The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models

DOI: 10.1002/sim.3157
Handle: http://hdl.handle.net/1942/8343

STATISTICS IN MEDICINE

Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models

S. Litière*, A. Alonso and G. Molenberghs

Hasselt University, Center for Statistics, Agoralaan Gebouw D, BE-3590 Diepenbeek, Belgium.

SUMMARY

Estimation in generalized linear mixed models is often based on maximum likelihood theory, assuming that the underlying probability model is correctly specified. However, the validity of this assumption is sometimes difficult to verify. In this paper we study, through simulations, the impact of misspecifying the random-effects distribution on the estimation and hypothesis testing in generalized linear mixed models. It is shown that the maximum likelihood estimators are inconsistent in the presence of misspecification. The bias induced in the mean structure parameters is generally small, as far as the variability of the underlying random-effects distribution is small as well. However, the estimates of this variability are always severely biased. Given that the variance components are the only tool to study the variability of the true distribution, it is difficult to assess whether problems in the estimation of the mean structure occur. The Type I error rate and the power of the commonly used inferential procedures are also severely affected. The situation is aggravated if more than one random effect is

*Correspondence to: Hasselt University, Center for Statistics, Agoralaan Gebouw D, BE-3590 Diepenbeek, Belgium. E-mail: saskia.litiere@uhasselt.be

included in the model. Further, we propose to deal with possible misspecification by way of sensitivity analysis, considering several random-effects distributions. All the results are illustrated using data from a clinical trial in schizophrenia.

KEY WORDS: consistency; heterogeneity model; Kullback-Leibler Information Criterion; non-normal random effects; power; type I error.

1. INTRODUCTION

When dealing with non-Gaussian data with multiple sources of variation, a commonly used subject-specific model is the generalized linear mixed model (GLMM; [1, 2]). In this framework, the (vector-valued) outcome variable Y_i for a subject i is assumed to satisfy $Y_i|b_i \sim F_i(y_i|\varphi, b_i)$, i.e., conditionally on random effects b_i, Y_i follows a pre-specified distribution F_i, parameterized through a vector φ of unknown parameters, common to all subjects. The vector of subject-specific parameters b_i is assumed to follow a distribution G which may also depend on a vector δ of unknown parameters. The parameter estimates are commonly calculated by maximizing the marginal likelihood, obtained by integrating out the random effects. Due to software limitations, the analysis is often restricted to the setting in which the random-effects distribution is normal with mean zero and variance-covariance matrix D. Since random effects are non-measurable, the validity of this assumption is difficult to check. The question naturally arising is what the impact is of misspecifying the random-effects distribution on the maximum likelihood estimators.

For linear mixed models (LMM; [3, 4]), it has been shown that the maximum likelihood estimators, obtained under the assumption of normally distributed random effects, are consistent and asymptotically normally distributed, even when the true random-effects

distribution is not normal [5]. Nevertheless, research carried out in recent years illustrates that similar results do not hold for GLMM. For instance, Neuhaus *et al.* [6] showed that the maximum likelihood estimators of a logistic–normal model with misspecified random-effect distribution are inconsistent but that the magnitude of the bias is typically small. Simulations by Chen *et al.* [7] also indicate that the estimation of the regression coefficients may be subject to only negligible bias under misspecification of the random-effects distribution. According to Agresti *et al.* [8] the choice of the random-effects distribution seems to have, in most situations, little effect on the maximum likelihood estimators. However, when there is a severe polarization of subjects, e.g., by omitting an influential binary covariate, this can affect the predictive quality for characteristics involving the random effects as well as the fixed effects. Similarly, Heagerty and Kurland [9] found substantial bias while using a random-intercept model when the random-effects distribution depends on measured covariates.

These results clearly illustrate the wide range of opinions which can be found in the literature regarding the impact of misspecifying the random-effects distribution on the maximum likelihood estimators in GLMM. However, it is important to note that all these simulation studies were performed using a limited number of distributions and, in all of them, only small variances for the random-effects were considered. As we will show in the subsequent sections, the magnitude of this variance can have an important effect on the bias induced by the misspecification, where larger biases associate with larger variances. Moreover, as will be seen from the case study, small variances may not be realistic in some important practical settings. Further, we will illustrate that the situation worsens when more than one random effect is included in the model. Another issue which has not received attention in the previous studies concerns the impact of the misspecification on commonly used inferential procedures

such as the Wald test. We will study the impact of such misspecification on the power and Type I error rate of frequently used tests for the mean-structure parameters. Therefore, the first main objective of the present work is to use a wide set of simulations so as to formulate some practical guidelines for handling random-effects misspecification in GLMM.

It is also important to point out that some alternative approaches have been suggested to deal with this misspecification. For instance, one could replace the normal random-effects distribution by a non-parametric distribution [10, 11, 12, 13, 14]. Although it is an appealing approach in many settings, there can be some loss of efficiency when using a non-parametric approach, compared to parametric assumptions close to the true distribution [8]. Additionally, model comparison can be difficult as standard asymptotic theory does not apply. Finally, a non-parametric approach is definitely not appropriate when the distribution of the random effects is of primary interest like, for example, in surrogate marker evaluation, the evaluation of the psychometric properties of rating scales, or when one wants to predict individual profiles or evolutions. Chen *et al.* [7] suggested a semi-parametric random-effects distribution, allowing the random-effects density to be skewed, multi-modal, fat- or thin-tailed, and including the normal as a special case. Lee and Thompson [15] used Markov Chain Monte Carlo methods to fit models with random effects following a t distribution, and skew extensions to the normal and the t distribution. In the present work, we will study another approach which consists in replacing the normal random-effects distribution by a finite mixture of normals [16, 17, 18, 2]. This allows one to cover a wide range of shapes for the random-effects density, including unimodal as well as multimodal, and symmetric as well as very skewed distributions. Our simulation studies with this model show that although using more flexible families of distributions can be a valid strategy in some settings, these more general families are not fully

robust either.

Therefore, we propose to incorporate the previous approach into a more general sensitivity analysis framework. In this scenario, different distributions are considered for the random effects. If the estimates of the parameters of interest and the associated inferential procedures are similar, irrespective of the distribution used to obtain them, the user can feel relatively confident about his/her results. On the other hand, if the results vary considerably, then they are obviously sensitive to the distributional assumptions for the random effects, and caution is needed. One could also use some known model selection criteria to select the distribution that fits the data best. Note that a sensitivity analysis is, perhaps, the most appropriate approach to deal with a missing data problem, whereby random effects can essentially be seen as missing observations. We start by introducing, in Section 2, the case study that motivated the present work. In Section 3, various aspects of the maximum likelihood estimators are investigated through extensive simulations. Next, in Section 4 some alternative approaches are suggested and applied to the case study.

2. CASE STUDY - THE SCHIZOPHRENIA DATA

The case study comprises of individual patient data from a randomized clinical trial, comparing the effect of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia [19]. Several measures can be used to assess a patient's global condition. The Clinical Global Impression (CGI) is generally accepted as a subjective but useful clinical measure of change. It is a 7-grade scale used to characterize a subject's mental condition. Our binary response variable y is a dichotomous version of this scale which equals 1 for patients classified as normal to mildly ill, and 0 for patients classified as moderately to severely ill.

Treatment was administered for 8 weeks and the outcome was measured at 6 fixed time points: at the beginning of the study and after 1, 2, 4, 6 and 8 weeks. In total, 128 patients were included in the trial, from which 64 were randomly assigned to receive risperidone and the rest to an active control. Figure 1 summarizes the probability of being classified as normal to mildly ill ($P(Y = 1)$) by time point and treatment group.

FIGURE 1 – About here

We analyzed these data using a random-intercept model, considering different link functions and mean structures. The random intercept was always assumed to follow a normal distribution with mean zero and variance σ_b^2. In the model building exercise, a total of nine models were fitted. These were constructed as combinations of three link functions, the logit, complementary log-log, and probit links, and three different mean structures: i) intercept, treatment, time, and treatment-by-time interaction, ii) as in (i) but without treatment-by-time interaction, and iii) as in (i) but without treatment. The AIC criterion was used to select the best fitting combination. Our final model takes the form

$$\text{logit}\{P(y_{ij} = 1|b_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_i, \tag{1}$$

where $z_i = 1$ (0) denotes the treatment (control) group, t_j denotes the occasion of measurement and b_i refers to the random intercept. Adding a random slope to Model (1) did not improve the fit. All the previous models were fitted using the SAS procedure NLMIXED with adaptive Gaussian quadrature and 20 quadrature points. The maximum likelihood estimates for Model (1) are given in the first part of Table I.

Table I – About here

Figure 1 also displays the plot of the fitted values obtained from our final model against the observed probability of being classified as normal to mildly ill ($P(Y = 1)$) by time point and treatment group. The fitted probabilities are calculated by numerically integrating out the random effect for each subject. Note that in Figure 1 the two groups seem to be equal at the start of the study, but a very rapid onset of the treatment is observed at week 1, and this difference remains constant throughout the rest of the study. Our final model, which includes treatment as an additive constant, seems to have some problems in describing this very rapid manifestation of the treatment during the first week. However, it does capture the general general trend in the data afterwards. Arguably, this capacity of the model to describe the long term evolution over time is very desirable and relevant from a clinical point of view. Indeed, until week 4 there seems to be a reasonable agreement between the fitted and the observed values. Nevertheless, some discrepancy is also observed in the last two measurement occasions. It is important to point out that the proportion of dropouts is significantly high for these two measurements, specially in week 8 for the control group (50%). In presence of missing data such a discrepancy is not necessarily an evidence of lack of fit [2]. However, we prefer not want to enter here in a comprehensive discussion of the missing data problem; so we will assume that the missing data generating mechanism is *missing at random* (MAR), rendering our likelihood approach a valid option [20, 4, 2, 21].

Note that, even though Model (1) emerged as the best fitting model among all those considered, it produces relatively extreme estimates for the intercept and the variance component. We believe this is the result of some extreme response pattern in the data. For example, in the control group, a high proportion of the patients (75%) have a response pattern of nothing but zeros whereas in the treatment group a more variable pattern of responses

is observed. There, only 56% of the patients have a response pattern of all zeros. Therefore, the large estimate for the variance of the random component could be explained by the high within-subject correlation that these data carry. Allowing the random-effect variance to vary among treatment groups did not improve the fit. Indeed, this analysis resulted in a random-effect variance of 20.00 (s.e. 7.93) for the treatment group and 22.61 (s.e. 10.69) for the control group. These high variances hint on a very strong and similar within-subject correlation, in each treatment group.

Arguably, these circumstances could render the assumption of a normal distribution for the random effects questionable. However, this situation should not be considered exceptional or infrequent. Indeed, in a typical placebo-controlled clinical trial such an extreme pattern of all zeros could be expected in the placebo group, whereas a more variable pattern would be expected in the responses of the treated group. The problem is aggravated by the random effects being unobserved latent variables, which renders difficult the evaluation of the associated distributional assumptions. Nevertheless, the conventional belief among data analysts seems to be that the choice of the random-effects distribution is not crucial for the quality of the inferences related to the regression coefficients, even though, as will be shown in what follows, this does not always hold.

3. RANDOM-EFFECTS MISSPECIFICATION IN GENERALIZED LINEAR MIXED MODELS

Let us consider a random variable Y that follows a distribution h. In practice, we assume that h belongs to a family of densities $\mathfrak{F} = \{f(y;\gamma) : \gamma \in \Gamma\}$, indexed by a parameter γ. If there exists a $\gamma_0 \in \Gamma$ such that $\mathfrak{F}$ contains the true distribution (i.e. $h(y) = f(y, \gamma_0)$), then the

maximum likelihood estimator $\widehat{\gamma}_n$ of γ_0 is consistent and asymptotically normal. Note that in the GLMM, γ would correspond to the vector containing the parameters φ and δ.

Since h is unknown, it is difficult to check whether it belongs to $\mathfrak{F}$ or not. White [22] derived that, under general regularity conditions, the maximum likelihood estimator $\widehat{\gamma}_n$ will (strongly) converge to the value of γ, denoted by γ^*, which minimizes the so-called Kullback-Leibler Information Criterion (KLIC):

$$I(h : f, \gamma) = E\left\{\log \frac{h(Y)}{f(Y, \gamma)}\right\}, \tag{2}$$

where the expectation in (2) is taken with respect to the true distribution. Additionally, he showed that $\widehat{\gamma}_n$ is asymptotically normal with mean γ^*. If the model for Y is correctly specified, then the information criterion will attain its unique minimum at $\gamma^* = \gamma_0$. Of interest now is whether γ^* still equals γ_0 when the random-effects distribution is misspecified and, if not, it is relevant to know the magnitude of the difference between γ^* and γ_0. We further study these issues via simulation.

3.1. Consistency of the ML estimators

In this first simulation study, binary response data were generated using Model (1). This model includes a binary covariate, z_i, denoting the treatment (randomly assigned to 0 or 1 with equal probability) and a within-cluster covariate t_j, with values 0, 1, 2, 4, 6, and 8. For the mean structure, values close to the estimates in Table I were chosen: $\beta_0^0 = -8$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. Further, 9 different random-effects distributions, each with variances $\sigma_{0b}^2 = 1$, 4, 16, and 32, were included in the study. These distributions were a mean-zero normal density, a uniform distribution, an exponential distribution, a chi-square distribution, a lognormal distribution, a power function distribution, a discrete distribution with equal probability at two support

points, and finally both a symmetric and an asymmetric mixture of two normal densities. In some cases, the distributions were transformed such that the zero mean condition was satisfied, and the corresponding variances equaled the prespecified value σ^2_{0b}.

The distributions considered here cover a wide range of densities varying from very symmetric to very skewed; with potentially very heavy tails. Note that on the one hand, the variances $\sigma^2_{0b} = 16$ and 32 of the random effects will help us to investigate scenarios with variances in the same order of magnitude as the one observed in the case study. On the other hand, the smaller values considered for σ^2_{0b} should allow us to study the performance of the maximum likelihood estimators in less extreme settings. In this way, we cover a wide range of practically relevant situations.

The simulations were performed with 7 different sample sizes of 25, 50, 100, 200, 400, 800, and 1600 subjects. For each setting, 500 data sets were generated and Model (1) was then fitted to the generated data assuming normally distributed random effects. All analyses were carried out using the SAS procedure NLMIXED with adaptive Gaussian quadrature.

Consistency was studied through the evolution of the median maximum likelihood estimates, over increasing sample sizes. Table II displays the results for σ^2_b.

TABLE II – ABOUT HERE

Table II, as well as all other tables discussed in this manuscript, displays only the results from the converging analyses. A lack of convergence occurred mainly for small σ^2_{0b}, in combination with small sample sizes. The proportion of non-converging analyses could be as high as 30%, when the data were generated using a power function distribution with $\sigma^2_{0b} = 1$ and only 25 subjects. However, this rate quickly drops to 7% and less, as of 100 subjects, or when σ^2_{0b} increased to 4.

The results displayed in Table II show that the estimates of the variance component are severely affected by the misspecification in most settings. In general, substantial bias can occur even for small variance of the random effects, especially for the power function distribution and the asymmetric mixture of two normals. Additionally, the direction of the bias can change depending on the true underlying distribution. For most of the distributions considered here, the variance component is overestimated. However, in the case of the power function distribution and the asymmetric mixture, we observe serious underestimation of the variance of the random effects.

The parameters of the mean structure seem to be less affected by the misspecification. The median estimates of, for example, the treatment effect are shown in Table III.

Table III – About here

From this table it can be seen that the bias related to the estimation of β_1 is generally small when the variance of the random effects is small. However, as of $\sigma^2_{0b} = 16$, bias of 20% and more can occur, even for relatively big sample sizes of 400 subjects. Given that the estimate of the variance component is the only tool to study the variability of the true random-effects distribution, this highly biased estimate makes it difficult to evaluate whether or not problems can occur in the mean structure as well.

The relative bias of the within-cluster coefficient, i.e., the time effect, remained under 5% in all scenarios considered (results not shown here). This concurs with results obtained by Heagerty and Kurland [9] and Chen *et al.* [7]. The latter argue that, since the estimation of the treatment effect and the random intercept are subject to between-individual variation, we could expect misspecification of the random-effects distribution to affect the quality of these estimates. However, a covariate which changes within subjects, would be roughly orthogonal

to between-individual effects and therefore less affected by the misspecification.

Further, to study the extent to which the results obtained from a logistic-normal model generalize to wider scenarios, we also generated binary responses using the model given by

$$\text{logit}\{P(y_{ij}=1|\boldsymbol{b}_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_{0i} + b_{1i}t_j, \tag{3}$$

which now includes a random slope for time. For the mean structure parameters, we considered $\beta_0^0 = -6$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. The random effects were generated from two multivariate distributions, including a multivariate normal $\boldsymbol{b}_i \sim N(\mathbf{0}, V)$ and a symmetric mixture of two multivariate normals $\boldsymbol{b}_i \sim \frac{1}{2}N(\boldsymbol{\mu}, D) + \frac{1}{2}N(-\boldsymbol{\mu}, D)$, where

$$D = \begin{pmatrix} d & d_{12} \\ d_{12} & d \end{pmatrix}.$$

In the case of the mixture, $\boldsymbol{\mu} = (4,4)^T$, $d = 1, 4$ and d_{12} was chosen such that $\rho = \text{corr}(b_{0i}, b_{1i}) = 0.5, 0.9$. This results in the following overall covariance matrices $V = \{\sigma_{ij}\}_{i,j=1,2}$ for $\boldsymbol{b}_i$

$$V_1 = \begin{pmatrix} 5 & 4.5 \\ 4.5 & 5 \end{pmatrix}, V_2 = \begin{pmatrix} 5 & 4.9 \\ 4.9 & 5 \end{pmatrix}, V_3 = \begin{pmatrix} 8 & 6 \\ 6 & 8 \end{pmatrix}, V_4 = \begin{pmatrix} 8 & 7.6 \\ 7.6 & 8 \end{pmatrix}.$$

These same covariance matrices were also used to generate the multivariate normal random effects. The simulations were performed with 50 and 100 subjects. For each setting, 500 data sets were generated and the model given by (3) was fitted to the generated data under the assumption of normally distributed random effects. The medians of the corresponding maximum likelihood estimates are shown in Table IV.

Table IV – About here

Clearly, including a further random effect increased the impact of the misspecification. Even though the variances used for the random effects in this setting were small, considerable bias

is observed for all parameters in the mean structure. Interestingly and unlike the results previously obtained, now the time effect is severely underestimated. Obviously, adding a random time effect induced a bias in the corresponding fixed effect under misspecification. Like before, the estimates of the variance components were most affected in this scenario, where a large bias was observed for all elements of the variance covariance matrix.

We should note that, when the random effects were generated from a mixture, we observed a high proportion of non-converging analyses (ranging between 57% and 72% of the total of 500 runs). It is also possible that, as a result of the misspecification, in some of the simulations the procedure to maximize the likelihood had converged to an ill-conditioned maximum, leading to some extreme estimates. In any case, these results illustrate that the impact of the random-effects misspecification is even worse in the presence of complicated covariance structures.

3.2. Hypothesis testing

In many situations, data analysts consider test statistics and corresponding p-values to evaluate, for example, whether or not a drug has a significant influence. Even though consistency has been studied to some extent in the literature, there does not seem to be much research done on the behaviour of the test statistics. Therefore, additional simulations with the logistic-normal model given by (1) were carried out for different values of the treatment effect β_1^0 to investigate the robustness of the inferential procedures. These simulations were performed for 3 different sample sizes (25, 100, and 400 subjects) and a total of 5 different β_1^0 values (0, 0.5, 1, 2, and 5). For each setting, 500 data sets were generated and the proportion of cases in which the procedure detected a treatment effect different from zero (on a 5% significance level) was determined. When there is no treatment effect, this proportion corresponds to the

type I error; for the other values of β_1^0, this proportion represents the power of the analysis. The results of these simulations are summarized in Figure 2.

FIGURE 2 – ABOUT HERE

The display is limited to a few distributions, including the normal, the power function, the discrete, and the asymmetric mixture of two normals. It is clear from these graphs that misspecification can severely affect the power of the analysis, depending on the shape and the variance of the real random-effects distribution. Actually, the power can be seriously affected even in settings where the random intercept accounts for a small variability. For example, let us consider in Figure 2 the graphs corresponding to a sample of 100 patients, where $\beta_1^0 = 1$. Even with $\sigma_{0b}^2 = 1$, the power to detect a significant treatment effect can drop as low as 20% for the power function distribution, whereas for the correctly specified model, we observed a value around 70%. On the other hand, in some cases, like the power function distribution and the asymmetric mixture, the misspecification can lead to an increase of the test's power. A closer analysis of this issue shows that this phenomenon is associated with underestimation of the parameter estimators' variability. Note that both the power function distribution and the asymmetric mixture considered in our simulations are severely skewed, which makes us speculate that this increase of power could be related to the skewness of the underlying random-effects distribution and/or apply to specific alternatives only. Obviously, this idea is based on empirical evidence only, and further research will be necessary before a more founded conclusion can be reached.

Interestingly, the Type I error rate (presented in the first panel of Table V) rarely exceeded the specified 5% level of significance in all the scenarios displayed in Figure 2.

TABLE V – ABOUT HERE

These findings concur with results obtained in Neuhaus *et al.* [6]. Indeed, these authors showed for a similar logistic-normal model that when $\beta_1^0 = 0$, the corresponding maximum likelihood estimator consistently estimates zero. It is possible to prove that in this situation the type I error rate will be preserved and therefore, a significant treatment effect could be considered reliable even though caution may be needed in the interpretation of the point estimates. As a consequence, we could be fairly confident about the borderline significant treatment effect observed in the analysis of the schizophrenia data.

However, this does not hold for all parameters in the model. For instance, in an additional simulation study the binary response variable was generated using Model (1), now with $\beta_0^0 = 0$ (and $\beta_1^0 = 2$, $\beta_2^0 = 1$). The findings of this study can be seen in the second panel of Table V. Here, we analyze the performance of the Wald test associated with the intercept parameter β_0. The results illustrate that the Type I error rate is severely affected by the misspecification. Even when the variance of the random intercept is small, e.g., when $\sigma_{0b}^2 = 1$, the Type I error rate can be dramatically inflated, up to 16% in one scenario. This suggests that the Type I error rate could be robust for those parameters in the fixed effect structure which do not have an associated random effect, and severely affected for the others.

4. ALTERNATIVE APPROACHES

The simulation studies presented in the previous section clearly show the need for alternative approaches when facing possible random-effects misspecification. In the introduction some of these approaches were already mentioned. In what follows we introduce and study the performance of another plausible alternative, the heterogeneity model, which consists of

replacing the normal random-effects distribution by a finite mixture of normals.

4.1. The heterogeneity model

The heterogeneity model is an extension of the GLMM, obtained by sampling the random effects $\boldsymbol{b}_i$ from a mixture of k normal distributions with mean vectors $\boldsymbol{\mu}_r$ and covariance matrix D, i.e., $\boldsymbol{b}_i \sim \sum_{r=1}^{k} \pi_r N(\boldsymbol{\mu}_r, D)$. The probability for a subject to belong to component r is π_r, with $\sum_{r=1}^{k} \pi_r = 1$. Note that each component has the same covariance matrix D. This constraint is necessary to avoid unbounded likelihoods [23].

Let $\boldsymbol{\pi}' = (\pi_1, ..., \pi_k)$ and $\boldsymbol{\gamma}$ be the vector containing the remaining parameters, i.e., the vector $\boldsymbol{\varphi}$ of unknown parameters common to all subjects, as well as all parameters in $\boldsymbol{\mu}_r$ and D. The joint density function of $\boldsymbol{y}_i$ can then be written as $f_i(\boldsymbol{y}_i) = \sum_{r=1}^{k} \pi_r f_{ir}(\boldsymbol{y}_i|\boldsymbol{\gamma})$ where

$$f_{ir}(\boldsymbol{y}_i|\boldsymbol{\gamma}) = \int f_i(\boldsymbol{y}_i|\boldsymbol{\varphi}, \boldsymbol{b}_i)\phi_r(\boldsymbol{b}_i)d\boldsymbol{b}_i.$$

Note that $\phi_r(\boldsymbol{b}_i)$ refers to the multivariate normal with mean $\boldsymbol{\mu}_r$ and covariance matrix D. Estimation is now based on the maximization of

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{N} \ln \left\{ \sum_{r=1}^{k} \pi_r f_{ir}(\boldsymbol{y}_i|\boldsymbol{\gamma}) \right\},$$

where $\boldsymbol{\theta}' = (\boldsymbol{\gamma}', \boldsymbol{\pi}')$, using the Expectation-Maximization (EM) algorithm described in Laird (1978).

In principle many distributions can be approximated with high precision by finite mixtures of normal densities what makes this approach theoretically appealing. Still, little research has been done to explore the actual performance of this model in practice.

4.2. The heterogeneity model: a simulation study

To further explore the potential of this model we consider again the generated data, described in Section 3.1, when the random intercept was sampled from a mean zero normal density, a uniform distribution, a lognormal distribution, a power function distribution and an asymmetric mixture of two normal densities. We limited this study to 50, 100 and 200 subjects, and to 100 data sets per setting. Now, Model (1) was fitted to the generated data but assuming a mixture of two normals for the random effects.

The first panel of Table VI shows the median maximum likelihood estimates $\widehat{\beta}_1$ obtained from fitting a two-component mixture.

Table VI – About here

When comparing these estimates with the estimates of the GLMM displayed in Table III, we observe that the difference between the two models is negligible when the variance of the random effect is small. However, as the variance increases, the heterogeneity model seems to perform better in the estimation of this parameter, especially when the sample size is small. Also in these simulations (results not included here) we found that the heterogeneity model is robust to the random-effects misspecification when estimating the time effect. The relative bias remained under 5% in all scenarios considered, even for $\sigma^2_{0b} = 32$. However, we still observed substantial bias when estimating the variance of the random effects, especially for the lognormal and the power function distribution (see the second panel in Table VI). As expected, using the heterogeneity model considerably improved the bias in the case of the asymmetric mixture of normals. Nevertheless, the variance of the random effects is still considerably underestimated. Further, observe that the model appears to be less efficient than the GLMM when the random

effects are generated from a normal distribution.

Additional simulations have shown that, as in the classical GLMM, the Type I error rate associated with the treatment effect is not affected by the choice of the random-effects distribution. To study whether the Type I error rate associated with β_0 improves using the heterogeneity model, we repeated the simulations described in Section 3.2, with random effects generated from a mean zero normal density, a uniform distribution, a lognormal distribution, a power function distribution and an asymmetric mixture of two normal densities. We considered 50, 100 and 200 subjects, and for each of these settings 100 data sets were generated. The results of these analyses are shown in Table VII.

Table VII – About here

In this table we study the performance of the Wald test associated with β_0. The results clearly illustrate that here, like before, the Type I error rate is severely affected by the random-effects misspecification when GLMM are used, however, it remains below the specified significance level when the data are analyzed using the heterogeneity model.

Although we cannot provide a clear indication that the model is fully robust against misspecification, we have seen from the simulations that the heterogeneity model tends to perform slightly or considerably better, depending on the setting, than the generalized linear mixed model, especially for small sample sizes. Further, due to computational and time constraints, it is difficult to assess the full power of the heterogeneity model. For instance, we have limited the simulations to two components with different means but equal variances. One could wonder if considering two or more components with equal means and different variances for the random effects could significantly improve the performance in some cases. Therefore, we believe the heterogeneity model is indeed worthy of consideration.

4.3. A sensitivity analysis

The previous studies seem to lead to two relevant remarks: i) misspecification of the random effects distribution can induce a large bias in the estimation of the variance components as well as a severe bias in the estimation of the mean structure in GLMM, ii) more robust alternatives like the heterogeneity model can represent a significant improvement in some scenarios but can still suffer from severe bias in others. Therefore, in this section we propose to include these alternatives within a sensitivity analysis framework, considering different distributions for the random effects. In the case of our example, this means that we will extend the analysis with some of the distributions used in the simulations, including the exponential, the chi-square, the uniform and the lognormal. It is important to point out that recent research has shown that such analysis can easily be carried out in standard statistical packages like SAS procedure NLMIXED using probability integral transformations [24]. We further consider a heterogeneity model with two and three components and a nonparametric approach using two and three support points. The corresponding parameter estimates are displayed in Table I. Note that the points estimates obtained from the models, considering unimodal non-normal random-effects distributions, are similar, especially if we exclude the ones that used the uniform and lognormal distributions and produced the highest AIC values. The inferential results were similar in all the cases as well. For instance, the treatment effect was significant in all the models except for the ones with the lognormal and exponential random effects which produced borderline p-values of 0.054 and 0.094 respectively. The variance of the random effect was rather large in all the scenarios considered, with a median value around 19. Therefore, all the models consistently hint on a very strong within-subject association. We could use the AIC to find the best fitting model, which in this case is the logistic-normal model.

Further, the task of choosing the number of components k in the heterogeneity model is not an easy one. One approach consists in fitting models with increasing numbers of components, and comparing them using likelihood ratio tests. For instance, we can consider testing $H_0 : k = 1$ versus $H_A : k = 2$. However, this null hypothesis can also be expressed as $H_0 : \pi_2 = 0$, which is clearly on the boundary of the parameter space. In this case, bootstrap simulations are required to derive the distribution of the likelihood ratio test statistic under the null [23]. An alternative ad-hoc approach consists in increasing the number of components k until some of the subpopulations reflect very small weights π_r, or until some subpopulations coincide [4]. In this manuscript, we have opted for the AIC to select the best fitting model. This resulted, again, in a preference for the one-component logistic-normal model.

As stated before, when the random-effects distribution is not of primarily interest, one can resort to a nonparametric approach. In this case the random-effects distribution can be approximated by a discrete distribution with a finite number of support points. The results of this analysis are shown in the final part of Table I. The estimates stabilized with 3 support points, and with this model we obtained values very close to the ones from the classical GLMM. For instance, the estimate of the treatment effect was close to the one obtained with the logistic-normal model and it was significantly different from zero. As a conclusion of this analysis, we can be rather confident about the results obtained from the logistic-normal model, and about the presence of a moderate treatment effect on the CGI scores of the schizophrenic patients.

5. CONCLUDING REMARKS

A commonly encountered perception among data analysts is that the choice of the random-effects distribution is not crucial for the quality of the inferences related with model parameters in GLMM. However, this is not a generally valid truth. Using a logistic-normal model we found that, the maximum likelihood estimators are no longer consistent. The bias is generally negligible for the mean-structure parameters, as far as the variance of the random effects is sufficiently small. This was the case for $\sigma^2_{0b} = 1$ and 4 in our simulations. However, caution is necessary when variances of 16 or higher are obtained. Note that large random-effects variances are not exceptional in clinical trials where little variability in the response is expected, for instance, when a placebo control group is used, whereas a more variable outcome pattern is expected in the treated group. In such a scenario, the mean parameters, including the treatment effect parameters, could be subject to considerable bias under misspecification. When more than one random effect is considered serious bias can appear even in the small variance settings. A cautionary remark is in place regarding the magnitude of the variances. While one might want to consider these in relative terms, against the background of measurement error, the presence of a mean-variance link, and hence a measurement error function that varies with covariate values, renders such a relative presentation cumbersome. Moreover, since the residual variance function is bounded between 0 and 0.25 in the binary case, we prefer to retain an absolute presentation.

On the other hand, the power and Type I error rate can be affected in important ways by such misspecification, regardless the variance of the random effects. We did, however, observe that the type I error seems to be maintained for some parameters in the mean structure like, for example, the treatment effect in the schizophrenia data. This would imply that we could be

confident about the presence of the treatment effect in this case, however, we should be careful not to put too much emphasis on the estimated size of the effect. We speculate that in general the Type I error rate could be maintained for those parameters in the mean structure that do not have an associated random effect. However, additional research is needed to determine whether this robustness holds.

Finally, note that the estimates of the variance components are always heavily biased, even for small variances of the underlying random-effects distribution. Given that these estimates are the only available tool to study the variability of the true distribution, the observed bias can make it difficult to evaluate whether problems in the mean structure occur. This bias in the variance components can also have severe consequences in applications in which the main interest is in the association structure. This is the case, for instance, in fields like surrogate marker validation, evaluation of the reliability of rating scales, or studies to analyze the criterion and predictive validity of psychiatric scales. It could also provoke misleading results when we want to predict the subject-specific trajectories or use the subject-specific parameters for classification.

In the light of our findings, it is therefore clear that research efforts should be geared towards models robust against this type of misspecification. The heterogeneity model could be, in certain circumstances, a plausible choice, especially when dealing with small sample sizes. However, our simulations showed that the model can be unstable and convergency can heavily depend on initial values. Even though the heterogeneity model performs slightly better than the GLMM, we still observed serious bias under certain model misspecifications.

Alternatively, we proposed to incorporate the heterogeneity model into a more general sensitivity analysis, considering different random-effects distributions and comparing the point

estimates and inferences obtained. Note that our sensitivity analysis is not a robust alternative to the classical GLMM, but rather an alternative to the lack of robustness of the GLMM. Indeed, given that we consider different choices for the random-effects distribution, one would expect the outcome to be optimal when the true random-effects distribution is very similar to one of these distributions. The idea is then to see how sensitive are our conclusions with respect to the distributional assumptions for the random effects. Similar results obtained under different assumptions will increase our confidence, different ones will increase our caution. Therefore, simulation studies are not of great value to evaluate the performance of such an approach.

In this work we have confined attention to the impact of misspecifying the random-effects distribution. However, misspecifications of other model aspects, such as the choice of link function, omitting an important covariate in the mean structure, using a random-intercept model when the random-effects distribution depends on measured covariates, and so forth, deserve a great deal of research attention too. It is becoming clear that there probably will not be a general easy answer on how to deal with model misspecification. Perhaps in some specific situations, good alternative models can be found by using e.g., random-effects distributions conjugate to the distribution of the outcome [25]. Still, an important topic for future research will be the development of diagnostic tools for detecting the lack of consistency. These tools, together with the ability to consider several random-effects distributions, would allow for a useful and, arguably, necessary sensitivity analysis.

ACKNOWLEDGEMENT

Financial support from the IAP research network # P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

REFERENCES

1. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data.* New York: Oxford University Press, 2002.
2. Molenberghs G, Verbeke G.*Models for discrete longitudinal data.* New York: Springer, 2005.
3. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982; **38**, 963–974.
4. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data.* New York: Springer-Verlag, 2000.
5. Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis.* 1997; **53**, 541–556.
6. Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika.* 1992; **79**, 755–762.
7. Chen J, Zhang D, Davidian M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random-effects distribution. *Biostatistics.* 2002; **3**, 347–360.
8. Agresti A, Caffo B, Ohman-Strickland P. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis.* 2004; **47**(3), 639–653.
9. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika.* 2001; **88**, 973–985.
10. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics.* 1999; **55**, 117–128.
11. Aitkin M. Meta-analysis by random effect modelling in generalised linear models. *Statistics in Medicine.* 1999;**18**, 2343–51.
12. Burr D, Doss H, Cooke GE, Goldschmidt-Clermont PJ. A meta-analysis of studies on the association of the platelet PlA polymorphism of Glycoprotein IIIa and risk of coronary heart disease. *Statistics in Medicine.* 2003; **22**, 1741–60.
13. Burr D, Doss H. A Bayesian semi-parametric model for random effects meta-analysis. *Journal of American Statistical Association.* 2005; **100**, 242–51.

14. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi-parametric models: applications in institutional comparisons. *Statistics in Medicine.* 2007; **26**, 2088–112.
15. Lee KJ, Thompson SG. Flexible parametric models for random effects distributions. *Statistics in Medicine.* 2007; DOI:10.1002/sim.2897
16. Marshall EC, Spiegelhalter DJ. Comparing institutional performance using Markov chain Monte Carlo methods. In: Everitt BS and Dunn G, ed. *Statistical analysis of medical data: New developments.* London: Arnold; 1998: 229–49.
17. Arends LR, Hoes AW, Lubsen J, Deiderik EG, Stijnen T. Baseline risk as a predictor of treatment benefit: three clinical meta-reanalysis. *Statistics in Medicine.* 2000;**19**, 3497–518.
18. Fieuws S, Spiessens B, Draney K. Mixture Models. In: De Boeck P, Wilson M ed. *Explanatory Item Response models: A Generalized Linear and Nonlinear Approach.* Springer: New York, 2004: 317–340
19. Alonso A, Geys H, Molenberghs G, Kenward MG, Vangeneugden T. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics.* 2004; **60**(4), 845–853.
20. Rubin DB. Inference and missing data. *Biometrika.* 1976; **63**, 581–592.
21. Kenward MG, Molenberghs G. *Missing Data in Clinical Studies*, Chichester: Wiley, 2007.
22. White H. Maximum likelihood estimation of misspecified models. *Econometrica.* 1982; **50**, 1–25.
23. Böhning D. *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others.* Monographs on Statistics and Applied Probability **81**, London: Chapman & Hall/CRC, 1999.
24. Nelson KP, Lipsitz SR, Fitzmaurice GM, Ibrahim J, Parzen M, Strawderman R. Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics.* 2006; **15**(1), 39–57.
25. Lee Y, Nelder JA. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B.* 1996; **58**, 619–678.

Table I. Case study - parameter estimates and standard errors using a logistic random-intercept model with the random effect (RE) assumed to follow a normal distribution, a mixture of 2 or 3 normals, a chi-square, an exponential, a uniform, a lognormal, and finally a nonparametric (NP) distribution with 2 or 3 support points.

Model	$\widehat{\beta}_0$ (s.e.)	$\widehat{\beta}_1$ (s.e.)	$\widehat{\beta}_2$ (s.e.)	$\widehat{\sigma}_b^2$ (s.e.)	AIC
GLMM	-7.37 (1.18)	2.14 (1.08)	0.65 (0.096)	21.01 (6.81)	391.9
RE, χ^2	-6.99 (1.18)	1.92 (0.88)	0.66 (0.096)	18.20 (6.07)	392.5
RE, exp.	-6.35 (1.00)	1.70 (0.88)	0.64 (0.098)	10.71 (2.76)	397.3
RE, uni.	-5.47 (0.75)	1.38 (0.64)	0.56 (0.082)	9.54 (1.93)	408.3
RE, logn.	-5.17 (0.78)	1.20 (0.71)	0.57 (0.086)	19.34 (7.52)	409.6
Mixture, $k = 2$	-7.88 (1.23)	1.99 (0.94)	0.67 (0.096)	28.03 (8.66)	395.1
Mixture, $k = 3$	-7.77 (4.28)	2.70 (0.85)	0.68 (0.094)	20.20 (31.6)	396.0
NP, $k = 2$	-4.84 (0.59)	1.28 (0.33)	0.52 (0.085)	-	-
NP, $k = 3$	-5.67 (0.72)	2.06 (0.54)	0.59 (0.097)	-	-

Table II. Median of the maximum likelihood estimates $\widehat{\sigma}_b^2$ obtained from fitting Model (1) to the data generated using the different random-effects distributions, sample sizes and values for σ_{0b}^2.

n	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
	Normal distribution				Uniform distribution			
25	1.34	3.43	14.70	30.38	1.38	3.86	16.51	39.84
50	1.05	3.85	15.68	31.99	1.02	4.12	16.79	41.61
100	1.01	3.74	15.65	32.52	0.98	4.36	17.16	41.66
200	0.98	3.98	15.72	31.45	0.91	4.25	17.29	41.37
400	1.00	3.89	15.83	32.38	0.94	4.21	17.12	42.06
800	1.00	4.03	15.76	32.11	0.97	4.21	17.19	40.95
1600	0.99	3.98	15.93	32.41	0.96	4.22	17.05	40.62
	Exponential distribution				Chi-square distribution			
25	1.77	5.19	20.36	46.93	1.85	5.01	18.18	37.48
50	1.22	5.12	21.20	46.75	1.42	5.18	19.01	37.88
100	1.33	5.28	21.78	45.45	1.36	5.26	20.12	38.54
200	1.21	5.39	21.53	45.08	1.36	5.39	20.60	37.27
400	1.29	5.44	21.69	46.25	1.45	5.53	20.20	38.39
800	1.32	5.50	21.44	45.25	1.46	5.50	20.38	38.17
1600	1.30	5.47	21.67	44.62	1.46	5.47	20.43	37.90
	Lognormal distribution				Power function distribution			
25	1.39	3.80	9.34	14.11	1.03	2.33	6.27	10.96
50	1.31	4.22	9.36	13.85	0.75	2.03	6.15	11.21
100	1.29	4.07	10.00	13.91	0.68	1.90	6.22	11.93
200	1.28	4.30	9.56	13.85	0.69	1.97	6.18	11.58
400	1.30	4.25	9.69	14.23	0.66	1.98	6.18	11.50
800	1.32	4.40	9.83	14.15	0.67	2.01	6.22	11.61
1600	1.34	4.35	9.81	14.04	0.66	1.99	6.21	11.55
	Discrete distribution				Symmetric mixture of two normals			
25	1.10	4.17	19.94	37.16	1.23	3.89	17.66	41.80
50	1.03	4.27	19.40	38.61	0.96	4.34	18.22	41.06
100	1.04	4.31	20.08	39.34	0.92	4.00	18.86	40.09
200	0.98	4.33	19.79	39.22	0.99	4.22	18.44	40.85
400	0.98	4.33	19.87	40.16	1.00	4.23	18.91	40.71
800	1.00	4.40	19.81	39.88	0.98	4.22	18.50	40.46
1600	1.01	4.39	19.82	39.61	1.01	4.19	18.62	40.46
	Asymmetric mixture of two normals							
25	1.14	2.17	5.99	8.00				
50	1.00	2.16	5.93	8.36				
100	0.81	2.10	6.10	8.39				
200	0.80	2.11	6.31	8.43				
400	0.79	2.12	6.24	8.32				
800	0.81	2.16	6.32	8.35				
1600	0.79	2.16	6.26	8.36				

Table III. Median of the maximum likelihood estimates $\widehat{\beta}_1$ obtained from fitting Model (1) to the data generated using the different random-effects distributions, sample sizes and values for σ^2_{0b} (note that $\beta^0_1 = 2$).

n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
	Normal distribution				Uniform distribution			
25	2.15	2.04	2.10	2.11	2.43	2.32	2.92	0.79
50	2.12	2.06	2.10	2.11	2.05	1.91	1.96	2.57
100	2.01	2.03	2.04	2.04	1.93	1.79	1.76	2.33
200	2.02	2.05	2.03	1.98	1.98	1.90	1.95	2.38
400	2.01	1.97	1.98	2.02	1.99	1.90	1.87	1.86
800	2.01	2.01	1.98	2.02	1.99	1.97	1.92	1.85
1600	1.99	1.99	1.99	2.02	1.98	1.93	1.86	1.83
	Exponential distribution				Chi-square distribution			
25	2.33	2.47	2.65	4.36	2.40	2.66	1.96	2.31
50	2.05	2.07	1.97	2.96	2.17	2.26	2.17	2.43
100	2.04	2.14	2.29	2.64	2.13	2.21	2.08	2.09
200	2.06	2.20	2.42	2.15	2.10	2.21	2.16	2.11
400	2.12	2.23	2.53	2.89	2.09	2.24	2.12	2.05
800	2.06	2.20	2.36	2.61	2.09	2.22	2.10	2.09
1600	2.05	2.19	2.47	2.60	2.10	2.25	2.13	2.08
	Lognormal distribution				Power function distribution			
25	2.24	2.34	2.48	2.57	2.12	1.89	1.58	1.34
50	2.07	2.26	2.46	2.64	2.04	1.95	1.77	1.93
100	2.10	2.20	2.46	2.60	2.00	1.91	1.74	1.81
200	2.05	2.20	2.37	2.60	1.97	1.90	1.75	1.97
400	2.06	2.20	2.38	2.56	1.95	1.84	1.64	1.63
800	2.08	2.23	2.42	2.55	1.95	1.86	1.73	1.71
1600	2.08	2.21	2.39	2.55	1.94	1.84	1.68	1.70
	Discrete distribution				Symmetric mixture of two normals			
25	2.25	2.18	1.81	2.88	2.19	1.86	1.98	1.93
50	2.09	1.94	1.55	1.36	2.05	2.00	1.82	1.90
100	2.07	1.98	1.64	1.69	2.02	1.99	1.73	2.01
200	2.03	2.03	1.75	1.34	2.03	1.96	1.72	1.91
400	2.07	2.12	1.99	2.04	2.02	1.95	1.72	1.93
800	2.04	2.04	1.79	1.90	1.98	1.96	1.79	1.89
1600	2.07	2.07	1.89	1.52	2.00	1.96	1.76	1.92
	Asymmetric mixture of two normals							
25	2.09	2.11	1.78	1.67				
50	2.10	1.92	1.79	1.78				
100	1.97	1.94	1.72	1.78				
200	1.97	1.92	1.80	1.70				
400	1.97	1.92	1.80	1.70				
800	1.97	1.92	1.76	1.72				
1600	1.96	1.93	1.76	1.70				

Table IV. Median of the maximum likelihood estimates obtained from fitting Model (3) to the data with random effects generated from a multivariate normal and a symmetric mixture of two multivariate normal distributions.

	Real		Normal		Mixture	
	value	V	$n = 50$	$n = 100$	$n = 50$	$n = 100$
Fixed effects						
β_0	-6	V_1	-6.45	-6.14	-10.34	-10.52
		V_2	-6.34	-6.33	-9.76	-9.28
		V_3	-6.46	-6.19	-9.65	-9.37
		V_4	-6.71	-6.37	-10.06	-9.49
β_1	2	V_1	2.09	2.04	3.27	2.52
		V_2	2.10	2.09	3.19	2.43
		V_3	2.03	2.10	3.17	2.04
		V_4	2.11	2.01	3.69	2.73
β_2	1	V_1	1.04	1.04	-0.63	-0.13
		V_2	0.99	1.02	-0.14	-0.25
		V_3	0.98	0.88	-0.09	0.15
		V_4	0.99	0.94	-0.23	0.03
Variance structure						
σ_{11}	5	V_1	4.98	5.00	48.60	57.38
		V_2	5.32	5.89	45.43	49.18
	8	V_3	8.63	8.12	41.97	49.32
		V_4	8.87	8.32	49.16	59.76
σ_{22}	5	V_1	6.84	5.84	967.80	500.26
		V_2	6.41	6.37	969.68	597.18
	8	V_3	9.23	8.21	912.06	415.04
		V_4	11.17	10.20	962.09	453.80
σ_{12}	4.5	V_1	4.20	3.64	189.05	125.19
	4.9	V_2	4.10	4.82	174.16	134.14
	6	V_3	6.15	5.60	167.51	115.42
	7.6	V_4	7.76	7.36	176.39	131.02

Table V. Type I error for detecting a significant treatment effect when $\beta_1^0 = 0$, and a significant intercept when $\beta_0^0 = 0$ in the logistic random-intercept Model (1). Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.

		$\beta_1^0 = 0$				$\beta_0^0 = 0$			
Distribution	n	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
Normal	25	0.012	0.025	0.029	0.025	0.014	0.035	0.016	0.023
	100	0.041	0.052	0.050	0.026	0.042	0.048	0.040	0.034
	400	0.050	0.046	0.052	0.058	0.060	0.046	0.054	0.050
Power function	25	0.008	0.023	0.036	0.016	0.019	0.031	0.028	0.022
	100	0.041	0.040	0.050	0.028	0.043	**0.164**	**0.320**	**0.370**
	400	0.046	0.064	**0.076**	0.050	**0.158**	**0.682**	**0.946**	**0.962**
Discrete	25	0.023	0.012	0.014	0.004	0.021	0.046	**0.087**	0.073
	100	0.032	0.016	**0.084**	0.018	0.040	0.060	**0.136**	**0.156**
	400	0.048	**0.080**	0.024	**0.088**	**0.080**	**0.252**	**0.594**	**0.604**
Asymmetric mixture	25	0.014	0.014	0.018	0.038	0.015	0.025	0.011	0.045
	100	0.053	0.066	0.036	0.038	0.030	**0.328**	**0.408**	**0.886**
	400	0.053	0.057	0.036	0.032	**0.076**	**0.924**	**0.986**	**1.000**

Table VI. Median of the maximum likelihood estimates $\widehat{\beta}_1$ and $\widehat{\sigma}_b^2$ obtained from fitting Model (1) with the random intercept assumed to follow a mixture of two normal distributions, to the data generated using the different random-effects distributions, sample sizes and values for σ_{0b}^2.

		$\widehat{\beta}_1$			$\widehat{\sigma}_b^2$		
	n	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
Normal	50	2.02	2.06	1.56	3.92	14.93	24.50
	100	1.97	1.94	1.72	3.80	12.96	25.27
	200	2.02	1.86	1.95	3.68	14.32	27.17
Uniform	50	2.20	1.99	2.64	4.98	15.65	38.26
	100	2.21	2.16	2.36	4.02	15.12	33.92
	200	2.15	1.99	2.38	4.50	14.53	35.53
Lognormal	50	2.19	2.29	2.03	3.53	7.15	9.87
	100	2.15	2.16	2.11	3.50	6.89	9.21
	200	2.13	1.98	2.13	3.91	6.70	9.45
Power function	50	2.21	2.33	2.55	3.33	8.49	13.91
	100	2.13	2.08	2.17	3.16	6.92	12.07
	200	1.99	2.00	2.17	2.79	7.46	13.60
Asymmetric mixture	50	2.16	2.09	1.92	3.50	11.51	14.96
	100	2.15	2.05	2.05	3.04	12.36	21.08
	200	1.99	1.94	2.05	2.64	11.92	21.24

Table VII. *Type I error of the heterogeneity and the GLMM for detecting a significant intercept when* $\beta_0^0 = 0$. *Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		Heterogeneity model			GLMM		
Distribution	n	$\sigma^2_{0b} = 4$	$\sigma^2_{0b} = 16$	$\sigma^2_{0b} = 32$	$\sigma^2_{0b} = 4$	$\sigma^2_{0b} = 16$	$\sigma^2_{0b} = 32$
Normal	50	0.000	0.021	0.095	0.040	0.000	0.070
	100	0.000	0.020	0.030	0.080	0.040	0.040
	200	0.000	0.000	0.010	0.060	0.030	0.040
Uniform	50	0.000	0.011	0.079	0.000	0.030	0.020
	100	0.000	0.010	0.020	0.060	0.030	0.080
	200	0.000	0.030	0.040	0.050	0.040	0.060
Lognormal	50	0.000	0.000	0.019	**0.160**	**0.440**	**0.620**
	100	0.000	0.000	0.028	**0.180**	**0.540**	**0.880**
	200	0.000	0.000	0.039	**0.360**	**0.930**	**1.000**
Power function	50	0.000	0.000	0.084	0.050	**0.160**	**0.120**
	100	0.033	0.010	0.010	0.060	**0.250**	**0.320**
	200	0.010	0.000	0.000	**0.290**	**0.580**	**0.760**
Asymmetric mixture	50	0.000	0.000	0.000	**0.160**	**0.113**	**0.429**
	100	0.026	0.010	0.025	**0.370**	**0.370**	**0.910**
	200	0.000	0.040	0.011	**0.660**	**0.790**	**1.000**

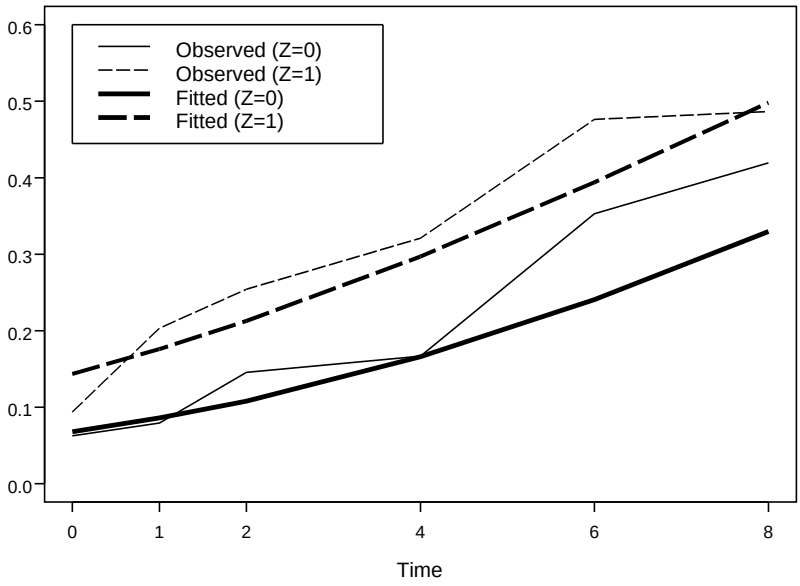
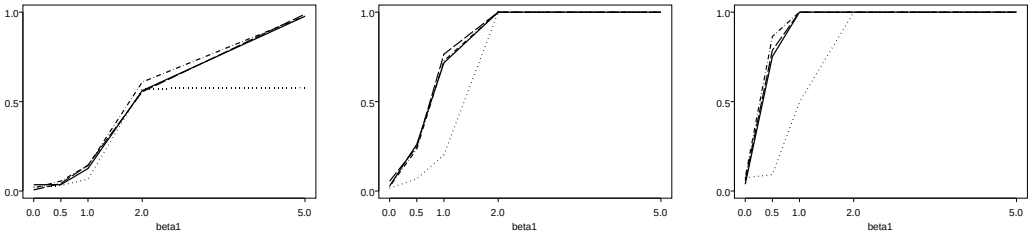
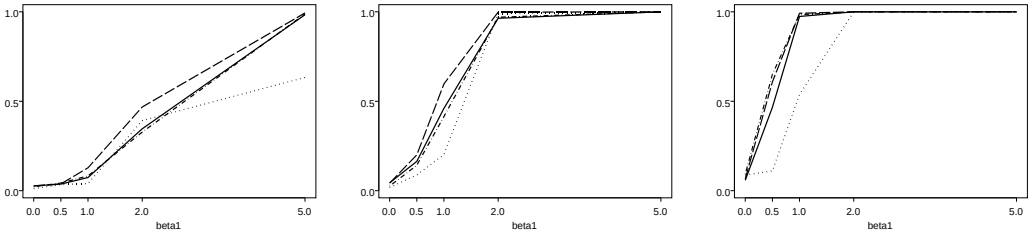


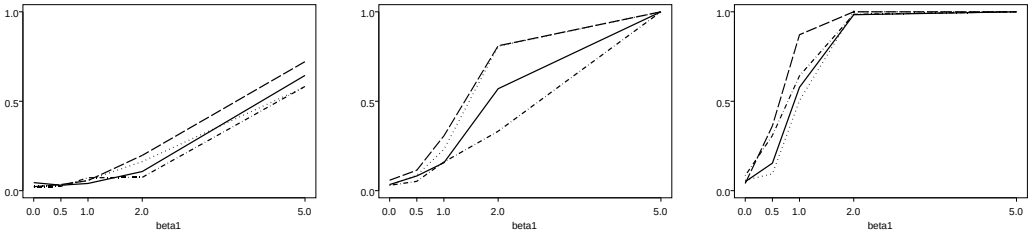
Figure 1. Evolution of the observed and fitted (using Model (1)) probabilities to be classified as a normal to mildly ill patient by treatment group. Here $Z = 1$ (0) denotes the treatment (control) group.



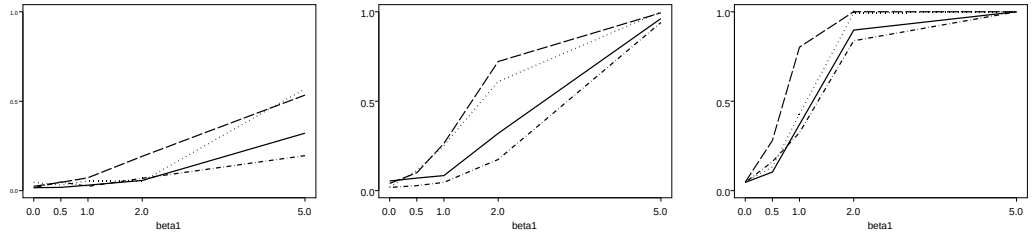
(a) $\sigma^2_{b0} = 1$, $n = 25$ (b) $\sigma^2_{b0} = 1$, $n = 100$ (c) $\sigma^2_{b0} = 1$, $n = 400$



(d) $\sigma^2_{b0} = 4$, $n = 25$ (e) $\sigma^2_{b0} = 4$, $n = 100$ (f) $\sigma^2_{b0} = 4$, $n = 400$



(g) $\sigma^2_{b0} = 16$, $n = 25$ (h) $\sigma^2_{b0} = 16$, $n = 100$ (i) $\sigma^2_{b0} = 16$, $n = 400$



(j) $\sigma^2_{b0} = 32$, $n = 25$ (k) $\sigma^2_{b0} = 32$, $n = 100$ (l) $\sigma^2_{b0} = 32$, $n = 400$

Figure 2. Power of the analysis of the logistic random-effects Model (1) to detect a significant treatment effect over a range of possible β_1^0 values, for 4 random-effects distributions: normal (solid line), power function (dotted line), discrete (dash-dotted line) and asymmetric mixture (dashed line).