Research Article

Statistics in Medicine

Received 9 July 2014, Accepted 19 January 2015 Published online 9 February 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6451

Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE

Shahab Jolani,[a*†‡] Thomas P. A. Debray,[b‡] Hendrik Koffijberg,[b] Stef van Buuren[c] and Karel G. M. Moons[b]

Individual participant data meta-analyses (IPD-MA) are increasingly used for developing and validating multivariable (diagnostic or prognostic) risk prediction models. Unfortunately, some predictors or even outcomes may not have been measured in each study and are thus systematically missing in some individual studies of the IPD-MA. As a consequence, it is no longer possible to evaluate between-study heterogeneity and to estimate study-specific predictor effects, or to include all individual studies, which severely hampers the development and validation of prediction models.

Here, we describe a novel approach for imputing systematically missing data and adopt a generalized linear mixed model to allow for between-study heterogeneity. This approach can be viewed as an extension of Resche-Rigon's method (Stat Med 2013), relaxing their assumptions regarding variance components and allowing imputation of linear and nonlinear predictors.

We illustrate our approach using a case study with IPD-MA of 13 studies to develop and validate a diagnostic prediction model for the presence of deep venous thrombosis. We compare the results after applying four methods for dealing with systematically missing predictors in one or more individual studies: complete case analysis where studies with systematically missing predictors are removed, traditional multiple imputation ignoring heterogeneity across studies, stratified multiple imputation accounting for heterogeneity in predictor prevalence, and multilevel multiple imputation (MLMI) fully accounting for between-study heterogeneity.

We conclude that MLMI may substantially improve the estimation of between-study heterogeneity parameters and allow for imputation of systematically missing predictors in IPD-MA aimed at the development and validation of prediction models.

Keywords: multiple imputation; prediction research; multilevel model; IPD meta-analysis; missing data

1. Introduction

An important aim in diagnostic and prognostic research is the development of clinical prediction models. These models aim to predict for an individual whether a certain outcome is present (diagnosis) or will occur (prognosis), respectively based on multiple predictors observed in the individual. These predictors may range from individual characteristics, signs and symptoms, to results of more invasive or costly measures such as imaging, electrophysiology, blood, urine, coronary plaque, or even genetic markers [1–3]. The development of a novel prediction model, diagnostic or prognostic, typically requires a set with so-called individual participant data (IPD). This dataset contains for each study participant the observed predictor values and outcomes to be predicted, and is ideally obtained from a prospective cohort study.

[a] *Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, The Netherlands*

[b] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*

[c] *The Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands*

**Correspondence to: Shahab Jolani, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Sjoerd Groenmangebouw, Padualaan 14, 3584CH Utrecht, The Netherlands.*

[†] *E-mail: S.Jolani@uu.nl*

[‡] *Equal Contribution*

However, during the past decades, the popularity of prediction research has increased and international collaboration has become more commonplace. Multiple individual participant datasets are therefore frequently combined when developing or validating a novel prediction model. This strategy is also known as IPD meta-analysis (IPD-MA) [4–9]. Other types of IPD-MA may arise in large cohort studies where subjects are clustered within the study centers.

A key issue in every IPD-MA is the presence of between-study heterogeneity, that is, systematic differences in populations from which samples are included. Heterogeneity in an IPD-MA for prediction modeling research typically manifests as differences in baseline risks across the individual studies, that is, the outcome prevalences (for diagnostic models) or outcome incidences (for prognostic models), or as differences in the predictor-outcome associations. Recently, we proposed a framework for developing, implementing, and validating a risk prediction model when IPD from multiple studies are available [8]. This framework introduces internal–external cross-validation to investigate the generalizability of a prediction model during its development and proposes to pursue homogeneity in predictor-outcome associations to improve the model's generalizability across different but related populations. The presence of between-study heterogeneity should therefore routinely be assessed when performing an IPD-MA for developing a prediction model.

Unfortunately, within an IPD-MA, in individual studies, often different predictors or tests may be measured, for example, because of budget constraints or lack of medical equipment, or local habits, such that some predictors are not measured in each individual dataset. When combining the individual study datasets, some predictors are no longer complete in the IPD-MA set and have become systematically missing in part of the IPD-MA dataset [10, 11]. As a consequence, researchers often choose to exclude entire studies with one or more missing predictors from the IPD-MA [5]. Alternatively, predictors with systematically missing values in one or more studies are ignored or excluded from the model development [12]. It is clear that both approaches are undesirable as available evidence is not optimally used, certainly if the individual studies are too large or important to be excluded, or the ignored predictors are known to be important. Moreover, by ignoring or excluding evidence, it becomes difficult or even impossible to evaluate the presence of between-study heterogeneity in all (potentially) relevant predictor effects, which may lead to models with decreased predictive performance [8]. It also becomes more difficult to evaluate the model's performance across different studies of the IPD-MA, thereby reducing the model's potential generalizability [8, 13]. For this reason, imputation strategies are needed to account for systematically missing data in an IPD-MA aimed at developing or validating prediction models.

Previously, the Fibrinogen studies collaboration proposed a bivariate random effects meta-analysis model to investigate the association between a certain exposure and disease in an IPD-MA where some confounders are systematically missing [12]. Their (two-stage) approach calculates a pooled estimate of the fully adjusted association by borrowing strength from partially adjusted associations and bears similarities with the adaptation method from Steyerberg *et al.* [14, 15]. Unfortunately, this approach requires all relevant confounders to be included in the statistical model, which may not be desirable when developing a prediction model. Furthermore, it is unclear how the approach can be extended to provide pooled estimates of the fully adjusted confounders (i.e., other predictors) and to estimate their between-study covariance. Finally, because the bivariate model does not generate imputed datasets, its implementation becomes particularly problematic when applying validation techniques such as internal–external cross-validation. For this reason, Resche-Rigon *et al.* proposed a one-stage approach for imputing systematically missing continuous predictors in individual studies of an IPD-MA [16]. Their approach adopts linear mixed effects (multilevel) models with random intercept terms and random slopes to account for heterogeneity across the included studies in the IPD-MA. Its implementation, however, requires knowledge about the standard errors around the estimated between-study covariance parameters. Unfortunately, the likelihood function of nonlinear mixed effects models often does not have a closed-form expression. Therefore second-order derivatives and standard error estimates may become unreliable [17]. In addition, the usefulness of standard errors as measure of uncertainty around between-study covariance parameters can be challenged because this uncertainty tends to be heavily skewed (even when log-transformed). This is one of the major reasons why some software packages for fitting nonlinear mixed effects models (e.g., *lme4* in R) do not provide estimates of standard errors around between-study covariance parameters. As a consequence, other approaches are needed to impute systematically missing predictors in an IPD-MA, certainly when these predictors do not have continuous values.

We here describe a novel imputation method that extends the approach taken by Resche-Rigon *et al.* to allow imputation of both continuous and non-continuous predictors that are systematically missing in one or more individual studies of an IPD-MA. Hereto, we adopt generalized linear mixed effects models and

departure from using error variance as estimates of uncertainty around between-study covariance parameters. Although we focus on the imputation of systematically missing binary predictors, the described methodology can directly be applied to many other types of missing data (e.g., continuous, ordinal, and count data). We begin by describing the statistical methodology for imputing linear (e.g., continuous) and nonlinear (e.g., binary) systematically missing predictors. We subsequently illustrate its implementation in an empirical example with a large-scale IPD-MA aimed at developing and validating a prediction model for the diagnosis of deep venous thrombosis. Subsequently, we conduct a simulation study to compare the performance of the novel imputation approach with that of approaches that either exclude entire studies with missing predictors or completely ignore the presence of between-study heterogeneity during imputation. We end by discussing the results and providing general recommendations for properly dealing with systematically missing predictors in IPD-MA.

2. Methods

In general, the presence of missing data can be described by three mechanisms with different assumptions about the probability of missingness. When this probability is identical for all subjects, predictors are missing completely at random (MCAR). Conversely, missing at random (MAR) occurs when the probability of missingness depends on the observed information. Finally, missing not at random occurs when the probability of missingness depends on the predictor itself or on other predictors that have not been measured conditional on observed data [18].

In the presence of missing data it is common to assume MAR and to apply multiple imputation. This approach generates several copies of the original dataset and replaces missing values by values drawn from a multivariate distribution (joint modeling) [19] or from a set of conditional densities (fully conditional specification) [20–22]. In this article, we assume that a multivariate distribution exists and that draws from it can be generated by iteratively sampling from the conditional distributions. Although these assumptions are less attractive from a theoretical point of view, they facilitate the implementation of more advanced analysis models (as compared with joint modeling) [21]. Subsequently, we adopt multivariate imputation by chained equations (MICE) [23] to impute non-continuous systematically missing predictors in an IPD-MA. In particular, we use generalized linear mixed effects models for specifying the conditional distributions and use the Wishart distribution for deriving estimates of uncertainty around between-study covariance parameters. Finally, we analyze the imputed datasets and pool the resulting model estimates using Rubin's rule [24]. We denote the corresponding imputation strategy as *multilevel multiple imputation*, or shortly, MLMI. An overview of symbols used in this article is presented in Appendix A.

2.1. Complete data model

Consider an IPD-MA of $i = 1, \dots, \mathrm{N}$ studies with $j = 1, \dots, \mathrm{N}_i$ subjects in the ith study. We denote the observed outcome y for subject j in study i as y_{ij}. Furthermore, we denote the vector of $k = 1, \dots, \mathrm{K}$ candidate predictors for subject j in study i as $\mathbf{x}_{ij} = \left[x_{ij1}, \dots, x_{ij\mathrm{K}}\right]$. Let $\mathbf{x}_{ij}$ be associated with a K-dimensional vector of fixed effect parameters $\boldsymbol{\beta}$, and let $\mathbf{u}_i$ be an L-dimensional (L < K) vector of random effects for the ith study. Finally, let $\mathbf{v}_{ij}$ represent a vector of the variables associated with $\mathbf{u}_i$ (typically a subset of $\mathbf{x}_{ij}$). A generalized linear mixed model for the complete data model that accounts for the study-specific predictor effects can then be defined in the exponential class with the form

$$\begin{aligned} f_1(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi) &= \exp\{\phi^{-1}[y_{ij}\zeta_{ij} - a(\zeta_{ij})] + b(y_{ij}, \phi)\} \\ \zeta_{ij} &= \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{v}_{ij}^{\mathrm{T}}\mathbf{u}_i \\ \mathbf{u}_i &\sim \mathrm{MVN}\,(\mathbf{0}, \mathbf{T}) \end{aligned} \tag{1}$$

where ϕ is a scalar dispersion parameter and $a(\cdot)$ represents the link function. The predictor values $\mathbf{x}_{ij}$ are assumed to be independent, and the random effects $\mathbf{u}_i$ are assumed to follow a multivariate normal (MVN) distribution with mean vector 0 and variance–covariance matrix $\mathbf{T}$. Finally, the functions a and b determine a particular family in the exponential class, such as binomial, normal, or Poisson. For instance, logistic regression and Poisson regression assume that $\phi = 1$ such that $b(y_{ij}, \phi) = b(y_{ij})$, and $f_1(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi) = f_1(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta})$ (Table I).

Model (1) can be estimated in R using the *lme4* package [25]. The estimated fixed effects terms, that is, $\hat{\boldsymbol{\beta}}$, can then be extracted using *fixef(object)* or *getME(object, "fixef")*. The estimated variance–covariance

Table I. Examples of linear mixed models.

Family	Data model	Parameterization				Probability density or mass function
		ζ_{ij}	ϕ	$a(\zeta_{ij})$	$b(y_{ij},\phi)$	$f_1(y_{ij}\|\mathbf{u}_i,\boldsymbol{\beta},\phi)$
Gaussian	$y_{ij}\sim\mathcal{N}(\mu_{ij},\sigma^2)$	μ_{ij}	σ^2	$\frac{1}{2}\zeta_{ij}^2$	$\frac{y_{ij}^2}{2\phi}-\frac{1}{2}\ln(2\pi\phi)$	$f_1(y_{ij}\|\mu_{ij},\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_{ij}-\mu_{ij})^2}{2\sigma^2}\right)$
Binomial	$y_{ij}\sim\text{Binom}(n_{ij},\pi_{ij})$	$\ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$	1	$n_{ij}\ln(1+\exp(\zeta_{ij}))$	$\ln\binom{n_{ij}}{y_{ij}}$	$f_1(y_{ij}\|n_{ij},\pi_{ij})=\binom{n_{ij}}{y_{ij}}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{n_{ij}-y_{ij}}$
Poisson	$y_{ij}\sim\text{Poisson}(\lambda_{ij})$	$\ln(\lambda_{ij})$	1	$\exp(\zeta_{ij})$	$-y_{ij}!$	$f_1(y_{ij}\|\lambda_{ij})=\frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!}\exp(-\lambda_{ij})$

Note: Parameterization of ζ_{ij} differs according to the adopted family in the complete data model (see also Equation 1). For instance, for Gaussian y_{ij}, we have $f_1(y_{ij}|\mu_{ij},\sigma^2)$ with $\mu_{ij}=\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{v}_{ij}^{\mathrm{T}}\mathbf{u}_i$ and $\mathbf{u}_i\sim\text{MVN}(\mathbf{0},\mathbf{T})$.

matrix of the fixed effect terms, that is, $\text{Var}(\hat{\boldsymbol{\beta}})$, can be extracted using *vcov(object)*. Finally, the estimated random effects parameter estimates $\hat{\mathbf{T}}$ can be extracted using *getME(object,"theta")*.

2.2. *Imputation model*

Suppose a subset of predictors is systematically missing for some studies in an IPD-MA. This implies that their predictor values are missing for all subjects within the affected studies. Here, we propose a generalization that can accommodate for many types of predictor values (e.g., binary data, categorical data, and count data) and adopts a generalized linear mixed model. When a predictor k is systematically missing in study i, all elements in $\left[x_{i1k},\ldots,x_{iN_ik}\right]$ are unknown. We assume the imputation model for each $x_{ijk} \in \left[x_{i1k},\ldots,x_{iN_ik}\right]$ takes the form

$$\begin{aligned} f_2(x_{ijk}|\boldsymbol{b}_{ik},\boldsymbol{\gamma}_k,\varphi_k) &= \exp\{\varphi_k^{-1}[x_{ijk}\eta_{ijk} - c\left(\eta_{ijk}\right)] + d(x_{ijk},\varphi_k)\} \\ \eta_{ijk} &= \mathbf{z}_{ijk}^{\mathrm{T}}\boldsymbol{\gamma}_k + \mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{b}_{ik} \\ \boldsymbol{b}_{ik} &\sim \text{MVN}\left(\mathbf{0},\boldsymbol{\Psi}_k\right) \end{aligned} \tag{2}$$

where $\mathbf{z}_{ijk}$ is a P-dimensional vector of covariates associated with the fixed effect parameters $\boldsymbol{\gamma}_k$ included in the imputation model. These covariates typically represent the remaining predictors x_{ijs} ($s \neq k$), variables that may have influenced the occurrence of missing data, variables that explain variance in the candidate predictors, and the outcome y_{ij} (or a function of it). Furthermore, φ_k is the scale dispersion parameter associated with covariate x_{ijk}, and $g(\cdot)$ is the corresponding link function. The Q-dimensional vector $\boldsymbol{b}_{ik}$ represents the random effects in the imputation models and is associated with the vector of subject-level covariates $\mathbf{w}_{ijk}$.

In general, two important issues should be considered when specifying model (2) for a certain missing data scenario. Firstly, the composition of $\boldsymbol{\gamma}_k$ should be defined as such it increases the plausibility of the MAR assumption. Hereto, the imputation model may consider to include certain predictors that are not of interest in the eventual analysis model. Secondly, the imputation model should be more general than the analysis model to allow inferential congeniality [26]. This implies that all predictors and outcomes from model (1) should be included in model (2) and that $\boldsymbol{\Psi}_k$ should be equally or less restricted (e.g., in terms of independence) than $\mathbf{T}$. As a consequence, if the complete data model aims to investigate correlation between the random effects of certain predictors (i.e., non-diagonal entries of $\mathbf{T}$), the imputation model should minimally estimate all entries of $\boldsymbol{\Psi}_k$ that involve the predictors of the analysis model.

The unknown parameters from model (2) are denoted as $\boldsymbol{\theta}_k$ and include the fixed effect parameters $\boldsymbol{\gamma}_k$, the between-study covariance $\boldsymbol{\Psi}_k$, and possible dispersion parameters φ_k resulting from the link function. Here, we define $\boldsymbol{\Xi}_k$ as $\left\{\boldsymbol{\Psi}_k,\varphi_k\right\}$. Note that the scale dispersion parameter $\varphi_k = 1$ for binary or count cases, such that $\boldsymbol{\Xi}_k$ then collapses to $\boldsymbol{\Psi}_k$.

2.2.1. Binary missing predictors. Suppose that the systematically missing predictor follows a Bernoulli distribution with success probability $Pr(x_{ijk} = 1) = \pi_{ijk}$. Choosing the logit transformation of the probability of success, $\eta_{ijk} = \ln\left(\pi_{ijk}/(1-\pi_{ijk})\right)$, as the link function leads to

$$\begin{aligned} x_{ijk} &\sim \text{Bernoulli}\left(\pi_{ijk}\right) \\ \pi_{ijk} &= \frac{1}{1+\exp(-\mathbf{z}_{ijk}^{\mathrm{T}}\boldsymbol{\gamma}_k - \mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{b}_{ik})} \\ \boldsymbol{b}_{ik} &\sim \text{MVN}(0,\boldsymbol{\Psi}_k) \end{aligned} \tag{3}$$

The unknown parameters from model (3) are $\boldsymbol{\gamma}_k$ and $\boldsymbol{\Psi}_k$. There is no dispersion parameter such that $\boldsymbol{\theta}_k = \left\{\boldsymbol{\gamma}_k,\boldsymbol{\Xi}_k\right\} = \left\{\boldsymbol{\gamma}_k,\boldsymbol{\Psi}_k\right\}$.

2.2.2. Continuous missing predictors. The systematically missing predictor x_{ijk} is assumed to be normally distributed with mean $\mathbf{z}_{ijk}^{\mathrm{T}}\boldsymbol{\gamma}_k + \mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{b}_{ik}$ and variance $\sigma_k^2 = \varphi_k$ and the identity link function, such that model (2) becomes

$$\begin{aligned} x_{ijk} &= \mathbf{z}_{ijk}^{\mathrm{T}}\boldsymbol{\gamma}_k + \mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{b}_{ik} + \boldsymbol{\epsilon}_{ijk} \\ \boldsymbol{b}_{ik} &\sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Psi}_k\right) \\ \boldsymbol{\epsilon}_{ijk} &\sim \mathcal{N}\left(0, \sigma_k^2\right) \end{aligned} \tag{4}$$

The unknown parameters from model (4) are $\boldsymbol{\gamma}_k$, Ψ_k, and σ_k^2. Mixed effect models typically assume that the error variance σ_k^2 and the between-study variance $\boldsymbol{\Psi}_k$ are independent. For sake of simplicity, we assume these parameters are a priori independent. The unknown parameters can then be denoted as $\boldsymbol{\theta}_k = \left\{\boldsymbol{\gamma}_k, \boldsymbol{\Xi}_k\right\}$ where

$$\boldsymbol{\Xi}_k = \begin{bmatrix} \boldsymbol{\Psi}_{11k} & \cdots & \boldsymbol{\Psi}_{1\mathrm{Q}k} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\Psi}_{\mathrm{Q}1k} & \cdots & \boldsymbol{\Psi}_{\mathrm{QQ}k} & 0 \\ 0 & \ldots & 0 & \sigma_k^2 \end{bmatrix}$$

2.3. The imputation procedure

We distinguish between the imputation of univariate and multivariate systematically missing predictors. For both types of missing data, MLMI consists of three main steps. First, model (2) is fitted to M studies where x_{ijk} is observed. Second, random draws of $\boldsymbol{\theta}_k^*$ are generated in sequence, using $\hat{\boldsymbol{\gamma}}_k$, $\hat{\boldsymbol{\Psi}}_k$, and $\hat{\sigma_k^2}$ (if applicable). Finally, $\boldsymbol{\theta}_k^*$ is used to generate imputations for systematically missing predictor x_{ijk}. The following algorithms were implemented as an extension for *mice* in R (see Appendix B for imputing binary data).

2.3.1. Univariate systematically missing predictor (one predictor is systematically missing in some studies). Suppose x_{ijk} is the only systematically missing predictor in an IPD-MA. Without loss of generality, assume x_{ijk} is fully observed in M studies (M < N and M > 1 to allow for mixed effects models), and let x_{ijk}^{obs} and x_{ijk}^{mis} be the fully observed and systematically missing subsets of x_{ijk}, respectively. The missing values are said to be MAR if $Pr(R_{ijk} = 1|x_{ijk}^{\mathrm{obs}}, x_{ijk}^{\mathrm{mis}}, \boldsymbol{\theta}_k) = Pr(R_{ijk} = 1|x_{ijk}^{\mathrm{obs}}, \boldsymbol{\theta}_k)$, where R_{ijk} denotes the response indicator of predictor k for subject j in study i. We consider the situation where MAR represents a reasonable assumption. The objective is to draw x_{ijk}^{mis} from its posterior distribution under model (2). This posterior distribution can be written as

$$Pr(x_{ijk}^{\mathrm{mis}}|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}}) = \int_{\boldsymbol{\theta}_k} Pr(x_{ijk}^{\mathrm{mis}}|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}}, \boldsymbol{\theta}_k)\, Pr(\boldsymbol{\theta}_k|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}})\, \partial\boldsymbol{\theta}_k \tag{5}$$

Following Rubin [24], we propose to draw D values of $\boldsymbol{\theta}_k$ from the posterior distribution $Pr(\boldsymbol{\theta}_k|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}})$. For each corresponding draw $\boldsymbol{\theta}_k^*$, we then draw a value of x_{ijk}^{mis} from the conditional posterior distribution $Pr(x_{ijk}^{\mathrm{mis}}|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}}, \boldsymbol{\theta}_k = \boldsymbol{\theta}_k^*)$.

Drawing from the posterior distribution $Pr(\boldsymbol{\theta}_k|\mathbf{z}_{ijk}, \mathbf{w}_{ijk}, x_{ijk}^{\mathrm{obs}})$ requires a prior distribution for $\boldsymbol{\theta}_k$. We use the suitable diffuse prior $p\left(\boldsymbol{\gamma}_k\right) \propto 1$; that is, the density of $\boldsymbol{\gamma}_k$ is uniform between $-\infty$ and $+\infty$. Also, we apply the standard reference prior $p\left(\boldsymbol{\Psi}_k^{-1}\right) \propto |\boldsymbol{\Psi}_k^{-1}|^{-(Q+1)/2}$. For the continuous case, we further assume the prior distribution of σ_k^2 has density proportional to σ_k^{-2}. Under these standard priors, the posterior distributions of the parameters of a (generalized) linear mixed model are not available in closed from. We therefore use a large sample approximation. The detailed steps of the imputation procedure are as follows:

(1) Obtain the estimates of the parameters $\boldsymbol{\gamma}_k$ and $\boldsymbol{\Xi}_k$ by the maximum likelihood (ML) estimator using the M studies where x_{ijk} is observed.
(2) Draw $\boldsymbol{\gamma}_k^*$ from MVN $\left(\hat{\boldsymbol{\gamma}}_k, \mathrm{var}\left(\hat{\boldsymbol{\gamma}}_k|\hat{\boldsymbol{\Xi}}_k\right)\right)$.

(3) For studies where x_{ijk} is observed, obtain $\boldsymbol{b}_{ik}$ and calculate the Q × Q dimensional matrix $\boldsymbol{\Lambda}_k = \sum_{i=1}^{M} \boldsymbol{b}_{ik}\boldsymbol{b}_{ik}^{\mathrm{T}}$.
(4) Draw $\boldsymbol{\Psi}_k^{*^{-1}}$ from its posterior distribution, that is, a Wishart distribution with M degrees of freedom and a scale matrix parameter equal to $\boldsymbol{\Lambda}_k^{-1}$.
(5) For (N − M) studies where x_{ijk} is systematically missing, draw $\boldsymbol{b}_{ik}^{*}$ from MVN $\left(\mathbf{0}, \boldsymbol{\Psi}_k^{*}\right)$.
(6) For a binary predictor x_{ijk}

   (a) Draw x_{ijk}^{*} for each study i with systematically missing binary predictor x_{ijk} using $\mathrm{logit}^{-1}(\mathbf{z}_{ijk}\boldsymbol{\gamma}_k^{*} + \mathbf{z}_{ijk}^{*}\boldsymbol{b}_{ik}^{*})$.

(7) For a continuous predictor x_{ijk}

   (a) Calculate $\sigma_k^{2^{*}} = (df\ \hat{\sigma}_k{}^2)/d$ by drawing d from a χ^2 distribution with $df = \sum_{l=1}^{\mathrm{M}} \mathrm{N}_l - \mathrm{P}$ degrees of freedom.
   (b) Draw x_{ijk}^{*} for each study i with systematically missing continuous predictor x_{ijk} using $x_{ijk}^{*} = \mathbf{z}_{ijk}\boldsymbol{\gamma}_k^{*} + \mathbf{z}_{ijk}^{*}\boldsymbol{b}_{ik}^{*} + \epsilon_{ijk}^{*}$ where $\epsilon_{ijk}^{*} \sim \mathcal{N}\left(0, \sigma_k^{2^{*}}\right)$.

The algorithm proposed here is similar to the algorithm previously proposed by Resche-Rigon *et al.* [16]. However, we do not assume that $\boldsymbol{\Xi}_k$ is (log-)normally distributed and instead adopt a Wishart distribution to accommodate for the asymmetry in the distribution of the estimator. Finally, our algorithm is not limited to a continuous predictor and can be applied to other types of systematically missing predictors such as binary, ordinal, nominal, or count variables.

2.3.2. Multivariate systematically missing predictors (two or more predictors are systematically missing in some studies). Suppose L predictors of $\mathbf{x}_{ij}$ ($2 \leqslant \mathrm{L} < \mathrm{K}$) are systematically missing in some studies. According to the MICE algorithm, each systematically missing predictor is imputed in turn using the latest imputed values of the other predictors. The procedure is then iterated for a sufficient number of times.

For each systematically missing predictor x_{ijl} with $l = 1, \ldots, \mathrm{L}$, a value of $\boldsymbol{\theta}_l$ is first drawn. Afterwards, the missing part of x_{ijl} is imputed using the drawn value of the corresponding parameters. Similar to the univariate case, the fully observed and systematically missing parts of x_{ijl} are denoted respectively by x_{ijl}^{obs} and x_{ijl}^{mis}. Starting from an initial imputation step, the tth iteration involves successively drawing from

$$\begin{aligned}
\boldsymbol{\theta}_1^{*(t)} &\sim Pr(\theta_1 | \mathbf{z}_{ij1}^{(t-1)}, x_{ij1}^{\mathrm{obs}}) \\
x_{ij1}^{\mathrm{mis}(t)} &\sim Pr(x_{ij1}^{\mathrm{mis}} | \mathbf{z}_{ij1}^{(t-1)}, x_{ij1}^{\mathrm{obs}}, \boldsymbol{\theta}_1^{*(t)}) \\
&\vdots \\
\boldsymbol{\theta}_{\mathrm{L}}^{*(t)} &\sim Pr(\theta_{\mathrm{L}} | \mathbf{z}_{ijL}^{(t)}, x_{ij\mathrm{L}}^{\mathrm{obs}}) \\
x_{ij\mathrm{L}}^{\mathrm{mis}(t)} &\sim Pr(x_{ijL}^{\mathrm{mis}} | \mathbf{z}_{ij\mathrm{L}}^{(t)}, x_{ij\mathrm{L}}^{\mathrm{obs}}, \boldsymbol{\theta}_{\mathrm{L}}^{*(t)})
\end{aligned}$$

where $\mathbf{z}_{ijl}$ typically consists of the predictors x_{ijs} ($s \neq l$) that were imputed in the previous steps, the outcome y_{ij}, and other variables that may have influenced the occurrence of missing data or explain variability in the predictor values. Executing the cycle repeatedly for a sufficient number of iterations creates one set of the completed data. Repeating the whole procedure D times produces D versions of the imputed datasets for post-imputation analyses.

2.4. Combining the repeated complete data estimates and variances

After imputation, the resulting D versions of the completed data are analyzed by complete data model (1). Suppose $\hat{\boldsymbol{\beta}}^{(d)}$ denotes the estimated values of $\boldsymbol{\beta}$ in model (1) from dth imputed dataset, and $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(d)})$ denotes their associated variance–covariance matrix. The combined estimate of $\boldsymbol{\beta}$ and its variance can be obtained using Rubin's rule [24, 27]. The overall estimate of $\boldsymbol{\beta}$ is simply the average

$$\bar{\boldsymbol{\beta}} = \frac{1}{\mathrm{D}} \sum_{d=1}^{\mathrm{D}} \hat{\boldsymbol{\beta}}^{(d)} \tag{6}$$

with total variance

$$\text{Var}(\bar{\boldsymbol{\beta}}) = \frac{1}{\text{D}}\sum_{d=1}^{\text{D}} \text{Var}\left(\hat{\boldsymbol{\beta}}^{(d)}\right) + \left(\frac{\text{D}+1}{\text{D}^2-\text{D}}\right)\sum_{d=1}^{\text{D}}\left(\hat{\boldsymbol{\beta}}^{(d)}-\bar{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}^{(d)}-\bar{\boldsymbol{\beta}}\right)^{\text{T}} \tag{7}$$

Similarly, the overall estimate of the variance–covariance matrix of random effects parameters is given as

$$\bar{\mathbf{T}} = \frac{1}{\text{D}}\sum_{d=1}^{\text{D}} \hat{\mathbf{T}}^{(d)} \tag{8}$$

Equation 7 indicates $\text{Var}(\bar{\boldsymbol{\beta}})$ that approximates the average of within-impuation covariances $\text{Var}(\hat{\boldsymbol{\beta}}^{(d)})$ as $\text{D} \to \infty$. In general, the efficiency of an estimate based on D imputations is approximately

$$\left(1+\frac{\gamma}{\text{D}}\right)^{-1} \tag{9}$$

where γ is the fraction of missing information for the quantity being estimated [24]. The fraction γ quantifies how much more precise the estimate might have been if no data had been missing.

3. Empirical example

In order to illustrate the approaches, we describe a clinical example involving the diagnosis of deep vein thrombosis (DVT) presence. DVT is a blood clot that forms in a vein in the body and may lead to pulmonary embolism, preventing oxygenation of the blood and potentially causing death. Clinical DVT diagnosis is not straightforward. For this reason, multivariable diagnostic prediction models have been developed to predict the probability of presence of DVT in suspected patients. A well-known example is the model developed by Oudega *et al.*, which includes the results from history taking, physical examination, and D-dimer testing for ruling out DVT in primary care [28]. It is, however, unclear to which extent this model is generalizable, as it is possible that some of its included predictor effects differ across study populations.

In this case study, we aimed to investigate the overall strength of association and the presence of between-study heterogeneity in the predictors of the *Oudega* model. Hereto, we used an IPD meta-analysis of 13 studies conducted for diagnosing DVT in patients with a suspected DVT (Table II). The IPD-MA contains a total of 10,002 subjects of which 1864 (18.6%) truly have DVT as established by

Table II. Pattern of the missing covariates in the empirical example.

Study		1	2	3	4	5	6	7	8	9	10	11	12	13	Model
Size		1028	814	153	1756	791	1075	429	325	1295	436	541	550	809	
Country	Type	NL	NL	CA	NL	NL	CA	CA	SE	NL	US	CA	CA	CA	
sex	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
malign	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
par	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
surg	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
tend	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
leg	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
pit	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
vein	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
adiag	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
cdif3	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
age	c	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
coag	d	✓			✓	✓	✓		✓		✓	✓			
notraum	d	✓		✓	✓	✓	✓			✓		✓			✓
oachst	d	✓			✓	✓				✓					✓
ddimd	d	✓	✓		✓	✓	✓		✓	✓		✓			✓

Note: The ✓-symbol indicates whether the corresponding predictor has been measured in a certain study and whether it was included in the *Oudega* model. Systematically missing predictor variables are printed in bold. Countries: NL, The Netherlands; CA, Canada; SE, Sweden; US, United States. Variable types: *d*, dichotomous; *c*, continuous.

the reference standard [29]. The following 11 predictors were measured in all studies: age, male gender (*sex*), active malignancy (*malign*), paresis (*par*), recent surgery or bedridden (*surg*), localized tenderness deep venous system (*tend*), entire leg swollen (*leg*), difference in calf circumference >= 3 cm (*cdif3*), pitting edema (*pit*), vein distension (*vein*), and alternative diagnosis present (*adiag*). Furthermore, four binary predictors were systematically missing in one or more studies: D-dimer positive (*ddimd*), family history of thrombofilia (*coag*), no leg trauma present (*notraum*), and oral contraceptive use (*oachst*). The *Oudega* model consists of a combination of fully observed (*sex*, *malign*, *surg*, *vein*, and *cdif3*) and systematically missing (*notraum*, *oachst*, and *ddimd*) predictors.

3.1. Dealing with missing data

We adopted four approaches to investigate the presence of between-study heterogeneity in the predictors of the *Oudega* model. First, we performed a complete case analysis (CCA) where studies with systematically missing predictors were removed. This approach assumes MCAR for systematically missing predictors. Secondly, we implemented a naive imputation approach that assumes MAR and completely ignores heterogeneity across studies. Hereto, we used the *logreg* imputation procedure from *mice* in R [23, 30]. We denote this approach as *traditional* multiple imputation (TMI). An alternative, but similar approach, may allow for heterogeneous intercept terms in the imputation model by dummy-coding (rather than discarding) the study identification number [11]. This approach is denoted as stratified multiple imputation (SMI). Because study-specific intercept terms of the imputation model are unidentifiable for studies with systematically missing predictors, the average of the observed study-specific (stratified) intercept terms is used in such cases. Finally, we adopted MLMI as an alternative approach that accounts for heterogeneity across studies. This approach allowed for joint random effects on all unknown parameters of the imputation model (Section 2) and was implemented in the R-package *mice* [30]. For TMI, SMI, and MLMI, the systematically missing data were multiply imputed using information from all 15 predictors and the outcome. Hereby, we allowed for 20 iterations of the Gibbs algorithm and generated 20 imputed replications of the original dataset.

The MLMI approach allowed for joint random effects on all input variables to ensure congeniality with subsequent analyses. This resulted in imputation models with 120 unknown heterogeneity and 15 unknown fixed effects parameters (MLMI^1). We also explored a simplified version of MLMI where random effects were removed for *sex*, *surg*, *vein*, and *oachst*. The resulting imputation models (MLMI^2) then consisted of 10 heterogeneity and 15 fixed effects parameters.

3.2. Analysis models

We estimated two mixed effects models in the completed data using all predictors from the *Oudega* model. In the first analysis model (Analysis 1), independent random effects were placed on the intercept term and all predictors (model D in [31]). This corresponded to a mixed effect model with nine heterogeneity and nine fixed effects parameters. In the second analysis model (Analysis 2), independent random effects were placed on the intercept term and a subset of predictors. In particular, homogeneity was assumed for regression coefficients where $\hat{\tau}$ was relatively small for all approaches. The analysis model then consisted of five heterogeneity and nine fixed effects parameters. Because all analysis models were rather complex, we allowed for a maximum of 20,000 function evaluations in the optimization process of the likelihood functions. Estimates for the regression coefficients ($\hat{\beta}$), corresponding standard errors ($\text{SE}(\hat{\beta})$), and between-study standard deviation ($\hat{\tau}$) were obtained using the *glmer* procedure from the R-package *lme4* 1.1-7 [25]. All analyses were performed in R 3.1.1 on a 64-bit operating system with linux 3.13.0-24-generic.

3.3. Results (Analysis 1)

Results demonstrate that estimation of the analysis models was problematic, leading to poor convergence rates for CCA, TMI, SMI, and MLMI (Table III). Furthermore, substantial computation time was needed for applying MLMI and analyzing the resulting datasets. Because heterogeneity was not observed in all predictors, we decided to remove random effects for *sex*, *surg*, *vein*, and *oachst* in the complete data model of Analysis 2 and in the imputation model of MLMI^2.

Table III. Results from the empirical example.

		Analysis 1				Analysis 2				
Approach		CCA	TMI	SMI	MLMI[1]	CCA	TMI	SMI	MLMI[1]	MLMI[2]
No. of studies		4	13	13	13	4	13	13	13	13
No. of subjects		4870	10,002	10,002	10,002	4870	10,002	10,002	10,002	10,002
(Intercept)	$\hat{\beta}$	−4.96	−5.00	−4.89	−4.42	−4.96	−5.00	−4.89	−4.42	−4.46
	SE($\hat{\beta}$)	0.24	0.21	0.20	0.28	0.26	0.21	0.20	0.28	0.29
	$\hat{\tau}$	0.29	0.46	0.40	0.77	0.29	0.46	0.40	0.77	0.81
sex	$\hat{\beta}$	0.56	0.47	0.44	0.45	0.56	0.47	0.44	0.45	0.45
	SE($\hat{\beta}$)	0.09	0.06	0.06	0.07	0.09	0.06	0.06	0.07	0.07
	$\hat{\tau}$	0.00	0.00	0.01	0.05	—	—	—	—	—
malign	$\hat{\beta}$	0.37	0.76	0.68	0.83	0.37	0.76	0.68	0.83	0.82
	SE($\hat{\beta}$)	0.13	0.15	0.14	0.16	0.13	0.16	0.14	0.16	0.16
	$\hat{\tau}$	0.00	0.36	0.31	0.41	0.00	0.36	0.31	0.42	0.41
surg	$\hat{\beta}$	0.41	0.36	0.35	0.37	0.41	0.37	0.35	0.37	0.37
	SE($\hat{\beta}$)	0.12	0.09	0.09	0.08	0.12	0.09	0.09	0.08	0.09
	$\hat{\tau}$	0.00	0.00	0.00	0.00	—	—	—	—	—
vein	$\hat{\beta}$	0.43	0.44	0.44	0.45	0.43	0.43	0.43	0.45	0.44
	SE($\hat{\beta}$)	0.10	0.09	0.09	0.10	0.10	0.08	0.08	0.08	0.08
	$\hat{\tau}$	0.00	0.09	0.09	0.13	—	—	—	—	—
notraum*	$\hat{\beta}$	0.53	0.54	0.56	0.40	0.53	0.54	0.56	0.40	0.41
	SE($\hat{\beta}$)	0.12	0.11	0.10	0.13	0.12	0.11	0.10	0.13	0.12
	$\hat{\tau}$	0.00	0.03	0.02	0.18	0.00	0.03	0.01	0.18	0.15
oachst*	$\hat{\beta}$	0.59	0.66	0.55	0.50	0.59	0.66	0.55	0.50	0.50
	SE($\hat{\beta}$)	0.17	0.15	0.17	0.17	0.17	0.15	0.17	0.15	0.18
	$\hat{\tau}$	0.00	0.00	0.00	0.13	—	—	—	—	—
ddimd*	$\hat{\beta}$	2.68	2.69	2.71	2.06	2.68	2.69	2.71	2.05	2.07
	SE($\hat{\beta}$)	0.18	0.15	0.15	0.34	0.19	0.15	0.15	0.34	0.33
	$\hat{\tau}$	0.17	0.26	0.26	1.07	0.17	0.26	0.26	1.07	1.09
cdif3	$\hat{\beta}$	1.09	1.12	1.11	1.15	1.09	1.12	1.11	1.16	1.15
	SE($\hat{\beta}$)	0.14	0.08	0.08	0.09	0.14	0.08	0.08	0.09	0.09
	$\hat{\tau}$	0.21	0.15	0.15	0.19	0.21	0.15	0.15	0.19	0.19
Comp. time[†]	Imputation	NA	5 m	6 m	1660 h	NA	5 m	6 m	1660 h	80 h
Comp. time[‡]	Analysis	23 s	31 m	29 m	25 m	23 s	9 m	9 m	11 m	11 m
Convergence	Analysis	0/1	5/20	4/20	8/20	1/1	19/20	20/20	18/20	20/20

Note: Estimates are based on a mixed effect model with independent random effects for all regression coefficients.
* Corresponding variable was systematically missing in one or more studies.
† Total computation time needed for generating 20 imputed datasets.
‡ Total computation time needed for estimating the analysis model in each (imputed) dataset and combining the results using Rubin's rule (if applicable).
[1] Joint random effects in the imputation model were placed on all model parameters.
[2] Joint random effects in the imputation model were placed on the intercept term, *malign*, *cdif3*, *notraum*, and *ddimd*.
s, seconds; m, minutes; h, hours; NA, not applicable; CCA, complete case analysis; TMI, traditional multiple imputation; SMI, stratified multiple imputation; MLMI, multilevel multiple imputation; SE, standard error.

3.4. Results (Analysis 2)

In this second analysis, convergence issues basically disappeared, while results remained fairly similar even when the imputation model of MLMI was further simplified (MLMI[2]).

Estimates of regression coefficients (representing the predictor effects) were quite similar for all methods, except for *malign* (0.37 for CCA vs. 0.76 for TMI and 0.83 for MLMI) and *ddimd* (2.05 for MLMI vs. 2.68 for CCA and 2.69 for TMI). The strongest similarities were found between TMI and SMI, which also yielded similar estimates of between-study heterogeneity and error variance. Surprisingly, MLMI and CCA achieved similar errors of estimated regression coefficients, except for *ddimd* where it increased from 0.19 (CCA) to 0.34 (MLMI). It is likely that the analysis models of MLMI did not fully benefit from the additional IPD because the underlying imputation models were very complex (in this example, they involved estimating 45 heterogeneity parameters and 9 fixed effects parameters).

Complete data on the predictor variables from the *Oudega* model were only available for studies 1, 4, 5, and 9. The remaining studies were typically initiated by different investigators in other countries (including Canada, Sweden, and USA, Table II) . It is therefore plausible that the presence of systematically missing predictors did not occur completely at random and that CCA may have led to a more homogeneous set of studies. This effect is also illustrated in Table III, where estimates of between-study heterogeneity were lowest for CCA and substantially larger for TMI, SMI, and MLMI. For instance, the between-study standard deviation for *malign* increased from 0.00 (CCA) to 0.36 (TMI) and 0.41 (MLMI) in Model 1. Although the estimated predictor effects from CCA are likely to be representative for the Netherlands (because of the low degree of between-study heterogeneity and the relatedness of the remaining studies), they may not generalize well towards new study populations. Unfortunately, this potential deficiency could not be identified with CCA. The multiple imputation approaches revealed that the predictor effects of the *Oudega* model substantially vary when all studies are included. Researchers aiming to develop a novel prediction model for diagnosing DVT should therefore carefully consider whether geographical adjustments are needed.

4. Simulation study

A set of simulation studies was conducted to evaluate the performance of MLMI in the presence of systematically missing predictors. The number of Monte Carlo simulations was set to 500 for each scenario. In the following, we describe the several stages involved.

4.1. Data generation

In each simulation study, we began by generating complete datasets according to

$$\text{logit}\{Pr(y_{ij}=1)\} = -2.321 + 1.112\,x_{ij1} + 1.375\,x_{ij2} + u_{i0} + u_{i1}x_{ij1} + u_{i2}x_{ij2}$$
$$\mathbf{u}_i \sim \text{MVN}\left(\mathbf{0}, \begin{bmatrix} \tau_0^2 & \tau_{01}\tau_{02} \\ \tau_{01} & \tau_1^2\tau_{12} \\ \tau_{02} & \tau_{12}\tau_2^2 \end{bmatrix}\right).$$

where $\tau_0 = 0.573$, $\tau_1 = 0.389$, $\tau_2 = 0.186$, $\tau_{01} = -0.192$, $\tau_{02} = -0.037$, and $\tau_{12} = 0.039$. These parameter values were taken from the DVT data with *malign* and *cdif3* as predictors x_1 and x_2, respectively. For each study, the predictors x_{ij1} and x_{ij2} were simulated from a bivariate binary process with marginal probability $\pi_i = (\pi_{ij1}, \pi_{ij2})^T$ and $\text{Corr}(x_{ij1}, x_{ij2}) = \rho_i$. These parameter values were also taken from the DVT data and correspond to low (x_{ij1}) and medium (x_{ij2}) success probabilities (Appendix C). Finally, the complete dataset was simulated for two scenarios with small ($N = 6$) and medium ($N = 13$) number of studies. We generated a fixed number of 500 subjects for each study, leading to a total sample size of 3000 ($N = 6$) and 6500 ($N = 13$).

We defined additional scenarios by considering different patterns of systematically missing predictor variables. First, we considered a univariate pattern where x_1 was systematically missing across some studies. We also implemented a bivariate pattern where either x_1 or x_2 was systematically missing. For each scenario, we evaluated a low rate (20%) and a medium rate (50%) of missingness where the corresponding predictor(s) were chosen to be systematically missing under an MCAR mechanism. For all scenarios, we calculated computation times for generating one imputed dataset (Appendix D).

Table IV. Estimates of the fixed effect parameters in the simulation study.

	(Intercept) $\beta_0 = -2.321$				Predictor 1 $\beta_1 = 1.112$				Predictor 2 $\beta_2 = 1.375$			
Parameter	Est.	RB	RMSE	CR	Est.	RB	RMSE	CR	Est.	RB	RMSE	CR
One systematically missing predictor (x_1)												
Scenario 1: 6 studies, 1 study missing (≈20%)												
REF	−2.343	0.009	0.265	88	1.116	0.004	0.230	93	1.391	0.012	0.135	94
CCA	−2.351	0.013	0.301	85	1.124	0.011	0.257	93	1.394	0.014	0.149	93
TMI	−2.341	0.009	0.266	88	1.113	0.001	0.253	92	1.392	0.012	0.136	95
SMI	−2.339	0.008	0.266	88	1.121	0.008	0.254	91	1.391	0.012	0.136	94
MLMI	−2.347	0.011	0.266	89	1.127	0.014	0.253	93	1.393	0.013	0.137	95
Scenario 2: 6 studies, 3 studies missing (≈50%)												
REF	−2.347	0.011	0.246	89	1.112	0.000	0.240	93	1.381	0.004	0.135	95
CCA	−2.372	0.022	0.358	79	1.128	0.015	0.323	94	1.392	0.013	0.189	95
TMI	−2.348	0.012	0.245	90	1.106	0.005	0.329	88	1.384	0.007	0.136	95
SMI	−2.345	0.010	0.244	90	1.109	0.003	0.328	88	1.382	0.005	0.136	95
MLMI	−2.368	0.020	0.261	92	1.124	0.011	0.359	91	1.386	0.008	0.140	96
Scenario 3: 13 studies, 3 studies missing (≈20%)												
REF	−2.322	0.001	0.161	93	1.108	0.003	0.146	94	1.371	0.003	0.090	95
CCA	−2.333	0.005	0.180	92	1.114	0.002	0.164	94	1.374	0.000	0.102	95
TMI	−2.320	0.000	0.161	93	1.098	0.012	0.164	93	1.373	0.001	0.091	95
SMI	−2.315	0.002	0.161	93	1.107	0.004	0.166	93	1.373	0.002	0.091	96
MLMI	−2.325	0.002	0.161	93	1.114	0.002	0.167	95	1.372	0.002	0.091	96
Scenario 4: 13 studies, 7 studies missing (≈50%)												
REF	−2.326	0.002	0.168	91	1.113	0.001	0.156	94	1.374	0.000	0.091	94
CCA	−2.323	0.001	0.251	86	1.113	0.001	0.238	91	1.385	0.007	0.139	93
TMI	−2.320	0.000	0.171	89	1.085	0.024	0.245	84	1.379	0.003	0.094	95
SMI	−2.316	0.002	0.171	89	1.100	0.011	0.241	86	1.376	0.001	0.094	96
MLMI	−2.330	0.004	0.175	91	1.109	0.002	0.251	90	1.380	0.004	0.096	95
Two systematically missing predictors (x_1 and x_2)												
Scenario 5: 13 studies, 3 studies missing (≈20%)												
REF	−2.327	0.003	0.169	92	1.113	0.001	0.153	93	1.381	0.005	0.093	93
CCA	−2.329	0.004	0.190	91	1.116	0.004	0.177	94	1.379	0.003	0.109	92
TMI	−2.324	0.001	0.170	91	1.106	0.005	0.166	91	1.379	0.003	0.099	93
SMI	−2.324	0.001	0.170	91	1.110	0.002	0.165	92	1.383	0.006	0.099	93
MLMI	−2.329	0.003	0.171	91	1.113	0.001	0.168	94	1.383	0.006	0.101	94
Scenario 6: 13 studies, 7 studies missing (≈50%)												
REF	−2.317	0.002	0.168	94	1.097	0.013	0.154	95	1.381	0.004	0.095	94
CCA	−2.324	0.001	0.244	87	1.105	0.006	0.228	91	1.379	0.003	0.137	95
TMI	−2.308	0.006	0.170	92	1.082	0.027	0.192	87	1.372	0.002	0.108	93
SMI	−2.313	0.003	0.171	93	1.087	0.022	0.198	88	1.382	0.006	0.107	93
MLMI	−2.319	0.001	0.173	93	1.091	0.019	0.202	91	1.383	0.006	0.110	95

Note: REF indicates the results that were obtained *before* missingness was introduced and can be viewed as a benchmark for comparing the performance of methods that are applied *after* missingness is introduced: complete case analysis (CCA), traditional multiple imputation (TMI), stratified multiple imputation (SMI), and multilevel multiple imputation (MLMI). The following values are given: mean of estimates (Est.), relative bias (RB), root mean squared error (RMSE), and the coverage rate (CR) of the nominal 95% CI.

4.2. *Data analysis*

The generated datasets with systematically missing predictors were completed using CCA, TMI, SMI, or MLMI. To ensure that imputations could be generated within a reasonable amount of time, we allowed for 10 cycles in the imputation algorithms and chose $D = 5$. The completed datasets were then analyzed by estimating a mixed effect model with joint random effects on all regression coefficients. This model

Table V. Estimates of the random effect parameters in the simulation study.

	(Intercept) $\tau_0 = 0.573$		Predictor 1 $\tau_1 = 0.389$		Predictor 2 $\tau_2 = 0.186$	
Parameter	Mean	Median	Mean	Median	Mean	Median
One systematically missing predictor (x_1)						
Scenario 1:	6 studies, 1 study missing (≈20%)					
REF	0.505	0.494	0.362	0.356	0.204	0.170
CCA	0.481	0.469	0.354	0.342	0.201	0.162
TMI	0.499	0.486	0.334	0.316	0.204	0.171
SMI	0.500	0.488	0.334	0.317	0.204	0.172
MLMI	0.504	0.494	0.372	0.343	0.211	0.172
Scenario 2:	6 studies, 3 studies missing (≈50%)					
REF	0.492	0.484	0.352	0.341	0.207	0.176
CCA	0.403	0.384	0.319	0.263	0.227	0.156
TMI	0.475	0.460	0.269	0.240	0.210	0.170
SMI	0.477	0.463	0.270	0.241	0.208	0.170
MLMI	0.506	0.484	0.459	0.320	0.242	0.197
Scenario 3:	13 studies, 3 studies missing (≈20%)					
REF	0.529	0.521	0.363	0.355	0.192	0.173
CCA	0.524	0.514	0.368	0.367	0.200	0.172
TMI	0.521	0.517	0.322	0.314	0.191	0.173
SMI	0.521	0.515	0.329	0.319	0.190	0.174
MLMI	0.524	0.519	0.358	0.345	0.193	0.173
Scenario 4:	13 studies, 7 studies missing (≈50%)					
REF	0.521	0.521	0.368	0.365	0.193	0.168
CCA	0.482	0.463	0.356	0.341	0.200	0.155
TMI	0.497	0.497	0.255	0.238	0.192	0.164
SMI	0.497	0.498	0.259	0.238	0.192	0.167
MLMI	0.507	0.508	0.368	0.307	0.199	0.175
Two systematically missing predictors (x_1 and x_2)						
Scenario 5:	13 studies, 3 studies missing (≈20%)					
REF	0.541	0.535	0.368	0.368	0.110	0.178
CCA	0.530	0.524	0.367	0.372	0.105	0.169
TMI	0.532	0.530	0.337	0.330	0.111	0.169
SMI	0.533	0.527	0.341	0.333	0.112	0.169
MLMI	0.537	0.533	0.360	0.350	0.112	0.174
Scenario 6:	13 studies, 7 studies missing (≈50%)					
REF	0.531	0.527	0.354	0.354	0.109	0.170
CCA	0.481	0.479	0.344	0.332	0.086	0.163
TMI	0.510	0.510	0.288	0.275	0.098	0.148
SMI	0.511	0.508	0.295	0.288	0.099	0.149
MLMI	0.521	0.517	0.345	0.327	0.110	0.162

Note: REF indicates the results that were obtained *before* missingness was introduced and can be viewed as a benchmark for comparing the performance of methods that are applied *after* missingness is introduced: complete case analysis (CCA), traditional multiple imputation (TMI), stratified multiple imputation (SMI), and multilevel multiple imputation (MLMI). The following values are given: mean and median of the estimates.

was fitted using the *glmer* function of *lme4* in R [25], by adopting ML estimation. Finally, the parameter estimates from each imputed dataset were pooled using Rubin's rule [24].

We subsequently assessed the distinctive and relative merits of each imputation procedure by reporting the empirical mean of the parameter estimates, the relative bias, the root mean squared error (RMSE), and the coverage rate (CR) of nominal 95% CI for all model parameters of the analysis model. Because ML estimation is known to yield downwardly biased estimates of variance parameters when relatively

few studies are at hand [32, 33], the simulation setup values may not always reflect a realistic benchmark for comparing the performance of the imputation models. In particular, performance issues in the analysis model may not necessarily be related to the adopted imputation approach. For this reason, we also include an analysis model that is based on the full data; that is, the generated data *before* missingness was introduced (REF).

4.3. Results

4.3.1. Fixed effects (Table IV). No substantial bias was found in estimates of $\boldsymbol{\beta}$. In particular, the relative bias was below 3% for all imputation approaches and similar to REF (i.e., the mixed effect model that was estimated before missingness was introduced). We noticed lowered coverage rates for CCA, TMI, and SMI particularly in scenarios where few studies with full observations were at hand (e.g., scenarios 1, 2, 4, and 6). Coverage issues are also reflected by the RMSE, which was largest for CCA and tended to increase with an increase in missingness rate. For $\hat{\beta}_0$ and fully observed predictor $\hat{\beta}_2$, the RMSE of TMI, SMI, and MLMI tended to agree with the RMSE of REF. However, for systematically missing predictor $\hat{\beta}_1$, the RMSE substantially inflated for all approaches when x_1 was missing in many studies (scenarios 2, 4, and 6). In these scenarios, the CR of TMI declined to around 85%, while it was still above 90% for MLMI. In the other scenarios (scenarios 1, 3, and 5), the CR of MLMI achieved the nominal level and was always higher than the other approaches.

4.3.2. Random effects (Table V). As anticipated, some biases were found in estimates of τ_0, τ_1, and τ_2 when applying REF. This is likely related to the fact that ML estimation was used for fitting the statistical models, which is known to yield downwardly biased estimates of variance parameters. We therefore considered the estimates of REF as gold standard for comparing estimates of CCA, TMI, SMI, and MLMI.

For heterogeneity in the intercept term (i.e., τ_0), we noted a downward bias when adopting CCA, TMI, or SMI. This bias tended to increase with larger missingness rates (scenarios 2, 4, and 6). No evidence of such bias was found for MLMI, and corresponding estimates were very close to REF. Similar results were obtained for estimates of heterogeneity in β_1 (i.e., τ_1). For example, in scenario 4, TMI yielded an estimate for τ_1 of 0.255 as compared with 0.368 (REF), 0.356 (CCA), and 0.368 (MLMI). As an exception, τ_1 tended to be severely overestimated by MLMI (0.459 vs. 0.352 for REF) when few stud-

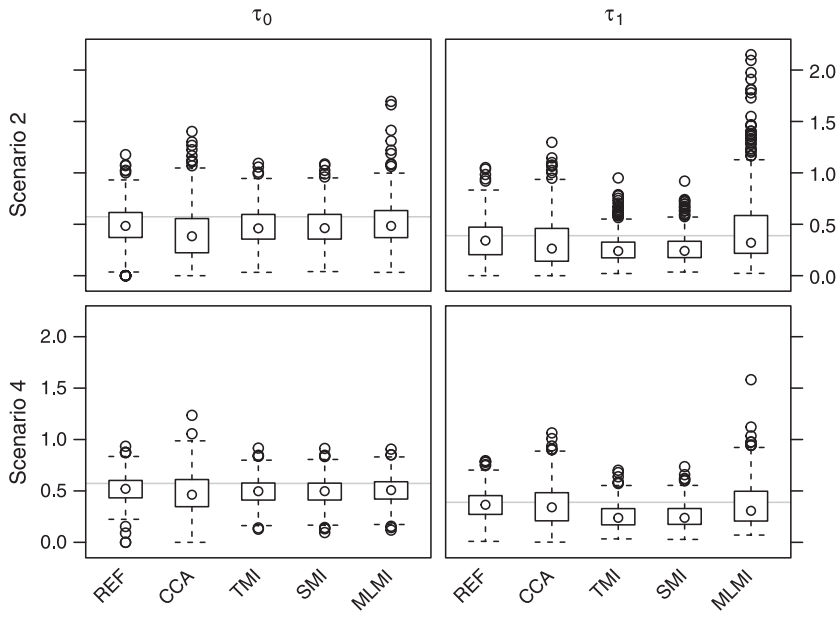

Figure 1. Boxplots of estimated random effects parameters. All boxplots are based on 500 simulation runs. The horizontal line indicates the true value of between-study heterogeneity. REF indicates the results that were obtained *before* missingness was introduced and can be viewed as a benchmark for the multiple imputation approaches: complete case analysis (CCA), traditional multiple imputation (TMI), stratified multiple imputation (SMI), and multilevel multiple imputation (MLMI).

ies were available (scenario 2). This discrepancy, however, disappeared when comparing the medians (0.320 for MLMI vs. 0.341 for REF) across all simulations of scenario 2. MLMI may thus sometimes lead to very extreme estimates of between-study heterogeneity for systematically missing predictors when few studies are available. This is also illustrated in Figure 1 where the skewness of random effects estimates for MLMI decreases when more studies are at hand (scenario 4). It is likely that for scenario 2, the imputation and analysis models were over-parameterized as they involved estimation of six heterogeneity parameters (i.e., $\mathbf{T}$) using information from merely three studies. Finally, for heterogeneity parameter τ_2, the bias was relatively low and only lead to substantial overestimation in scenario 2 (CCA and MLMI).

4.4. Conclusion

In general, none of the investigated methods resulted in biased estimates of fixed effect parameters. However, the CI coverage rates of the fixed effects estimates were marginally lower for the naive methods CCA, TMI, and SMI. For these methods, the random effect parameters were also underestimated, whereas MLMI maintained satisfactory performance. MLMI required, however, substantially more computation time to generate an imputed dataset as compared with TMI and SMI. In summary, simulations demonstrated that the implementation of MLMI for imputing systematically missing predictors lead to increased performance at the cost of computation time.

5. Discussion

Meta-analysis of multiple individual participant datasets has become increasingly popular for developing and validating risk prediction models. In general, IPD-MA may substantially improve the predictive performance of a developed model across different participant populations and allows to validate a developed model directly in a variety of different datasets [8, 9]. Unfortunately, the individual studies of an IPD-MA often provide to varying extents different predictors such that they may become systematically missing in one or more individual studies of the IPD-MA. Although several imputation strategies have been proposed for dealing with missing data within a single dataset [23], few solutions currently exist to accommodate for missing data in an entire study of an IPD-MA [12, 16].

Using both empirical and simulation data, we compared three different strategies to account for systematically missing predictors in an IPD-MA. The first strategy assumes MCAR for systematically missing predictors and performs CCA where studies with systematically missing predictors are excluded. The second strategy implements a naive imputation approach that assumes MAR and does not account for heterogeneity across studies (TMI). The third strategy also assumes MAR and allows for heterogeneity in predictor prevalences (SMI). The final strategy adopts our MLMI approach that fully accounts for heterogeneity across studies. In contrast to previously proposed approaches [16], our generalized approach for dealing with systematically missing predictors in an IPD-MA can directly be implemented to impute linear (e.g., continuous) and nonlinear (e.g., binary or count data) predictors. Furthermore, it no longer requires (typically unreliable) estimates of standard errors around between-study covariance parameters. Hereby, it also avoids the need to obtain second-order derivatives allowing the implementation of more efficient estimation procedures that require less computational resources (e.g., method-of-moments estimators).

5.1. Complete case analysis

In general, the implementation of CCA is only justified when data are MCAR. However, as we have demonstrated in our simulation study, results from CCA are suboptimal even when the MCAR assumption does hold. The main reason for this pitfall is that estimating the analysis model for CCA is difficult when few studies remain in the final dataset (see, for instance, simulation scenarios 1, 2, 4, and 6). In such scenarios, the analysis model may become over-parameterized and suffer from convergence issues. Although this problem could partially be resolved by adopting restricted ML or penalized quasi-likelihood

estimation [34, 35], these strategies are unlikely to fully overcome the lack of data, particularly when many studies are excluded.

As anticipated, the CCA approach became completely unreliable when the MCAR assumption did not hold. In our empirical example, it leads to a small subset of studies that were more related and did not show evidence of potential between-study heterogeneity. As a consequence, the generalizability of estimated predictor effects was limited, and the presence of between-study heterogeneity for systematically missing predictors could no longer be evaluated. In addition, the analysis model failed to converge, casting doubt on the validity of the obtained results.

In summary, the use of CCA is strongly discouraged when the missingness of predictors depends on observed data and may substantially hamper the development or validation of a prediction model. The use of CCA can also be unfavorable when missingness occurs completely at random, as the removal of many studies ($\geqslant$50%) may substantially hamper the estimation of standard errors and between-study heterogeneity.

5.2. Naive multiple imputation (TMI and SMI)

Although TMI and SMI adopt different assumptions about the clustering of subjects within studies, both approaches lead to similar results in our empirical example and simulation studies. We found that they tend to mask the actual degree of between-study heterogeneity and may lead to overoptimistic standard errors of predictor effects in the analysis models. These performance issues can be viewed as a direct consequence of ignoring (most elements of) between-study heterogeneity during imputation, leading to uncongeniality between the imputation and analysis model. In particular, the imputation models assume fixed effects for all regression coefficients (TMI) except the intercept term (SMI), such that imputed datasets become more homogeneous. Because SMI allows for heterogeneity in the prevalence of a missing predictor, it slightly outperforms TMI, which completely disregards between-study heterogeneity.

In summary, the use of SMI and particularly TMI is discouraged and may lead to a detrimental selection of important predictors, particularly when homogeneity of predictor effects is pursued. During model validation, they may lead to estimates of model performance that show little variation across studies (as heterogeneity of predictor effects is masked) and therefore incorrectly promote a model's generalizability [13].

5.3. Multilevel multiple imputation

We found that MLMI was the optimal approach in terms of coverage (predictor effects) and bias (between-study variability of predictor effects). Even when the MCAR assumption is justified (and CCA is a reasonable approach), MLMI still outperformed all other approaches. When the MCAR assumption is no longer justified, MLMI preserves a strong degree of between-study heterogeneity in predictor effects, and allows for more complicated (and congenial) analysis models. For this reason, MLMI models are crucial to safeguard the development and validation of generalizable prediction models when some predictors are systematically missing across individual studies in an IPD-MA.

The implementation of our proposed approach, that is, MLMI, however, also has some limitations. First, multilevel models involve many unknown parameters, particularly when random effects are assumed for all explanatory variables. As a consequence, it is possible that some imputation models run into convergence problems, leading to improper imputation (which may introduce bias in subsequent analyses). The number of random effects is particularly problematic when the number of studies with complete data on all predictors is limited (as seen in scenarios 2 and 4 in the simulation study). For this reason, Resche-Rigon *et al.* previously suggested to place random effects on the intercept term and additional random effects on the exposure(s) of interest. Further simplifications are possible by assuming independent random effects in the imputation model. Alternatively, it is possible to only impute those predictors for which at least, say, three studies are available that have included it. A second limitation of MLMI is

its reliance on ML estimation, which is known to yield downwardly biased estimates of variance parameters when few studies are included by the IPD-MA [32, 33]. This implies that multiply imputed datasets may not fully capture all relevant uncertainties and that subsequent analyses could still underestimate error variance and between-study heterogeneity. MLMI could therefore further be improved by adopting restricted ML estimation [34, 35]. Unfortunately, alternatives for ML estimation of nonlinear mixed effects models are still unavailable for many common software packages [33]. A third limitation of MLMI arises when imputing continuous systematically missing predictors. Because the study-specific error variance σ^2_{ik} cannot be estimated for studies where predictor k is systematically missing, MLMI assumes a common error variance term σ^2_k across all studies. This reduced flexibility of the covariance structure may degrade the coverage properties of MLMI [36], as was observed in the simulation study. Although the coverage rate of MLMI was always higher than TMI or SMI, and better than CCA in most scenarios, achieved levels were sometimes around 90%. It may therefore be advantageous to further extend MLMI and allow more flexibility in its covariance structure. The estimation of σ^2_{ik} could, for instance, be facilitated by assuming a relationship with study-level characteristics. Further research is needed to investigate whether such approach would yield accurate estimates of σ^2_{ik} and decrease the amount of noise in imputed datasets. Problems may also arise when imputing other types of systematically missing predictors. For instance, models for imputing binary predictors with very low success probabilities may suffer from sparse data biases [37–39] and degrade coverage rates. This problem was observed in the simulation study, where systematically missing predictor x_2 had a relatively low success probability (around 5–10%) and the coverage rate of $\hat{\boldsymbol{\beta}}_2$ substantially degraded. A fourth limitation of multilevel imputation models is that their estimation requires substantial computational power. In our simulation study, MLMI was about 100–150 times slower as compared with TMI. As a consequence, it may not always be feasible to generate a large number of imputed datasets. Although five imputations should suffice in many applications, it has previously been recommended to allow for 20–100 imputations [40]. This becomes particularly relevant when there is a large fraction of missing information or when imputation models are very complex. In order to facilitate the implementation of MLMI in real applications, it may be necessary to allow for parallelization. This can be achieved fairly straightforward with MLMI, as imputed datasets can be generated independently of one another. We applied this strategy in our empirical example, where 20 central processing units were used to simultaneously generate 20 imputations. A fifth limitation is that we did not employ a full Gibbs sampler for drawing parameters in MLMI. Instead, we used large sample approximations to the posterior distributions as these allow to substantially reduce the required computational time. A sixth limitation of MLMI is that an appropriate (ideally multivariate) distribution for the random effects must be chosen. Although it is quite common and computationally efficient to adopt an MVN distribution, this approach may not always be appropriate, particularly when random effects are skewed. Finally, implementation of the described multilevel model is only justified under the MAR and MCAR mechanisms. Further research is needed to develop imputation models that can be used when the probability of systematic missingness depends on unmeasured variables (i.e., when predictors are missing not at random).

5.4. Conclusions and recommendations

In conclusion, MLMI is a valuable addition to the current toolkit of approaches for dealing with missing data particularly when multiple individual datasets are being used such as in IPD-MA. We recommend its use when performing an IPD-MA with systematically missing predictors that are nonlinear and unlikely to be MCAR, rather than excluding individual studies with unmeasured predictors or applying traditional imputation methods that do not (fully) account for between-study heterogeneity. In situations when the number of studies is limited, or computational power is low, MLMI may still be feasible if the number of (joint) random effects is reduced. We tentatively suggest to generate 20–50 imputations and recommend the implementation of penalized estimation strategies when imputation models need to be applied in sparse data.

Appendix A: Symbols used

Symbol	Description	Dimension
$\boldsymbol{b}_{ik}$	Vector of random effects in study i when imputing predictor k (imputation model)	$Q \times 1$
M	Number of studies where x_{ijk} is fully observed	
N	Number of studies	
N_i	Number of subjects in study i	
$\mathbf{T}$	Variance–covariance matrix of random effects parameters (complete data model)	$L \times L$
$\mathbf{u}_i$	Vector of random effects in study i (complete data model)	$L \times 1$
$\mathbf{v}_{ij}$	Vector of variables associated with $\mathbf{u}_i$ (complete data model)	$L \times 1$
$\mathbf{w}_{ijk}$	Vector of variables associated with $\boldsymbol{b}_{ik}$ (imputation model)	$Q \times 1$
$\mathbf{x}_{ij}$	Vector of predictor values for subject j in study i	$K \times 1$
x_{ijk}	Value for subject j in study i for predictor k	
y_{ij}	Outcome for subject j in study i	
$\mathbf{z}_{ijk}$	Vector of covariates for imputing x_{ijk} (imputation model)	$P \times 1$
$\boldsymbol{\beta}$	Vector of fixed effects parameters (complete data model)	$K \times 1$
$\boldsymbol{\gamma}_k$	Vector of fixed effects parameters when imputing predictor k (imputation model)	$P \times 1$
$\boldsymbol{\theta}_k$	Collection of unknown parameters of the imputation model when imputing predictor k,with $\boldsymbol{\theta}_k = \{\boldsymbol{\gamma}_k, \boldsymbol{\Xi}_k\}$	
$\boldsymbol{\Lambda}_k$	Scale matrix parameter for generating samples of $\boldsymbol{\Psi}_k^{*-1}$ when imputing predictor k	$Q \times Q$
$\boldsymbol{\Xi}_k$	Matrix of (co)variance parameters in the imputation model when imputing predictor k.	
	Imputation of binary predictor ($\boldsymbol{\Xi}_k$ collapses to $\boldsymbol{\Psi}_k$).	$Q \times Q$
	Imputation of continuous predictor	$(Q+1) \times (Q+1)$
σ_k^2	Error variance when imputing continuous predictor k (imputation model)	
$\boldsymbol{\Psi}_k$	Variance–covariance matrix of random effects parameters when imputing predictor k (imputation model)	$Q \times Q$

Appendix B: R code

All source code of the simulation studies is available as Supplementary Material. We provide the main script of MLMI below.

Listing 1: mice.impute.2l.bin.r

```
require(lme4)
require(MASS)
require(mvtnorm)

mice.impute.2l.bin <- function(y, ry, x, type) {
  # the main code
  x <- data.frame(cbind(1, as.matrix(x)))
  names(x) <- paste("V",1:ncol(x),sep="")

  type <- c(2, type)

  clust <- names(x)[type==(-2)]
  rande <- names(x)[type==2]
  fixe <- names(x)[type>0]

  n.class <- length(unique(x[,clust]))
  x[,clust] <- factor(x[,clust], labels=1:n.class)
  lev <- levels(x[,clust])

  X <- as.matrix(x[,fixe])
```

```
Z <- as.matrix(x[,rande])
xobs <- x[ry,]
yobs <- y[ry]
Xobs <- as.matrix(X[ry,])
Zobs <- as.matrix(Z[ry,])

randmodel <- paste("yobs ~ ", paste(fixe[-1],
    collapse="+"), "+ (1 +", paste(rande[-1],collapse="+"), "|", clust, ")")
    #[-1] to remove intercept

suppressWarnings(fit <- try(glmer(formula(randmodel),
    data = data.frame(yobs,xobs), family =
    binomial),silent=T))
if(!is.null(attr(fit,"class"))){
  if(attr(fit,"class")=="try-error"){
    warning("glmer cannot be run, sorry!")
    return(y[!ry])
  }
}

# draw beta*
beta <- fixef(fit)
rv <- t(chol(vcov(fit)))
beta.star <- beta + rv %*% rnorm(ncol(rv))

# calculate psi*
rancoef <- as.matrix(ranef(fit)[[1]])
lambda <- t(rancoef)%*%rancoef
temp <- lambda
temp <- ginv(temp)
ev <- eigen(temp)
if(sum(ev$values<0)>0)
{
  ev$values[ev$values<0]<-0
  temp <- ev$vectors%*%diag(ev$values)%*%t(ev$vectors)
}
deco <- (ev$vectors)%*%sqrt(diag(ev$values))
temp.psi.star <- rWishart(1, nrow(rancoef),
    diag(nrow(lambda)))[,,1]
psi.star <- ginv(deco%*%temp.psi.star%*%t(deco))

#### psi.star positive definite?
if (!isSymmetric(psi.star)) psi.star <- (psi.star +
    t(psi.star))/2
valprop<-eigen(psi.star)
if(sum(valprop$values<0)>0)
{
  valprop$values[valprop$values<0]<-0
  psi.star <-
      valprop$vectors%*%diag(valprop$values)%*%t(valprop$vectors)
}

# the main imputation task
misindicator <- which((unique(x[,clust]) %in%
    unique(xobs[,clust])) == F)
```

```
68    for (i in misindicator){
69      suppressWarnings(bi.star <- t(rmvnorm(1,mean =
            rep(0,nrow(psi.star)), sigma = psi.star,
            meth="chol"))) # draw bi
70      logit <- X[!ry & x[,clust]==i,] %*% beta.star + Z[!ry
            & x[,clust]==i,]%*% bi.star
71      y[!ry & x[,clust]==i] <- rbinom(nrow(logit), 1,
            as.vector(1/(1 + exp(-logit))))
72    }
73    return(y[!ry])

74  }
```

As an example, consider a dataset with a binary outcome y and two binary systematically missing predictors x_1 and x_2. We denote the corresponding object as *data*. In the following example, *data* consist of four columns (*x1*, *x2*, *y*, and *study*) and 5000 rows (one row per subject). We can generate 20 imputed versions of *data* as follows:

```
1  library(mice)
2  source("mice.impute.2l.bin.r")
3
4  #MLMI
5  imp0 <- mice(data, print = F, maxit=0)
6  pred <- imp0$pred
7  pred[pred == 1] <- 2
8  pred[pred[,"study"] != 0,"study"] <- -2
9  imp.mlmi <- mice(data, print = F, pred=pred,
       method="2l.bin", m=20)
10
11 #SMI
12 data[,"study"] <- as.factor(data[,"study"])
13 data <- as.data.frame(data)
14 imp.smi  <- mice(data, print = F, method="logreg", m=20)
15
16 #TMI
17 imp0 <- mice(data, print = F, maxit = 0)
18 pred <- imp0$pred
19 pred[,"study"] <- 0
20 imp.tmi <- mice(data, print = F, pred = pred, meth =
       "logreg", m=20)
```

Appendix C: Simulation setup

For each scenario, we generated 13 studies with $N_1 = N_2 = \ldots = N_{13} = 500$ subjects. The binary outcomes y_{ij} were calculated using the following statistical model:

$$\begin{aligned}\text{logit}\{Pr(y_{ij}=1)\} &= -2.320748 + 1.111842\,x_{ij1} + 1.374697\,x_{ij2} \\ &\quad + u_{i0} + u_{i1}x_{ij1} + u_{i2}x_{ij2} \\ \mathbf{u}_i &\sim \text{MVN}\left(\mathbf{0}, \begin{bmatrix} 0.3282893 & -0.19195017 & -0.03652540 \\ -0.1919502 & 0.15152094 & 0.03935728 \\ -0.0365254 & 0.03935728 & 0.03472415 \end{bmatrix}\right)\end{aligned}$$

For each study, the predictors x_{ij1} and x_{ij2} were simulated from a bivariate binary process with marginal probability $\pi_i = (\pi_{ij1}, \pi_{ij2})^T$ and $\text{Corr}(x_{ij1}, x_{ij2}) = \rho_i$.

i	π_{ij1}	π_{ij2}	ρ_i
1	0.05252918	0.3025292	0.08223687
2	0.10565110	0.4336609	0.13471700
3	0.04575163	0.3856209	0.21211730
4	0.12756260	0.2425968	0.08226884
5	0.04804046	0.4083439	0.05391694
6	0.05116279	0.2818605	0.10789640
7	0.10955710	0.2237762	0.09817456
8	0.03692308	0.2861538	0.09264411
9	0.06254826	0.4293436	0.09163356
10	0.05963303	0.1513761	0.08280471
11	0.18299450	0.2994455	0.01414159
12	0.09090909	0.2072727	0.10354450
13	0.06798517	0.1891224	0.03258127

Appendix D: Required computation time

In the following are the absolute (abs., in seconds) and relative (rel.) computation times (averaged over 100 replications) needed by the CPU for generating one imputated dataset in the simulation studies. Results are based on a system with the following properties:

- Processor: Intel, Core i5-4670 CPU @ 3.40GHz
- RAM: 8 GB
- System type: 64-bit

Scenario	N	M	Approach	Time	
				abs.	rel.
1	6	5	TMI	0.0264	1.00
			SMI	0.0721	2.73
			MLMI	4.4948	170.26
2	6	3	TMI	0.0243	1.00
			SMI	0.0666	2.74
			MLMI	3.1522	129.72
3	13	10	TMI	0.0537	1.00
			SMI	0.2332	4.34
			MLMI	7.7229	143.82
4	13	6	TMI	0.0451	1.00
			SMI	0.2098	4.65
			MLMI	5.3681	119.03
5	13	10	TMI	1.56	1.00
			SMI	3.90	2.50
			MLMI	139.39	89.35
6	13	6	TMI	1.00	1.00
			SMI	2.46	2.46
			MLMI	136.60	136.60

Times represent *user* times, that is, the CPU time charged for the execution of the corresponding R scripts. TMI, traditional multiple imputation; SMI, stratified multiple imputation; MLMI, multilevel multiple imputation.

Acknowledgements

We thank M. Resche-Rigon and I. White for their valuable comments during the preparation of this manuscript. Furthermore, we gratefully acknowledge the following authors for sharing of the anonymized individual participant data from the deep vein thrombosis (DVT) studies for this pure methodological exercise: A. J. Ten Cate-Hoek, R. Oudega, R. E. G. Schutgens, D. R. Anderson, P. S. Wells, R. A. Kraaijenhagen, D. B. Toll, C. Kearon, J. L. Elf, S. M. Stevens, and S. M. Bates. Finally, we thank the anonymous reviewers and associate editor of Statistics in Medicine for their constructive feedback on this paper. This study was supported by the Netherlands Organisation for Scientific Research (9120.8004, 918.10.615, and 916.11.126).

References

1. Riley RD, Hayden JA, Steyerberg EW, Moons Karel GM, Abrams K, Kyzas PA, Malats N, Briggs A, Schroter S, Altman DG, Hemingway H. Prognosis research strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine* 2013; **10**(2):e1001380.
2. Steyerberg EW, Moons Karel GM, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* 2013; **10**(2):e1001381.
3. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; **98**(9):683–690.
4. Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine* 2004; **23**(6):907–926.
5. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, Maas AIR. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine* 2008; **5**(8):e165.
6. Schuit E, Kwee A, Westerhuis MEMH, Van Dessel HJHM, Graziosi GCM, Van Lith JMM, Nijhuis JG, Oei SG, Oosterbaan HP, Schuitemaker NWE, Wouters MGAJ, Visser GHA, Mol BWJ, Moons KGM, Groenwold RHH. A clinical prediction model to assess the risk of operative delivery. *BJOG: An International Journal of Obstetrics and Gynaecology* 2012; **119**(8):915–923.
7. Phillips RS, Sutton AJ, Riley RD, Chisholm JC, Picton SV, Stewart LA, the PICNICC Collaboration. Predicting infectious complications in neutropenic children and young people with cancer (IPD protocol). *Systematic Reviews* 2012; **1**(1):8.
8. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* 2013; **32**(18):3158–3180.
9. Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Medical Research Methodology* 2014; **14**(1):3.
10. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* 2008; **27**(5):625–650.
11. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine* 2013; **32**(26):4499–4514.
12. The Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine* 2009; **28**(8):1218–1237.
13. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; **68**(3):279–289.
14. Debray TPA, Koffijberg H, Lu Difei, Vergouwe Y, Steyerberg EW, Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Medical Research Methodology* 2012; **12**(1):121.
15. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* 2000; **19**(2):141–160.
16. Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* 2013; **32**(28):4890–4905.
17. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**(1):12–35.
18. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: Hoboken, N.J., 2002.
19. Schafer JL. *CRC Monographs on Statistics & Applied Probability* 1st ed. Chapman & Hall: London, 1997.
20. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specifications in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; **72**(12):1049–1064.
21. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007; **16**(3):219–242.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 2011; **30**(4):377–399.
23. van Buuren S, Oudshoorn K. Flexible multivariate imputation by MICE, Technical Report PG 99.054, TNO Prevention and Health, 1999.
24. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons: New York, 1987.

25. Bates D, Maechler M, Bolker B, Walker S. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. http://CRAN.R-project.org/package=lme4 [Accessed on November 2014].
26. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**(4): 538–558.
27. Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; **8**(1):3–15.
28. Oudega R, Moons KGM, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. *Thrombosis and Haemostasis* 2005; **94**(1):200–205.
29. Geersing GJ, Zuithoff NPA, Kearon C, Anderson DR, Ten Cate-Hoek AJ, Elf JL, Bates SM, Hoes AW, Kraaijenhagen RA, Oudega R, Schutgens REG, Stevens SM, Woller SC, Wells PS. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: individual patient data meta-analysis. *British Medical Journal* 2014; **348**:g1340.
30. van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**(3):1–67.
31. Debray TPA, Moons KGM, Abo-Zaid GMA, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?. *PLoS One* 2013; **8**(4):e60650.
32. Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *International Journal of Biostatistics* 2010; **6**(1).
33. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 2009; **24**:127–135.
34. Meza C, Jaffrzic F, Foulley J-L. REML estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm. *Biometrical Journal* 2007; **49**(6):876–888.
35. Noh M, Lee Y. REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis* 2007; **98**(5): 896–915.
36. Hox JJ, Roberts JK. *Handbook of Advanced Multilevel Analysis*. Routledge: New York, 2011.
37. Owen AB. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research* 2007; **8**:761–773.
38. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine* 2006Dec; **25**(24):4216–4226.
39. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology* 2000; **151**(5):531–539.
40. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 2007; **8**(3):206–213.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.