A goodness-of-fit test for the random-effects distribution in mixed models

Achmad Efendi[1] **Reza Drikvandi**[1] **Geert Verbeke**[1,2] **Geert Molenberghs**[2,1]

Interuniversity Institute for Biostatistics and statistical Bioinformatics

[1] *Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[2] *Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

Abstract

In this paper, we develop a simple diagnostic test for the random-effects distribution in mixed models. The test is based on the gradient function, a graphical tool proposed by Verbeke and Molenberghs[1] to check the impact of assumptions about the random-effects distribution in mixed models on inferences. Inference is conducted through the bootstrap. The proposed test is easy to implement and applicable in a general class of mixed models. The operating characteristics of the test are evaluated in a simulation study, and the method is further illustrated using two real data analyses.

Some Keywords: Bootstrap, Goodness-of-fit, Gradient function, Mixed models, Random effects.

1 Introduction

Repeated measures data are common in many areas of research, including medicine, economics, and social sciences. A common modeling approach used for the analysis of such data is mixed models. The approach is flexible and easy-to-use software implementations are widely available. Reviews of mixed models can be found in the book by Verbeke and Molenberghs[2] for linear mixed models and the book by Molenberghs and Verbeke[3] for generalized (non-)linear mixed models. An important aspect of mixed models is the assumption that part of the variability observed in the data can be modeled using so-called random effects, unit-specific parameters that are sampled from some pre-specified distribution, known as random-effects distribution or mixing distribution. For likelihood inferences, the marginal distribution of the response is obtained by integrating out the conditional density over the random effects.

It is common to assume the random effects to follow a normal distribution. Various authors have studied the impact of this assumption for marginal inferences. Neuhaus[4] examined the performance of the mixed-effects logistic models with misspecified mixing distribution and reported that the magnitude of the asymptotic bias in the estimated regression coefficients is typically small. Verbeke and Lesaffre[5], showed that, for linear mixed models, misspecification of the mixing distribution does not affect the consistency of the maximum likelihood estimators. Recently, McCulloch and Neuhaus[6] reported a large degree of robustness of maximum likelihood methods for fitting a generalized linear mixed model when misspecifying the distribution of the random effects. On the other hand, Heagerty and Zeger[7] observed that regression parameters in random-effects models have bias, which is more

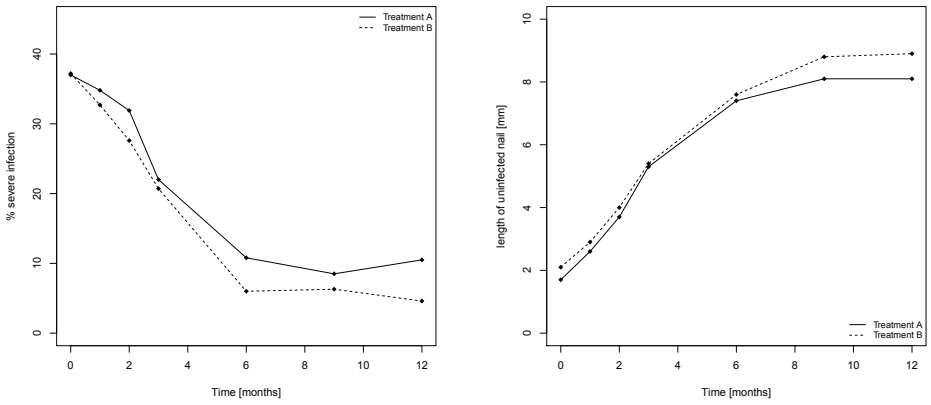
sensitive to the random-effects assumption than their counterpart in the corresponding marginal models. Heagerty and Kurland[8] showed that substantial asymptotic relative bias occurs from incorrect assumptions about the random-effects distribution, using a random-intercept model and when assuming normal whereas the true distribution is gamma, for the random effects. Various examples in which misspecification of the random-effects distribution reduces efficiency were noted by Agresti, Caffo, and Ohman-Strickland[9]. Moreover, Litière, Alonso, and Molenberghs[10] showed, for generalized linear mixed models, that the maximum likelihood estimators are inconsistent when the random-effects distribution is misspecified and the problem is more severe as the number of random effects in the model increases.

Checking distributional assumptions about the random effects is far from straightforward, and several proposals have been made in the statistical literature. Agresti, Caffo, and Ohman-Strickland[9] suggested comparing results from parametric and non-parametric approaches. Substantial differences suggest results from the parametric model should be interpreted with extreme caution. Alternatively, several efforts have been made in relaxing the parametric assumption about random-effects distribution. Tsonaka, Verbeke, and Lesaffre[11] used semi-parametric maximum likelihood estimation for the distribution of random shared parameters in dropout models. Subsequently, Ghidey, Lesaffre, and Verbeke[12] reviewed four methods of smoothly estimating the random-effects distribution in linear mixed models.

As also reviewed in Verbeke and Molenberghs[1], tests for misspecification in mixed models have been available so far. Ritz[13] developed goodness-of-fit tests based on comparison between distributions of the predicted random effects, the standardized estimated best linear unbiased predictors (EBLUPs), and of its expected values. Similarly, Pan and Lin[14] developed methods by comparing the residuals and the predicted values of the response variable under the assumed model. Another diagnostic test was developed by Tchetgen and Coull[15] by comparing the marginal maximum likelihood and conditional maximum likelihood estimators of a subset of the fixed effects in the model. Huang[16] proposed a diagnostic method by comparing inferences based on the original and on derived outcomes. Additionally, Alonso, Litière, and Molenberghs[17,18] developed diagnostic tools by comparing model-based and robust inferences. Apart from some advantages of the aforementioned methods, there are limitations, for example, they are restricted to specific forms of mixed models, such as generalized linear mixed models for binary data and linear mixed models for continuous data. Besides, they require considerable implementation efforts (e.g., Monte Carlo simulation), or test overall goodness-of-fit rather than focusing on misspecification of the random-effects distribution.

Verbeke and Molenberghs[1] recently proposed to use the gradient function as a simple exploratory graphical tool to check goodness-of-fit of the random-effects distribution in mixed models. Their technique does not require any calculations in addition to the computations needed to fit the model, and can be applied in various families of mixed models, including linear and generalized linear mixed models. And in case of any evidence for misspecification, their method indicates how the parametric model can be improved to better describe the observed data. An additional advantage of their method is that it indicates how a parametric model can be improved in case of misspecification. On the other hand, the tool is informal, and should not be interpreted as a formal testing procedure for the random-effects distributional assumptions in mixed models. In this paper, the gradient function will serve as basis for the construction of a formal test.

In Section 2, we present two motivating case studies where a goodness-of-fit test for the random effects would be extremely helpful in formulating an appropriate mixed model. A brief overview of mixed models is given in Section 3. Section 4 describes the gradient function and how Verbeke and



(a) Binary outcomes (b) Continuous outcomes

Figure 1: *Toenail Data. Evolution of the percentage of severe toenail infections and the average unaffected naillength for both treatment groups separately.*

Molenberghs[1] advocates to use it as a diagnostic tool. In Section 5, a formal testing procedure will be developed based on the gradient function. In Section 6, the proposed test will be evaluated and illustrated using simulations. We analyze the two real data examples using our method in Section 7 and compare our results with the diagnostic tests proposed by Alonso, Litière, and Molenberghs[17,18]. Finally, a general discussion will be presented in Section 8.

2 Case Studies

2.1 Toenail Dermatophyte Onychomyosis

This data set results from a randomized, doubled-blind, parallel group, multicenter study for the comparison of two oral treatments (coded as A and B) for toenail dermatophyte onychomycosis (TDO). TDO is a common toenail infection, difficult to treat, affecting more than 2% of the population[19]. The aim of the present study was to compare the efficacy and safety of 12 weeks of continuous therapy with one of two treatments (A and B). In total, 2 x 189 patients were randomized, distributed over 36 centers. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail that is followed over time. We will restrict our analysis to only those patients for which the target nail was one of the two big nails. This reduces our sample under consideration to 146 and 148 subjects, in group A and group B, respectively. The outcomes considered here are the binary infection severity (0: not severe, 1: severe), and the continuous unaffected naillength (expressed in mm). Interest is in studying the evolution over time and differences in evolution between both treatments. More details about the study can be found in the work by De Backer et al.[20], and the extensive analyses using linear and generalized linear mixed models have been reported by Verbeke and Molenberghs[2] and Molenberghs and Verbeke[3]. A graphical representation of the data considered here is given in Figure 1.

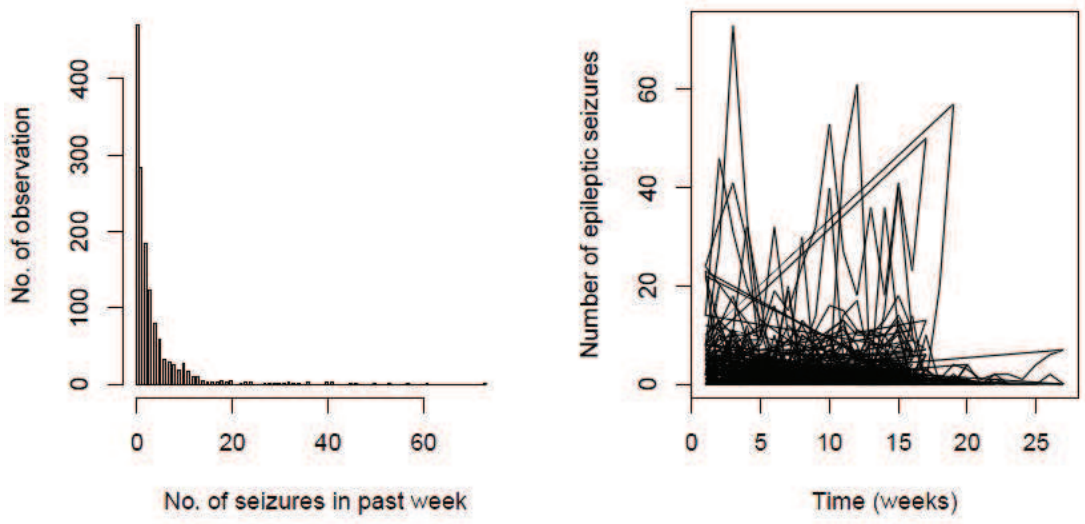


Figure 2: *Epilepsy Data. Frequency plot and individual profiles.*

2.2 Epileptic Seizures

The epileptic seizure data are obtained from a randomized, double-blind, parallel group multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in the work by Faught et al.[21]. The randomization of epilepsy patents took place after a 12-week baseline period that served as a stabilization period for the use of AED's, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were measured weekly, and followed (double-blind) during 16 weeks, after which they were entered into a long-term open-extension study. The outcome of interest is the number of epileptic seizures experienced during the last week, i.e., since the last time the outcome was measured. Of interest is to compared the evolution over time between the two treatment groups. A frequency plot as well as the individual profiles are shown in Figure 2.

3 The general mixed model

Let Y_{ij} be the jth measurement for subject i, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, and let $\boldsymbol{Y_i}$ represent the vector of n_i repeated measurements for subject i. Throughout this paper, the elements in $\boldsymbol{Y_i}$ can be of any type (continuous, binary, count, etc.). When repeated measures are analyzed using mixed models, it is assumed that the association between the observations Y_{ij} of subject i is modeled by a q-dimensional vector $\boldsymbol{b_i}$ of random effects, shared by all measurements of the subject. Let $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$ denote the density function of $\boldsymbol{y_i}$, conditional on $\boldsymbol{b_i}$, possibly depending on a vector of unknown parameters $\boldsymbol{\theta}$. Likelihood-based inference for $\boldsymbol{\theta}$ is usually based on the marginal distribution

$$f_i(\boldsymbol{y_i}|G) \quad = \quad \int f_i(\boldsymbol{y_i}|\boldsymbol{b}) dG(\boldsymbol{b}) \tag{1}$$

of $\boldsymbol{Y_i}$, obtained from integrating out the random effects $\boldsymbol{b_i}$ over a pre-specified distribution G, often called the mixing distribution. Assuming subjects to be independent of each other, the corresponding

log-likelihood function equals

$$\ell(G) \quad = \quad \sum_{i=1}^{N} \ln[f_i(\boldsymbol{y_i}|G)]. \tag{2}$$

The mixing distribution G is often assumed to belong to a specific parametric family, characterized by a vector ψ of unknown parameters, and likelihood-based inference for $\boldsymbol{\theta}$ and ψ jointly follows from (2). Linear and generalized linear mixed models with normal mixing distribution are discussed in full detail by Verbeke and Molenberghs[2] and Molenberghs and Verbeke[3], respectively. It immediately follows from (2) that the choice of G potentially affects inference for the parameters of interest. Verbeke and Molenberghs[1] have proposed the gradient function to graphically check whether the log-likelihood can be increased substantially by replacing the assumed mixing distribution by another one, indicating that the model has been misspecified.

4 The gradient function

Without loss of generality, it will be assumed from now on that the mixing distribution is continuous. Also, in order to simplify notation, we will assume $q = 1$. Let $\widehat{G}$ denote the fitted mixing distribution obtained from maximizing (2). Note that, if G is assumed to belong to some parametric family, estimation of G is equivalent to estimating the unknown parameter vector ψ which characterizes G. Verbeke and Molenberghs[1] suggested the gradient function as

$$\Delta(\widehat{G}, \boldsymbol{b}) = \frac{1}{N} \sum_{i=1}^{N} \frac{f_i(\boldsymbol{y_i}|\boldsymbol{b})}{f_i(\boldsymbol{y_i}|\widehat{G})},$$

and showed that in case the likelihood cannot be maximized further by replacing the fitted random-effects distribution by any other mixing distribution H, the gradient as a function of $\boldsymbol{b}$ does not exceed 1 and reaches 1 in all support points of the fitted random-effects distribution $\widehat{G}$. Under normality for the random effects, this implies that the gradient function equals one on the entire real line. Therefore, severe deviations from one can be used as evidence against the assumed mixing distribution. Moreover, it can be shown that $\Delta(\widehat{G}, \boldsymbol{b})$ does not need to be studied over the entire real line[1], but that attention can be restricted to any closed interval I that contains all values $\boldsymbol{b}$ for which $f_i(y_i|\boldsymbol{b})$ is maximized, $i = 1, \ldots, N$. Hence, once a mixed model has been fitted, goodness-of-fit of the random-effects distribution can easily be assessed by quantifying the deviation of the implied gradient function $\Delta(\widehat{G}, \boldsymbol{b})$ from one. In the next section, this will serve as basis for the construction of a formal testing procedure.

5 The testing procedure

As explained above, severe deviations of $\Delta(\widehat{G}, \boldsymbol{b})$ from one, within the interval I provide evidence against the assumed mixing distribution. We therefore propose the following formal testing procedure. Let $\{b_k, k = 1, \ldots, K\}$ be a sufficiently fine grid in I, and define the test statistic

$$T \quad = \quad \frac{1}{K} \sum_{k=1}^{K} \left| \widehat{\Delta}(\widehat{G}, b_k) - 1 \right|. \tag{3}$$

Table 1: *Parameterization and selected parameter values for the models used in the simulation study*

Model	Distribution	Parameterization	Parameter values
Linear	$Y_{ij}\|b_i \sim \text{Normal}(\mu_{ij}, \sigma_e^2)$	$\mu_{ij} = \beta_0 + \beta_1 t_{ij} + b_i$	$\beta_0 = 0$, $\beta_1 = 0.05$, $\sigma_e = 1$
Logistic	$Y_{ij}\|b_i \sim \text{Bernoulli}(\pi_{ij})$	$\ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + b_i$	$\beta_0 = 0$, $\beta_1 = 0.05$
Count	$Y_{ij}\|b_i \sim \text{Poisson}(\lambda_{ij})$	$\ln(\lambda_{ij}) = \beta_0 + \beta_1 t_{ij} + b_i$	$\beta_0 = 0$, $\beta_1 = 0.05$

Note that our notation $\widehat{\Delta}$ explicitly acknowledges the fact that the unknown parameters $\boldsymbol{\theta}$ in $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$ have been replaced by their estimators $\widehat{\boldsymbol{\theta}}$. Obviously, T quantifies the deviation of $\Delta(\widehat{G}, \boldsymbol{b})$ from one, within the interval I. The null-distribution of T, needed to formally test whether the assumed mixing distribution G is appropriate, can be obtained using parametric bootstrap. The following steps are then required in order to perform the bootstrap test:

1. Based on the observed data, fit the mixed model under consideration, with a particular assumption for the mixing distribution G, i.e., maximize $\ell(G)$ with respect to the vector $\boldsymbol{\omega}' = (\boldsymbol{\theta}', \boldsymbol{\psi}')$ of unknown parameters which completely characterizes the marginal density $f_i(\boldsymbol{y_i}|G)$.

2. Construct the gradient function and compute the resulting observed value T_a for the test statistic T.

3. For $s = 1, \ldots, S$, repeat the following steps:

   (a) Sample a new vector $\boldsymbol{\omega}^s$ of parameter values from a multivariate normal distribution with mean $\widehat{\boldsymbol{\omega}}$ and covariance matrix equal to the inverse Fisher information matrix for the fitted model.

   (b) Sample random effects $\boldsymbol{b_i}^s$, $i = 1, \ldots, N$, from G in which $\boldsymbol{\psi}$ has been replaced by $\boldsymbol{\psi}^s$.

   (c) Sample new observations $\boldsymbol{Y_i}^s$, $i = 1, \ldots, N$, from $f_i(\boldsymbol{y_i}|\boldsymbol{b_i}^s)$ in which $\boldsymbol{\theta}$ has been replaced by $\boldsymbol{\theta}^s$. Note that the data set should have the same structure as the original data set (covariates, number of measurements, etc.)

   (d) Fit the mixed model under consideration based on the sampled data $\boldsymbol{Y_i}^s$, $i = 1, \ldots, N$.

   (e) Construct the gradient function and compute the resulting observed value T^s for the test statistic T

4. Calculate the p-value as the proportion of values T^s exceeding T_a.

Note that, in our bootstrap procedure, the interval I changes with each bootstrap sample because the construction of interval I depends on the observations. In fact, the interval is determined from knowing the minimum and maximum of the unique modes of all $f_i(y_i|b)$ as functions of b. The unique modes are calculated through maximizing each $f_i(y_i|b)$ (model fitting by subject/cluster) with parameter estimates from maximizing $f(\mathbf{y}|b)$ set as offsets except the one related to b. Note also that, in case of binary data, subjects with all observations equal to zero or to one lead to modes equal to minus or plus infinity, respectively. In order to be able to study the gradient function on a closed finite interval, those subjects are excluded from the calculation of the interval I, as suggested by Verbeke and Molenberghs[1].

Table 2: *Random-intercepts distributions used in the simulation study.*

Model	Distribution
Normal	$b_i \sim N(0, 2^2)$
Symmetric mixture of normals	$b_i \sim \frac{1}{2}N(-1.9, 0.6245^2) + \frac{1}{2}N(1.9, 0.6245^2)$
Asymmetric mixture of normals	$b_i \sim \frac{3}{10}N(-3, 0.3779^2) + \frac{7}{10}N(\frac{9}{7}, 0.3779^2)$
Shifted log-normal distribution	$b_i \sim 2[\exp N(3, 1) - 33.1154]/43.40881$

Table 3: *Simulated Type I error rates to test normality of random intercepts in three different mixed models and for three sample sizes.*

Model	$N = 50$	$N = 100$	$N = 200$
Linear	0.041	0.042	0.045
Logistic	0.011	0.032	0.060
Count	0.036	0.048	0.047

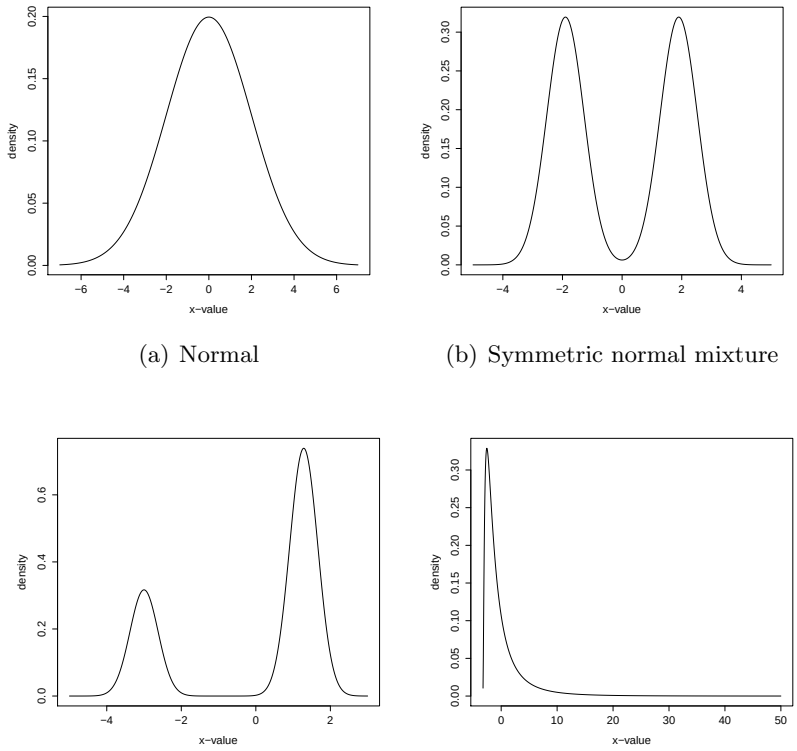
6 Simulation

We conducted a small-scale simulation study to evaluate the operating characteristics of the proposed bootstrap test. The models considered are the linear mixed model for continuous data, the logistic mixed model for binary data, and the Poisson mixed model for count data. All models were random-intercepts models. The formal parameterization of the various models, as well as the parameter values used in the simulations, are shown in Table 1. The four distributions considered for the random intercepts b_i are presented in Table 2, and the densities are shown in Figure 3, and they have been selected such that they all have mean 0 and variance 4.

For each combination of model and random-effects distribution, 500 data sets were simulated for $N = 50$, $N = 100$, and $N = 200$ clusters, respectively, with 10 repeated measurements per cluster. Each time the gradient test for normality of the random effects was performed, as discussed in Section 5. The number of bootstrap runs B was set equal to 200, and the test statistic (3) was based on K grid points that is obtained from the range I divided by a small value h, i.e. $h = 0.1$.

The scenarios where the true random-effects distribution is normal is used to evaluate the Type I error rate, estimated as the proportion of times, out of the 500 simulated data sets, that the test leads to a rejection of the normality assumption at the 5% level of significance. The results for all three mixed models considered, and for the three sample sizes, are summarized in Table 3. The simulated Type I error rates are relatively close and get closer to 5% as the sample sizes increase. The same phenomenon is observed for all types of outcomes.

The scenarios where the true random-effects distribution is not normal is used to evaluate the power of the test, estimated as the proportion of times, out of the 500 simulated data sets, that the test leads to a rejection of the normality assumption at the 5% level of significance. The results for all the models considered, for all three alternatives, and for the three sample sizes, are summarized in Table 4. The simulated power of the proposed test is reasonably large and increases with the sample size, as expected. Moreover, the power to detect skewness (asymmetric mixture and log-normal) is



(a) Normal (b) Symmetric normal mixture

(c) Asymmetric normal mixture (d) Log-normal

Figure 3: *True random-intercepts distributions, used in the simulation study.*

higher than to detect multi-modality (symmetric mixture).

7 Applications

In this section, we apply our methodology to test normality of random effects in mixed models for the analysis of the two real data sets introduced in Section 2.

7.1 Toenail Dermatophyte Onychomyosis

We first analyze the binary outcome, i.e., infection severity. Let Y_{ij} be the binary outcome indicating the severity of the toenail infection for patient i at measurement j. The model used by Verbeke and Molenberghs[1] is given by

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Bernoulli}(\pi_{ij}), \\ \text{logit}(\pi_{ij}) &= \beta_0 + b_i + \beta_1 \text{treat}_i + \beta_2 t_{ij} + \beta_3 \text{treat}_i t_{ij}, \end{aligned} \tag{4}$$

where treat_i is the treatment indicator for patient i, t_{ij} is the time-point (in months) at which the jth measurement is taken for the ith patient, and b_i is a random subject-specific intercept. Verbeke and Molenberghs[1] provided evidence that the random-effects distribution is multi-modal and hence not normal. We will check if this is confirmed by the testing procedure developed here.

Table 4: *Simulated power values to detect deviations from normality for random intercepts in three different mixed models, for three different alternative distributions, and for three sample sizes.*

Model	Random-intercepts distribution	$N = 50$	$N = 100$	$N = 200$
Linear	Symmetric Mixture	0.021	0.096	0217
	Asymmetric Mixture	0.494	0.546	0.669
	Log-normal	0.824	0.996	1.000
Logistic	Symmetric Mixture	0.083	0.484	0.801
	Asymmetric Mixture	0.765	0.992	1.000
	Log-normal	0.365	0.784	0.993
Count	Symmetric Mixture	0.024	0.061	0.213
	Asymmetric Mixture	0.736	0.964	0.990
	Log-normal	0.768	0.995	1.000

Maximum likelihood estimates and associate standard errors, assuming normality for the random effects, are presented in Table 5, and the implied gradient function for this model is shown in panel (a) of Figure 4. The gradient suggests non-normality of the random-intercepts distribution, which has been confirmed by our testing procedure. The test-statistic, based on $K = 69$ grid points equals $T_a = 0.1962$, which is significant with $p = 0.001$, based on $B = 200$ bootstrap samples. For comparison, other diagnostic tests proposed by Alonso, Litière, and Molenberghs[17,18] provide similar results, the determinant test with a test statistic of 4425.83 produces $p < 0.001$, and the determinant-trace test with a test statistic of 4100.39 gives $p < 0.001$. We need to point out here that our proposed test has additional advantage of providing the nature of misspecification through the gradient function plot to improve the model fit thence to better describe the observed data.

Verbeke and Molenberghs[1] suggested that a mixture of normals for the random effects might lead to a much better model in terms of log-likelihood. Mixtures not only can handle skewness, they can also account for multi-modality in the random-effects distribution[2,9,22]. Verbeke and Molenberghs[1] suggested to replace the normality assumption for the random effects by

$$b_i \sim \pi_1 N(\mu_1, \sigma_b^2) + \pi_2 N(\mu_2, \sigma_b^2) + \pi_3 N(\mu_3, \sigma_b^2),$$

with $\pi_1 + \pi_2 + \pi_3 = 1$, and with the additional restriction of $\pi_1\mu_1 + \pi_2\mu_2 + \pi_3\mu_3 = 0$ in order for the random effects to have mean zero. Parameter estimates and associated standard errors under this extended model are also included in Table 5. The corresponding gradient function is shown in panel (b) of Figure 4. It no longer provides evidence for misspecification of the random-intercepts distribution, and this is confirmed by our testing procedure ($T_a = 0.0975$, $p = 0.345$, based on $K = 66$ grid points and $B = 200$ bootstrap samples).

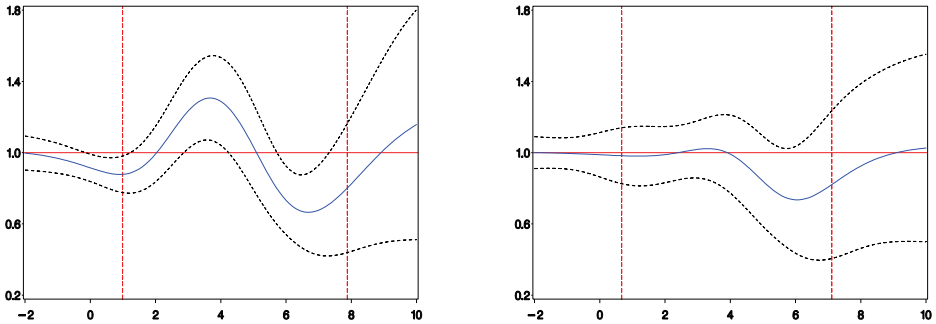
For the continuous outcome, let Y_{ij} be the unaffected naillength (in mm), for patient i at measurement j. The linear mixed model considered is

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Normal}(\mu_{ij}, \sigma_e^2) \\ \mu_{ij} &= \beta_0 + b_i + \beta_1 \text{treat}_i + \beta_2 t_{ij} + \beta_3 \text{treat}_i t_{ij}, \end{aligned} \tag{5}$$

with treat_i and t_{ij} as before. Table 5 shows parameter estimates and associated standard errors assuming $b_i \sim N(0, \sigma^2)$. The implied gradient function is shown in Figure 5 and does not reveal

Table 5: *Toenail Data: Estimates and associated standard errors for the generalized linear and linear mixed models fitted to the binary and continuous outcome, respectively.*

		Generalized linear		Linear
		Normal b_i	Mixture b_i	Normal b_i
Effect	Parameter	Estimate(s.e.)	Estimate(s.e.)	Estimate(s.e.)
Intercept	β_0	-1.6306(0.4345)	-1.5160(0.4854)	2.5165(0.2465)
Treat	β_1	-0.1146(0.5852)	0.4479(0.4306)	0.2488(0.3463)
Time	β_2	-0.4041(0.0459)	-0.3992(0.0466)	0.5608(0.0226)
Treat×Time	β_3	-0.1613(0.0718)	-0.1562(0.0758)	0.0474(0.0314)
s.d. b_i	σ_b	4.0133(0.3763)	0.8561(0.1889)	2.5467(0.1233)
s.d. error	σ_e			2.6343(0.0471)
Prob-1	π_1		0.5770(0.0422)	
Prob-2	π_2		0.3779(0.0426)	
Prob-3	π_3		0.0451(0.0129)	
Mean-1	μ_1		-2.5617(0.4831)	
Mean-2	μ_2		2.7744(0.3146)	
Mean-3	μ_3		9.5282(1.2788)	
-2 log-likelihood		1247.8	1219.5	9414.8



(a) Normal RE (b) Mixture Normals RE

Figure 4: *Toenail data: Gradient function and 95% pointwise confidence bands for the generalized linear mixed model (4) for two different random-intercepts distributions. The region I is indicated by two vertical lines.*

any evidence against normality for the random effects. Our formal test confirms this ($T_a = 0.3387$, $p = 0.155$, based on $K = 148$ grid points and $B = 200$ bootstrap samples). But, both the determinant test and the determinant-trace test of Alonso, Litière, and Molenberghs[17,18] reject the normality assumption of the random effects with the same p-value of $p < 0.001$. As pointed out by these authors, a significant result with their tests will not necessarily imply that there is a problem with the random-effects distribution. We conjecture that, for this model, another type of misspecification

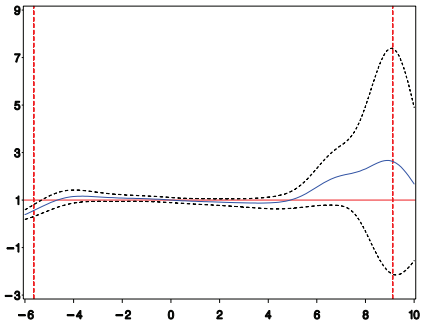


Figure 5: *Toenail data: Gradient function and 95% pointwise confidence bands for linear mixed model (5) model with normal random intercepts. The region I is indicated by two vertical lines.*

Table 6: *Epilepsy data: Estimates and associated standard errors for parameters in Poisson mixed model (6) assuming normal random intercepts.*

Effect	Parameter	Estimate (s.e)
Intercept Placebo	β_{00}	0.8182(0.1677)
Slope Placebo	β_{10}	-0.0143(0.0044)
Intercept Treatment	β_{01}	0.6475(0.1701)
Slope Treatment	β_{11}	-0.0120(0.0043)
Variance b_i	σ_b^2	1.1564(0.1843)
-2 log-likelihood		6271.9

such as a missing covariate or random part has a large effect on their tests to detect misspecification in the random-effects distribution. This phenomenon was explicitly demonstrated in the Discussion of their paper. To investigate this further as well as the sufficiency of the model (5), we added random slopes for the time effect t_{ij}, and after fitting the model with both random intercepts and random slopes we observed that the random slopes are significant. Therefore, the random structure of model (5) may not be appropriate when considering only random intercepts.

7.2 Epileptic Seizures

Let Y_{ij} be the number of epileptic seizures patient i experienced during week j. Furthermore, let t_{ij} denote the time-point at which Y_{ij} has been measured. The following Poisson mixed model is considered:

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Poisson}(\lambda_{ij}) \\ \ln(\lambda_{ij}) &= \begin{cases} \beta_{00} + \beta_{10} t_{ij} + b_i & \text{if placebo,} \\ \beta_{01} + \beta_{11} t_{ij} + b_i & \text{if active treatment.} \end{cases} \end{aligned} \tag{6}$$

Table 6 shows parameter estimates and associated standard errors assuming the random intercepts b_i to be normally distributed with mean zero and variance σ_b^2. The corresponding gradient function presented in Figure 6 does not suggest any misspecification in the random-intercepts distribution.

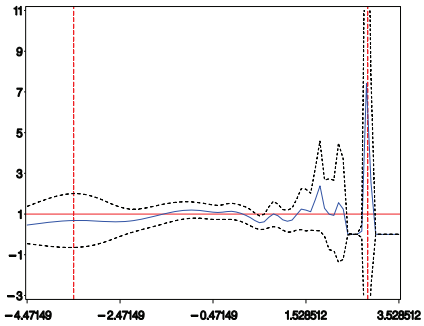


Figure 6: *Epilepsy data: Gradient function and 95% pointwise confidence bands for Poisson mixed model (6) assuming normal random intercepts. The region I is indicated by two vertical lines.*

This has been confirmed by our test ($T_a = 0.3824$, $p = 0.615$, based on $K = 63$ grid points and $B = 200$ bootstrap samples), while both the determinant test and the determinant-trace test of Alonso, Litière, and Molenberghs[17,18] reject the normality assumption of the random effects with the same p-value of $p < 0.001$. As discussed in the previous data analysis, a possible reason for this result might be another type of misspecification such as a missing covariate or random part has a large effect on their tests to detect misspecification in the random-effects distribution. We added random slopes for the time effect t_{ij} in model (6) as also suggested by Molenberghs and Verbeke[3], and after fitting the model with both random intercepts and random slopes we observed that the random slopes are significant, indicating misspecification in the structure of random part of model (6). Moreover, Molenberghs, Verbeke, Demétrio, and Vieira[23] discussed that an overdispersion model should be considered for the epilepsy data.

8 Concluding Remarks

In this paper, a formal test procedure for checking the appropriateness of random-effects assumptions in mixed models has been developed based on the graphical tool proposed by Verbeke and Molenberghs[1]. A bootstrap method is used to assess the null-distribution of the proposed test statistic. A small-scale simulation study with some promising results has been performed to study the operating characteristics of the new test in a number of scenarios. The proposed test has several advantages. First, computations are relatively straightforward once the mixed model under consideration has been fitted. Calculation of the test-statistic only requires evaluation of the gradient function in a dense grid. Second, the test can be used to assess the random-effects distribution in a very wide class of mixed models, including linear mixed models, generalized linear mixed models, and non-linear mixed models. The SAS code used for one of the test implementation in Section 7 is available on the website www.ibiostat.be/software. Third, while most emphasis has been on detecting non-normality of random effects, the procedure can be used to check appropriateness of any mixing distribution, as has been illustrated in Section 7.1. Fourth, while all examples have been in the context of mixed models with a single random effect, the procedure can be generalized to multivariate random effects in a straightforward way.

Finally, we emphasize that in our bootstrap procedure the interval I changes with each bootstrap sample because it is constructed using the observations. In a small simulation study, not reported

here, we found that the size of the test would be highly inflated, if the intervals I were fixed in bootstrap samples.

Acknowledgement

Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

References

[1] Verbeke G, Molenberghs G. The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics* 2013; **14**(3):477-90.

[2] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer; 2000.

[3] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York: Springer; 2005.

[4] Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution specification when fitting mixed-effects logistic models. *Biometrika* 1992; **79**:755–762.

[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; **23**:542–556.

[6] McCulloch CE, Neuhaus JM. Misspecifying the shape of a random-effects distribution: why getting it wrong may not matter. *Statistical Science* 2011; **26**(3):388–402.

[7] Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000; **15**(1):1–26.

[8] Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models, *Biometrika* 2001; **88**(4):973–985.

[9] Agresti A, Caffo B, Ohman-Strickland, P. Examples in which misspecification of random-effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis* 2004; **47**:639–653.

[10] Litière S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed model. *Statistics in Medicine* 2008; **27**:3125–3144.

[11] Tsonaka R, Verbeke G, Lesaffre E. A semiparametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* 2009; **65**:81–87.

[12] Ghidey W, Lesaffre E, Verbeke G. A comparison of methods for estimating the random-effects distribution of a linear mixed model. *Statistical Methods in Medical Research* 2010; **19**:575–600.

[13] Ritz C. Goodness-of-fit tests for mixed models. *Scandinavian Journal of Statistics* 2004; **31**:443–458.

[14] Pan Z, Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 2005; **61**:1000–1009.

[15] Tchetgen EJ, Coull BA. A diagnostic test for the mixing distribution in a generalized linear mixed model. *Biometrika* 2006; **93**:1003–1010.

[16] Huang X. Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics* 2009; **65**:361–368.

[17] Alonso A, Litière S, Molenberghs G. A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. *Computational Statistics and Data Analysis* 2008; **52**:4474–4486.

[18] Alonso A, Litière S, Molenberghs G. Testing for misspecifications in generalized linear mixed models. *Biostatistics* 2010; **11**(4):771–786.

[19] Roberts DT. Prevalence of dermatophyte onychomicosis in the United Kingdom: Results of an omnibus survey. *British Journal of Dermatology* 1992; **126** Suppl 39:23–27.

[20] De Backer M, De Keyser P, De Vroey C, Lesaffre E. A 12-week treatment for dermatophyte toe onychomycosis: terninafine 250mg/day vs. itraconazole 200mg/day - a double-blind comparative trial. *British Journal of Dermatology* 1996; **143**:16–17.

[21] Faught E, Wilder BJ, Ramsay RE, Reife RA, Kramer LD, Pledger GW, et al. Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosage. *Neurology* 1996; **46**:1684–1690.

[22] Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association* 1996; **91**:217–221.

[23] Molenberghs G, Verbeke G, Demétrio C, Vieira A. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* 2010; **5**:325–347.