schafer and Yucel, 2002): JM imputation by linear mixed effects models assuming homoscedasticity,
- FCS-2lnorm (van Buuren, 2011): FCS imputation by linear mixed effects models assuming heteroscedasticity,

- FCS-noclust (Schafer, 1997): FCS imputation by normal or logistic regression,
- FCS-fixclust: FCS imputation by normal or logistic regression with fixed intercept to account for the second-level,
- FCS-fixclustPMM (Little, 1988): FCS imputation by predictive mean matching with fixed intercept to account for the second-level.

JM-pan and FCS-2lnorm only allow imputation of continuous data (Section 2.1.2). For this reason, binary variables are treated as continuous, without applying any rounding strategy (Allison, 2002). Furthermore, because of the heteroscedasticity assumption, the parameters of FCS-2lnorm are not identifiable in the presence of systematically missing values. We address this issue by imputing sporadically missing values and systematically missing values separately from each other. In particular, clusters without systematically missing data are used to fit the imputation model and to impute clusters with sporadically missing values. Afterwards, parameters obtained from the first clusters without systematically missing data are used to impute the remaining clusters with systematically missing data. The FCS-noclust, FCS-fixclust and FCS-fixclustPMM methods use fixed intercepts, implying nonidentifiability of the intercept with systematically missing variables. This issue is addressed by centring the dummy variables, so that clusters with systematically missing values are imputed using the observed average across the remaining clusters.

3.1.4 *Performance measures.* The primary parameters of interest are $\beta^{(1)}$, $\beta^{(2)}$, ψ_{00} and ψ_{11} in model (13) or (14). The performance of the methods in estimating these parameters is assessed by the bias, the root mean squared error (RMSE), the root mean square of estimated standard error (Model SE), the empirical Monte Carlo standard error (Emp SE) and the coverage of the associated confidence interval (Morris, White and Crowther, 2017). The average time required to multiply impute one dataset is also reported.

3.1.5 *Implementation.* Simulations are performed with R software (R Core Team, 2016). Multiple imputation with the JM-jomo method is performed with the R package *jomo* (Quartagno and Carpenter, 2016b). The number of iterations for the burn-in step is set to 2000, and 1000 iterations are run between imputed datasets. Convergence is checked from an incomplete dataset simulated from the base-case configuration by checking the stationarity of the parameters of the imputation model.

Multiple imputation with FCS methods is performed using the R package *mice* (van Buuren and Groothuis-Oudshoorn, 2011) with 5 cycles. Convergence is checked from an incomplete dataset simulated from the base-case configuration by checking the stationarity of marginal quantities (means and standard deviations). For both FCS approaches, conditional imputation models contain all available covariates, which are included in the fixed and random design matrices ($\mathbf{Z}_k = \mathbf{W}_k$).

In both cases, MI is performed using 5 imputed datasets. Each imputed dataset is analysed using the R package *nlme* (Pinheiro et al., 2016) for a continuous outcome and using the *glmer* package for a binary outcome (Bates et al., 2015). Calculation was performed on an Intel® Xeon® CPU E7530 1.87 GHz. The R code used to perform simulations is given in the Supplementary Materials (Audigier et al., 2018).

3.2 Results

3.2.1 *Base-case configuration.* Table 3 describes the simulation study results for the base-case configuration. Overall, all methods yield satisfactory estimates for the coefficient $\beta^{(2)}$ of the binary variable. In particular, biases are smaller than 2%, and coverage of the confidence intervals are close to their nominal level. Performance differences mainly occur for estimation of the coefficient $\beta^{(1)}$ of the continuous variable and for estimation of the variance of random effects ψ_{00} and ψ_{11}. In particular, the variance of $\widehat{\beta^{(1)}}$ is generally underestimated, leading to confidence intervals that do not reach their nominal level.

As expected, ad-hoc methods suffer from several deficiencies. First of all, they underestimate the variance of $\widehat{\beta^{(1)}}$. This is likely related to the use of imputation models with homoscedastic error terms (FCS-noclust, FCS-fixclust, FCS-fixclustPMM and JM-pan) and to not properly modelling heterogeneity between clusters (FCS-noclust, FCS-fixclust and FCS-fixclustPMM). For FCS-2lnorm, underestimation of the variance is likely also caused by ignoring uncertainty on the random coefficients for systematically missing values. A second problem of the ad-hoc methods is that they yield severely biased estimates for the variance of random effects (ψ_{00} and ψ_{11}). Finally, FCS-noclust also introduces bias in the fixed effect coefficients. In particular, by ignoring the multilevel data structure, imputed values of FCS-noclust are biased towards the overall mean and thereby affect corresponding covariate-outcome associations.

The primary methods of interest, JM-jomo, FCS-GLM and FCS-2stage, provide inferences that are

TABLE 3

Simulation study results from the base-case configuration. Point estimate, relative bias, model standard error, empirical standard error, 95% *coverage and RMSE for analysis model parameters and for several methods* (*Full data, Complete-case analysis, FCS-noclust, FCS-fixclust, FCS-fixclustPMM, JM-pan, FCS-2lnorm, JM-jomo , FCS-GLM, FCS-2stage with REML estimator, FCS-2stage with moment estimator*)*. Criteria are based on* 500 *incomplete datasets. Average time to multiply impute one dataset is indicated in minutes. Criteria related to the continuous* (*resp. binary*) *covariate are in light* (*resp. dark*) *grey. True values are* $\beta^{(1)} = -0.11$, $\beta^{(2)} = 0.03$, $\sqrt{\psi_{00}} = 0.088$, $\sqrt{\psi_{11}} = 0.02$

	Full	**CC**	**FCS-noclust**	**FCS-fixclust**	**FCS-fixclustPMM**	**JM-pan**	**FCS-2lnorm**	**JM-jomo**	**FCS-GLM**	**FCS-2stageREML**	**FCS-2stageMM**
$\beta^{(1)}$ est	−0.1101	−0.1104	−0.1102	−0.1102	−0.1039	−0.1088	−0.1098	−0.1087	−0.1100	−0.1089	−0.1090
$\beta^{(1)}$ rbias (%)	0.1	0.3	0.2	0.2	−5.5	−1.1	−0.2	−1.2	−0.0	−1.0	−0.9
$\beta^{(1)}$ model se	0.0047	0.0070	0.0043	0.0043	0.0042	0.0044	0.0056	0.0068	0.0047	0.0059	0.0059
$\beta^{(1)}$ emp se	0.0048	0.0071	0.0057	0.0058	0.0066	0.0056	0.0063	0.0057	0.0057	0.0058	0.0058
$\beta^{(1)}$ 95% cover	93.8	92.2	86.0	85.6	60.4	87.6	92.2	97.6	91.1	95.0	95.0
$\beta^{(1)}$ rmse	0.0048	0.0071	0.0057	0.0058	0.0090	0.0058	0.0063	0.0058	0.0057	0.0059	0.0059
$\beta^{(2)}$ est	0.0301	0.0299	0.0300	0.0300	0.0290	0.0295	0.0301	0.0297	0.0297	0.0295	0.0297
$\beta^{(2)}$ rbias (%)	0.2	−0.4	0.2	−0.0	−3.2	−1.7	0.2	−1.1	−1.1	−1.6	−1.0
$\beta^{(2)}$ model se	0.0029	0.0053	0.0044	0.0043	0.0043	0.0044	0.0064	0.0069	0.0046	0.0054	0.0049
$\beta^{(2)}$ emp se	0.0030	0.0053	0.0043	0.0042	0.0044	0.0042	0.0055	0.0049	0.0042	0.0044	0.0044
$\beta^{(2)}$ 95% cover	94.2	94.4	95.2	95.4	93.6	95.6	95.4	97.2	94.2	97.0	96.2
$\beta^{(2)}$ rmse	0.0030	0.0053	0.0043	0.0042	0.0045	0.0042	0.0055	0.0049	0.0043	0.0045	0.0044
$\sqrt{\psi_0}$ est	0.0859	0.0835	0.0747	0.0745	0.0712	0.0806	0.0816	0.0941	0.0783	0.0869	0.0863
$\sqrt{\psi_0}$ rbias (%)	−2.1	−4.8	−14.9	−15.1	−18.8	−8.2	−7.0	7.3	−10.8	−0.9	−1.6
$\sqrt{\psi_0}$ rmse	0.0154	0.0240	0.0189	0.0191	0.0213	0.0146	0.0166	0.0150	0.0171	0.0148	0.0150
$\sqrt{\psi_1}$ est	0.0193	0.0184	0.0138	0.0138	0.0135	0.0137	0.0174	0.0224	0.0152	0.0197	0.0194
$\sqrt{\psi_1}$ rbias (%)	−3.7	−8.1	−30.9	−31.1	−32.3	−31.4	−13.2	12.0	−24.2	−1.3	−2.9
$\sqrt{\psi_1}$ rmse	0.0041	0.0073	0.0073	0.0074	0.0075	0.0074	0.0053	0.0043	0.0066	0.0046	0.0048
Time			1.8	1.2	0.9	1.4	3.3	8.0	114.7	2.5	1.0

more satisfying as compared to the ad-hoc methods. In particular, biases are smaller and confidence intervals are closer to their nominal level. Nevertheless, some important differences are identified. First of all, the JM-jomo method tends to overestimate the variance of the estimators $\widehat{\beta^{(1)}}$ and $\widehat{\beta^{(2)}}$. Conversely, FCS-GLM tends to underestimate this variance for $\widehat{\beta^{(1)}}$, similar to ad-hoc methods assuming homoscedasticity. Another drawback of FCS-GLM is that its implementation required substantially more time to generate an imputed dataset. Finally, the FCS-2stage method provided satisfactory inferences with both versions (REML and MM). In particular, it is the only method to provide unbiased estimates for variance components ψ_{00} and ψ_{11}.

3.2.2 *Robustness to the proportion of systematically missing values.* To assess the influence of the proportion of systematically missing values, we modify this proportion to 0.1, 0.25 and 0.4 (configurations 2, 1, 3 in Table 9 in Appendix B.1.1, respectively). The proportion of sporadically missing values are modified accordingly to keep the same proportion of missing values in expectation.

For the three methods of primary interest, bias remains stable regardless of the proportion of systematically missing values (see Appendix B.2.1). The relative bias for the standard error estimates is reported in Figure 1. For JM-jomo, the standard error estimate for $\beta^{(1)}$ tends to deviate from the empirical standard error, becoming upwardly biased as the relative extent of systematically missing data increases. This issue is likely related to the use of (informative) conjugate prior distributions, as their influence on the posterior is substantial when the proportion of systematically missing variables is large. Because the FCS methods use other modellings less sensitive to the prior distributions (some prior distributions are derived from the data for FCS-GLM, large sample approximation are used for FCS-2stage), they were less sensitive to overestimation of standard errors.

3.2.3 *Robustness to the number of clusters.* Influence of the number of clusters is assessed by restricting the generated datasets to their K first clusters, and varying K in $\{7, 14, 28\}$. Note that as consequence the total sample size also increases with K (2139, 4256 and 11,685, respectively). Figure 2 describes the impact of the number of clusters on the bias. The impact on the variance estimate is reported in Appendix B.2.2.

For all estimands, the bias obtained from the JM-jomo method is substantial when the number of clusters is small, but decreases when as the number of clusters increases. On the contrary, the FCS methods provide more robust estimates for the variance of the random effects when the number of clusters is small. This behaviour is again likely related to the choice of the prior distributions.

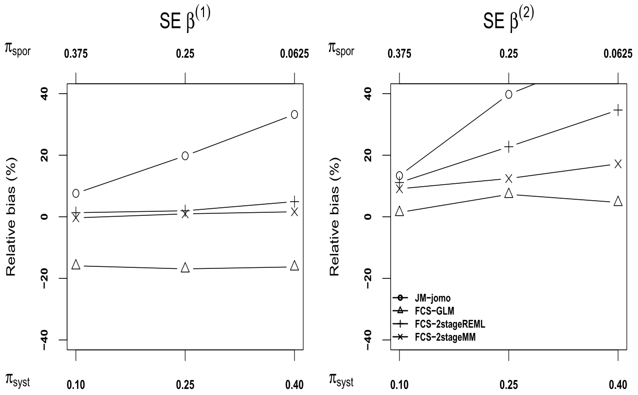

FIG. 1. *Robustness to the proportion of systematically missing values: estimate of the relative bias for the SE estimate for $\widehat{\beta^{(1)}}$ (left), $\widehat{\beta^{(2)}}$ (right) according to π_{syst} for each MI method. The estimated relative bias is calculated by the difference between the model SE and the empirical SE, divided by the empirical SE. The proportion of sporadically missing values is accordingly modified to keep a constant proportion of missing values (in expectation).*

3.2.4 *Robustness to the cluster size.* To explore the robustness of the inferences provided by the MI methods with respect to the size of the clusters, we extend the simulation study to generate clusters with equal sizes varying in $\{15, 25, 50, 100, 200, 400\}$. Relative biases are reported in Figure 3.

The biases obtained by the FCS-2stage methods are large for small clusters, but decrease when the cluster size increases. This behaviour was expected since the posterior distribution for the conditional imputation model's parameters are based on asymptotic properties (cf. Section 2.2). For the JM-jomo method, the bias mainly depends on the sample size for $\beta^{(1)}$ only. On the contrary, the FCS-GLM method is fairly stable for $\beta^{(1)}$ and $\beta^{(2)}$ across the cluster sizes. Note that the bias on ψ_{11} observed in the base-case configuration disappears with small clusters as well as the undercoverage issue because of a better estimate of the standard error (see Figure 6 in Appendix B.2.3) making the method relevant in such a case.

3.2.5 *Robustness to the type of imputed variables.* The results in Table 4 demonstrate that the type of imputed variable also affects the performance of the imputation methods. In general, we found that JM-jomo

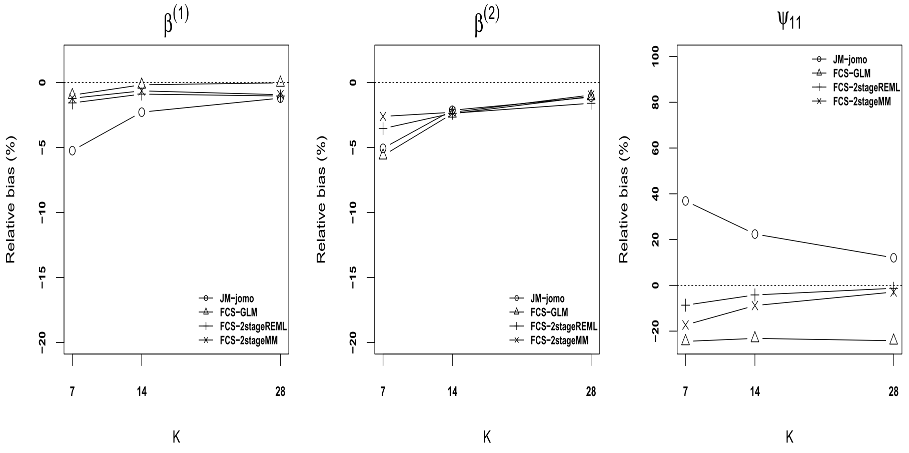

FIG. 2. *Robustness to the number of clusters*: *relative bias for the estimate of* $\beta^{(1)}$ (*left*), $\beta^{(2)}$ (*middle*) *and* ψ_{11} (*right*) *according to K for each MI method*.

provides smaller bias for $\beta^{(1)}$, $\beta^{(2)}$ and ψ_{11} as compared to FCS-GLM and FCS-2stage. This is likely related to the fact that FCS-GLM is not tailored for imputing sporadically missing data in binary variables, and that the two-stage estimator used in FCS-2stage is known to be biased in the presence of small clusters.

3.2.6 *Robustness to the variance of random effects*. Table 5 provides inference results when the covariance matrix of the random effects is multiplied by a factor 2. The biases reported for the variance of the random effects are less than 2% for JM-jomo, while they reached 11% in the base-case configuration. Such behaviour can be explained by the smaller influence of the prior distribution for random effects when the effect of random effects is stronger. On the contrary, the bias increases for the FCS-2stage methods.

3.2.7 *Other configurations*. Other configurations that have been investigated are presented in Appendix B.1.1. These configurations consider the nature of the outcome (configuration 5), the missing data mechanism for sporadically missing values (configurations 6, 7, 8), the complexity of the analysis model (configuration 9), the number of individuals with unequal cluster sizes (configuration 11, 12), the intra-class correlation (configurations 13, 14), the correlation between random intercepts generating variables x_1, x_2, x_3 (configuration 15), the correlation between continuous variables in each cluster (configuration 16), the covariance matrix of the random effects (configuration 17, 18), and the use of a probit link for generating binary covariates (configuration 19). Figure 7 in Appendix B.2.4 reports the distribution of the relative bias over all

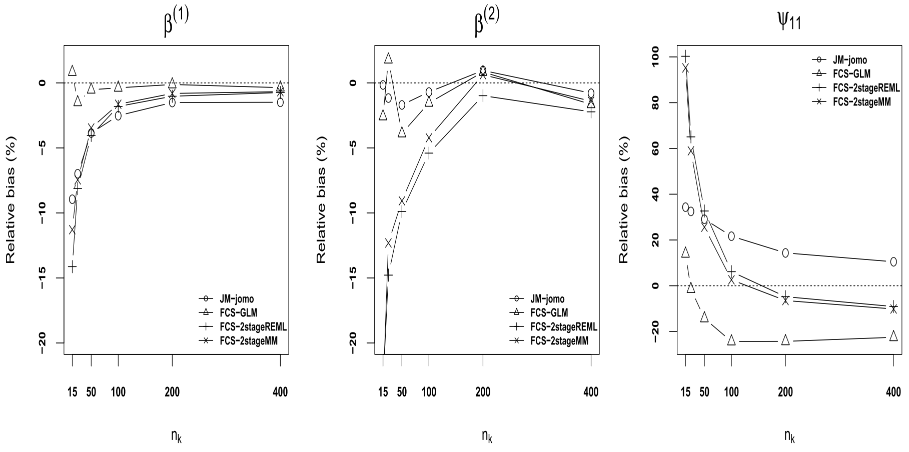

FIG. 3. *Robustness to the cluster size*: *relative bias for the estimate of* $\beta^{(1)}$ (*left*), $\beta^{(2)}$ (*middle*) *and* ψ_{11} (*right*) *according to* n_k *for each MI method. Criteria are based on* 500 *incomplete datasets*.

TABLE 4

Binary covariates. Point estimate, relative bias, model standard error, empirical standard error, 95% *coverage and RMSE for analysis model parameters and for several methods (Full data, JM-jomo, FCS-GLM, FCS-2stage with REML estimator, FCS-2stage with moment estimator). Criteria are based on* 500 *incomplete datasets. Average time to multiply impute one dataset is indicated in minutes. True values are* $\beta^{(1)} = -0.11$, $\beta^{(2)} = 0.03$, $\sqrt{\psi_{00}} = 0.088$, $\sqrt{\psi_{11}} = 0.02$

	Full	**JM-jomo**	**FCS-GLM**	**FCS-2stageREML**	**FCS-2stageMM**
$\beta^{(1)}$ est	−0.1098	−0.1091	−0.1081	−0.1080	−0.1083
$\beta^{(1)}$ rbias (%)	−0.2	−0.8	−1.7	−1.8	−1.5
$\beta^{(1)}$ model se	0.0050	0.0074	0.0057	0.0063	0.0056
$\beta^{(1)}$ emp se	0.0049	0.0064	0.0059	0.0060	0.0061
$\beta^{(1)}$ 95% cover	95.0	97.0	92.0	94.0	90.4
$\beta^{(1)}$ rmse	0.0049	0.0064	0.0062	0.0063	0.0063
$\beta^{(2)}$ est	0.0303	0.0298	0.0295	0.0294	0.0295
$\beta^{(2)}$ rbias (%)	1.0	−0.6	−1.8	−2.1	−1.6
$\beta^{(2)}$ model se	0.0029	0.0072	0.0044	0.0051	0.0045
$\beta^{(2)}$ emp se	0.0028	0.0047	0.0043	0.0044	0.0043
$\beta^{(2)}$ 95% cover	95.0	98.6	95.2	96.2	95.0
$\beta^{(2)}$ rmse	0.0028	0.0047	0.0044	0.0044	0.0044
$\sqrt{\psi_{00}}$ est	0.0876	0.0861	0.0807	0.0834	0.0827
$\sqrt{\psi_{00}}$ rbias (%)	−0.2	−1.9	−8.0	−4.9	−5.7
$\sqrt{\psi_{00}}$ rmse	0.0123	0.0122	0.0137	0.0129	0.0131
$\sqrt{\psi_{11}}$ est	0.0198	0.0232	0.0151	0.0185	0.0164
$\sqrt{\psi_{11}}$ rbias (%)	−0.8	16.2	−24.3	−7.7	−17.8
$\sqrt{\psi_{11}}$ rmse	0.0046	0.0053	0.0069	0.0057	0.0068
Time		5.6	95.1	1.4	0.6

configurations, while tables gathering inference results are available in the Supplementary Materials (Audigier et al., 2018). We found similar results as compared to the base-case configuration.

4. APPLICATION TO GREAT DATA

The MI methods are applied to the GREAT data (Appendix A). Model (13) is the analysis model, with x_1 representing the variable BNP, x_2 the variable AFIB and y the variable LVEF. Although only three variables are included in the analysis model, the imputation models are based on nine variables to render the MAR assumption more credible (Enders, 2010, Schafer, 1997).

The results in Table 6 indicate that the missing data mechanism in the GREAT application is not likely to be missing completely at random. In particular, complete-case analysis yielded estimates for the fixed coefficient β_{BNP} farther away from null as compared to the MI methods.

As expected, standard errors obtained by CC were larger than those obtained from the base-case configuration of the simulation study. In general, standard errors for fixed effects estimates were smaller for the MI methods as compared to CC. An exception occurred for the variable β_{BNP} when using FCS-2stageMM or FCS-GLM. Possibly, this is related to convergence issues resulting from the limited number of iterations and relatively small number of imputed data sets. For instance, when we allowed for 50 iterations (rather than 10) and generated 50 imputed data sets (instead of 20), FCS-2stageMM yielded a standard error of 0.0091 for β_{BNP}.

Remarkably, JM-jomo yielded the smallest standard errors for fixed effect parameters. This situation did not arise in the simulation studies, and is likely related to over-parametrisation of the FCS methods. In particular, the conditional imputation models assume random effects for all covariates, which leads to a substantial increase of the number of parameters as the number of covariates increases, hence inflating the variance around imputed values.

Finally, in agreement with the simulation studies, the FCS-GLM method requires substantially more computation time than the other MI methods. This limitation

TABLE 5

Higher variance for $\boldsymbol{\Psi}$. *Point estimate, relative bias, model standard error, empirical standard error,* 95% *coverage and RMSE for analysis model parameters and for several methods (Full data, JM-jomo , FCS-GLM, FCS-*2*stage with REML estimator, FCS-*2*stage with moment estimator). Criteria are based on* 500 *incomplete datasets. Average time to multiply impute one dataset is indicated in minutes. Criteria related to the continuous (resp. binary) covariate are in light (resp. dark) grey. True values are* $\beta^{(1)} = -0.11$, $\beta^{(2)} = 0.03$, $\sqrt{\psi_{00}} = 0.124$, $\sqrt{\psi_{11}} = 0.028$

	Full	JM-jomo	FCS-GLM	FCS-2stageREML	FCS-2stageMM
$\beta^{(1)}$ est	−0.1102	−0.1087	−0.1099	−0.1089	−0.1090
$\beta^{(1)}$ rbias (%)	0.2	−1.1	−0.1	−1.0	−0.9
$\beta^{(1)}$ model se	0.0061	0.0075	0.0059	0.0070	0.0070
$\beta^{(1)}$ emp se	0.0063	0.0071	0.0072	0.0073	0.0073
$\beta^{(1)}$ 95% cover	93.8	95.0	90.1	93.0	93.8
$\beta^{(1)}$ rmse	0.0063	0.0073	0.0072	0.0073	0.0073
$\beta^{(2)}$ est	0.0301	0.0297	0.0295	0.0294	0.0296
$\beta^{(2)}$ rbias (%)	0.2	−0.8	−1.6	−2.0	−1.3
$\beta^{(2)}$ model se	0.0029	0.0070	0.0046	0.0055	0.0050
$\beta^{(2)}$ emp se	0.0030	0.0048	0.0042	0.0044	0.0044
$\beta^{(2)}$ 95% cover	94.2	98.2	94.2	97.4	96.0
$\beta^{(2)}$ rmse	0.0030	0.0048	0.0043	0.0044	0.0044
$\sqrt{\psi_0}$ est	0.1220	0.1225	0.1089	0.1172	0.1166
$\sqrt{\psi_0}$ rbias (%)	−1.7	−1.3	−12.2	−5.6	−6.0
$\sqrt{\psi_0}$ rmse	0.0198	0.0175	0.0240	0.0203	0.0205
$\sqrt{\psi_1}$ est	0.0275	0.0279	0.0220	0.0262	0.0260
$\sqrt{\psi_1}$ rbias (%)	−2.8	−1.5	−22.1	−7.3	−8.2
$\sqrt{\psi_1}$ rmse	0.0048	0.0043	0.0083	0.0057	0.0059
Time		7.7	102.5	2.2	0.9

is somewhat problematic, as the number of incomplete variables was rather limited in the GREAT application. As a result, checking convergence of the distribution of missing values to their posterior distribution becomes very difficult.

5. DISCUSSION

As international collaboration becomes more common and access to large shared datasets increases, researchers increasingly face incomplete multilevel data.

TABLE 6

*GREAT data: Point estimate and model standard error for the parameters of a linear mixed effects model for several methods (Complete-case analysis, JM-jomo, FCS-GLM, FCS-*2*stage with REML estimator, FCS-*2*stage with moment estimator).* 20 *imputed data sets are used for MI methods.* 10 *iterations are used for FCS methods. Time to multiply impute the dataset is indicated in minutes. Criteria related to the continuous (resp. binary) covariate are in light (resp. dark) grey*

		CC	JM-jomo	FCS-GLM	FCS-2stageREML	FCS-2stageMM
β_{BNP}	est	−0.1132	−0.0891	−0.1002	−0.0854	−0.1009
	model se	0.0108	0.0078	0.0163	0.0099	0.0112
β_{AFIB}	est	0.0268	0.0216	0.0218	0.0215	0.0273
	model se	0.0071	0.0046	0.0066	0.0040	0.0045
ψ_{00}	est	0.1112	0.1075	0.1232	0.1220	0.1189
ψ_{BNP}	est	0.0290	0.0306	0.0348	0.0351	0.0332
Time (min)			94.0	30,819.5	361.3	31.8

Thus, handling systematically missing values becomes inevitable (Debray et al., 2015a, 2015b). In this work, we compared three recent multiple imputation methods for addressing this issue: JM-jomo, FCS-GLM and FCS-2stage. We also considered several extensions to better handle continuous and binary data in the presence of sporadically and systematically missing values. Such extensions are available in the R packages mice (van Buuren and Groothuis-Oudshoorn, 2011) and micemd (Audigier and Resche-Rigon, 2017). We highlighted the relevance of using these methods, and demonstrated their superiority over ad-hoc strategies through extensive simulation studies. Although the differences between the three imputation models are mainly technical, their properties may substantially differ according to the considered dataset.

In general, we found that JM-jomo tends to be conservative. This behaviour is in line with simulation study presented in Quartagno and Carpenter (2016a), and is related to the use of inverse-Wishart prior distributions for modelling the covariance matrices. Although this distribution avoids convergence issues of the Gibbs sampler (Schafer and Yucel, 2002), its use is not necessarily appropriate for the setting in hand. Furthermore, the prior distributions for the parameters of the inverse Wishart distribution appear to be rather influential. In particular, by sampling the covariance matrices using few degrees of freedom, too much variability is introduced for the within-cluster covariance matrices. As a result, fixed effects estimates vary too much across imputed datasets, leading to over-estimation of the variance components. Note that the influence of the prior distributions of the imputation model parameters have also been recently demonstrated in the context of continuous data (Kunkel and Kaizar, 2017).

Bias is observed for JM-jomo when the number of individuals and/or clusters is small and/or variance of random effects is small. In such situations, the Inverse-Wishart prior distributions become very informative (Gelman, 2006). Furthermore, because the Inverse-Wishart distribution tends to generate too much variability, its use may lead to shrinkage of regression coefficients when imputing continuous covariates ($\beta^{(1)}$ in the simulation study). This issue is less problematic for binary covariates ($\beta^{(2)}$) because the diagonal terms of the within covariance matrices are constrained to be equal to one. However, inference for GLMM models with few clusters is challenging, even without missing data (McNeish and Stapleton, 2016). When the number of clusters is large, substantial improvements can be obtained by increasing the degrees of freedom of the chi-squared distribution.

For FCS-GLM, we found that imputations were quite accurate for continuous variables. However, FCS-GLM is limited by the homoscedasticity assumption (van Buuren, 2011, Resche-Rigon and White, 2016). In particular, by fitting a homoscedastic model to heteroscedastic data, standard errors tend to be underestimated, even in the absence of missing data. As a result, the FCS-GLM method cannot fully propagate the sampling variability, leading to an underestimation of the variance of the parameters of the analysis model. This results in confidence intervals that are too narrow. However, the homoscedastic assumption becomes an advantage with small clusters since it avoids overfitting issues: as a result, the standard errors become well estimated for continuous covariates. For this reason, FCS-GLM is an appropriate method to use with small clusters. Another current problem of FCS-GLM is the time required for generating imputed datasets, particularly in large datasets. These results are in line with the simulation study of Resche-Rigon and White (2016) comparing FCS-GLM and FCS-2stage.

FCS-2stage does not present any recurrent trend. Simulation study results suggest that using the lognormal distribution as an approximation for the posterior distribution of the error variance outperformed modelling through Inverse-Wishart distributions (as in JM-jomo and FCS-GLM). Although the MM estimator of FCS-2stage is known to underestimate relevant variance components, similar inferences were obtained using REML (Langan, Higgins and Simmonds, 2017). However, FCS-2stage may be problematic when imputing binary covariates in small clusters, as the maximum likelihood estimator used in stage one is known to yield biased estimates in such circumstances. For this reason, we investigated the use of Firth's correction (Firth, 1993), but this did not yield substantial improvements. Further research is warranted to investigate how FCS-2stage may be improved when applied to datasets with small clusters. In the GREAT application, we found that JM-jomo and FCS-2stage produced similar point estimates, but that the former yielded smaller standard errors. This may reflect the ability of JM-jomo to borrow information about the study-specific covariance matrix across studies, and/or the appropriateness of using the Inverse Wishart model in the GREAT data.

A key issue in all MI methods is the use of *congenial* imputation models (Meng, 1994). Congeniality means that there is a joint model which implies the imputation model and the analysis model as submodels. Some results have been obtained for continuous variables when the analysis model does not include a random slope (Quartagno and Carpenter, 2016a, Resche-Rigon and White, 2016). However, as raised in Grund, Lüdtke and Robitzsch (2016), with a random slope, these imputation models are uncongenial. Indeed, considering model (1), the outcome depends on a product of two random variables: the random effect (b_k) and the associated covariate ($\mathbf{W}_k$). Consequently, the marginal distribution of the outcome becomes highly complex, whereas a joint imputation model like the one used in JM-jomo [without covariates in the right-hand side of model (2)] assumes simpler Gaussian marginal distributions in each cluster. In the same way, for FCS methods, the conditional distribution of one covariate is no longer analytically tractable. This implies that the distribution of the covariates given the outcome cannot be written as a GLMM model. Thus, imputation models are misspecified whatever the imputation method used. Nevertheless, our simulation study shows that this is a minor practical issue since the biases remain very small for fixed coefficients and variance of random effects (see also Appendix B.2.5). Recent development of imputation models ensuring congeniality even in complex settings (Bartlett et al., 2015) seems promising for the multilevel setting.

Another source of misspecification is the choice of random and fixed effects in the imputation models. In particular, selecting each conditional imputation model is tricky in practice for FCS approaches, particularly with a lot of variables, but is needed to avoid overparametrisation. More generally, finding conditional imputation models with few parameters in FCS approaches is a current topic of research (Zhao and Long, 2016) in the one-level case, and appears even more challenging in the two-level case. In this paper, we used the default model for each method: for JM-jomo, all variables are in the response part of model (2), which corresponds to normal marginal distributions with random intercept, whereas FCS approaches include all covariates in fixed and random effects, making such marginal distribution more complex. These differences between the imputation models could explain some differences between JM and FCS approaches.

An additional difficulty for all MI methods is the imputation of binary variables. JM-jomo overcomes this quite well by considering a fully Bayesian multilevel modelling approach with a probit link function, but FCS-GLM and FCS-2stage are less tailored for such variables: FCS-GLM because it uses a logit link making it difficult to handle sporadically missing values, and FCS-2stage because it draws inferences separately on each cluster, making samples too small to provide accurate inferences. For these reasons, both FCS methods could be improved: by adopting a probit link function for FCS-GLM, and by applying bias correction for variances and point estimates for FCS-2stage. Note that in contrast to FCS-2stage, FCS-GLM and JM-jomo can also handle nominal and count variables. FCS-2stage could be further extended by considering additional link functions in the regression models used at stage 1. Finally, although FCS-2lnorm provides encouraging performance for imputing missing continuous and binary multilevel data, it does not *properly* reflect (Schafer, 1997, page 105) the variability of random coefficients. For this reason, its usefulness remains limited in the presence of systematically missing values.

Although we only considered a 2-level setting in this study, extensions of the presented MI methods to a higher hierarchical structure are relatively straightforward. At this moment, only JM-jomo can handle such situations. Note that missing values may also occur at level-2. JM-jomo naturally handles this setting, but suitable FCS approaches have also been developed (van Buuren and Groothuis-Oudshoorn, 2011). In addition, we did not focus on longitudinal data, or more generally on data with very few observations per cluster, as often found in educational research. However, systematically missing values are also frequent in such cases and raise additional overfitting issues of the imputation models.

Our study focuses on the use of GLMM models to analyse multilevel data, which facilitates the use of MI. Direct maximum likelihood inference is another possible strategy to address missing data, but its implementation becomes difficult when dealing with multilevel variables (Longford, 2008, Schafer, 1997). However, many other statistics than the parameters of a GLMM model can be of interest. For instance, Curran and Hussong, Curran et al. (2009, 2008) proposed using item response theory to fit measurement models.

From a general point of view, whatever the imputation method used, accurate inferences for a GLMM model can be expected only with a high (or moderate) number of clusters. Heteroscedastic MI methods perform better than homoscedastic methods, which should be reserved with few individuals only. Methods based on conjugate prior distributions should be used with caution when the proportion of missing values is very high. Multiple imputation of binary variable is challenging, and all methods can have drawbacks in this case. Specifically, JM-jomo could be recommended when the number of incomplete binary variables is large and when the number of observed clusters is large. FCS-2stage performs quite well, but should be avoided when clusters are small, or equivalently, when the proportion of sporadically missing values is large. This method is particularly relevant compared to the others when the number of clusters with systematically missing variables is large. The MM version offers a quick solution to give an initial overview of the inference results. Finally, FCS-GLM appears advantageous when clusters are small.

We believe that the topic of inference for multilevel incomplete data needs strengthened theoretical underpinnings to improve the fit of imputation models, as well as some developments to broaden the scope of the evaluated methods. Among these, congeniality has been recently discussed, but need more attention for analysis models with random slopes. Machine learning methods offering more flexibility could be considered for this purpose. In addition, solutions to handle missing data without assuming the ignorability of the missing data mechanism need to be investigated. Furthermore, in the big data era, MI methods handling a large number of variables need to be studied. Finally, proposing imputation models for nominal or ordered variables avoiding informative prior distributions is an important line of research.

APPENDIX A: DESCRIPTION OF GREAT DATA

The GREAT Network performed an IPD meta-analysis to explore risk factors associated with short-term mortality in acute heart failure (AHF) (Global Research on Acute conditions Team (GREAT) Network, 2013). Their dataset consists of 28 studies: 8 were carried out in Western Europe (2 in Italy, 2 in Spain, and 1 in each of France, Finland, Switzerland, Netherlands), 13 in Central Europe (12 in Czech Republic and 1 in Austria), 3 in America (2 in the United States and 1 in Argentina), 3 in Asia (China, Japan, Korea), and 1 in Africa (Tunisia) (Mebazaa et al., 2013). The principal investigators of each study provided the original data collected for each patient, including a list of patient characteristics and potential risk factors (Lassus et al., 2013).

One biomarker of interest was brain natriuretic peptide (BNP), which is known to be elevated in acute heart failure. Since measuring the left ventricular ejection fraction (LVEF) requires an ultrasound examination, one objective is to explain LVEF from biomarkers that are easier to measure, such as BNP or electrocardiographic characteristics such as the atrial fibrillation (AFIB). The generalized linear mixed effects model (GLMM) (Pinheiro and Bates, 2000, Lee, Nelder and Pawitan, 2006) is a suitable statistical model to achieve this goal.

The dataset contains two binary variables (AFIB and Gender) and 7 continuous variables [BMI, Age, Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Heart rate (HR), LVEF and BNP]. Variables are described in Table 8 in Appendix A. The total number of individuals is 11,685 and study sizes range from 18 to 1834.

Each study is incomplete, leading to sporadically missing values on all variables except gender and LVEF. However, BNP measurement is a recent technique, so this variable has been collected on 10 studies only, leading to systematically missing values. Four other variables are systematically missing on some studies (Table 7), notably the binary variable AFIB.

TABLE 7
GREAT data: percentages of missing values by variable and study

k	n_k	Gender	BMI	Age	SBP	DBP	HR	BNP	AFIB	LVEF
1	410	0	36	<1	1	2	3	57	<1	0
2	567	0	19	0	2	3	1	10	0	0
3	210	0	43	0	1	2	1	0	100	0
4	375	0	2	0	1	1	2	4	42	0
5	107	0	1	0	0	0	0	100	0	0
6	267	0	100	0	100	100	<1	100	0	0
7	203	0	<1	0	1	2	1	<1	0	0
8	354	0	44	1	16	16	19	12	22	0
9	137	0	1	0	0	0	0	0	0	0
10	48	0	100	0	0	0	4	100	0	0
11	208	0	24	0	0	<1	0	100	0	0
12	622	0	27	0	<1	<1	1	100	0	0
13	78	0	60	0	0	0	0	100	100	0
14	670	0	77	<1	1	1	2	100	<1	0
15	1000	0	13	0	2	2	2	82	<1	0
16	1093	0	0	0	0	0	0	100	0	0
17	18	0	6	0	0	0	0	22	0	0
18	1834	0	19	0	1	1	<1	92	<1	0
19	358	0	7	0	0	0	0	99	0	0
20	54	0	6	0	2	2	2	100	2	0
21	588	0	10	0	<1	<1	0	97	<1	0
22	651	0	24	0	2	2	2	73	2	0
23	455	0	2	0	0	0	<1	86	<1	0
24	294	0	4	0	<1	<1	<1	81	0	0
25	397	0	1	0	0	0	0	100	0	0
26	295	0	11	0	0	0	0	66	0	0
27	303	0	11	0	<1	<1	0	79	0	0
28	89	0	0	0	0	0	0	38	0	0

TABLE 8
GREAT data: description of variables. Binary variables are presented by counts and percentages, while continuous variables by their median and quartiles

Variable	Value	Size	Summary
Gender	0	6865	58.65%
	1	4820	42.35%
AFIB	0	7704	69.18%
	1	3431	30.81%
BMI		9259	26.58 [23.66143862; 30.12]
Age		11,678	72.7 [62.7; 80]
SBP		11,278	130 [111; 153]
DBP		11,262	80 [68; 90]
HR		11,518	87 [72; 105]
LVEF		11,685	0.38 [0.27; 0.5]
BNP		2776	2.99 [2.66; 3.29]

APPENDIX B: SIMULATION

B.1 Simulation Design

B.1.1 *Investigated configurations.* Table 9 summarizes all configurations that have been investigated in the simulation study.

B.2 Complementary Results

B.2.1 *Robustness to the proportion of systematically missing values.* Figure 4 reports the influence of the proportion of systematically missing values in terms of bias on point estimates.

B.2.2 *Robustness to the number of clusters.* Figure 5 reports the influence of the number of clusters in terms of bias on standard error estimates.

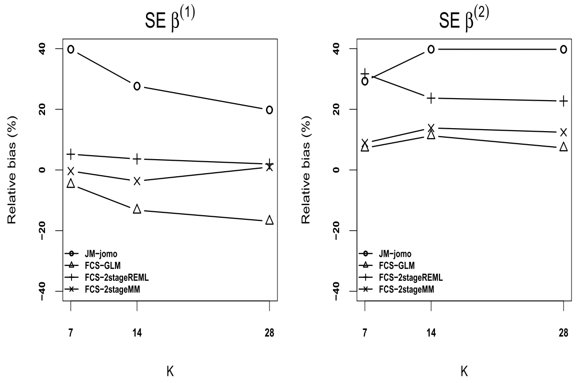

FIG. 5. *Robustness to the number of clusters: estimate of the relative bias for the SE estimate for $\widehat{\beta^{(1)}}$ (left), $\widehat{\beta^{(2)}}$ (right) according to K for each MI method. The estimated relative bias is calculated by the difference between the model SE and the empirical SE, divided by the empirical SE. Criteria are based on* 500 *incomplete datasets.*

B.2.3 *Robustness to the cluster size.* Figure 6 reports the influence of the cluster size in terms of bias on standard error estimates.

B.2.4 *Other configurations.* Figure 7 reports the distributions of the biases over all configurations.

B.2.5 *Influence of the random slope.* The multivariate version of model (1) used in JM-jomo requires that all missing variables are in the left-hand side of the model (2). However, if the analysis model includes a random slope in the right-hand side [like in the simulation study (Section 3.1)], then the imputation model is misspecified. To assess the influence of this misspecification on the random effect variance, we compare the estimates of ψ_{00} when the outcome of the model is generated according to an analysis model including a

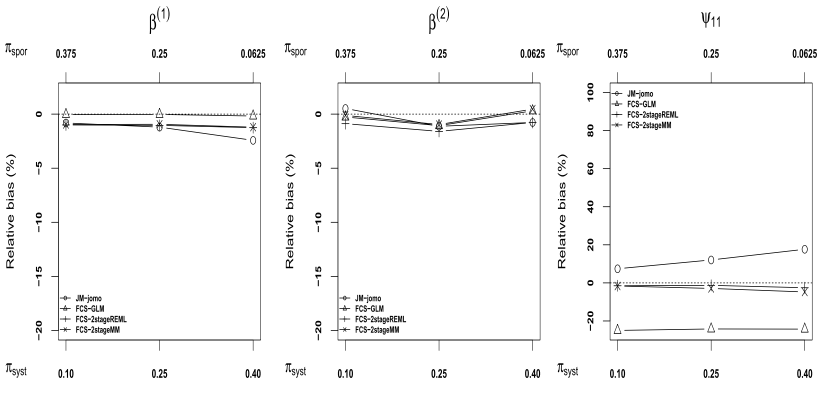

FIG. 4. *Robustness to the proportion of systematically missing values: relative bias for the estimate of $\beta^{(1)}$ (left), $\beta^{(2)}$ (middle) and ψ_{11} (right) according to π_{syst} for each MI method. The proportion of sporadically missing values is modified to keep a constant proportion of missing values (in expectation).*

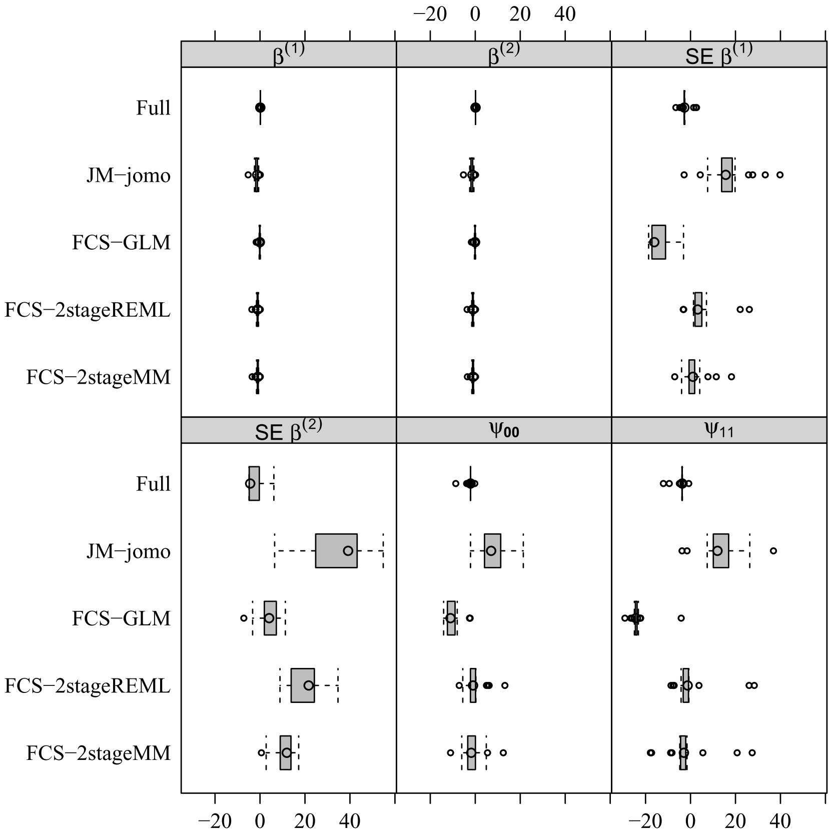

FIG. 7. *Distribution of the relative bias over the* 20 *configurations for several methods* (*Full*, *CC*, *JM-jomo*, *FCS-GLM*, *FCS-*2*stageREML*, *FCS-*2*stageMM*) *and several parameters of interest* ($\beta^{(1)}$, $\beta^{(2)}$, ψ_{00}, ψ_{11}, *SE* $\beta^{(1)}$, *SE* $\beta^{(2)}$). *One point represents the relative bias observed for one configuration.*

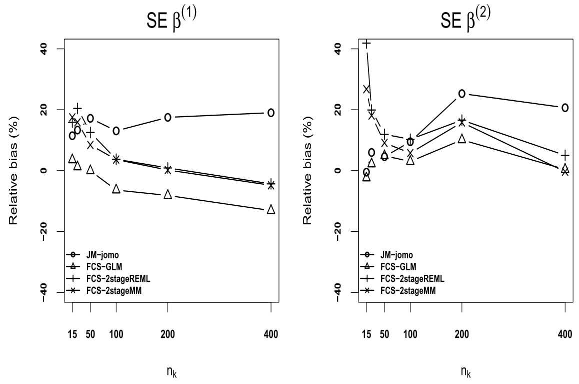

FIG. 6. *Robusteness to the cluster size*: *estimate of the relative bias for the SE estimate for* $\widehat{\beta^{(1)}}$ (*left*), $\widehat{\beta^{(2)}}$ (*right*) *according to* n_k *for each MI method. The estimated relative bias is calculated by the difference between the model SE and the empirical SE*, *divided by the empirical SE. Criteria are based on* 500 *incomplete datasets.*

random slope (base-case configuration), with the estimates of ψ_{00} when the outcome of the model is generated according to an analysis model with a random intercept only. Estimates are reported in Figure 8.

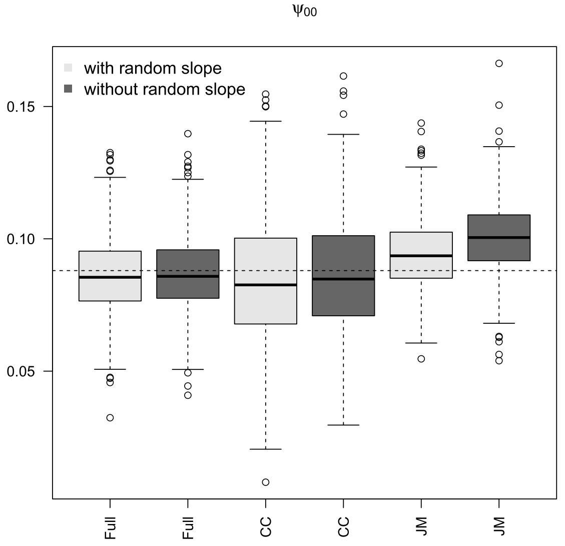

FIG. 8. *Distribution of the estimate of* $\sqrt{\psi_{00}}$ *over the* 500 *generated datasets for Full*, *CC and JM-jomo methods. Both configurations are considered*: *a one with a random slope* (*in grey*, *corresponding to the base-case configuration*) *and one without random slope* (*in blue*). *The red dashed line represents the true value of* $\sqrt{\psi_{00}}$.

TABLE 9

Tuning parameters for the several configurations: K the number of clusters, n the number of individuals, λ a scalar multiplying the variance of random effects, ρ_v the correlation between random intercepts generating variables x_1, x_2, x_3, ρ_b the correlation between random coefficients of the analysis model, the type of the outcome, the type of the covariates (x_1 or x_2), the proportion of systematically missing values on the covariates (x_1 or x_2), the proportion of sporadically missing values on the covariates (x_1 or x_2), the nature of the missing data mechanism R, the intra cluster correlation for continuous variables (ICC), ρ_{x_1,x_3} the correlation between x_1 and x_3 in each cluster, the presence of a random effect on the covariate x_2, the link function used to generate the binary covariate. Parameters varying from the base-case configuration are in boldface

						y	x_1						x_2					
Case	K	n	λ	ρ_v	ρ_b	**Type**	**Type**	π_{syst}	π_{spor}	R	**ICC**	$\rho(x_1, x_3)$	**Type**	π_{syst}	π_{spor}	R	**Random effect on x_2**	**Link function**
1 (base-case)	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
2	28	11,685	1	0.07	−0.87	cont	cont	**0.1**	**0.375**	MCAR	0.25	0.3	bin	**0.1**	**0.375**	MCAR	no	logistic
3	28	11,685	1	0.07	−0.87	cont	cont	**0.4**	**0.0625**	MCAR	0.25	0.3	bin	**0.4**	**0.0625**	MCAR	no	logistic
4	28	11,685	1	0.07	−0.87	cont	**bin**	0.25	0.25	MCAR			bin	0.25	0.25	MCAR	no	logistic
5	28	11,685	1	0.07	−0.87	**bin**	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
6	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	**0**	**0**		no	logistic
7	28	11,685	1	0.07	−0.87	cont	cont	**0**	**0**		0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
8	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	**MAR**	0.25	0.3	bin	0.25	0.25	**MAR**	no	logistic
9	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	**yes**	logistic
10	**14**	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
11	28	**5845**	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
12	28	**2923**	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
13	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	**0.5**	0.3	bin	0.25	0.25	MCAR	no	logistic
14	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	**0.1**	0.3	bin	0.25	0.25	MCAR	no	logistic
15	28	11,685	1	**0.3**	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
16	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	**0.5**	bin	0.25	0.25	MCAR	no	logistic
17	28	11,685	**2**	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
18	28	11,685	1	0.07	**−0.3**	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic
19	28	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	**probit**
20	**7**	11,685	1	0.07	−0.87	cont	cont	0.25	0.25	MCAR	0.25	0.3	bin	0.25	0.25	MCAR	no	logistic

A bias is observed even if the outcome is generated from a model with no random slope, indicating that misspecification of the imputation model is not the main reason for the observed bias in the base-case configuration. As shown in Section 3.2.6 it is more likely that the use of wrongly informative prior distributions biases the inference in presence of very small values for the level-2 variances.

ACKNOWLEDGEMENTS

We thank the Global Research on Acute Conditions Team (Global Research on Acute conditions Team (GREAT) Network, 2013) for allowing the use of their data. TPAD was supported by the Netherlands Organization for Scientific Research (91617050 and 91215058). IRW, MQ and JC were funded by the Medical Research Council (Unit Programmes U105260558 (IRW only) and MC_UU_12023/21). VA and MRR was supported by the Agence nationale de sécurité du médicament et des produits de santé (ANSM).

SUPPLEMENTARY MATERIAL

Supplement to "Multiple Imputation for Multilevel Data with Continuous and Binary Variables" (DOI: 10.1214/18-STS646SUPPA; .pdf). Technical details on the posterior distributions of imputation model parameters and inference results for all configurations that have not been discussed in detail in the main text.

Supplement to "Multiple Imputation for Multilevel Data with Continuous and Binary Variables" (DOI: 10.1214/18-STS646SUPPB; .zip). R code for the simulation study.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319

ALLISON, P. (2002). *Missing Data*. Sage, Thousand Oaks, CA.

ANDRIDGE, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom. J.* **53** 57–74. MR2767378

ASPAROUHOV, T. and MUTHÉN, B. (2010). Multiple imputation with Mplus. Technical report. Available at http://www.statmodel.com/download/Imputations7.pdf.

AUDIGIER, V. and RESCHE-RIGON, M. (2017). micemd: Multiple imputation by chained equations with multilevel data. R package version 1.2.0.

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T. P. A., QUARTAGNO, M., CARPENTER, J., VAN BUUREN, S. and RESCHE-RIGON, M. (2018). Supplement to "Multiple imputation for multilevel data with continuous and binary variables." DOI:10.1214/18-STS646SUPPA, DOI:10.1214/18-STS646SUPPB.

BARTLETT, J. W., SEAMAN, S. R., WHITE, I. R. and CARPENTER, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat. Methods Med. Res.* **24** 462–487. MR3372102

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.

BLOSSFELD, H.-P., GÜNTHER ROSSBACH, H. and VON MAURICE, J., eds. (2011). *Education as a Lifelong Process*: *The German National Educational Panel Study* (*NEPS*). VS Verlag für Sozialwissenschaften, Wiesbaden, Germany.

BOS, W., LANKES, E.-M., PRENZEL, M., SCHWIPPERT, K. and VALTIN, R., eds. (2003). *Erste Ergebnisse aus IGLU*: *Schülerleistungen Am Ende der Vierten Jahrgangsstufe Im Internationalen Vergleich* [*the First*]. Waxmann, Münster, Germany.

CARPENTER, J. and KENWARD, M. (2013). *Multiple Imputation and Its Application*, 1st ed. Wiley, New York.

CARRIG, M. M., MANRIQUE-VALLIER, D., RANBY, K. W., REITER, J. and HOYLE, R. H. (2015). A nonparametric, multiple imputation-based method for the retrospective integration of data sets. *Multivar. Behav. Res.* **50** 383–397.

CURRAN, P. J. and HUSSONG, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychol. Methods* **14** 81–100.

CURRAN, P. J., HUSSONG, A. M., CAI, L., HUANG, W., CHASSIN, L., SHER, K. J. and ZUCKER, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Dev. Psychol.* **44** 365–380.

DEBRAY, T., RILEY, R., ROVERS, M., REITSMA, J., MOONS, K. and ON BEHALF OF THE COCHRANE IPD META-ANALYSIS METHODS GROUP (2015b). Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: Guidance on their use. *PLoS Med.* **12** e1001886.

DEBRAY, T., MOONS, K., VAN VALKENHOEF, G., EFTHIMIOU, O., HUMMEL, N., GROENWOLD, R. and REITSMA, J. O. (2015a). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res. Synth. Methods* **6** 293–309.

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188.

DRECHSLER, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *J. Educ. Behav. Stat.* **40** 69–95.

ENDERS, C. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

ENDERS, C. K., KELLER, B. T. and LEVY, R. (2017). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychol. Methods*.

ENDERS, C., MISTLER, S. and KELLER, B. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods* **21** 222–240.

ERLER, N. S., RIZOPOULOS, D., VAN ROSMALEN, J., JADDOE, V. W. V., FRANCO, O. H. and LESAFFRE, E. M. E. H.

(2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Stat. Med.* **35** 2955–2974. MR3528236

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

Global Research on Acute conditions Team (GREAT) Network (2013). Managing acute heart failure in the ED—case studies from the acute heart failure academy. Available at http://www.greatnetwork.org.

Goldstein, H., Bonnet, G. and Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *J. Educ. Behav. Stat.* **32** 252–286.

Goldstein, H., Carpenter, J., Kenward, M. G. and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Stat. Model.* **9** 173–197. MR2756416

Graham, J. W. (2012). *Missing Data*: *Analysis and Design*. Springer, New York. MR2952499

Grund, S., Lüdtke, O. and Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behav. Res. Methods* **48** 640–649.

Hughes, R. A., White, I. R., Seaman, S., Carpenter, J., Tilling, K. and Sterne, J. (2014). Joint modelling rationale for chained equations. *BMC Med. Res. Methodol.* **14** 28.

Jackson, D., White, I. R. and Riley, R. D. (2013). A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. *Biom. J.* **55** 231–245. MR3045843

Jolani, S. (2018). Hierarchical imputation of systematically and sporadically missing data: An approximate Bayesian approach using chained equations. *Biom. J.* **60** 333–351.

Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S. and Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. *Stat. Med.* **34** 1841–1863. MR3334696

Kropko, J., Goodrich, B., Gelman, A. and Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Polit. Anal.* **22** 497–519.

Kunkel, D. and Kaizar, E. E. (2017). A comparison of existing methods for multiple imputation in individual participant data meta-analysis. *Stat. Med.* **36** 3507–3532. MR3696506

Langan, D., Higgins, J. P. T. and Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Res. Synth. Methods* **8** 181–198.

Lassus, J., Gayat, E., Mueller, C., Peacock, W., Spinar, J., Harjola, V., van Kimmenade, R., Pathak, A., Mueller, T. et al. (2013). Incremental value of biomarkers to clinical variables for mortality prediction in acutely decompensated heart failure: The multinational observational cohort on acute heart failure (MOCA) study. *Int. J. Cardiol.* **168** 2186–2194.

Lee, K. and Carlin, J. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.* **171** 624–632.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*: *Unified Analysis via H-Likelihood. Monographs on Statistics and Applied Probability* **106**. Chapman & Hall/CRC, Boca Raton, FL. With 1 CD-ROM (Windows). MR2259540

Little, R. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

Liu, J., Gelman, A., Hill, J., Su, Y.-S. and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101** 155–173. MR3180663

Longford, N. T. (2008). Missing data. In *Handbook of Multilevel Analysis* 377–399. Springer, New York. MR2412943

Mathew, T. and Nordström, K. (2010). Comparison of one-step and two-step meta-analysis models using individual patient data. *Biom. J.* **52** 271–287. MR2756877

McNeish, D. and Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivar. Behav. Res.* **51** 495–518.

Mebazaa, A., Gayat, E., Lassus, J., Meas, T., Mueller, C. et al. (2013). Association between elevated blood glucose and outcome in acute heart failure: Results from an international observational cohort. *J. Am. Coll. Cardiol.* **61** 820–829.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **10** 538–573.

Morris, T. P., White, I. R. and Crowther, M. J. (2017). Using simulation studies to evaluate statistical methods. ArXiv e-prints.

Mullis, I., Martin, M., Gonzalez, E. and Kennedy, A. (2003). Pirls 2001 international report: Iea's study of reading literacy achievement in primary school in 35 countries. Available at: https://timssandpirls.bc.edu/pirls2001i/pdf/p1_IR_book.pdf.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *J. Multivariate Anal.* **98** 896–915. MR2325413

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Pinheiro, J., Bates, D., DebRoy, S. and Sarkar, D. (2016). nlme: Linear and nonlinear mixed effects models. R package version 3.1-128.

Quartagno, M. and Carpenter, J. R. (2016a). Multiple imputation for IPD meta-analysis: Allowing for heterogeneity and studies with missing covariates. *Stat. Med.* **35** 2938–2954. MR3528235

Quartagno, M. and Carpenter, J. (2016b). jomo: A package for multilevel joint modelling multiple imputation. R package version 2.2-0.

Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–96.

Reiter, J., Raghunathan, T. E. and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Surv. Methodol.* **32** 143.

Resche-Rigon, M. and White, I. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat. Methods Med. Res.* DOI:10.1177/0962280216666564.

RESCHE-RIGON, M., WHITE, I. R., BARTLETT, J. W., PETERS, S. A. E., THOMPSON, S. G. and GROUP, P. S. (2013). Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat. Med.* **32** 4890–4905. MR3127183

RILEY, R. D., LAMBERT, P. C., STAESSEN, J. A., WANG, J., GUEYFFIER, F., THIJS, L. and BOUTITIE, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat. Med.* **27** 1870–1893. MR2420350

RILEY, R. D., ENSOR, J., SNELL, K. I. E., DEBRAY, T. P. A., ALTMAN, D. G., MOONS, K. G. M. and COLLINS, G. S. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353** i3140.

ROBERT, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. Springer, New York. MR2723361

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability* **72**. Chapman & Hall, London. MR1692799

SCHAFER, J. L. and YUCEL, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.* **11** 437–457. MR1938143

SIMMONDS, M., HIGGINS, J., STEWART, L., TIERNEY, J., CLARKE, M. and THOMPSON, S. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clin. Trials* **2** 209–217.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing. Version* 3.3.0. R Foundation for Statistical Computing, Vienna, Austria.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16** 219–242. MR2371007

VAN BUUREN, S. (2011). Multiple imputation of multilevel data. In *The Handbook of Advanced Multilevel Analysis* (J. J. Hox, ed.) 173–196. Routledge, New York.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data* (*Chapman & Hall/CRC Interdisciplinary Statistics*). Chapman & Hall/CRC, London.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN BUUREN, S., BRAND, J. P. L., GROOTHUIS-OUDSHOORN, C. G. M. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76** 1049–1064. MR2307507

VINK, G., LAZENDIC, G. and VAN BUUREN, S. (2015). Partitioned predictive mean matching as a multilevel imputation technique. *Psychol. Test Assess. Model.* **57** 577–594.

WAGSTAFF, D. and HAREL, O. (2011). A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom. *Stata J.* **11** 403–419.

YUCEL, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Stat. Model.* **11** 351–370. MR2906705

ZHAO, Y. and LONG, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Stat. Methods Med. Res.* **25** 2021–2035. MR3553324

ZHAO, E. and YUCEL, R. (2009). Performance of sequential imputation method in multilevel applications. In *Proceedings of the Survey Research Methods Section* (*JSM* 2009) 2800–2810. Amer. Statist. Assoc., Alexandria, VA.

ZHU, J. and RAGHUNATHAN, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *J. Amer. Statist. Assoc.* **110** 1112–1124. MR3420688