Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data

Peer-reviewed author version

VAN DER ELST, Wim; HERMANS, Lisa; VERBEKE, Geert; Kenward, Michael G.; NASSIRI, Vahid & MOLENBERGHS, Geert (2015) Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data. In: JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 86 (11), p. 2123-2139.

DOI: 10.1080/00949655.2015.1103738
Handle: http://hdl.handle.net/1942/20862

Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data

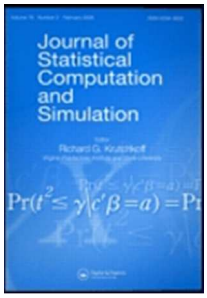
Article *in* Journal of Statistical Computation and Simulation · October 2015
DOI: 10.1080/00949655.2015.1103738

6 authors, including:

Wim Van der Elst
Janssen Pharmaceutica
78 PUBLICATIONS **1,625** CITATIONS
SEE PROFILE

Lisa Hermans
Hasselt University
6 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Geert Verbeke
KU Leuven
362 PUBLICATIONS **12,397** CITATIONS
SEE PROFILE



Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data

Journal:	*Journal of Statistical Computation and Simulation*
Manuscript ID	GSCS-2015-0050.R2
Manuscript Type:	Original Paper
Areas of Interest:	
2010 Mathematics Subject Classification:	62J99, 62P10

To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–22

Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Convergence problems often arise when complex linear mixed-effects models are fitted. Previous simulation studies (see e.g., [1, 2]) have shown that model convergence rates were higher (i) when the number of available clusters in the data increased, and (ii) when the size of the between-cluster variability increased (relative to the size of the residual variability).

The aim of the present simulation study is to further extend these findings by examining the effect of an additional factor that is hypothesized to affect model convergence, i.e., imbalance in cluster size. The results showed that divergence rates were substantially higher for datasets with unbalanced cluster sizes – in particular when the model at hand had a complex hierarchical structure. Further, the use of Multiple Imputation to restore 'balance' in unbalanced datasets reduces model convergence problems.

Keywords: simulation study, model convergence, mixed-effects model, multiple imputation, unbalanced data

AMS Subject Classification: 62J99; 62P10

1. Introduction

Fitting a linear mixed-effects model is typically done using Newton-Raphson or quasi-Newton based procedures (for details, see [3]). Based on some starting values for the parameters at hand, these procedures iteratively update the parameter estimates until sufficient convergence is achieved. Unfortunately, non-converging iteration processes often occur when complex linear mixed-effects models are considered. This means that the iterative process does not converge at all, or that it converges to values that are close to or outside the boundary of the parameter space (i.e., variances that are close to zero or negative, which may lead to a non-positive definite variance-covariance matrix of the random effects $\mathbf{D}$). Such problems mainly occur in complex models with many covariance components [4].

Motivating setting. As an example of a complex model with many covariance components where convergence is difficult to attain, consider the so-called surrogate endpoint evaluation setting. In a clinical trial, a surrogate endpoint is a replacement outcome for the true endpoint (i.e., the most credible indicator of treatment efficacy) that is useful when the latter endpoint is difficult to measure (e.g., infrequent, expensive, invasive, and/or distant in time) [1, 5, 6]. Obviously, the surrogate endpoint (S) can only replace the true endpoint (T) when it has been formally evaluated. Nowadays, the meta-analytic framework is commonly used to statistically evaluate the appropriateness of a candidate S [1, 5]. In this approach, it is assumed that information regarding S and T is available from multiple clinical trials (or from multiple other relevant units in which the patients

are clustered, such as hospitals or countries; [7]). Based on these data, the following linear mixed-effects model is fitted:

$$\begin{cases} S_{ij} = \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} + \varepsilon_{Sij} \\ T_{ij} = \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} + \varepsilon_{Tij} \end{cases}, \quad (1)$$

where S_{ij}, T_{ij} refer to the (jointly normally distributed) surrogate and true endpoints for subject j in cluster i; Z_{ij} is the binary treatment indicator for subject j in cluster i; μ_S, μ_T are the fixed intercepts for S and T; m_{Si}, m_{Ti} are the corresponding random intercepts; α, β are the fixed treatment effects for S and T; a_i, b_i are the corresponding random effects, and ε_{Sij}, ε_{Tij} are the error terms for S and T, respectively. It is assumed that $(m_{Si}, m_{Ti}, a_i, b_i) \sim N(\mathbf{0}, \mathbf{D})$ and $(\varepsilon_{Sij}, \varepsilon_{Tij}) \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{D}$ and $\mathbf{\Sigma}$ are unstructured variance-covariance matrices.

The appropriateness of a candidate surrogate is quantified by two metrics. First, the coefficient of trial-level surrogacy (R^2_{trial}), which essentially quantifies the strength of the association between the random treatment effects on S and T based on the variance-covariance matrix of the random effects $\mathbf{D}$, i.e., $R^2_{trial} = \frac{d^2_{ab}}{d_{aa}d_{bb}}$. Second, the coefficient of individual-level surrogacy (R^2_{indiv}), which quantifies the treatment- and trial-corrected strength of association between S and T based on the variance-covariance matrix of the residuals $\mathbf{\Sigma}$, i.e., $R^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}$.

Earlier simulation studies. In a surrogate evaluation context, the (S, T) endpoints (level 1) are nested within patients (level 2), and the patients are nested within clinical trials (or other relevant clustering units; level 3). Given the relatively complex hierarchical structure of the data, it is hardly surprising that convergence problems are frequently encountered in a surrogate evaluation context.

To gain more insight into the factors that affect model convergence, [1] and [2] conducted a number of simulation studies. Their conclusion was that model convergence rates were higher when the number of available clusters increased and when the size of the between-cluster variability $\mathbf{D}$ increased relative to the residual variability $\mathbf{\Sigma}$ (in particular the d_{aa}, d_{bb} components relative to the σ_{SS}, σ_{TT} components). Other factors, such as the number of patients per cluster, the normality assumption (for S and T), and the strength of the correlation between the random treatment effects had no substantial impact on model convergence.

Aim of the present study. The aim of the present study is to further extend the results of the earlier simulation studies [1, 2] by evaluating the effect of an additional factor that may affect model convergence, i.e., imbalance in cluster size. Indeed, in the earlier simulation studies, *balanced* datasets were considered, i.e., all clusters had exactly the same number of observations. However, in real-life datasets, it is nearly always the case that the cluster sizes are unbalanced. In fact, even when a balanced design was initially planned, the actually collected data will often be unbalanced due to, for example, missingness.

To understand why balance in cluster size may be a relevant factor to consider, recall that the key computational difficulty in fitting mixed-effects models is in the estimation of the covariance parameters [4]. Iterative numerical optimization of the log-likelihood functions using (RE)ML estimation is conducted, subject to constraints that are imposed on the model parameters to ensure positive definiteness of the $\mathbf{D}$ and $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{\Sigma}_i$ matrices (where $\mathbf{Z}_i$ are matrices of known covariates associated with the random

effects). Notice that positive definiteness of both $\mathbf{D}$ and $\mathbf{V}_i$ is not needed when one is merely interested in the marginal model. In the latter case, the only requirement for valid inference based on the marginal model is that the overall $\mathbf{V}$ matrix is positive definite (see Discussion). To maximize complicated likelihoods or to find good starting values that can subsequently be used in the Newton-Raphson algorithm, the Expectation Maximization (EM) algorithm is often used [8]. When *unbalanced* data are considered, the E-step involves, at least conceptually, the creation of a 'balanced' dataset (or a 'complete' dataset in a missing data context [9]) based on a hypothetical scenario where it is assumed that data have been obtained from a balanced design (or from a study in which there were no missing values in a missing data context [4, 10]). Based on the 'balanced' data, an objective function is constructed and maximized in the M-step, and the parameter estimates are subsequently iteratively updated. In essence, the underlying assumption behind the EM algorithm is that the optimization of the balanced (complete) data log-likelihood function is easier than the optimization of the unbalanced (observed) data log-likelihood [10]. In the same spirit, it can be expected that model convergence issues will occur more frequently when the actually observed data are unbalanced in cluster size, compared to the setting where the actually observed data are balanced. The first aim of the present study is to examine this hypothesis.

The second aim of the present study is to examine whether the convergence rates of unbalanced data could be increased by using Multiple Imputation (MI) prior to fitting the mixed-effects model. Based on the reasoning in the previous paragraph, it was expected that the use of MI (to make an unbalanced dataset 'balanced') would lead to higher convergence rates.

Organization of the paper. The remainder of this paper is organized as follows. In Section 2, the impact of imbalance in cluster size on the convergence rates of mixed-effects models is evaluated. In Section 3, the impact of using MI to introduce balance in unbalanced datasets on model convergence rates is examined and the statistical properties of the estimators are evaluated. In Section 4, a case study (the age-related macular degeneration trial) is analyzed. Finally, Section 5 summarizes the results and discusses some of the limitations of the present paper.

2. Unbalanced data and model convergence

Three mixed-effects models with an increasing level of complexity were considered: (i) a random-intercept model, (ii) a reduced surrogate evaluation model (i.e., a simplification of Model (1) where the fixed- and random-treatment effects are discarded), and (iii) a surrogate evaluation model (Model (1)). The idea is to gradually build-up the (hierarchical) complexity of the model, so that it can be examined at which level of complexity the impact of an imbalance in cluster size on model convergence becomes apparent.

2.1. *Outcomes of interest*

The key outcome of interest in the simulations was model convergence. Three model convergence categories were distinguished: (i) proper convergence, i.e., the model converged and the variance-covariance matrix of the random effects ($\mathbf{D}$) and the final Hessian ($\mathbf{H}$), used to compute the standard errors of the covariance parameters, were positive definite (PD); (ii) the model converged but $\mathbf{D}$ or $\mathbf{H}$ was not PD; and finally, (iii) divergence.

In addition, the number of required iterations to achieve convergence was recorded and analyzed.

2.2. *Simulation design*

2.2.1. Random-intercept model

Consider the following random-intercept model:

$$S_{ij} = \mu_S + m_{Si} + \varepsilon_{Sij}, \tag{2}$$

where S_{ij} is a (normally distributed) endpoint for patient j in cluster i, μ_S is the fixed intercept, m_{Si} is the corresponding random intercept, and ε_{Sij} is the error term. It is assumed that $m_{Si} \sim N(0,\, d)$ and $\varepsilon_{Sij} \sim N(0,\, \sigma_{SS})$.

In all simulations, $\mu_S = 450$, $\sigma_{SS} = 300$, and the mean sample size per cluster $M(n_i) = 20$. Three conditions were varied. First, the number of clusters $i = 1, 2, ..., N$, with $N = \{5,\, 10,\, 20,\, 50\}$. Second, the level of imbalance in cluster size (n_i). In the balanced scenario, all cluster sizes were equal, i.e., $n_i = n = 20$. In the two unbalanced scenarios, $\widetilde{n}_i$ was determined based on a draw from a normal distribution and rounded to the nearest integer (i.e., $n_i = \text{round}\,(\widetilde{n}_i)$). In the low imbalance scenario, $\widetilde{n}_i \sim N\left(20,\, 2.5^2\right)$. In the high imbalance scenario, $\widetilde{n}_i \sim N\left(20,\, 5^2\right)$. Third, the between-cluster variability ($d = \gamma(1000)$), which is either large ($\gamma = 1$) or small ($\gamma = 0.1$) relative to the residual variability ($\sigma_{SS} = 300$). There were thus a total of 24 possible scenarios, for each of which 1000 datasets were generated. These dataset were subsequently analyzed by fitting Model (2) using the mixed procedure in SAS, and model convergence and the required number of iterations were recorded (see Section 2.1).

2.2.2. Reduced surrogate evaluation model

Consider the following linear mixed-effects model:

$$\begin{cases} S_{ij} = \mu_S + m_{Si} + \varepsilon_{Sij} \\ T_{ij} = \mu_T + m_{Ti} + \varepsilon_{Tij} \end{cases}, \tag{3}$$

where S_{ij} and T_{ij} are (normally distributed) endpoints for patient j in cluster i (e.g., a surrogate and true endpoint); μ_S, μ_T are the fixed intercepts for S and T; m_{Si}, m_{Ti} are the corresponding random intercepts; and ε_{Sij}, ε_{Tij} are the error terms. It is assumed that $(m_{Si},\, m_{Ti}) \sim N\,(\mathbf{0},\, \mathbf{D})$ and $(\varepsilon_{Sij},\, \varepsilon_{Tij}) \sim N\,(\mathbf{0},\, \mathbf{\Sigma})$, where $\mathbf{D}$ and $\mathbf{\Sigma}$ are unstructured 2 by 2 variance-covariance matrices. As can be seen, Model (3) is a simplification of Model (1), where the fixed- and random- treatment effects are omitted.

Using Model (3), data were simulated. In all simulations, $\mu_S = 450$, $\mu_T = 500$, and

$$\mathbf{\Sigma} = \begin{pmatrix} 300 & 212.132 \\ 212.132 & 300 \end{pmatrix},$$

yielding $\text{corr}\,(\varepsilon_{Sij},\, \varepsilon_{Tij})^2 = 0.5$. Three conditions were varied in the simulations. First, the number of clusters $N = \{5,\, 10,\, 20,\, 50\}$. Second, the level of imbalance in n_i (see Section 2.2.1). Third, the between-cluster variability ($\mathbf{D}$), which is either large ($\gamma = 1$) or small ($\gamma = 0.1$) relative to the residual variability ($\mathbf{\Sigma}$):

$$\mathbf{D} = \gamma \begin{pmatrix} 1000 & 400 \\ 400 & 1000 \end{pmatrix}.$$

For each of the 24 settings a total of 1000 datasets were generated, and model convergence and the required number of iterations were recorded (see Section 2.1).

2.2.3. Surrogate evaluation model

Using the linear mixed-effects Model (1) detailed above, data were simulated. Notice that Z_{ij} was coded as -1 = control treatment and 1 = experimental treatment (rather than as 0 = control treatment and 1 = experimental treatment), and thus the fixed treatment effects on S and T are 2α and 2β, respectively. This was done because a 0/1 coding, for a positive-definite $\mathbf{D}$ matrix, forces the variability in the experimental arm to be greater than or equal to the variability in the control arm. A $-1/1$ coding on the other hand ensures that the same components of variability operate in both treatment arms [5].

In all simulations, $\mu_S = 450$, $\mu_T = 500$, $\alpha = 300$, $\beta = 500$, and

$$\mathbf{\Sigma} = \begin{pmatrix} 300 & 212.132 \\ 212.132 & 300 \end{pmatrix},$$

yielding $R^2_{indiv} = \text{corr}\left(\varepsilon_{Sij},\, \varepsilon_{Tij}\right)^2 = 0.5$. Again, three conditions were varied in the simulations. First, the number of clusters $N = \{5,\, 10,\, 20,\, 50\}$. Second, the level of imbalance in n_i (see Section 2.2.1). Third, the between-cluster variability ($\mathbf{D}$), which is either large ($\gamma = 1$) or small ($\gamma = 0.1$) relative to the residual variability ($\mathbf{\Sigma}$):

$$\mathbf{D} = \gamma \begin{pmatrix} 1000 & 400 & 0 & 0 \\ 400 & 1000 & 0 & 0 \\ 0 & 0 & 1000 & 707.107 \\ 0 & 0 & 707.107 & 1000 \end{pmatrix},$$

yielding $R^2_{trial} = \text{corr}\left(a_i,\, b_i\right)^2 = 0.5$. Further, in the balanced scenario, treatment (Z) is also balanced within a cluster. In the unbalanced scenarios, treatment allocation is determined based on a binomial distribution with success probability 0.50. A total of 1000 datasets were generated for each of the 24 settings. The generated data were subsequently analyzed by fitting Model (1) in SAS. Two different parametrizations for the $\mathbf{D}$ matrix were considered. First, a completely general (unstructured; UN) $\mathbf{D}$ matrix that is parameterized directly in terms of variances and covariances. Second, a non-diagonal factor-analytic structure with 4 factors (FA0(4)). The latter structure specifies a Cholesky root parametrization for the 4×4 unstructured blocks in $\mathbf{D}$. This leads to a substantial simplification of the optimization problem, i.e., the problem now changes from a constrained one to an unconstrained one. The FA0(4) structure has $\frac{q}{2}\left(2t - q + 1\right)$ covariance parameters, where q refers to the number of factors and t is the dimension of the matrix. In the present setting, the FA0(4) structure thus has a total of 10 parameters. These parameters are used to compute the components in $\mathbf{D}$, i.e., the $(i,\, j)^{th}$ element of $\mathbf{D}$ is computed as $\Sigma_{k=1}^{\min(i,j,k)} \lambda_{ik}\lambda_{jk}$. The Cholesky root parametrization ensures that $\mathbf{D}$ (and $\mathbf{V}_i$) is positive definite during the entire estimation process [10]. Model convergence and

the required number of iterations were recorded (see Section 2.1).

2.3. *Results*

As shown in Table 1, the rates of proper convergence exceeded 0.960 and 0.841 in the various scenarios for the random-intercept and reduced surrogate models, respectively. Overall convergence (i.e., proper convergence or convergence but non-PD **D** or **H** matrix) was 100% for the random-intercept model and $\geq 97.8\%$ for the reduced surrogate model. The rates of proper and overall convergence were similar in all scenarios, irrespective of the level of imbalance in cluster size (n_i). However, a larger level of imbalance in cluster size was associated with a higher mean number of iterations that were required to achieve proper convergence for both the random-intercept and the reduced surrogate models (see Table 2). This suggests that the optimization of the log-likelihood function is more difficult in the unbalanced scenarios, even when relatively simple (in terms of their hierarchical structure) mixed-effects models are considered.

When an unstructured (UN) **D** matrix was used in the surrogate evaluation models, overall convergence exceeded 99.7% when cluster sizes were balanced (see Table 1). The overall convergence rates were, however, substantially lower in the unbalanced scenarios, in particular when N and γ were small. For example, when $N = 5$ and $\gamma = 0.1$, the model divergence rates were as high as 65.5% and 77.3% in the small and large imbalance scenarios, respectively (compared to only 0.3% in the balanced scenario). At the same time, the impact of level of imbalance on proper convergence was small, i.e., proper convergence rates were quite similar in all scenarios irrespective of the level of imbalance in the data. When a non-diagonal factor analytic structure with 4 factors (FA0(4)) was used for the **D** matrix in the surrogate evaluation models, the rates of proper convergence exceeded 71.0% in all scenarios (see Table 1) and were thus substantially higher compared to those that were observed in the UN scenario. Further, divergence rates were substantially lower in the unbalanced FA0(4) scenarios compared to those in the unbalanced UN scenarios. As was also observed for the random-intercept and the reduced surrogate models, a higher level of imbalance in cluster size was associated with a larger mean number of required iterations to achieve proper convergence for both the UN and FA0(4) surrogate evaluation models (see Table 2). A noteworthy observation is that proper convergence was *always* achieved after 1 iteration when the cluster size was balanced for the random-intercept, reduced surrogate, and UN surrogate evaluation models (Table 2). The reverse also holds approximately, i.e., when a model converged after 1 iteration, in more than 99.9% of the cases there was proper convergence. In contrast to what was the case for the UN surrogate evaluation models, proper convergence was not always achieved after 1 iteration for the FA0(4) surrogate evaluation models. Nonetheless, the number of required iterations to achieve proper convergence was also substantially lower in the balanced FA0(4) surrogate evaluation models compared to what was the case in the unbalanced FA0(4) surrogate evaluation models.

Table 1. Convergence rates for the random-intercept models, reduced surrogate models, and surrogate models as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability.

Model			Balanced (equal n_i)				Small imbalance $\widetilde{n}_i \sim N\left(20,\, 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\, 5^2\right)$			
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Random intercept	Proper convergence	Small (0.1)	0.974	0.999	1	1	0.969	0.999	1	1	0.960	0.998	1	1
		Large (1)	0.999	1	1	1	0.997	1	1	1	0.989	1	1	1
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.026	0.001	0	0	0.031	0.001	0	0	0.040	0.002	0	0
		Large (1)	0.001	0	0	0	0.003	0	0	0	0.011	0	0	0
	Divergence	Small (0.1)	0	0	0	0	0	0	0	0	0	0	0	0
		Large (1)	0	0	0	0	0	0	0	0	0	0	0	0
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Reduced surrogate model	Proper convergence	Small (0.1)	0.853	0.991	1	1	0.853	0.990	1	1	0.841	0.987	1	1
		Large (1)	0.990	1	1	1	0.992	1	1	1	0.995	1	1	1
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.147	0.009	0	0	0.141	0.010	0	0	0.137	0.013	0	0
		Large (1)	0.010	0	0	0	0.007	0	0	0	0.004	0	0	0
	Divergence	Small (0.1)	0	0	0	0	0.006	0	0	0	0.022	0	0	0
		Large (1)	0	0	0	0	0.001	0	0	0	0.001	0	0	0
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Surrogate model, UN	Proper convergence	Small (0.1)	0.112	0.826	0.999	1	0.090	0.816	1	1	0.091	0.800	0.999	1
		Large (1)	0.555	0.998	1	1	0.566	0.999	1	1	0.540	0.996	1	1
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.885	0.174	0.001	0	0.255	0.171	0	0	0.136	0.155	0.001	0
		Large (1)	0.444	0.002	0	0	0.163	0.001	0	0	0.119	0.001	0	0
	Divergence	Small (0.1)	0.003	0	0	0	0.655	0.013	0	0	0.773	0.045	0	0
		Large (1)	0.001	0	0	0	0.271	0	0	0	0.341	0.003	0	0
Surrogate model, FA0(4)	Proper convergence	Small (0.1)	0.745	0.984	1	1	0.717	0.976	1	1	0.710	0.976	0.999	1
		Large (1)	0.931	1	1	1	0.931	0.994	0.997	0.998	0.935	0.992	0.998	0.999
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.068	0.007	0	0	0.054	0.016	0	0	0.044	0.015	0	0
		Large (1)	0.030	0	0	0	0.027	0.001	0.002	0	0.018	0.004	0.001	0.001
	Divergence	Small (0.1)	0.187	0.009	0	0	0.229	0.008	0	0	0.246	0.009	0.001	0
		Large (1)	0.039	0	0	0	0.042	0.005	0.001	0.002	0.047	0.004	0.001	0

Note. UN = unstructured; FA0(4) = factor analytic.

Table 2. Mean (SD) number of iterations per convergence category for the random-intercept models, the reduced surrogate models, and the surrogate models as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability.

Model			Balanced (equal n_i)				Small imbalance $\widetilde{n}_i \sim N\left(20,\, 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\, 5^2\right)$			
	Convergence category	Between-cluster variability γ	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Random intercept	Proper convergence	Small (0.1)	1 (0)	1 (0)	1 (0)	1 (0)	2.18 (0.85)	2.22 (0.80)	2.09 (0.73)	1.88 (0.63)	2.70 (1.06)	2.71 (0.95)	2.52 (0.87)	2.32 (0.78)
		Large (1)	1 (0)	1 (0)	1 (0)	1 (0)	2.91 (1.31)	2.95 (1.14)	2.82 (1.02)	2.56 (0.84)	3.70 (1.89)	3.85 (1.85)	3.64 (1.68)	3.13 (1.04)
	Convergence but non-PD $\mathbf{D}/\mathbf{H}$ matrix	Small (0.1)	1 (0)	1 (−)	–	–	1 (0)	1 (−)	–	–	1 (0)	1 (0)	–	–
		Large (1)	1 (−)	–	–	–	1 (0)	–	–	–	1 (0)	–	–	–
	Divergence	Small (0.1)	–	–	–	–	–	–	–	–	–	–	–	–
		Large (1)	–	–	–	–	–	–	–	–	–	–	–	–
	Convergence category	Between-cluster variability γ	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Reduced surrogate model	Proper convergence	Small (0.1)	1 (0)	1 (0)	1 (0)	1 (0)	2.67 (0.78)	2.59 (0.70)	2.35 (0.58)	2.08 (0.44)	3.51 (1.02)	3.37 (0.89)	3.06 (0.77)	2.66 (0.65)
		Large (1)	1 (0)	1 (0)	1 (0)	1 (0)	4.06 (1.78)	4.12 (1.72)	3.84 (1.35)	3.32 (0.79)	5.36 (2.30)	5.42 (2.32)	5.11 (2.36)	4.28 (1.51)
	Convergence but non-PD $\mathbf{D}/\mathbf{H}$ matrix	Small (0.1)	4.28 (1.61)	3.22 (0.67)	–	–	4.38 (1.47)	4.10 (0.88)	–	–	4.46 (1.44)	4.15 (1.46)	–	–
		Large (1)	4.20 (1.40)	–	–	–	4.14 (1.35)	–	–	–	5.50 (2.53)	–	–	–
	Divergence	Small (0.1)	–	–	–	–	36.17 (6.31)	–	–	–	50.46 (58.29)	–	–	–
		Large (1)	–	–	–	–	44.00 (−)	–	–	–	36.00 (−)	–	–	–
	Convergence category	Between-cluster variability γ	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model, UN	Proper convergence	Small (0.1)	1 (0)	1 (0)	1 (0)	1 (0)	4.99 (1.14)	4.30 (0.85)	3.70 (0.73)	3.11 (0.58)	5.34 (1.12)	4.72 (0.86)	4.15 (0.70)	3.52 (0.67)
		Large (1)	1 (0)	1 (0)	1 (0)	1 (0)	7.68 (2.60)	6.73 (2.49)	5.84 (2.43)	4.69 (1.75)	8.21 (2.40)	7.35 (2.67)	6.67 (2.63)	5.54 (2.32)
	Convergence but non-PD $\mathbf{D}/\mathbf{H}$ matrix	Small (0.1)	9.16 (9.06)	3.97 (1.37)	4 (−)	–	6.85 (2.13)	5.34 (1.34)	–	–	6.73 (1.83)	5.56 (1.32)	5 (−)	–
		Large (1)	6.35 (3.50)	3.50 (2.12)	–	–	9.90 (2.71)	8.00 (−)	–	–	9.57 (2.61)	8.00 (−)	–	–
	Divergence	Small (0.1)	309.33 (306.91)	–	–	–	50.10 (16.51)	49.00 (12.17)	–	–	47.66 (15.01)	48.81 (16.05)	–	–
		Large (1)	205.00 (−)	–	–	–	50.79 (19.71)	–	–	–	49.40 (15.39)	44.00 (2.00)	–	–
Surrogate model, FA0(4)	Proper convergence	Small (0.1)	4.90 (3.90)	1.66 (1.87)	1 (0.09)	1 (0)	6.00 (2.99)	4.37 (3.16)	2.87 (0.62)	2.36 (0.49)	6.45 (3.31)	4.94 (3.38)	3.29 (1.20)	2.77 (0.52)
		Large (1)	2.24 (1.19)	1.01 (0.16)	1 (0)	1 (0)	9.54 (7.69)	6.36 (3.42)	5.67 (3.33)	4.65 (2.27)	10.87 (8.57)	6.98 (4.54)	6.27 (3.08)	5.40 (2.61)
	Convergence but non-PD $\mathbf{D}/\mathbf{H}$ matrix	Small (0.1)	6.12 (1.94)	5.71 (2.21)	–	–	6.91 (2.05)	5.69 (1.45)	–	–	6.55 (1.30)	5.80 (1.93)	–	–
		Large (1)	5.13 (1.68)	–	–	–	11.85 (7.48)	20 (−)	23.50 (0.71)	–	13.39 (14.14)	24 (16.67)	23 (−)	38 (−)
	Divergence	Small (0.1)	21.96 (13.84)	30.78 (12.13)	–	–	19.41 (11.75)	18.25 (9.77)	–	–	20.10 (12.88)	32.33 (18.32)	49 (0)	–
		Large (1)	31.67 (11.43)	–	–	–	29.57 (13.46)	27 (11.98)	11 (−)	17 (5.66)	27.87 (11.82)	29.50 (21.61)	9 (−)	–

Note. UN = unstructured; FA0(4) = factor analytic.

Table 3. Hypothetical dataset. Number of observations per cluster as a function of treatment (Z), before and after imputation.

	Before imputation		After imputation	
Cluster	$Z=-1$	$Z=1$	$Z=-1$	$Z=1$
1	5	11	18	18
2	13	8	18	18
3	10	18	18	18
4	9	5	18	18
5	9	11	18	18

3. Multiple imputation

3.1. *Simulation design*

The same unbalanced datasets that were generated in the surrogate evaluation model scenario described above (see Section 2.2.3) were considered here. In these unbalanced datasets, multiple imputation (MI) was used to introduce balance in terms of cluster size and treatment allocation (Z). As an example of what is meant by this, consider the hypothetical dataset with 5 clusters shown in Table 3. As can be seen, the maximum number of patients for each of the cluster by treatment (Z) groups is 18. Thus for all cluster $\times$ treatment groups having less than 18 observations, MI is used to restore balance. For example, in cluster 1 there were 5 observations for S and T in treatment group $Z=-1$ and 11 observations in treatment group $Z=1$. Thus, the data of 13 and 7 patients are imputed in cluster 1 for $Z=-1$ and $Z=1$, respectively.

Proper multivariate imputations were conducted using the default Markov chain Monte Carlo method [11] in SAS with a non-informative prior (Jeffreys). The imputation model included S, T, and Z. The imputation model was run 'by cluster'. This is a valid approach here, provided that the number of clusters increases sufficiently slowly relative to the number of subjects per cluster. A total of 200 burn-in iterations were used (i.e., the number of initial iterations before the first iteration for a chain), and the number of iterations between the imputations in a chain equalled 100. A total of 3 imputations were conducted for each dataset. Thus in total, 24000 datasets were considered in the analyses (i.e., 4 (number of N) $\cdot$ 2 (number of γ) $\cdot$ 1000 (number of runs) $\cdot$ 3 (number of imputations)) for both the small imbalance and the large imbalance scenarios.

The 'balanced' data were subsequently analyzed by fitting Model (1) using the UN and FA0(4) parametrizations for the $\mathbf{D}$ matrix (see Section 2.2.3). The outcomes of interest were model convergence and the number of required iterations to achieve convergence (see Section 2.1). In addition, the bias, efficiency (standard deviation of the estimate) and Mean Squared Errors of the estimates of the R^2_{trial} and R^2_{indiv} metrics were evaluated for the surrogate evaluation models with MI using an unstructured (UN MI) and non-diagonal factor analytic structure with 4 factors (MI FA0(4)) for the $\mathbf{D}$ matrix. The focus was on R^2_{trial} and R^2_{indiv} rather than on the fixed-effects and variance components because the coefficients of determination are the main quantities of interest in a surrogate evaluation context.

3.2. *Results*

Table 4 shows the convergence rates for the MI UN and MI FA0(4) models. Compared to what was the case for the surrogate evaluation models in the unbalanced non-MI scenarios, the rates of proper and overall convergence were substantially higher in both the MI UN and MI FA0(4) scenarios – and this was particularly so when N was small. The use of MI to make the unbalanced data balanced was thus a successful strategy to improve convergence and reduce divergence. In line with the results discussed in Section 2.3, proper convergence was *always* achieved after 1 iteration in the MI UN scenario but not in the MI FA0(4) scenario (see Table 5). Nonetheless, the number of required iterations to achieve convergence was substantially reduced in the unbalanced MI FA0(4) scenarios compared to what was the case in the non-MI unbalanced scenarios (compare Tables 2 and 5).

Tables 6 and 7 show the bias, efficiency, and MSE of the estimates of R^2_{indiv} and R^2_{trial}, respectively, in the non-MI and MI settings that properly converged. As expected, the bias, efficiency and MSE in the estimation of both R^2_{indiv} and R^2_{trial} improved when the number of clusters increased. With respect to the estimation of R^2_{indiv}, the bias was low in all scenarios but the efficiency and MSE were poorer in the MI scenarios compared to the non-MI scenarios. In contrast to R^2_{indiv}, the bias, efficiency and MSE were of similar magnitude in the MI and non-MI scenarios. Only when $N = 5$ did the bias in the estimation of R^2_{trial} tend to be substantially higher in the MI scenario compared to the non-MI scenario. Note that the bias was negative in all scenarios, indicating that the true R^2_{trial} tends to be somewhat underestimated.

Further, the bias and MSE in the estimation of R^2_{indiv} was smaller compared to what was observed for R^2_{trial}, and the efficiency somewhat lower, because there is less replication than for the individual level quantity.

4. Case study

The results in Section 3 indicated that the use of MI to balance an unbalanced dataset (prior to fitting the mixed-effects model) reduces model convergence issues. In this section, this method is applied to a real-life dataset, the age-related macular degeneration trial.

4.1. *An age-related macular degeneration trial*

Age-related macular degeneration (ARMD) is a condition in which patients progressively lose vision [12]. In the ARMD trial, patients were randomly allocated to two treatment conditions: placebo and the experimental treatment (interferon-α). Treatment efficacy was evaluated using changes in visual acuity over time. Visual acuity was measured as the total number of letters that were correctly read using standardized vision charts.

The ARMD trial was analyzed earlier in a surrogate evaluation context [5, 13]. The idea is to examine whether the change in visual acuity after 24 weeks is an appropriate surrogate for the change in visual acuity after 52 weeks. The ARMD data are included in the R library *Surrogate*, which can be downloaded at `http://cran.r-project.org/web/packages/Surrogate/index.html`.

Table 4. Convergence rates for the surrogate MI UN (unstructured) and MI FA0(4) (factor analytic) models as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability.

Model			Small imbalance $\widetilde{n}_i \sim N\left(20,\ 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\ 5^2\right)$			
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	5	10	20	50
Surrogate model, MI UN	Proper convergence	Small (0.1)	0.164	0.941	1	1	0.189	0.946	0.999	0.999
		Large (1)	0.617	0.999	1	1	0.616	0.999	0.999	1
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.811	0.059	0	0	0.770	0.054	0.001	0.001
		Large (1)	0.365	0.001	0	0	0.343	0.001	0.001	0
	Divergence	Small (0.1)	0.025	0	0	0	0.042	0	0	0
		Large (1)	0.018	0	0	0	0.041	0	0	0
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	5	10	20	50
Surrogate model, MI FA0(4)	Proper convergence	Small (0.1)	0.813	0.998	1	1	0.817	0.996	0.999	0.999
		Large (1)	0.954	0.999	1	1	0.950	1	0.999	1
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	0.077	0.002	0	0	0.072	0.003	0	0
		Large (1)	0.012	0.001	0	0	0.022	0	0.001	0
	Divergence	Small (0.1)	0.109	0.002	0	0	0.111	0.001	0.001	0
		Large (1)	0.029	0.001	0	0	0.028	0	0	0

Table 5. Mean (*SD*) number of iterations per convergence category for the surrogate MI UN (unstructured) and MI FA0(4) (factor analytic) models as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability. *Note.* –: quantity cannot be computed.

Model			Small imbalance $\widetilde{n}_i \sim N\,(M = 20,\ SD = 2.5)$				Large imbalance $\widetilde{n}_i \sim N\,(M = 20,\ SD = 5)$			
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	5	10	20	50
Surrogate model, MI UN	Proper convergence	Small (0.1)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
		Large (1)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	8.19 (4.61)	3.68 (1.35)	–	–	8.41 (17.51)	3.87 (1.51)	9.00 (–)	8.00 (–)
		Large (1)	6.20 (3.79)	5.00 (0)	–	–	6.15 (4.48)	5.00 (–)	2.50 (2.12)	–
	Divergence	Small (0.1)	66.93 (82.89)	–	–	–	56.49 (12.16)	–	–	–
		Large (1)	55.26 (8.29)	–	–	–	54.96 (17.06)	–	–	–
	Convergence category	Between-cluster variability γ	Number of clusters 5	10	20	50	5	10	20	50
Surrogate model, MI FA0(4)	Proper convergence	Small (0.1)	4.24 (2.91)	1.22 (1.08)	1 (0)	1 (0)	4.04 (2.66)	1.19 (0.99)	1 (0)	1 (0)
		Large (1)	2.06 (1.91)	1.00 (0.07)	1 (0)	1 (0)	2.04 (1.90)	1.00 (0.06)	1 (0)	1 (0)
	Convergence but non-PD **D**/**H** matrix	Small (0.1)	5.92 (1.94)	4.50 (1.38)	–	–	5.87 (1.99)	4.89 (1.17)	–	–
		Large (1)	4.92 (1.19)	4.00 (–)	–	–	4.55 (1.30)	–	13.50 (10.61)	–
	Divergence	Small (0.1)	21.25 (12.83)	33.60 (7.27)	–	–	22.18 (13.28)	38.75 (10.05)	19.00 (–)	56.00 (–)
		Large (1)	35.01 (12.00)	37.00 (0)	–	–	32.43 (12.87)	–	–	–

Table 6. Bias, efficiency and MSE of the estimates of R^2_{indiv} in the non-MI and MI surrogate evaluation models that properly converged as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability.

Model			Balanced (equal n_i)				Small imbalance $\widetilde{n}_i \sim N\left(20,\ 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\ 5^2\right)$			
	Measure	Between-cluster variability (γ)	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model, non-MI, UN	Bias	Small (0.1)	−0.001	−0.001	−0.002	0.001	−0.014	0.001	−0.001	0.001	0.009	0.001	−0.002	0.001
		Large (1)	−0.005	−0.001	−0.001	−0.001	−0.009	−0.001	0.001	−0.001	−0.009	−0.001	0.001	−0.001
	Efficiency	Small (0.1)	0.080	0.052	0.037	0.023	0.072	0.053	0.037	0.023	0.067	0.052	0.038	0.023
		Large (1)	0.072	0.053	0.037	0.024	0.071	0.052	0.037	0.024	0.073	0.052	0.037	0.024
	MSE	Small (0.1)	0.006	0.003	0.001	0.001	0.005	0.003	0.001	0.001	0.004	0.003	0.001	0.001
		Large (1)	0.005	0.003	0.001	0.001	0.005	0.003	0.001	0.001	0.005	0.003	0.001	0.001
	Measure	Between-cluster variability (γ)	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model, non-MI, FA0(4)	Bias	Small (0.1)	−0.001	−0.009	−0.002	0.001	−0.001	0.001	−0.001	0.001	−0.001	0.001	−0.002	0.001
		Large (1)	−0.006	−0.001	−0.001	−0.001	−0.007	−0.001	0.001	−0.001	−0.007	−0.001	0.001	−0.001
	Efficiency	Small (0.1)	0.074	0.052	0.037	0.023	0.076	0.053	0.037	0.023	0.072	0.051	0.038	0.023
		Large (1)	0.073	0.053	0.037	0.024	0.072	0.053	0.037	0.026	0.074	0.052	0.038	0.027
	MSE	Small (0.1)	0.006	0.003	0.001	0.001	0.006	0.003	0.001	0.001	0.005	0.003	0.001	0.001
		Large (1)	0.005	0.003	0.001	0.001	0.005	0.003	0.001	0.001	0.006	0.003	0.001	0.001
	Measure	Between-cluster variability (γ)	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model, MI UN	Bias	Small (0.1)	–	–	–	–	−0.005	0.001	−0.001	0.003	0.008	0.004	−0.001	0.004
		Large (1)	–	–	–	–	−0.005	0.002	0.002	0.001	−0.005	0.001	0.004	0.001
	Efficiency	Small (0.1)	–	–	–	–	0.089	0.066	0.047	0.030	0.095	0.074	0.057	0.040
		Large (1)	–	–	–	–	0.087	0.065	0.047	0.031	0.098	0.076	0.057	0.043
	MSE	Small (0.1)	–	–	–	–	0.008	0.004	0.002	0.001	0.009	0.005	0.003	0.002
		Large (1)	–	–	–	–	0.008	0.004	0.002	0.001	0.010	0.006	0.003	0.002
	Measure	Between-cluster variability (γ)	Number of clusters				Number of clusters				Number of clusters			
			5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model, MI FA0(4)	Bias	Small (0.1)	–	–	–	–	−0.001	0.001	−0.001	0.003	0.003	0.003	−0.001	0.004
		Large (1)	–	–	–	–	−0.005	0.002	0.002	0.001	−0.005	0.001	0.004	0.001
	Efficiency	Small (0.1)	–	–	–	–	0.090	0.066	0.047	0.030	0.097	0.074	0.057	0.040
		Large (1)	–	–	–	–	0.097	0.065	0.047	0.031	0.099	0.076	0.057	0.043
	MSE	Small (0.1)	–	–	–	–	0.008	0.004	0.002	0.001	0.009	0.006	0.003	0.002
		Large (1)	–	–	–	–	0.008	0.004	0.002	0.001	0.010	0.006	0.003	0.002

Note. UN = unstructured; FA0(4) = factor analytic.

Table 7. Bias, efficiency and MSE of the estimates of R^2_{trial} in the non-MI and MI surrogate evaluation models that properly converged as a function of balancedness of n_i, the number of clusters (5, 10, 20, 50) and the between-cluster variability.

Model			Balanced (equal n_i)				Small imbalance $\widetilde{n}_i \sim N\left(20,\ 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\ 5^2\right)$			
	Measure	Between-cluster variability (γ)	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Surrogate model, non-MI, UN	Bias	Small (0.1)	−0.152	−0.097	−0.041	−0.019	−0.168	−0.093	−0.036	−0.016	−0.174	−0.094	−0.034	−0.002
		Large (1)	−0.168	−0.075	−0.031	−0.013	−0.170	−0.074	−0.031	−0.012	−0.152	−0.074	−0.029	−0.009
	Efficiency	Small (0.1)	0.254	0.224	0.182	0.111	0.246	0.229	0.184	0.112	0.280	0.232	0.187	0.114
		Large (1)	0.248	0.219	0.158	0.095	0.239	0.222	0.157	0.096	0.248	0.221	0.157	0.095
	MSE	Small (0.1)	0.087	0.059	0.035	0.013	0.088	0.061	0.035	0.013	0.108	0.063	0.036	0.013
		Large (1)	0.089	0.053	0.026	0.009	0.086	0.055	0.026	0.009	0.084	0.054	0.025	0.009
	Measure	Between-cluster variability (γ)	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Surrogate model, non-MI, FA0(4)	Bias	Small (0.1)	−0.139	−0.090	−0.071	−0.035	−0.157	−0.088	−0.061	−0.032	−0.151	−0.082	−0.057	−0.018
		Large (1)	−0.099	−0.077	−0.060	−0.024	−0.109	−0.073	−0.063	−0.024	−0.076	−0.071	−0.058	−0.020
	Efficiency	Small (0.1)	0.267	0.233	0.203	0.121	0.269	0.238	0.197	0.122	0.273	0.237	0.201	0.125
		Large (1)	0.272	0.232	0.178	0.101	0.269	0.236	0.179	0.100	0.266	0.234	0.177	0.101
	MSE	Small (0.1)	0.090	0.062	0.046	0.016	0.096	0.064	0.043	0.016	0.097	0.063	0.044	0.016
		Large (1)	0.084	0.060	0.035	0.011	0.084	0.061	0.036	0.011	0.076	0.060	0.035	0.011
	Measure	Between-cluster variability (γ)	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Surrogate model, MI UN	Bias	Small (0.1)	–	–	–	–	−0.186	−0.080	−0.034	−0.016	−0.193	−0.087	−0.036	−0.022
		Large (1)	–	–	–	–	−0.165	−0.071	−0.031	−0.012	−0.163	−0.072	−0.032	−0.018
	Efficiency	Small (0.1)	–	–	–	–	0.255	0.226	0.172	0.105	0.257	0.229	0.174	0.128
		Large (1)	–	–	–	–	0.244	0.221	0.155	0.095	0.254	0.222	0.160	0.108
	MSE	Small (0.1)	–	–	–	–	0.099	0.057	0.031	0.011	0.103	0.060	0.031	0.017
		Large (1)	–	–	–	–	0.087	0.054	0.025	0.009	0.091	0.054	0.027	0.012
	Measure	Between-cluster variability (γ)	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50	Number of clusters 5	10	20	50
Surrogate model, MI FA0(4)	Bias	Small (0.1)	–	–	–	–	−0.258	−0.088	−0.034	−0.016	−0.266	−0.094	−0.036	−0.022
		Large (1)	–	–	–	–	−0.193	−0.071	−0.031	−0.012	−0.193	−0.072	−0.032	−0.018
	Efficiency	Small (0.1)	–	–	–	–	0.251	0.228	0.172	0.105	0.251	0.233	0.174	0.128
		Large (1)	–	–	–	–	0.253	0.221	0.155	0.095	0.261	0.222	0.160	0.108
	MSE	Small (0.1)	–	–	–	–	0.130	0.060	0.031	0.011	0.134	0.063	0.031	0.017
		Large (1)	–	–	–	–	0.101	0.054	0.025	0.009	0.106	0.054	0.027	0.011

Note. UN = unstructured; FA0(4) = factor analytic.

4.2. *Sample descriptives*

The ARMD trial enrolled a total of 181 patients from 36 centers. Here, the unit of analysis (i.e., the clustering variable) is center. Centers that enrolled less than 5 patients (19 centers in total) were discarded from the analyses to avoid problems during the MI phase (recall that the imputations are conducted for each center separately).

The data of 119 patients from 17 centers were analyzed. On average, there were 7 patients per center. A total of 6, 2, 4, 1, 2, and 1 centers had 5, 6, 4, 7, 8, 9, and 18 patients, respectively. In the center with 18 patients, 9 patients received placebo and 9 patients received interferon-α.

4.3. *Analysis*

The same procedure that was described in Section 3 to obtain balanced datasets (using MI) was employed here. Thus, in all center by treatment groups that had less than 9 patients, data were imputed to achieve balance. The imputations were conducted for each of the centers separately, using S, T, and Z in the imputation model.

A total of 1000 imputations were conducted. For each of the balanced datasets, Model (1) was fitted using S = change in visual acuity after 24 weeks and T = change in visual acuity after 52 weeks. Both the FA0(4) and UN covariance parametrizations for $\mathbf{D}$ were used.

The key outcome of interest was model convergence (using the same convergence categories as were defined in Section 2). In addition, the trial- and individual-level coefficients of surrogacy (R^2_{trial} and R^2_{indiv}) were computed for all datasets.

4.4. *Results*

4.4.1. Convergence rates

When Model (1) was fitted to the non-imputed data of the case study (both the entire ARMD dataset and the dataset that only included centers that enrolled at least 5 patients were considered), convergence issues occurred. In particular, the models that used the UN parametrization for the $\mathbf{D}$ matrix did not converge and the models that used the FA0(4) parametrization converged to a non-PD $\mathbf{D}/\mathbf{H}$ matrix.

Table 8 shows the convergence rates that were obtained when the MI-based approaches were used. Overall convergence was high and equaled 100% and 96.9% in the MI UN and MI FA0(4) scenarios, respectively. Further, the use of the MI FA0(4) strategy led to higher rates of proper convergence compared to the MI UN strategy (94.4% versus 70.1%, respectively), whereas the MI UN strategy led to lower divergence rates compared to the MI FA0(4) scenario (0% versus 4.1%, respectively). These results are fully in line with the results that were obtained in the simulation studies detailed above.

4.4.2. Coefficients of surrogacy

The densities of the trial- and individual-level surrogacy estimates ($\widehat{R}^2_{trial}$ and $\widehat{R}^2_{indiv}$, respectively) using the MI UN and MI FA0(4) strategies are shown in Figure 1.

The mean $\widehat{R}^2_{trial}$ and $\widehat{R}^2_{indiv}$ equalled 0.573 and 0.453 for the MI UN models, and 0.597 and 0.431 for the MI FA0(4) models, respectively. To establish a frame of reference against which these estimates can be compared, the two-stage equivalent of Model (1) was fitted to the non-imputed ARMD data (a simplified approach was used because Model (1) did

Table 8. Convergence rates for the ARMD data (using MI UN and MI FA0(4) to restore 'balance' in cluster size).

	MI UN	MI FA0(4)
Proper convergence	0.701	0.944
Convergence but non-PD **D**/**H** matrix	0.299	0.015
Divergence	0	0.041

Note. UN = unstructured; FA0(4) = factor analytic.

not converge). In particular, in the first stage the following bivariate model is fitted:

$$\begin{cases} S_{ij} = \mu_S + \alpha_i Z_{ij} + \varepsilon_{Sij} \\ T_{ij} = \mu_T + \beta_i Z_{ij} + \varepsilon_{Tij} \end{cases},$$

where μ_S, μ_T are the common intercepts for S and T, α_i, β_i are the fixed trial-specific treatment effects for S and T, and the other parameters are the same as defined above. In the second stage, $\widehat{\beta}_i$ is regressed on $\widehat{\alpha}_i$. The classical coefficient of determination of the fitted stage 2 model provides an estimate of R^2_{trial} and $R^2_{indiv} = \text{corr}\left(\varepsilon_{Sij}, \varepsilon_{Tij}\right)^2$ (for details on simplifying model-fitting strategies, see [14]). This analysis yielded $\widehat{R}^2_{trial} = 0.729$ with $CI_{95\%} = [0.487; 0.972]$ and $\widehat{R}^2_{indiv} = 0.512$ with $CI_{95\%} = [0.384; 0.639]$. The vertical solid and dashed lines in Figure 1 indicate these point estimates and their 95% confidence intervals, respectively.

Overall, the results indicate that there was an acceptable agreement between the trial- and individual-level surrogacy estimates that were obtained in the MI-based and non-MI based approaches, though it should be noted that the variability of the MI-based $\widehat{R}^2_{trial}$ and $\widehat{R}^2_{indiv}$ values was large (see also Figure 1). For example, the 2.5th and 97.5th percentile values of $\widehat{R}^2_{trial}$ equalled $Pc_{2.5} = 0.078$, $Pc_{97.5} = 0.941$ and $Pc_{2.5} = 0.069$, $Pc_{97.5} = 0.985$ in the MI UN and MI FA0(4) scenarios, respectively. The large variability of these estimates should be evaluated in light of the relatively small number of clusters and patients in the ARMD dataset. In addition, there were large imbalances in the cluster sizes in the ARMD dataset. For example, 7 out of the 17 centers that were available for analysis had only 5 patients and thus the ratio of the available data relative to the data that had to be imputed in these centers was small (5 versus 13 patients). It seems reasonable to assume that the variability of $\widehat{R}^2_{indiv}$ and $\widehat{R}^2_{trial}$ will be smaller when this ratio is higher, though additional analyses are needed to substantiate this claim. Further, the results showed that the use of the MI UN and MI FA0(4) strategies yielded nearly identical estimates for both coefficients of surrogacy.

5. Discussion

In line with earlier research [1, 2], the convergence rates of the mixed-effects models were found to be substantially higher when the number of available clusters increased and when the size of the between-cluster variability **D** was large relative to the residual variability $\boldsymbol{\Sigma}$. The present simulation study further extend these findings by showing that

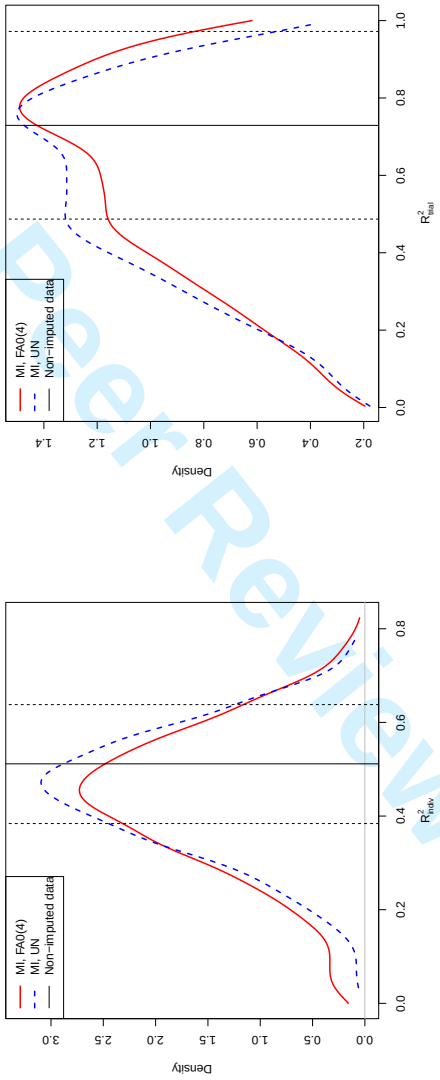


Figure 1. Densities of $\widehat{R}^2_{indiv}$ (left) and $\widehat{R}^2_{trial}$ (right) in the ARMD datasets using the MI UN and MI FA0(4) strategies. The vertical full lines are the point estimates for $\widehat{R}^2_{indiv}$ and $\widehat{R}^2_{trial}$ in the non-imputed dataset (using a two-stage modeling approach) and the vertical dashed lines are the 95% confidence intervals.

an imbalance in cluster size was associated with more model convergence issues. This was particularly the case when the model at hand had a complex hierarchical structure. The divergence rates were higher when the imbalance in cluster size was larger, and the use of MI to make the unbalanced datasets balanced reduced model convergence issues. Bias in the estimation of R^2_{indiv} was similar in the non-MI and MI scenarios, but the use of MI led to a decreased efficiency and increased MSE. With respect to the estimation of R^2_{trial}, bias, efficiency, and MSE were comparable in all scenarios where there were more than 5 clusters available.

The combination of good convergence properties and satisfactory statistical properties of the estimators, as follows from our simulation study, also suggests that there is little risk of convergence to local optima. In addition, our experience from past data analyses and simulations (see, for example, [5]) is in line with this finding.

With due caution, in scenarios where the convergence properties of the maximum likelihood estimator are poor (e.g., when N and γ are small), a forthright recommendation is to use multiple imputation with the Cholesky decomposition formulation of the variance-covariance matrix. Of course, there are secondary issues pertaining to e.g., the target of inference, which may or may not place emphasis on the variance components. For example, in the present simulations, three model convergence categories were distinguished, i.e., (i) proper convergence, (ii) convergence but a non-PD **D** or **H** matrix, and (iii) divergence. The relevance of distinguishing between categories (i)–(ii) depends on the substantive research question at hand. In a surrogate validation context, the distinction is important because one is mainly interested in the variance components (e.g., **D** should be PD to guarantee that $\widehat{R}^2_{trial}$ is within the unit interval). If one is merely interested in the fixed-effects (the marginal model), this distinction is unimportant because the marginal model can be used to make valid inferences regarding the fixed-effect parameters as long as the overall **V** matrix is PD [4, 10]. Thus, in practice, a researcher who is mainly interested in making inferences regarding the random effects may opt for the strategy that leads to the highest rates of proper convergence (e.g., MI with FA0(4) for the **D** matrix), whereas a researcher who is mainly interested in the marginal model may opt for the strategy that leads to the highest rates of overall convergence (e.g., MI with UN for the **D** matrix).

Several alternative imputation models are potentially of use here – though any feasible model needs to be compatible with the analysis model. For example, hierarchical versions could be considered that take into account all three levels [16]. In our specific context, where typically there is a relatively small number of trials, with a good amount of replication per trial, also a trial-specific strategy is viable. The method we have proposed is both computationally convenient and has good convergence and statistical properties.

Some comments and suggestions for future research are in place. First, the convergence rates were relatively high for all models in all scenarios. For example, the proper convergence rates were close to 100% when the number of clusters exceeded 20 for all models in all scenarios (see Tables 1 and 4). Obviously, the convergence rates that are obtained in a simulation study depend on the choice of the parameters that are used to generate the data. For example, in the present simulations the variance components in the **D** matrix that were used to generate the data were all relatively large. This choice was made to avoid model convergence problems that arise by hitting the boundary of the parameter space. When lower-valued **D** matrices were used (keeping all other parameters constant), the convergence rates substantially decreased but the global pattern of the results in terms of the impact of an imbalance in cluster size on model divergence rates and the effect of the use of MI remained the same. This is further illustrated in Table A1 in the Appendix (which shows the convergence rates for a scenario where $\gamma = 0.01$, keeping all other parameters that were used to generate the data identical to the ones provided in

Section 2) and in the Supplementary Materials.

In the Supplementary Materials, lower-valued $\mathbf{D}$ matrices were used and a number of additional scenarios to introduce imbalance in cluster size and/or treatment allocation were considered. Further, the (mean) cluster size that was used in the analyses in the Supplementary Materials was substantially lower and closer to the mean cluster size in the ARMD dataset (i.e., mean $n = 10$ instead of $n = 20$). The results that were obtained in these additional scenarios were similar to the main results discussed above.

Note also that the focus of the current study was on linear mixed effects models only. Further simulation studies would be needed to examine this issue in non-linear mixed-effects models.

Second, missing data frequently arise in a surrogate evaluation setting (i.e., the measurement of T is by definition cumbersome, otherwise there would be no need for a surrogate) and in many other research settings. An advantage of the MI-based strategy proposed above is that it provides a natural framework to deal with unbalanced cluster sizes and missingness at the same time. This can be done in a flexible way, e.g., it is straightforward to include covariates, such as the age of the patient or a post-randomization non-compliance measure in the imputation model whilst at the same time keeping the standard substantive Model (1) [9].

Finally, MI was used to augment the data and Newton-Raphson was used to conduct the optimization of the log-likelihood functions, but other choices are viable as well. For example, future studies may consider the use of EM to augment the data and/or Fisher scoring for the optimization. Further, all simulation results discussed in the present paper were obtained using SAS, and it would be useful to evaluate whether similar results are obtained when other software tools (e.g., R, Stata, MLwiN) are used.

Funding

Financial support from the IAP research network #$P7/06$ of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. Wim Van der Elst acknowledges funding from the European Seventh Framework programme $FP7$ $2007 - 2013$ under grant agreement Nr. 602552. Geert Molenberghs acknowledges funding from Intel, Janssen Pharmaceutica and by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT).

Supplemental material

Supplementary Materials where the results of some additional simulation studies are discussed are available on the journal's website.

References

[1] Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 2000;1,49–67.
[2] Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. Biometrical Journal. 2002;44, 921–935.
[3] Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effectss models for repeated-measures data. Journal of the American Statistical Association. 1988;83,1014–1022.
[4] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag; 2000.
[5] Burzykowski T, Molenberghs G, Buyse M. The Evaluation of Surrogate Endpoints. New York: Springer-Verlag; 2005.
[6] Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. Statistics in Medicine. 1989;8,431–440.
[7] Cortiñas Abrahantes J, Molenberghs G, Burzykowski T, Shkedy Z, Renard D. Choice of units of analysis and modeling strategies in multilevel hierarchical models. Computational Statistics and Data Analysis. 2004;47,537–563.
[8] Dempster AP, Laird NM, Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B. 1977;39;1–38.
[9] Molenberghs G, Kenward, M.G. Missing Data in Clinical Studies. New York: John Wiley & Sons; 2007.
[10] West BT, Welch KB, Galecki, AT. Linear mixed models: A Practical Guide Using Statistical Software. New York: Chapman & Hall/CRC; 2007.
[11] Schafer, JL. Analysis of Incomplete Multivariate Data. New York: Chapman & Hall; 1997.
[12] Pharmacological Therapy for Macular Degeneration Study Group. Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. Archives of Ophthalmology. 1997;115;865–872.
[13] Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. Biometrics. 2014.
[14] Tibaldi FS, Cortiñas Abrahantes J, Molenberghs G, Renard D, Burzykowski T, Buyse M, Parmar M, Stijnen T, Wolfinger R. Simplified hierarchical linear models for the evaluation of surrogate endpoints. Journal of Statistical Computation and Simulation. 2003;73;643–658.
[15] Diaz-Ordaz, K, Kenward, MG, Grieve, R. A comparison of multiple imputation models for bivariate hierarchical outcomes. Retrieved from http://arxiv.org/abs/1407.4703; 2014.
[16] Carpenter, J, Kenward MG. Multiple Imputation and its Application. New York: John Wiley & Sons; 2014.
[17] van Buuren S. Multiple imputation of multilevel data. In: The Handbook of Advanced Multilevel Analysis. Milton Park (UK): Routledge; 2011.

Appendix A.

Table A1. Convergence rates (using $\gamma = 0.01$) for the random-intercept models, reduced surrogate models, and surrogate evaluation models as a function of balancedness of n_i and the number of clusters (5, 10, 20, 50).

Model		Balanced (equal n_i)				Small imbalance $\widetilde{n}_i \sim N\left(20,\, 2.5^2\right)$				Large imbalance $\widetilde{n}_i \sim N\left(20,\, 5^2\right)$			
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Random	Proper convergence	0.656	0.797	0.910	0.982	0.656	0.812	0.910	0.982	0.651	0.798	0.903	0.981
intercept	Convergence but non-PD **D**/**H**	0.344	0.203	0.081	0.018	0.344	0.188	0.090	0.018	0.349	0.202	0.097	0.019
	Divergence	0	0	0	0	0	0	0	0	0	0	0	0
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Reduced surrogate	Proper convergence	0.280	0.512	0.735	0.946	0.281	0.511	0.747	0.944	0.242	0.484	0.710	0.945
	Convergence but non-PD **D**/**H**	0.720	0.488	0.265	0.054	0.654	0.487	0.253	0.056	0.614	0.503	0.289	0.055
	Divergence	0	0	0	0	0	0	0	0	0.144	0.013	0.001	0
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model	Proper convergence	0	0.032	0.170	0.648	0.001	0.019	0.171	0.672	0.001	0.019	0.149	0.624
non-MI, UN	Convergence but non-PD **D**/**H**	0.999	0.968	0.830	0.352	0.056	0.654	0.809	0.328	0.015	0.428	0.789	0.376
	Divergence	0.010	0	0	0	0.943	0.327	0.020	0	0.984	0.553	0.062	0
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model	Proper convergence	0.281	0.457	0.674	0.918	0.280	0.454	0.634	0.918	0.307	0.467	0.629	0.898
non-MI, FA0(4)	Convergence but non-PD **D**/**H**	0.064	0.063	0.047	0.023	0.052	0.047	0.050	0.024	0.054	0.044	0.059	0.028
	Divergence	0.655	0.480	0.279	0.059	0.668	0.499	0.316	0.058	0.639	0.489	0.312	0.074
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model	Proper convergence	–	–	–	–	0.008	0.291	0.867	0.999	0.014	0.388	0.932	0.999
MI UN	Convergence but non-PD **D**/**H**	–	–	–	–	0.963	0.709	0.133	0.001	0.942	0.612	0.068	0.001
	Divergence	–	–	–	–	0.029	0	0	0	0.044	0	0	0
	Convergence	Number of clusters				Number of clusters				Number of clusters			
	category	5	10	20	50	5	10	20	50	5	10	20	50
Surrogate model	Proper convergence	–	–	–	–	0.428	0.796	0.984	1	0.439	0.848	0.989	0.999
MI FA0(4)	Convergence but non-PD **D**/**H**	–	–	–	–	0.061	0.051	0.009	0	0.067	0.044	0.008	0.001
	Divergence	–	–	–	–	0.511	0.154	0.007	0	0.494	0.108	0	0

Note. UN = unstructured; FA0(4) = factor analytic.

To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–8

Supplementary Materials: Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

The results in [1] showed (i) that divergence rates of mixed-effects models were substantially higher for unbalanced datasets, and (ii) that the use of Multiple Imputation to restore 'balance' in unbalanced datasets reduces model convergence problems.

In these Supplementary Materials, the results of some additional simulation studies are discussed. In particular, lower-valued **D** matrices were used to generate the data and a number of additional scenarios to intoduce imbalance in cluster size and/or treatment allocation were considered. The global pattern of results of these analyses is in line with the results of [1].

Keywords: simulation study, model convergence, mixed-effects model, multiple imputation, unbalanced data

AMS Subject Classification: 62J99; 62P10

1. Surrogate model

1.1. *Notation and model*

The same model that was described in Section 2.2.3 of [1] was used to generate the data, using different parameter values for the fixed and random effects. In line with the simulation setting of [2], the following model was used to simulate the data (see also [3], pp. 104-106):

$$\begin{cases} S_{ij} = 45 + m_{Si} + (3 + a_i)\, Z_{ij} + \varepsilon_{Sij} \\ T_{ij} = 50 + m_{Ti} + (5 + b_i)\, Z_{ij} + \varepsilon_{Tij} \end{cases},$$

where $(m_{Si},\, m_{Ti},\, a_i,\, b_i) \sim N(\mathbf{0},\, \mathbf{D})$ with

$$\mathbf{D} = \gamma \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0.9 & 1 \end{pmatrix},$$

and $(\varepsilon_{Sij},\, \varepsilon_{Tij}) \sim N(\mathbf{0},\, \mathbf{\Sigma})$ with

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 2.4 \\ 2.4 & 3 \end{pmatrix}.$$

For the total number of clusters (i.e., N), the grid of values $G_1 = \{10,\, 20,\, 50\}$ was

considered. The number of subjects per cluster was fixed to 10. The γ parameter was set to 0.1 (small between-trial variability) or 1 (large between-trial variability).

Eight scenarios were used to generate the data:

- First, a balanced scenario in which each of the clusters contain the same number of observations (i.e., 10 observations per cluster). Furthermore, treatment (Z) is balanced within a cluster (i.e., out of the 10 patients per cluster, 5 patients receive the control and 5 patients receive the experimental treatment).
- Second, an unbalanced scenario where $\widetilde{n}_i$ (the number of observations in a particular cluster) was determined based on a draw from a normal distribution, i.e., $\widetilde{n}_i \sim N\left(10,\, 5^2\right)$. All obtained $\widetilde{n}_i$ values were rounded to the nearest integer (i.e., $n_i = \text{round}\left(\widetilde{n}_i\right)$) and when $\widetilde{n}_i < 2$ it is automatically rounded up to 2 (i.e., each cluster should contain at least two observations). Furthermore, treatment is balanced within a cluster (when n_i is an odd number, $\left(n_i - 1\right)/2$ patients receive the control and the experimental treatment, and the remaining patient is randomly allocated to the control or the experimental treatment).
- Third, another unbalanced scenario where $\widetilde{n}_i$ was determined based on a normal distribution with a smaller SD (i.e., $\widetilde{n}_i \sim N\left(10,\, 2.5^2\right)$) and treatment is balanced within a cluster.
- Fourth, another unbalanced scenario where n_i was determined based on a draw from a discrete uniform distribution (i.e., $n_i \sim \text{uniform}\left(1,\, 19\right)$) and treatment is balanced within a cluster.
- Scenarios 5-8 are identical to scenarios 1-4, but treatment (Z) is no longer balanced within a cluster (i.e., Z is drawn from a binomial distribution with success probability 0.50).

A total of $M = 1,000$ runs were conducted for each setting. A completely general (unstructured; UN) $\mathbf{D}$ matrix that is parameterized directly in terms of variances and covariances was used. The generated datasets were analyzed using the mixed procedure in SAS.

The key outcome of interest in all simulations was model convergence. The same model convergence categories as were used in [1] were considered, i.e., (i) proper convergence (the model converged and the variance-covariance matrix of the random effects ($\mathbf{D}$) and the final Hessian ($\mathbf{H}$) were positive definite); (ii) the model converged but $\mathbf{D}$ or $\mathbf{H}$ was not positive definite; and finally, (iii) divergence. In addition, the number of required iterations to achieve convergence was recorded and analyzed.

1.2. *Results*

Tables 1-3 show the convergence rates. The rates of overall convergence (i.e., proper convergence or convergence but non-PD $\mathbf{D}$ or $\mathbf{H}$ matrix) were 100% in all scenarios where the cluster size and Z (treatment) were balanced (see Table 3). In contrast, the overall convergence rates decreased substantially in the unbalanced scenarios – in particular when γ and the number of clusters were small. For example, the divergence rate was as high as 90.6% in the scenario where the cluster size is highly unbalanced ($n_i \sim \text{unif}(1,\, 19)$), $\gamma = 0.1$, Z is unbalanced, and $N = 10$.

Overall, the same pattern of results that was described in [1] is observed: models for which proper convergence could not be achieved tended to diverge when the cluster sizes and/or Z were unbalanced, whereas these models tended to converge to a solution with a non-PD $\mathbf{D}$ or $\mathbf{H}$ matrix when the cluster sizes and Z were balanced. Further, proper convergence was *always* achieved after 1 iteration when the cluster size and Z were

Table 1. Proportion of runs for which proper convergence was achieved (the model converged and the variance-covariance matrix of the random effects (**D**) and the final Hessian (**H**) were positive definite).

		Z balanced			Z unbalanced		
		Number of clusters			Number of clusters		
	γ	10	20	50	10	20	50
$n_i = n = 10$	0.1	0.003	0.023	0.098	0.004	0.017	0.095
	1	0.409	0.912	1	0.327	0.877	1
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 2.5^2\right)$	0.1	0.001	0.019	0.099	0.004	0.014	0.088
	1	0.381	0.888	0.999	0.325	0.848	0.999
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 5^2\right)$	0.1	0.002	0.012	0.087	0.001	0.014	0.101
	1	0.241	0.831	0.998	0.213	0.788	0.997
	γ	10	20	50	10	20	50
$n_i \sim \text{unif}(1,\ 19)$	0.1	0.005	0.021	0.106	0.002	0.018	0.114
	1	0.258	0.835	1	0.228	0.791	0.998

Table 2. Proportion of runs for which the model converged but **D** or **H** was not positive definite.

		Z balanced			Z unbalanced		
		Number of clusters			Number of clusters		
	γ	10	20	50	10	20	50
$n_i = n = 10$	0.1	0.997	0.977	0.902	0.567	0.919	0.902
	1	0.591	0.088	0	0.598	0.122	0
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 2.5^2\right)$	0.1	0.394	0.830	0.899	0.246	0.742	0.898
	1	0.445	0.111	0.001	0.399	0.139	0.001
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 5^2\right)$	0.1	0.160	0.547	0.871	0.097	0.444	0.839
	1	0.317	0.146	0.002	0.263	0.181	0.002
	γ	10	20	50	10	20	50
$n_i \sim \text{unif}(1,\ 19)$	0.1	0.175	0.637	0.888	0.092	0.542	0.866
	1	0.300	0.154	0	0.251	0.172	0.002

balanced (Table 4).

2. Surrogate model with multiple imputation

2.1. *Simulations*

2.1.1. Scenarios

The same model and scenarios that were described in Section 1 were considered, with the exception of the following issues:

Table 3. Proportion of runs for which the model diverged.

		Z balanced			Z unbalanced		
		Number of clusters			Number of clusters		
	γ	10	20	50	10	20	50
$n_i = n = 10$	0.1	0	0	0	0.429	0.064	0.003
	1	0	0	0	0.075	0.001	0
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 2.5^2\right)$	0.1	0.605	0.151	0.002	0.750	0.244	0.014
	1	0.174	0.001	0	0.276	0.013	0
	γ	10	20	50	10	20	50
$\widetilde{n}_i \sim N\left(10,\ 5^2\right)$	0.1	0.838	0.441	0.042	0.902	0.542	0.060
	1	0.442	0.023	0	0.524	0.031	0.001
	γ	10	20	50	10	20	50
$n_i \sim \text{unif}(1,\ 19)$	0.1	0.820	0.342	0.006	0.906	0.440	0.020
	1	0.442	0.011	0	0.521	0.037	0

- Only one unbalanced scenario was considered. In particular, only the scenario where $\widetilde{n}_i$ (the cluster size) was determined based on a draw from a normal distribution ($\widetilde{n}_i \sim N\left(10,\ 5^2\right)$) and where Z (treatment) was unbalanced within a cluster was considered.
- The same datasets that were used in Section 1 were used in the current analyses, though multiple imputation (MI) was used to 'fill in' the 'missing' data to restore balancedness for both n_i and Z. The imputation model included cluster, Z, S, and T. A total of 3 imputations were used for each dataset.
- Two different variance-covariance matrices were used for the random effect structure of the mixed model. First, a completely general (unstructured; UN) $\mathbf{D}$ matrix. Second, a non-diagonal factor-analytic structure with 4 factors (FA0(4)). The latter structure specifies a Cholesky root parametrization for the 4×4 unstructured blocks in $\mathbf{D}$.

2.1.2. Results

Table 5 shows the convergence rates in the MI UN and MI FA0(4) scenarios. The rates of *proper* convergence were substantially higher in the MI FA0(4) scenario ($\geq$ 78.0%) compared to what was the case in the MI UN scenario ($\geq$ 0.2%). With respect to *overall* convergence, the results were reversed as overall convergence rates were higher in the MI UN scenario ($\geq$ 99.9%) compared to what was the case in the MI FA0(4) scenario ($\geq$ 88.9%). In line with [1], it can be concluded that the use of MI reduces model divergence issues. Further, proper convergence was always achieved after 1 iteration in the MI UN scenario (see Table 6).

Table 4. Mean (*SD*) number of iterations to achieve proper convergence.

		Z balanced				Z unbalanced			
		Number of clusters				Number of clusters			
	γ	10	20	50		10	20	50	
$n_i = n = 10$	0.1	1 (–)	1 (–)	1 (–)		3.75 (0.50)	2.82 (0.64)	2.44 (0.71)	
	1	1 (–)	1 (–)	1 (–)		4.40 (1.03)	3.73 (0.72)	3.13 (0.54)	
	γ	10	20	50		10	20	50	
$\widetilde{n}_i \sim N\left(10,\, 2.5^2\right)$	0.1	5.00 (–)	3.50 (0.67)	2.83 (0.61)		8.00 (–)	3.64 (0.74)	3.11 (0.51)	
	1	4.80 (0.83)	4.32 (0.70)	3.85 (0.74)		5.28 (1.01)	4.66 (1.10)	4.04 (0.70)	
	γ	10	20	50		10	20	50	
$\widetilde{n}_i \sim N\left(10,\, 5^2\right)$	0.1	3.00 (–)	2.89 (0.74)	2.22 (0.44)		3.75 (0.96)	3.29 (0.61)	2.75 (0.59)	
	1	4.11 (0.78)	3.69 (0.67)	3.22 (0.56)		4.84 (1.02)	4.22 (0.70)	3.60 (0.68)	
	γ	10	20	50		10	20	50	
$n_i \sim \text{unif}(1,\, 19)$	0.1	4.20 (1.48)	3.29 (0.46)	2.69 (0.49)		3.50 (0.71)	3.61 (0.78)	3.00 (0.52)	
	1	4.77 (0.91)	4.31 (0.69)	3.72 (0.65)		5.23 (0.93)	4.59 (0.75)	3.97 (0.66)	

Table 5. Convergence rates in the MI UN (left column) and MI FA0(4) (right column) scenarios.

		MI, UN			MI, FA0(4)		
		Number of clusters			Number of clusters		
	γ	10	20	50	10	20	50
Proper	0.1	0.002	0.007	0.023	0.788	0.789	0.780
Convergence	1	0.077	0.308	0.764	0.864	0.937	0.990
	γ	10	20	50	10	20	50
Convergence but	0.1	0.997	0.993	0.977	0	0	0
non-PD $\mathbf{D}/\mathbf{H}$ matrix	1	0.923	0.692	0.236	0	0	0
	γ	10	20	50	10	20	50
Divergence	0.1	0.001	0	0	0.212	0.211	0.202
	1	0	0	0	0.136	0.063	0.010

Table 6. Mean (*SD*) number of iterations per convergence category in the MI UN (left column) and MI FA0(4) (right column) scenarios.

		MI, UN				MI, FA0(4)			
		Number of clusters				Number of clusters			
	γ	10	20	50		10	20	50	
Proper Convergence	0.1	1 (0)	1 (0)	1 (0)		9.83 (7.03)	10.06 (7.38)	9.81 (7.08)	
	1	1 (0)	1 (0)	1 (0)		6.97 (5.79)	5.01 (5.32)	2.16 (3.01)	
	γ	10	20	50		10	20	50	
Convergence but non-PD $\mathbf{D}/\mathbf{H}$ matrix	0.1	9.91 (2.86)	8.08 (2.49)	6.33 (2.20)		–	–	–	
	1	7.04 (2.55)	4.91 (1.83)	3.49 (1.34)		–	–	–	
	γ	10	20	50		10	20	50	
Divergence	0.1	3 (–)	–	–		23.10 (12.42)	23.61 (11.79)	24.75 (12.92)	
	1	–	–	–		24.66 (13.57)	30.04 (13.17)	36.16 (13.10)	

References

[1] Van der Elst W, Hermans L, Verbeke G, Kenward M, Nassiri V, Molenberghs G. Imbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data.
[2] Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 2000;1,49–67.
[3] Burzykowski T, Molenberghs G, Buyse M. The Evaluation of Surrogate Endpoints. New York: Springer-Verlag; 2005.