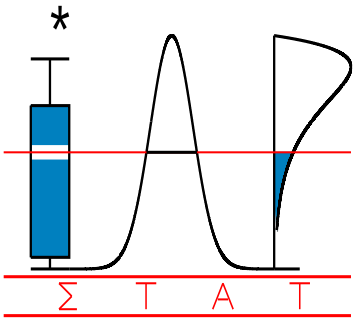
T E C H N I C A L

R E P O R T

06106

A SENSITIVITY ANALYSIS FOR RANDOM-EFFECTS MISSPECIFICATION IN GENERALIZED LINEAR MIXED MODELS

LITIERE, S., ALONSO, A. and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Sensitivity Analysis for Random-effects Misspecification in Generalized Linear Mixed Models

Saskia Litière[1], **Ariel Alonso**[2], **and Geert Molenberghs**[3]

Center for Statistics, Hasselt University,

Agoralaan, Building D, B3590 Diepenbeek, Belgium.

[1] *E-mail:* saskia.litiere@uhasselt.be

Tel: +32-(0)-11-268282

Fax: +32-(0)-11-268299

[2] *E-mail:* ariel.alonso@uhasselt.be

[3] *E-mail:* geert.molenberghs@uhasselt.be

Abstract

Recent research is showing that generalized linear mixed models (GLMM) may not be robust against certain model misspecifications. In this work we focus on misspecifying the random-effects distribution and its impact on maximum likelihood estimation. We propose to deal with possible misspecification by way of sensitivity analysis, considering several random-effects distributions. First, we analyze a case study using the heterogeneity model, i.e., a GLMM where the normal random-effects distribution is replaced by a finite mixture of normals. It is shown through simulations that this model performs slightly better in the presence of misspecification. We complete the sensitivity analysis of the case study with a Bayesian approach, where we fit logistic models with different distributions for the random effects. Here the Deviance Information Criterion (DIC) can be used as a criterion to select the most appropriate model.

Some Keywords: Bayesian Modeling, Heterogeneity model, Hierarchical models, Maximum likelihood, Misspecification, Random-effects.

Running Title: Misspecification in GLMM.

1 Introduction

When dealing with non-Gaussian longitudinal measurements, observations on a subject are unlikely to be independent. This dependence can be taken into account using subject-specific parameters.

A popular approach to handle this type of correlated data is the generalized linear mixed model (GLMM; Agresti, 2002; Diggle, Heagerty, Liang and Zeger, 2002; Fahrmeir and Tutz, 2001; Molenberghs and Verbeke, 2005). This model has been widely used in different areas like, e.g., toxicology (Molenberghs and Verbeke, 2005), epidemiology (Kleinman, Lazarus and Platt, 2004), dairy science (Tempelman, 1998), etc., and is easy to apply using software tools such as the SAS procedures NLMIXED and GLIMMIX. In this model, conditional on the random effects $\boldsymbol{b}_i$, the outcome variable $\boldsymbol{y}_i$ for subject i follows a pre-specified distribution $F_i(\boldsymbol{y}_i|\boldsymbol{\varphi}, \boldsymbol{b}_i)$, parameterized through a vector $\boldsymbol{\varphi}$ of unknown parameters common to all subjects. The subject-specific effects $\boldsymbol{b}_i$ are assumed to come from a distribution $G(\boldsymbol{b}_i|\boldsymbol{\delta})$, which may depend on a vector $\boldsymbol{\delta}$ of unknown parameters. Estimation is usually based on maximum likelihood, assuming that the underlying probability model is correctly specified.

A wide range of software tools are available for fitting these models. However, the analysis is often limited to the setting of Gaussian random effects. Since random effects are not observed, diagnostic tools to study the random-effects distribution are not straightforward. Indeed, one should be careful in using empirical Bayes estimates of the random effects to detect departures from normality. Even when the random effects are coming from a normal distribution, the empirical Bayes estimates will usually not be normally distributed. Therefore, it is relevant to assess the robustness of the parameter estimates with respect to this type of model misspecification.

Verbeke and Lesaffre (1997) showed that the maximum likelihood estimators for fixed effects and variance components in linear mixed models, obtained under the assumption of normally distributed random effects, are consistent and asymptotically normally distributed, even when the random-effects distribution is not normal. However, results obtained in recent years show that moving away from the realm of normality leads to qualitative differences. For instance, Neuhaus, Hauck and Kalbfleisch (1992) examined the performance of mixed-effects logistic regression models with misspecified random-effects distributions. They showed that the maximum

likelihood estimators of the model parameters are inconsistent but that the magnitude of the bias is typically small. Simulations by Chen, Zhang, and Davidian (2002) with a comparable model also indicate that the estimation of the regression coefficients may be subject to negligible bias only. According to Agresti, Caffo, and Ohman-Strickland (2004), 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 qualities of characteristics involving the random effects as well as the fixed effects. Similarly, Heagerty and Kurland (2001) found substantial bias while using a random-intercept logistic model, when the random-effects distribution depends on measured covariates.

Nevertheless, we should underscore 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. For example, in Agresti, Caffo, and Ohman-Strickland (2004) the largest random-effects variance used for simulations was equal to 1. As we will illustrate with our case study in Section 2, these small values may not always be realistic. Litière, Alonso and Molenberghs (2007b) found, using simulations with a random-intercept logistic model and a wide range of distributions for the random effect, that the estimates of the variance components are always subject to considerable bias when the random-effects distribution is misspecified. Although variance components are generally treated as nuisance parameters, this bias can have an important impact in studies where they are of main interest. This is the case, for instance, in fields like surrogate marker validation, reliability of rating scales, or studies of the criterion and predictive validity of psychiatric scales.

Furthermore, the bias induced in the estimates of the mean structure parameters appears to depend on the magnitude of the variance components, whereby large bias is associated with large random-effects variances. Additionally, Litière, Alonso, and Molenberghs (2007a) established that

the type I error and the power related to the tests of the mean structure parameters, can also be severely impacted. Clearly, in any practical situation, the bias present in the variance component estimators, under misspecification, will make it hard to distinguish between the two scenarios, i.e., small or larger variance components. Therefore, it can be difficult to determine how severe the impact on the mean parameters can be.

This overview illustrates the wide range of opinions that exist in the literature with respect to the impact of misspecifying the random-effects distribution. Nevertheless, it becomes clear that more research is needed to find models that are more robust against this type of misspecification. Some alternative robust approaches have been suggested. Butler and Louis (1992) proposed to replace the normal random-effects distribution by a non-parametric distribution. However, Agresti, Caffo, and Ohman-Strickland (2004) reported that there can be some loss of efficiency, when using a non-parametric approach, compared to a parametric assumption close to the real distribution. Additionally, model comparison can be difficult as standard asymptotic theory does not apply. Chen, Zhang and Davidian (2002) 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. These authors then used a Monte Carlo EM algorithm with a rejection sampling scheme to obtain estimates of the model parameters and the semi-parametric random-effects 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 (Fieuws, Spiessens and Draney, 2004; Molenberghs and Verbeke, 2005). This allows one to cover a wide range of shapes for the random-effects distribution, including unimodal as well as multimodal, and symmetric as well as very skewed distributions.

We also 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,

irrespectively of the distribution used to obtain them, we can feel relatively confident about our results. On the other hand, if the results vary considerably, then they are obviously sensitive to our distributional assumptions for the random effects, and caution is needed. We could also use some known model selection criteria to select that distribution which fits the data best.

Since Bayesian models are very flexible in the choice of the random-effects distribution, they have been used in the past to deal with possible random-effects misspecification. For instance, Kleinman and Ibrahim (1998) suggest extending the GLMM by allowing the random effect to have a non-parametric prior distribution. More recently, Kizilkaya, Carnier, Albera, Bittante and Tempelman (2003) discussed a hierarchical threshold mixed model based on a cumulative t-link specification for the analysis of ordinal data. Additionally, Kizilkaya and Tempelman (2005) proposed a general Bayesian approach to model heteroskedastic error in GLMM, in which linked functions of conditional means and residual variances were specified as separate linear combinations of fixed and random effects. As a part of the sensitivity analysis, a Bayesian approach might therefore be a good alternative to the classic frequentist methods.

We will start by analyzing, in Section 2, the motivating case study using a GLMM with normal random effects. In Section 3, we will introduce and apply the heterogeneity model to the example. Next, in Section 4, various aspects of the parameter estimates resulting from the heterogeneity model are investigated through extensive simulations. Finally, in Section 5, we will consider a Bayesian approach to fitting GLMM with non-Gaussian random effects and perform a sensitivity analysis for our case study within the Bayesian framework.

2 Case Study: The Schizophrenia Data

Our case study consists of individual patient data from a randomized clinical trial, comparing the effect of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia (Alonso, Geys, Molenberghs, Kenward and Vangeneugden, 2004). The response variable Y

Table 1: *Parameter estimates (standard errors) and p-values for the parameters in the logistic random-intercept model given by (1), fitted to the schizophrenia data. The SAS procedure NLMIXED with adaptive Gaussian quadrature and 20 quadrature points was used.*

Effect	Parameter	Estimate (s.e.)	p-value
Fixed effects			
Intercept	β_0	-7.37 (1.18)	< 0.001
Treatment effect	β_1	2.14 (1.08)	0.049
Time effect	β_2	0.65 (0.10)	< 0.001
Variance structure			
Random intercept variance	σ_b^2	21.01 (6.81)	

is a dichotomous version of the Clinical Global Impression (CGI) scale and equals 1 for patients classified as normal to mildly ill, and 0 for patients classified as moderately to severely ill. The treatment variable, Z, is set to 0 for the control group and 1 for the risperidone group. 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. One hundred twenty-eight patients were included in the study.

Previous data analysis has shown that an adequate model for these data is given by (Litière, Alonso and Molenberghs, 2007a)

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

where y_{ij} denotes the response for the i-th patient at time t_j and b_i denotes a random intercept assumed to follow a mean zero normal distribution with variance σ_b^2. The estimates of the model parameters are shown in Table 1.

They give evidence of certain treatment effect. Note that we observed a large random-intercept variance. This could be explained by the high proportion of patients (75%) in the control group that have a response pattern of nothing but zeros. Such a high inter-subject correlation is accommodated in the model through a large value of the random-effect variance. Clearly, a large

random-effects variance should not be considered rare in clinical trials where, for example little variability in the response is expected in the placebo group and considerably more variability in the treated group. As a result, this variance could imply a serious bias for the mean structure parameters, including the treatment effect, if the random-effects distribution is misspecified (Litière, Alonso and Molenberghs, 2007b). Arguably, these circumstances could render the assumption of a normal distribution for the random effects questionable. Nevertheless, Litière, Alonso and Molenberghs (2007a) introduced a theoretical result which states the conditions under which the type I error is robust to misspecification of the random-effects distribution. Applied to our case study, this theorem implies that the type I error corresponding to the treatment effect will not be affected by the choice of the random-effects distribution. Therefore, we can be fairly confident about the presence of a (borderline) significant treatment effect. However, we should be careful when interpreting the estimated size of the effect due to the bias that can be introduced by misspecification.

As mentioned in the introduction, a plausible alternative approach to the GLMM consists in replacing the Gaussian random-effects distribution by a finite mixture of normals. In the next section we will study this model in more detail.

3 The Heterogeneity Model

In GLMM, the random-effects $\boldsymbol{b}_i$ are assumed to be sampled from a normal distribution. The heterogeneity model is an extension of this model, obtained by sampling the random effects from a mixture of k normal distributions with mean vectors $\boldsymbol{\mu}_r$ and covariance matrix $\boldsymbol{D}$, i.e., $\boldsymbol{b}_i \sim \sum_{r=1}^{k} \pi_r N(\boldsymbol{\mu}_r, \boldsymbol{D})$. 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 $\boldsymbol{D}$. This constraint is necessary to avoid unbounded likelihoods (Böhning, 1999).

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 $\boldsymbol{D}$. 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 $\boldsymbol{D}$. 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).

3.1 Case Study: The Heterogeneity Model

To fit a heterogeneity model to the case study, some small changes have to be made to the model formulation. For example, since there are no restrictions on the $\boldsymbol{\mu}_r$, the expected value of the random effects is no longer fixed at zero. Therefore, to avoid overparameterization, we will not include an intercept

$$\text{logit}\{P(Y_{it} = 1|b_i)\} = \beta_1 Z_i + \beta_2 t + b_i, \tag{2}$$

however, note that β_0 can be estimated from $\sum_{r=1}^{k} \pi_r \mu_r$. We will consider a random-effects distributions with two components

$$b_i \sim \pi_1 N(\mu_1, d) + (1 - \pi_1)N(\mu_2, d),$$

as well as with three components

$$b_i \sim \pi_1 N(\mu_1, d) + \pi_2 N(\mu_2, d) + (1 - \pi_1 - \pi_2)N(\mu_3, d).$$

For these mixtures, the overall variance of the random-effects can be calculated using the following expression

$$\sigma_b^2 = \sum_{j=1}^{k} \pi_j \mu_j^2 - \left(\sum_{j=1}^{k} \pi_j \mu_j \right)^2 + d, \tag{3}$$

Table 2: *Parameter estimates (standard errors) and p-values for the parameters in the model given by (2) with 2 and 3 mixture components, fitted to the schizophrenia data*

		$k=2$		$k=3$	
Effect	Parameter	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value
	Fixed effects				
Intercept	β_0	-7.88 (1.23)	< 0.001	-7.77 (4.28)	0.070
Treatment	β_1	1.99 (0.94)	0.034	2.70 (0.85)	0.002
Time	β_2	0.67 (0.10)	< 0.001	0.68 (0.10)	< 0.001
	Variance structure				
	μ_1	-9.19 (1.33)	< 0.001	-10.76 (5.37)	0.045
	μ_2	-5.31 (0.95)	< 0.001	-6.51 (1.97)	0.001
	μ_3			-2.95 (2.03)	0.146
	π_1	0.66 (0.01)		0.51 (0.45)	
	π_2			0.25 (0.58)	
	d	24.66 (7.66)		9.53 (11.9)	
AIC			395.1		396.0

with $k = 2$ or 3, depending on the number of components in the mixture. The parameter estimates obtained for the case study, using these two models, are shown in Table 2.

The estimates of the fixed effects for both models are very similar, and they are also close to the results of the one-component or homogeneity model shown in Table 1. The overall variance of the random effects can be calculated using (3) and the estimates given in Table 2. This leads to $\hat{\sigma}_b^2 = 28.03$ for the two-component model, and $\hat{\sigma}_b^2 = 20.21$ for the three-component model. These results are in the same order of magnitude of $\hat{\sigma}_b^2 = 21.01$ reported in Table 1. Additionally, there is not so much difference between the AIC of these models and the AIC corresponding to the homogeneity model, given by 391.0. Therefore, the one-component model seems to be most appropriate for the data at hand. Evidently, this result increases the level of confidence in our previous findings.

As illustrated here, the heterogeneity model could be a plausible alternative or extension to the

classical GLMM. Still, little research has been done to study the actual performance of this model in the presence of random-effects misspecification. Therefore, in the next section, we will explore the capabilities of the model via simulations.

4 A Simulation Study

In this simulation study, binary data were generated using the model given by (1). For the fixed effects, values close to the estimates in Table 1 were chosen: $\beta_0^0 = -8$, $\beta_1^0 = 2$ and $\beta_2^0 = 1$. Further, 5 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, a lognormal distribution, a power function distribution and an asymmetric mixture of two normal densities. When needed, they were transformed to satisfy the mean zero condition of the random effects. 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 $\sigma_{0b}^2 = 16$ 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_{0b}^2 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. Further, we considered 3 different sample sizes: 50, 100 and 200, and for each of these settings 100 data sets were generated. Model (2) was then fitted to the generated data assuming a mixture of two normals for the random-effects distribution. At the same time, the generated data were also analyzed using the model given by (1), i.e., the one-component or homogeneity model, assuming normal random effects (using the SAS procedure NLMIXED with gaussian quadrature and 50 quadrature points). This allows us to study whether the heterogeneity model is more robust than the GLMM in the presence of random-effects misspecification.

4.1 Consistency

Consistency was studied through the evolution of the relative distance between the estimates and their real value, over increasing sample size. Let $\boldsymbol{\gamma}_0 = (-8, 2, 1, 32)'$ represent the vector of true parameter values and $\hat{\boldsymbol{\gamma}}_n = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_b^2)'$ the corresponding vector of maximum likelihood estimates. We can then define the relative distance between $\boldsymbol{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_n$ as

$$d_\gamma = \frac{||\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0||}{||\boldsymbol{\gamma}_0||},$$

where $||.||$ denotes the Euclidean distance. The relative distance between the treatment effect estimate and its real value is similarly given by

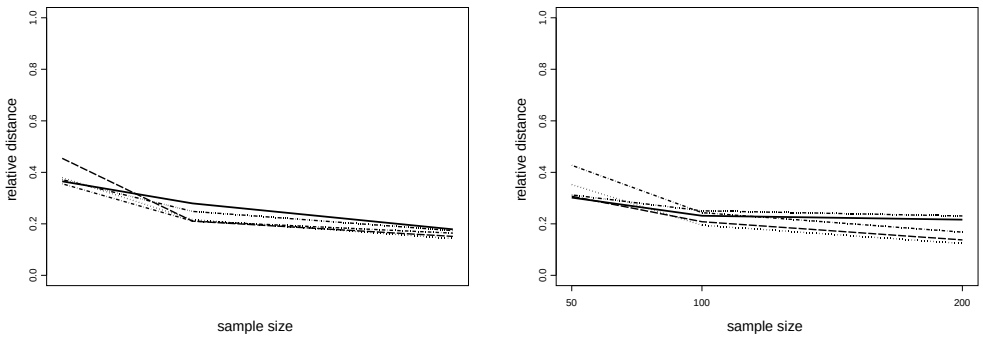
$$d_{\beta_1} = \left| \frac{\hat{\beta}_1 - \beta_1^0}{\beta_1^0} \right|.$$

If the estimators remain consistent after misspecifying the model then these relative distances should go to zero with increasing sample sizes. Figure 1 shows the evolution of the relative distance d_γ between $\boldsymbol{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_n$ over increasing variance and sample size, for both the homogeneity and the heterogeneity model.

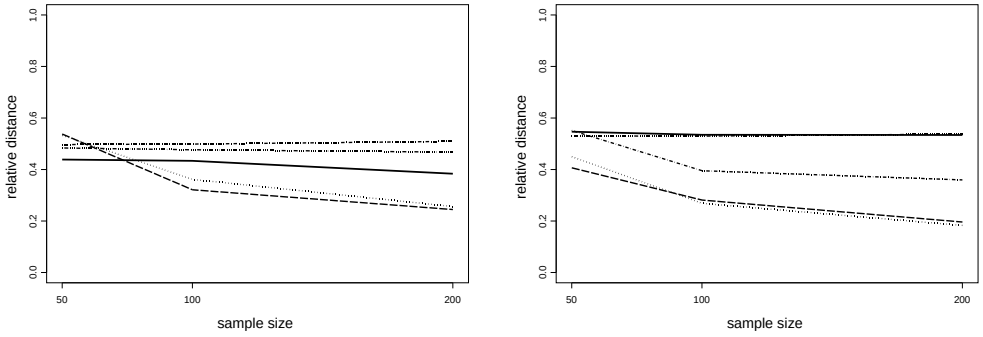
From these graphs it can be seen that, under the heterogeneity model, the smallest relative bias is observed for the correctly specified models, i.e., when the random-effects distribution is normal or an asymmetric mixture of normals. Additionally, note that, in the latter case, the heterogeneity clearly outperforms the homogeneity model, especially as σ_{0b}^2 increases.

Further, when the random-effects variance is small, the relative bias under both models is low. However, the relative bias of $\hat{\boldsymbol{\gamma}}_n$ can be as high as 50% when $\sigma_{0b}^2 = 16$ or 67% when $\sigma_{0b}^2 = 32$, as was the case for the data generated with lognormal random effects and analyzed using the heterogeneity model.

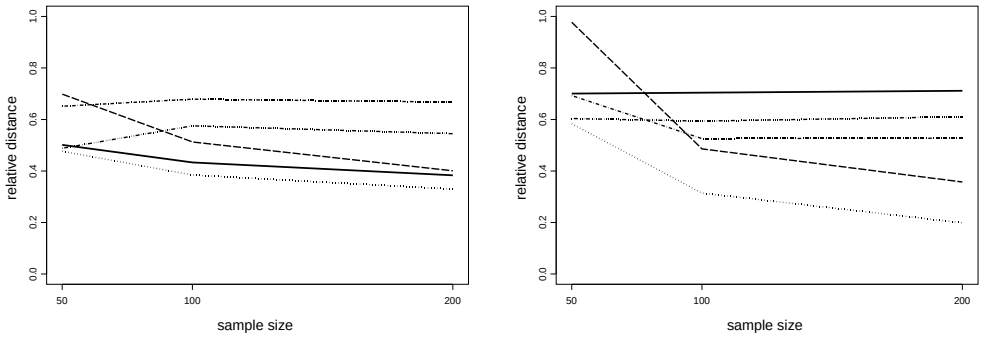
In general, the heterogeneity model seems to perform slightly better for smaller sample sizes. For example, Figure 1(e) clearly illustrates that when the sample size is 50 and $\sigma_{0b}^2 = 32$, the



(a) Heterogeneity model - $\sigma^2_{0b} = 4$ (b) Homogeneity model - $\sigma^2_{0b} = 4$



(c) Heterogeneity model - $\sigma^2_{0b} = 16$ (d) Homogeneity model - $\sigma^2_{0b} = 16$



(e) Heterogeneity model - $\sigma^2_{0b} = 32$ (f) Homogeneity model - $\sigma^2_{0b} = 32$

Figure 1: *Consistency of the parameter estimates under the homogeneity and heterogeneity model: the relative distance d_γ for each distribution over increasing σ^2_{0b} and sample size: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

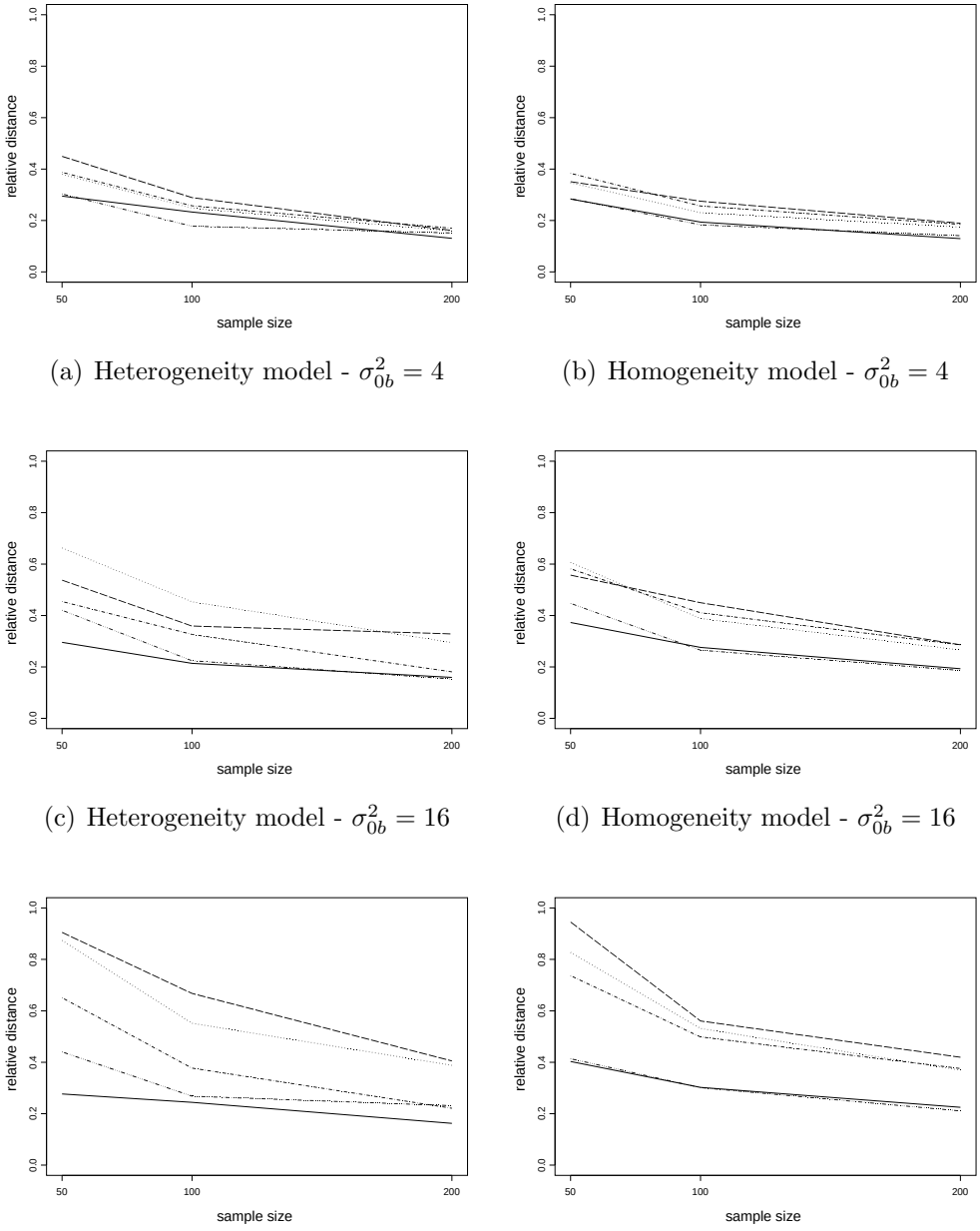
overall bias exceeds 60% only when the random effects are generated from a lognormal or uniform distribution. On the other hand, Figure 1(f) shows that for the homogeneity model the overall

bias is above 60% for all the misspecified random-effects distributions and it is near 100% when the random effects are generated from a lognormal distribution. The differences between both models are less dramatic for larger sample sizes.

Since interest often lies in the observed treatment effect, we also studied in more detail how its estimate can be influenced by the presence of random-effects misspecification. Figure 2 shows the relative distance d_{β_1} between the treatment effect and its estimate from fitting a heterogeneity and a homogeneity model respectively, for the different values of σ^2_{0b}. As before, 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 be more efficient in the estimation of this parameter, especially when the sample size is small.

Additionally, the simulations showed (results not included here) that the heterogeneity model is very 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 observed some substantial bias when estimating the variance of the random-effects (up to 57% for lognormal random effects, for sample size 200 and $\sigma^2_{0b} = 16$). As expected, using the heterogeneity model considerably improved the bias in the case of the asymmetric mixture of normals. For instance, even for $\sigma^2_{0b} = 32$ we observed a decrease from 73% bias under the homogeneity model down to less than 40% under the heterogeneity model.

Markedly, Figures 1(b), (d) and (f) show large relative distances for β_1, even under the correctly specified model with normal random effects. This result is counterintuitive since, under these conditions, the maximum likelihood method is expected to provide consistent estimates. However, it should be emphasized that consistency is an asymptotic result. After considerably increasing the sample size, the curve corresponding to the correctly specified (normal) model slowly decreases, while the other curves stabilize.



(a) Heterogeneity model - $\sigma^2_{0b} = 4$

(b) Homogeneity model - $\sigma^2_{0b} = 4$

(c) Heterogeneity model - $\sigma^2_{0b} = 16$

(d) Homogeneity model - $\sigma^2_{0b} = 16$

(e) Heterogeneity model - $\sigma^2_{0b} = 32$

(f) Homogeneity model - $\sigma^2_{0b} = 32$

Figure 2: *Consistency of the parameter estimates under the homogeneity and heterogeneity model: the relative distance d_{β_1} for each distribution over increasing σ^2_{0b} and sample size: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

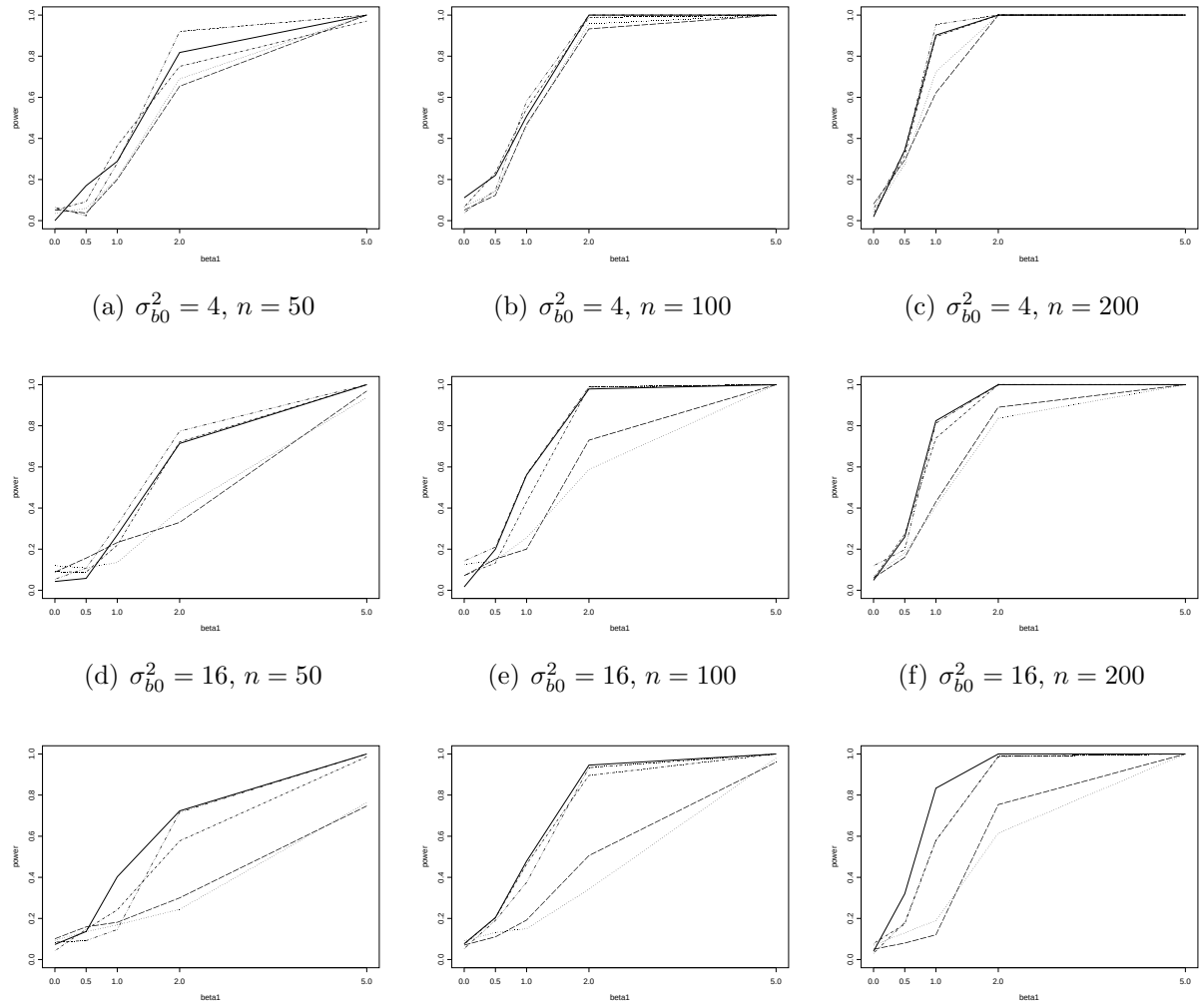
4.2 Hypothesis Testing

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. Litière, Alonso and Molenberghs

(2007a) showed that the type I error and the power can be seriously impacted by misspecifying the random-effects distribution. To study whether the heterogeneity model offers a solution to this problem, additional simulations were carried out for different values of the treatment effect β_1 to investigate the robustness of the inferential procedures. These simulations were performed for 3 different sample sizes (50, 100, and 200) and a total of 5 different β_1^0 values (0, 0.5, 1, 2, and 5). For each setting, 100 data sets were generated and the proportion of cases in which the heterogeneity and the homogeneity model 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 3 for the heterogeneity model, and for comparison in Figure 4 for the homogeneity model.

Also here we can see that there is not much difference between the performance of both models when σ_b^2 is small. However, as the variance increases, we can clearly see an increased power of the heterogeneity model to detect a significant treatment effect. For example, let us consider in Figures 3(e) and 4(e) the graphs corresponding to a sample of 100 patients, when $\beta_1 = 2$. By increasing the number of mixture components k in the heterogeneity model, the power to detect a significant treatment effect increases from 31% for $k = 1$ to 73% for $k = 2$. Similarly, when the random effects are generated from an asymmetric mixture, the use of two components increases the power from 79% to 97%. Clearly, a valuable result.

Further, the graphs also confirm the theoretical result presented in Litière, Alonso and Molenberghs (2007a), and briefly discussed in Section 2, which states that the type I error associated with the test for the presence of a covariate effect will not be affected by the (possibly wrong) choice of the random-effects distribution, as far as this covariate is not included in the random-effects structure. This is clearly the case for the treatment effect in our models. However, this is not the case for β_0. To study the type I error associated with β_0 under the heterogeneity model,



(a) $\sigma_{b0}^2 = 4$, $n = 50$ (b) $\sigma_{b0}^2 = 4$, $n = 100$ (c) $\sigma_{b0}^2 = 4$, $n = 200$

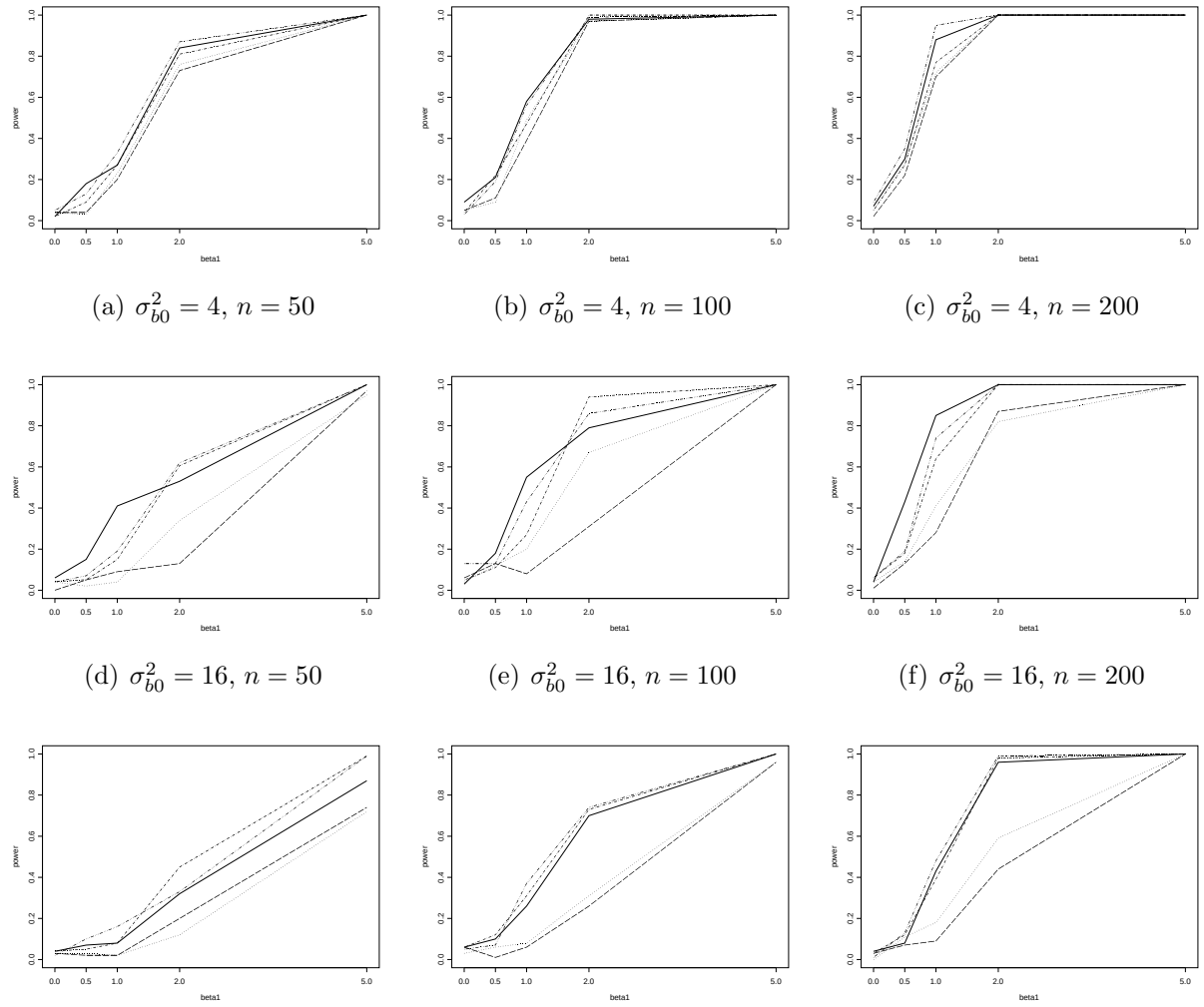
(d) $\sigma_{b0}^2 = 16$, $n = 50$ (e) $\sigma_{b0}^2 = 16$, $n = 100$ (f) $\sigma_{b0}^2 = 16$, $n = 200$

(g) $\sigma_{b0}^2 = 32$, $n = 50$ (h) $\sigma_{b0}^2 = 32$, $n = 100$ (i) $\sigma_{b0}^2 = 32$, $n = 200$

Figure 3: *Power of the heterogeneity model (2) to detect a significant treatment effect over a range of possible β_1^0 values, for 5 random-effects distributions: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

a third simulation study was organized, where 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 results using both a one-component and two-component mixture are shown in Table 3.

In this table we study the performance of the Wald test associated with β_0. The results clearly illustrate that the type I error is severely affected by the misspecification in the homogeneity



(a) $\sigma_{b0}^2 = 4$, $n = 50$ (b) $\sigma_{b0}^2 = 4$, $n = 100$ (c) $\sigma_{b0}^2 = 4$, $n = 200$

(d) $\sigma_{b0}^2 = 16$, $n = 50$ (e) $\sigma_{b0}^2 = 16$, $n = 100$ (f) $\sigma_{b0}^2 = 16$, $n = 200$

(g) $\sigma_{b0}^2 = 32$, $n = 50$ (h) $\sigma_{b0}^2 = 32$, $n = 100$ (i) $\sigma_{b0}^2 = 32$, $n = 200$

Figure 4: *Power of the homogeneity model (1) to detect a significant treatment effect over a range of possible β_1^0 values, for 5 random-effects distributions: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

model. Even when the variance of the random intercept is small, e.g., when $\sigma_b^2 = 4$, the type I error rate can be dramatically inflated, up to 66% in one scenario. However, this problem is clearly solved when 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

Table 3: *Type I error of the heterogeneity and the homogeneity model for detecting a significant treatment effect 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			Homogeneity Model		
Distribution		$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
Normal	$n = 50$	0.000	0.021	0.095	0.040	0.000	0.070
	$n = 100$	0.000	0.020	0.030	0.080	0.040	0.040
	$n = 200$	0.000	0.000	0.010	0.060	0.030	0.040
Asymmetric mixture	$n = 50$	0.000	0.000	0.000	**0.160**	**0.113**	**0.429**
	$n = 100$	0.026	0.010	0.025	**0.370**	**0.370**	**0.910**
	$n = 200$	0.000	0.040	0.011	**0.660**	**0.790**	**1.000**
Lognormal	$n = 50$	0.000	0.000	0.019	**0.160**	**0.440**	**0.620**
	$n = 100$	0.000	0.000	0.028	**0.180**	**0.540**	**0.880**
	$n = 200$	0.000	0.000	0.039	**0.360**	**0.930**	**1.000**
Uniform	$n = 50$	0.000	0.011	0.079	0.000	0.030	0.020
	$n = 100$	0.000	0.010	0.020	0.060	0.030	0.080
	$n = 200$	0.000	0.030	0.040	0.050	0.040	0.060
Power function	$n = 50$	0.000	0.000	0.084	0.050	**0.160**	**0.120**
	$n = 100$	0.033	0.010	0.010	0.060	**0.250**	**0.320**
	$n = 200$	0.010	0.000	0.000	**0.290**	**0.580**	**0.760**

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 imagine that considering two or more components with equal means and differing variances for the random effects could significantly improve the performance in some cases. Additionally, even though the heterogeneity model is very sensitive to the choice of starting values, in the analysis of a practical data set, one can try out different sets of starting values in order to improve convergence. Therefore, we believe the heterogeneity model is indeed a model worthy of consideration in the context of a sensitivity analysis, even though it may not be a fully robust alternative.

5 A Bayesian Approach

As stated in the previous section, the heterogeneity model, although a valuable tool, does not offer a comfortable degree of robustness against misspecification of the random-effects distribution. Actually, none of the alternatives proposed till now in the literature seem to be fully robust against this misspecification. In this work, we propose to approach the problem within a sensitivity analysis framework. The Bayesian paradigm provides a natural framework for this type of analysis. The Markov Chain Monte Carlo (MCMC) algorithm allows considerable flexibility and has made Bayesian models a popular and efficient tool in the analysis of hierarchical data (Gilks, Richardson and Spiegelhalter, 1996, and Gelman, Carlin, Stern and Rubin, 1995). In this framework, one combines *prior* information on the model parameters, independent from the evidence given by the data, with the support for different values of the parameter effects based on the available data, i.e., the *likelihood*. Together, these two sources of information lead to a *posterior* distribution for the parameters of interest. Samples drawn from the joint posterior distribution then allow to estimate characteristics of the joint and marginal posterior distribution like, e.g., posterior modes or means of the parameters of interest.

For a detailed exploration and application of Bayesian model formulation, model parametrization, choice of prior distribution, diagnosing convergence, comparison between models and model adequacy for binary longitudinal data, we refer to Albert and Jais (1998). In the next subsection we will continue the analysis of the case study by expanding the sensitivity analysis to include a wide range of non-normal random-effects distributions.

5.1 Case Study: The Bayesian Approach

In this section we will revisit the case study introduced in Section 2. We will use Bayesian methods as implemented in WinBUGS to fit several models, considering different random-effects distributions. We will include a mean zero normal with precision τ (note that $\tau = 1/\sigma^2$), an

exponential with parameter λ, a chi-square with k degrees of freedom, a lognormal with scale parameter 0 and precision parameter τ, a uniform with support between $-b$ and b, and a discrete distribution with unequal probability π at two support points X_i, $i = 1, 2$. Furthermore, we will choose vague priors for the parameters of the random-effects distributions. For example, a gamma distribution was chosen for τ, k, λ and b; the X_i are assumed to be sampled from a mean zero normal distribution; and the prior for the probability π is a uniform distribution with support between 0 and 1. For each setting, the Gibbs sampler had 3 chains with 10000 iterations as the burn-in period, plus 100000 additional iterations with a thinning interval of 10 for the normal, exponential and lognormal, 30 for the uniform, 60 for the chi-square and 75 for the discrete distributions.

Note that the considered random-effects distributions do not necessarily have mean zero. It has been suggested that this parametrization can improve convergency and stability of the samples by reducing the autocorrelations of the Gibbs chains (Albert and Jais, 1998). Also note that keeping the intercept in the model (in contrast to the approach in Section 3) improved the convergence of the models. To be able to compare the new estimates with the previously obtained results, we will use the following model

$$\text{logit}\{P(Y_{ij} = 1|b_i)\} = \alpha + \beta_1 Z_i + \beta_2 t_j + b_i, \tag{4}$$

such that $\beta_0 = \alpha + E(b)$.

Convergency of the models was studied through the trace plots of the sample values of the main parameters. As the number of iterations increases the trace plots should stabilize, varying randomly around a mean value. Additionally we have studied the Gelman-Rubin diagnostic tool (Gelman, Carlin, Stern and Rubin, 1996), which uses several parallel chains with widely dispersed starting values with respect to the true posterior distribution, to check convergence. This diagnostic tool compares the variability between- and within-chains by estimating a scale reduction factor. If the variance between the different chains is not larger than the variance within each

Table 4: *Parameter estimates (standard error and MC error between parenthesis) for the model given by (4) and the corresponding deviance information criterion (DIC) for the different random-effects distributions using MCMC in Winbugs.*

Distribution	β_0 (s.e., MCE)	β_1 (s.e., MCE)	β_2 (s.e., MCE)	σ_b^2 (s.e., MCE)	DIC
Normal	-7.91 (1.33, 0.02)	2.32 (1.21, 0.02)	0.68 (0.10, 0.001)	26.28 (9.58, 0.13)	275.72
Exponential	-6.45 (1.16, 0.02)	2.10 (1.20, 0.02)	0.66 (0.10, 0.001)	19.09 (7.72, 0.14)	275.27
Chi-square	-12.80 (3.08, 0.19)	2.39 (1.25, 0.04)	0.68 (0.10, 0.003)	10.86 (4.16, 0.26)	276.60
Uniform	-7.83 (1.43, 0.08)	1.51 (0.99, 0.01)	0.67 (0.10, 0.003)	27.04 (9.97, 0.61)	276.20
Lognormal	-4.31 (1.16, 0.01)	1.38 (0.82, 0.01)	0.56 (0.09, 0.008)	247.4 (1786, 11.2)	290.38
Discrete	-4.85 (0.56, 0.01)	1.25 (0.51, 0.01)	0.53 (0.08, 0.001)	60.45 (67.9, 4.24)	344.14

individual chain, then approximate convergence can be diagnosed.

The parameter estimates of fitting Model (4) are shown in Table 4. Noticeably, a lot of variability can be observed in the estimation of β_0 and σ_b^2. As one would intuitively expect, the variance components estimates seem to be very sensitive to the choice of the random-effects distribution. However, the estimates for the treatment and time effects are similar in all considered settings, and they are also similar to the results from the homogeneity and the heterogeneity model. Note that Table 4 also contains the Monte Carlo error, which can be used to assess the accuracy of the posterior estimates. This error decreases as the sample size used for posterior inference increases.

A useful tool to select the model that fits our data best, is given by the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin and van der Linde, 2002). It is similar to the Akaike Information Criterium (AIC), i.e., a compromise between the deviance and the number of parameters in the model, where smaller values are better. The DIC values of the models with the different random-effects assumptions are shown in the last column of Table 4. Based on the DIC, the models that assume a normal and exponential distribution for the random intercept seem to perform best and produce very similar estimates for the treatment effect. Therefore, we can still be very confident about the results obtained from the homogeneity model for the treatment

effect.

6 Discussion

In contrast to the conventional wisdom amongst data analysts, recent research is showing that the choice of the random-effects distribution can be crucial to the quality of inference about regression coefficients. Indeed, unlike for the linear mixed model, misspecifying the random-effects distribution in GLMM can lead to inconsistent estimators for both the mean and the covariance structure. At the present time fully robust alternatives are not available for the analyst and therefore we strongly suggest exploring the impact of such misspecifications using a sensitivity analysis. Along with these ideas, we have focused in the current work on two possible approaches to do a sensitivity analysis. 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.

Nevertheless, we should also add that in a concrete practical application, some choices not included in our simulations could be considered. Indeed, we could try to fit more than two components, components with different covariance structure, etc. The results obtained in the limited settings that were affordable in our simulation study, illustrate that there are potentials in this model that can be explored even further if some of the previous choices are also considered.

Alternatively, we proposed to incorporate the heterogeneity model into a more general sensitivity analysis, considering different random-effects distributions and comparing the obtained estimates and inference results. The Bayesian paradigm offers a nice framework to carry out such an analysis. Indeed, Bayesian models have become easy to apply in practice with the implementation of the MCMC algorithm in freely accessible software like WinBUGS. With its flexibility in the choice of

the random-effects distribution and the implementation of the DIC to choose the most appropriate model, this approach offers a natural way of implementing a sensitivity analysis, as illustrated in Section 5.

Note that our sensitivity analysis is not a robust alternative to the classical GLMM, but 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 best 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 results 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 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 (Lee and Nelder, 1996). 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.

Acknowledgment

The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Agresti, A. (2002) *Categorical Data Analysis*, (2nd Edn). Hoboken, N.J.: John Wiley & Sons.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004) Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* **47**, 639–653.

Albert, I. and Jais, J-P. (1998) Gibbs sampler for the logistic model in the analysis of longitudinal binary data *Statistics in Medicine*, **17**, 2905–2921.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M.G., Vangeneugden, T. (2004) Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics*, **60**, 845–853.

Böhning, D. (1999) *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others*. Monographs on Statistics and Applied Probability **81**, London: Chapman & Hall/CRC.

Butler, S.M. and Louis, T.A. (1992) Random effects models with non-parametric priors. *Statistics in Medicine*, **11**, 1981–2000.

Chen, J., Zhang, D., and Davidian, M. (2002) A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, **3**, 347–360.

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data*. New York: Oxford University Press.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models* Heidelberg: Springer.

Fieuws, S., Spiessens, B., and Draney, K. (2004) Mixture Models. In: P. De Boeck and M. Wilson (Eds.) *Explanatory Item Response models: A Generalized Linear and Nonlinear Approach*, pp. 317–340. New York: Springer.

Gelman, A. , Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis*. London: Chapman & Hall/CRC.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall/CRC.

Heagerty, P.J. and Kurland, B.F. (2001) Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, **88**, 973–985.

Kizilkaya, K., Carnier, P., Albera, A., Bittante, G., and Tempelman, R.J. (2003) Cumulative t-link threshold models for the genetic analysis of calving ease scores. *Genetics Selection Evolution*, **35**, 489–512.

Kizilkaya, K. and Tempelman R.J. (2005) A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genetics Selection Evolution*, textbf37, 31–56.

Kleinman, K., Lazarus, R., and Platt, R. (2004) A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology*, **159**, 217–224.

Laird, N.M. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–811.

Lee, Y. and Nelder, J.A. (1996) Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Litière, S., Alonso, A. and Molenberghs, G. (2007a) Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, **00**, 000–000. doi:10.1111/j.1541-0420.2007.00782.x.

Litière, S., Alonso, and A., Molenberghs, G. (2007b) The impact of a misspecified random-effects distribution on maximum likelihood estimation in generalized linear mixed models. *Submitted for publication*.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.

Natarajan, R. and Kass, R.E. (2000) Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, **95**, 227–237.

Neuhaus, J.M., Hauck, W.W., and Kalbfleisch, J.D. (1992) The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, **79**, 755–762.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Tempelman, R.J. (1998) Generalized linear mixed models in dairy cattle breeding. *Journal of Dairy Science*, **81**, 1428–1444.

Verbeke, G. and Lesaffre, E. (1997) The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, **53**, 541–556.

Zeger, S.L. and Karim M.R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.