Made available by Hasselt University Library in https://documentserver.uhasselt.be

Diagnosing Misspecification of the Random-Effects Distribution in Mixed Models

Peer-reviewed author version

Drikvandi, Reza; VERBEKE, Geert & MOLENBERGHS, Geert (2017) Diagnosing Misspecification of the Random-Effects Distribution in Mixed Models. In: BIOMETRICS, 73(1), p. 63-71.

DOI: 10.1111/biom.12551
Handle: http://hdl.handle.net/1942/24136

Diagnosing Misspecification of the Random-Effects Distribution in Mixed Models

Reza Drikvandi[1,2,*], **Geert Verbeke**[1,3], **and Geert Molenberghs**[1,3]

[1] I-BioStat, KU Leuven, Leuven, Belgium

[2] Department of Mathematics, Imperial College London, London, UK

[3] I-BioStat, Universiteit Hasselt, Hasselt, Belgium

[*] email: r.drikvandi@imperial.ac.uk

Abstract

It is traditionally assumed that the random effects in mixed models follow a multivariate normal distribution, making likelihood-based inferences more feasible theoretically and computationally. However, this assumption does not necessarily hold in practice which may lead to biased and unreliable results. We introduce a novel diagnostic test based on the so-called gradient function proposed by Verbeke and Molenberghs (2013) to assess the random-effects distribution. We establish asymptotic properties of our test and show that, under a correctly specified model, the proposed test statistic converges to a weighted sum of independent chi-squared random variables each with one degree of freedom. The weights, which are eigenvalues of a square matrix, can be easily calculated. We also develop a parametric bootstrap algorithm for small samples. Our strategy can be used to check the adequacy of any distribution for random effects in a wide class of mixed models, including linear mixed models, generalized linear mixed models, and non-linear mixed models, with univariate as well as multivariate random effects. Both asymptotic and bootstrap proposals are evaluated via simulations and a real data analysis of a randomized multicenter study on toenail dermatophyte onychomycosis.

Keywords: *Asymptotic distribution; Eigenvalues; Gradient function; Longitudinal data; Parametric bootstrap; Random effects.*

1. Introduction

Longitudinal and clustered studies produce correlated data with a complex structure. Mixed models are frequently used to analyze such data by incorporating random effects into the model to capture the heterogeneity among individuals or groups. Inferences are then usually based on the likelihood function after integrating out the random effects over their assumed distribution. To make likelihood-based inferences more feasible theoretically as well as to enable the use of standard software packages for fitting mixed models, it is common to assume a multivariate normal distribution for the random effects. However, this assumption may be violated in practice, which can result in misleading inferences.

In the literature there seems to be no general consensus about the impact of misspecifying the random-effects distribution in mixed models. The misspecification can affect inferences regarding two parts of the model: fixed and random components. For inferences about the fixed-effects parameters, some authors argued that the impact is minimal and the asymptotic bias in regression parameters is often small (see, e.g., Neuhaus et al., 1992; Chen et al., 2002; McCulloch and Neuhaus, 2011*a*,*b*). In contrast, some other authors have claimed a strong sensitivity to the normality assumption of random effects; see, for example, Heagerty and Kurland (2001), Agresti et al. (2004), Litière et al. (2008), and Alonso et al. (2010). They concluded that substantial bias in the maximum likelihood estimates of regression parameters can result when the random-effects distribution is misspecified. On the other hand, inferences about the

random effects themselves are more likely to be affected by misspecification of the random-effects distribution. For instance, the normality assumption often forces the predictions of random effects to reflect normality, even when the correct random-effects distribution is far from normal (Verbeke and Lesaffre, 1996).

Since random effects are latent, and hence unobservable quantities, it is conceptually difficult to evaluate their distributional assumptions. Some authors instead suggested to build more flexible distributional assumptions for the random effects to protect against misspecification. Examples include nonparametric maximum likelihood (Aitkin, 1999), smooth nonparametric fits (Zhang and Davidian, 2001), and mixtures of normal distributions (Verbeke and Lesaffre, 1996). However, such approaches often rely on complex optimization methods and their use in routine statistical practice is very limited due to the lack of standard software. Moreover, it may not be possible to explicitly explore the characteristics of random effects in their own right, if nonparametric or semiparametric methods are used (Huang, 2009).

Several diagnostic tools have been proposed so far for detecting misspecification of the random-effects distribution. Lange and Ryan (1989) suggested a generalized weighted normal plot to check the normality assumption in linear mixed models. Jiang (2001) presented a goodness-of-fit test for checking distributional assumptions in linear mixed models based on a test statistic similar to Pearson's goodness-of-fit statistic. Ritz (2004) and Pan and Lin (2005) proposed to compare distributions of residuals and/or predicted random effects with their expected distributions under the assumed model. Tchetgen and Coull (2006) suggested to compare the marginal and conditional maximum likelihood estimators of a subset of fixed-effects parameters to verify whether the assumed random-effects distribution is valid. Waagepetersen (2006) constructed a simulation-based test by generating random effects while conditioning on the observa-

tions. Alonso et al. (2008) provided a toolbox of tests to detect misspecification of the random-effects structure by comparing model-based and robust inferences. For linear mixed models, Claeskens and Hart (2009) suggested several tests of the hypothesis that the random effects and/or errors are normally distributed. Huang (2009) introduced a series of parametric diagnostic tools that make use of both the observed data and a reconstructed data set induced by the observed data. Apart from some advantages of the aforementioned methods, they all are restricted to very specific mixed models (e.g., models with a specific type of response and/or link function), they require considerable efforts for implementation, most of them were only developed to check the adequacy of a normal distribution for random effects, and they test overall goodness-of-fit rather than focus on the misspecification of the random-effects distribution.

More recently, Verbeke and Molenberghs (2013) proposed an exploratory diagnostic tool based on the so-called gradient function to graphically check the appropriateness of a specific parametric assumption about the random-effects distribution. Their method is easy to implement; however, a graphical tool based on visual judgment is informal and it is generally difficult to determine whether such a plot reveals misspecification or just random variability. Efendi et al. (2014) developed a simple bootstrap test using the gradient function, but their test was investigated via a simulation study. In this paper, we utilize the theoretical properties of the gradient function to develop a powerful test to assess the random-effects distribution. Using the Cramér-von Mises measure, we construct a test statistic based on the gradient function and we derive asymptotic properties of our proposal and also provide a parametric bootstrap algorithm for small samples. Beneficially, our method can be used for a general class of mixed models, with univariate as well as multivariate random effects.

2. The general mixed model

Let $Y_i = (Y_{i1}, ..., Y_{in_i})'$ denote the vector of n_i repeated measurements for individual or cluster i, $i = 1, ..., N$. Throughout this paper, the elements in Y_i could be continuous, discrete, or a combination thereof. Observations on the same individual are obviously correlated and there may be considerable heterogeneity among individuals. Random effects can be considered to take into account such correlation and variability in the analysis. It is assumed that, conditional on a q-dimensional vector b_i of random effects, the response vector Y_i has a pre-specified density $f_i(y_i|b_i)$ depending on covariates and parametrized through a vector θ of unknown parameters, common to all subjects. Hereafter, $f_i(y_i|b_i)$ will be referred to as the conditional distribution. The random effects b_i are also assumed to be sampled from a population of subject-specific parameters with distribution function G, parametrized by a vector α of unknown parameters. Under these assumptions, the marginal likelihood function of the model is given by

$$L(G) = \prod_{i=1}^{N} f_i(y_i|G) = \prod_{i=1}^{N} \int_{R^q} f_i(y_i|b)\, dG(b), \tag{1}$$

where we suppressed dependence on θ and α in the notation and instead emphasized dependence of the likelihood on the random-effects distribution G. Note that $f_i(y_i|G)$ is the marginal density of Y_i. The general mixed model considered here includes linear, generalized linear, and non-linear mixed models, among other models with random effects.

Let $\psi = (\theta', \alpha')'$ represent all unknown parameters in the model. Likelihood-based inferences about ψ are then based on maximizing marginal likelihood (1) for the observed data, and clearly the random-effects distribution G is crucial in the calculation of the likelihood function. When random effects b_i have a multivariate normal dis-

tribution, Gaussian quadrature (implemented in standard software packages, such as the SAS procedures NLMIXED and GLIMMIX and the SPlus/R function lme) are generally used to maximize marginal likelihood (1). Both adaptive and non-adaptive Gaussian quadrature methods are applicable; see Fitzmaurice et al. (2008) for a complete discussion on mixed models with normal random effects. Computations are more challenging in case of non-normal random effects. Nelson et al. (2006) and Liu and Yu (2008) presented two different transformations to apply Gaussian quadrature for mixed models with non-normal random effects.

3. A diagnostic test based on the gradient function

Herein we assume the conditional distribution $f_i\left(y_i|b_i\right)$ to be correctly specified, an assumption made by Verbeke and Molenberghs (2013) to develop the idea of the gradient function for assessing the distribution of random effects. This assumption is relaxed in Section 9 by a quasi-likelihood approach. Note that, because of the intangible nature of random effects as latent variables, the assumptions on random effects cannot be verified from data alone. For instance, Alonso et al. (2010) have shown that the correct specification of the conditional distribution is a necessary condition for the proper identifiability of the random-effects distribution.

We are interested in testing whether or not the assumed random-effects distribution G is correctly specified. In fact, the null hypothesis H_0 says G is correctly specified. Assuming the conditional distribution is correctly specified, Verbeke and Molenberghs

(2013) derived the so-called gradient function as

$$\Delta\left(G, b\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{f_i\left(y_i|b\right)}{f_i\left(y_i|G\right)}, \qquad b \in R^q.$$

For each point b, the gradient can be interpreted as an average of likelihood ratios, each ratio measuring how much more likely Y_i is to be observed for individual i if the corresponding random effect b_i equals b rather than it being sampled from G. Based on the theoretical results of Lindsay (1983*a*,*b*) and Böhning (1989), Verbeke and Molenberghs (2013) showed that if the random effects distribution G produces an adequate fit to the data in terms of marginal likelihood, then $\Delta\left(G, b\right) \leq 1$ for all $b \in R^q$ and additionally $\Delta\left(G, b\right) = 1$ for all b in the support of G. Alternatively, deviations of the gradient function from 1 in the support points of G indicate that the model can be improved by replacing G by some other random-effects distribution. They suggested to plot the gradient function versus points b in the support of G. Despite the simplicity of this approach, it is not clear how misspecification can generally be distinguished from random variability by such a plot. Furthermore, investigation of the operating characteristics of a diagnostic tool is impossible when merely using it as a graphical tool.

To provide a formal diagnostic tool based on the gradient function, we define $T(\psi) = \int_{R^q}\left(\Delta\left(G, b\right) - 1\right)^2 dG\left(b\right)$, which takes into account the deviation of the gradient from 1 for all possible b in the support of G. As discussed above, it is easy to show that $T(\psi) = 0$ under H_0, and hence the null hypothesis can be rejected for large values of $T(\psi)$. However, ψ is unknown and we must replace it with a suitable estimator. We use the maximum likelihood (ML) estimator $\hat{\psi}$ obtained under H_0, and then we get

$$T(\hat{\psi}) = \int_{R^q}\left(\hat{\Delta}(\hat{G}, b) - 1\right)^2 d\hat{G}\left(b\right), \tag{2}$$

where $\hat{G}$ is the estimated random-effects distribution and $\hat{\Delta}$ denotes the estimated gradient function based on $\hat{G}$ obtained by replacing the unknown parameters θ in $f_i(y_i|b)$ and $f_i(y_i|\hat{G})$ by their ML estimates $\hat{\theta}$.

We propose $T(\hat{\psi})$ as an appropriate test statistic for testing the null hypothesis for several reasons. First, $T(\hat{\psi})$ evaluates the gradient function at all possible b in the support of G. Second, the test statistic $T(\hat{\psi})$ considers a weight for each deviation of the gradient function from 1 for each point b. The weight is the estimated probability mass in point b. Third, large values of $T(\hat{\psi})$ indicate that G is not correctly specified and this leads to the rejection of H_0. We also note that $T(\hat{\psi})$ is constructed based on the Cramér-von Mises measure of distance between the gradient and 1.

To complete our diagnostic tool, we need the null distribution of the proposed test statistic (2). Since finding the exact distribution of $T(\hat{\psi})$ is difficult, we instead derive its asymptotic distribution under H_0. We further develop a parametric bootstrap procedure to approximate the finite-sample distribution of $T(\hat{\psi})$ under the null hypothesis.

4. Asymptotic results

In this section, we investigate the asymptotic properties of $T(\hat{\psi})$ under H_0. Suppose $\psi_0 = (\psi_{01}, ..., \psi_{0L})'$ is the true parameter vector.

Theorem 1. *Under general regularity conditions and provided that the model is correctly specified,* $T(\hat{\psi}) = \sum_{j=1}^{r} \lambda_j \chi_j^2 + o_p(1)$, *where* χ_j^2 $(j = 1, ..., r)$ *are independent* χ_1^2 *random variables, and* $\lambda_1 \geq ... \geq \lambda_r$ *are the eigenvalues of* $A'Q(\psi_0)A$, *in which* A *is the square root of the inverse Fisher information matrix of the model parameters and*

$Q(\psi_0)$ is the $L \times L$ matrix with (l, l')-th element

$$Q_{ll'}(\psi_0) = \int_{R^q} \left\{ \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E\left[\frac{\partial}{\partial \psi_{0l}} \frac{f_i(Y_i|b)}{f_i(Y_i|G)} \right] \right\} \left\{ \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E\left[\frac{\partial}{\partial \psi_{0l'}} \frac{f_i(Y_i|b)}{f_i(Y_i|G)} \right] \right\} dG(b). \tag{3}$$

For the detailed proof, see Web Appendix A. Note that the eigenvalues λ_j depend on the true parameter vector ψ_0, which is unknown. We can consistently estimate the eigenvalues since, based on the proof of Theorem 1, a consistent estimator for (3) is given by

$$\hat{Q}_{ll'}(\hat{\psi}) = \int_{R^q} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \psi_l} \frac{f_i(y_i|b)}{f_i(y_i|G)} \Big|_{\psi_l=\hat{\psi}_l} \right\} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \psi_{l'}} \frac{f_i(y_i|b)}{f_i(y_i|G)} \Big|_{\psi_{l'}=\hat{\psi}_{l'}} \right\} d\hat{G}(b). \tag{4}$$

The exact distribution of a weighted sum of independent chi-squared random variables has been derived by Imhof (1961, p. 422) using the inversion formula. Thus, critical values of the proposed test statistic can be computed analytically (see Web Appendix B). As a simple approximation, the test statistic $T(\hat{\psi})$ can be adjusted such that the asymptotic distribution of the modified statistic is approximately χ^2 with r degrees of freedom. Similar to Rao and Scott (1981) and Rotnitzky and Jewell (1990) we have the following result.

Corollary 1. *Let $T^* = T(\hat{\psi})/\bar{\lambda}$, where $\bar{\lambda}$ is the mean of eigenvalues. Then, under the conditions of Theorem 1, the adjusted test statistic T^* is asymptotically distributed ($\approx$) as χ^2 with r degrees of freedom.*

5. A parametric bootstrap algorithm for small samples

Asymptotic results may not always apply to small or moderate sample sizes. In those cases, we propose a parametric bootstrap procedure to approximate the finite-sample distribution of the test statistic $T(\hat{\psi})$ under the null. Our parametric bootstrap test is set up as follows:

1. Fit the model under H_0 to the original data to get $\hat{\psi}$ as the ML estimate of model parameters. Note that $\hat{\psi}' = (\hat{\theta}', \hat{\alpha}')$.
2. Calculate the test statistic (2) for the original data using the fitted model, and denote it by T_{obs}.
3. For $s = 1, ..., S$, repeat the following steps:
   (a) Generate random effects b_i^s, $i = 1, ..., N$, from G_α in which α is replaced by $\hat{\alpha}$.
   (b) Generate new observations y_i^s, $i = 1, ..., N$, from $f_\theta(y_i|b_i^s)$ in which θ is replaced by $\hat{\theta}$.
   (c) Fit the model under H_0 to the new observations y_i^s, $i = 1, ..., N$, and calculate the test statistic (2) for these observations using the corresponding fitted model, and denote it by T^s.
4. Compute the empirical p-value as the proportion of T^s exceeding T_{obs}.
5. Given the significance level δ, reject H_0 if δ is greater than the empirical p-value.

6. Implementation: A quasi-Monte Carlo integration

We use the SAS procedures NLMIXED and IML to implement our methodology. Indeed, calculation of the gradient function is straightforward since it only requires the calculation of the marginal and conditional distributions of all N individuals. We calculate the test statistic $T(\hat{\psi})$ by using a quasi-Monte Carlo (QMC) integration method as follows

$$T(\hat{\psi}) = \int_{R^q} \left(\hat{\Delta}(\hat{G}, b) - 1\right)^2 d\hat{G}\,(b) = \frac{1}{K}\sum_{k=1}^{K} \left(\hat{\Delta}(\hat{G}, b_k) - 1\right)^2, \tag{5}$$

where $b_k = \hat{G}^{-1}(c_k)$, in which $\{c_k : k = 1, ..., K\}$ are the quasi-Monte Carlo integration nodes over the unit cube $C^q = [0, 1)^q$. The QMC nodes c_k are deterministic and have the smallest discrepancy over the unit cube with respect to the Kolmogorov-Smirnov distance (see Fang and Wang, 1994). Interestingly, when $q = 1$, the QMC integration nodes are easily derived as $\{c_k = \frac{2k-1}{2K} : k = 1, ..., K\}$ with the discrepancy of $1/2K$. Because the QMC integration nodes are deterministic, no automatic random variation appears in the QMC integration approach, and further since the nodes have the smallest discrepancy over $[0, 1)^q$, it would be sufficient to increase K as much as possible to get a reliable approximation.

Using the QMC integration approach, we similarly approximate (4) as

$$\hat{Q}_{ll'}(\hat{\psi}) = \frac{1}{KN^2}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{i'=1}^{N} \left(\frac{\partial}{\partial\psi_l}\frac{f_i(y_i|b_k)}{f_i(y_i|G)}\Big|_{\psi_l=\hat{\psi}_l}\right)\left(\frac{\partial}{\partial\psi_{l'}}\frac{f_{i'}(y_{i'}|b_k)}{f_{i'}(y_{i'}|G)}\Big|_{\psi_{l'}=\hat{\psi}_{l'}}\right) \tag{6}$$

to calculate the eigenvalues $\lambda_1, ..., \lambda_r$. Specifically, we write the derivative of the ratio

of conditional and marginal distributions in (6) as

$$\frac{\partial}{\partial \psi_l} \frac{f_i(y_i|b_k)}{f_i(y_i|G)} = \left(\frac{\partial}{\partial \psi_l} \log f_i(y_i|b_k) - \frac{\partial}{\partial \psi_l} \log f_i(y_i|G) \right) \frac{f_i(y_i|b_k)}{f_i(y_i|G)},$$

which can be directly calculated using the SAS procedures NLMIXED and IML.

7. Simulation study

In this section, we evaluated the performance of the proposed diagnostic tests via simulations. We first examined the asymptotic proposal for large samples and then investigated the performance of the proposed bootstrap algorithm in small-sample situations. Because misspecifying the random-effects distribution is most problematic for binary data, we considered here a logistic generalized linear mixed model. We also compared our diagnostic tests with three recent tests proposed by Tchetgen and Coull (2006), Alonso et al. (2008), and Efendi et al. (2014), respectively.

7.1. Evaluation of the asymptotic test

For each combination of $N \in \{100, 200, 300, 500, 1000\}$ and $n \in \{10, 15\}$, we generated 1000 data sets from the logistic generalized linear mixed model

$$\text{logit}\left(P(Y_{ij} = 1|b_i)\right) = \beta_0 + \beta_1 x_{ij} + \beta_2 w_{ij} + b_i, \tag{7}$$

where Y_{ij} denotes the binary response for individual i at time point j, x_{ij} and w_{ij} represent two covariates, and b_i is a random effect with mean 0 and variance σ_b^2. We set $\beta_0 = 2$, $\beta_1 = -2$, and $\beta_2 = 1$. The covariates x_{ij} and w_{ij} were generated randomly from Uniform$(1, 5)$ and Uniform$(1, 2)$, respectively. We also generated the random effect b_i

from four distinct distributions: Normal$(0, 1)$, Chi-squared(2), Log-normal$(3, 1)$, and F$(1, 7)$. All the generated b_i's were shifted and rescaled such that each b_i has mean 0 and variance $\sigma_b^2 = 9$.

[Table 1 appears here]

The logistic mixed model (7) was fitted to each of the generated data sets under a normality assumption for the random effect b_i (the null hypothesis), and 1000 QMC integration nodes were used to calculate the test statistic $T(\hat{\psi})$ as in (5) and the eigenvalues $\lambda_1, ..., \lambda_r$ as in (6). Then, the p-value of our test was computed using the asymptotic distribution as in Theorem 1. The adjusted test statistic $T^* = T(\hat{\psi})/\bar{\lambda}$ was computed as well. For comparison, we calculated the test statistic of Tchetgen and Coull (2006) as well as the determinant-trace test statistic of Alonso et al. (2008).

For each simulation setting, we determined the proportion of cases in which a significant result was detected at the nominal level 0.05. When the true random-effects distribution was a normal distribution, this proportion corresponds to the Type I error rate, otherwise it represents the power of the test to detect misspecification. The simulation results of our asymptotic test, say T, the test based on the adjusted test statistic, say T^*, the determinant-trace test of Alonso et al. (2008), say δ_{dt}, and the test of Tchetgen and Coull (2006), say D, are presented in Table 1. The results indicate that both T and T^* show a Type I error rate smaller than the nominal level 0.05, while Type I error rates of δ_{dt} and D are closer to the nominal level. Type I error rates of our asymptotic test T and the adjusted test T^* get closer to the nominal level when N increases. Therefore, our asymptotic test is conservative in terms of Type I error when the sample size is not very large. From the results, it can be seen that our asymptotic test T is more powerful than the two tests of Tchetgen and Coull (2006) and Alonso et al. (2008). The power of our test is almost 1 for sample sizes

of 500 or larger. Furthermore, the simulated power associated with the adjusted test statistic T^* overestimates the power for sample sizes smaller than 300. Note also that the four random-effects distributions in Table 1 are sorted according to their skewness. The performance of the proposed test is better when skewness of the random-effects distribution is larger.

Heagerty and Kurland (2001) demonstrated other forms of misspecification related to the random-effects part, such as ignoring a random effect, group-specific variances, and autoregressive random effects, that could impact inference on fixed-effects parameters. We conducted further simulations to examine the power of our test in detecting such types of misspecification. The results, given in Web Appendix C, suggest that our diagnostic test has a good power to detect misspecification regarding the ignorance of some random effect from the model, however it is not powerful enough to detect autoregressive random effects. Both the test of Tchetgen and Coull (2006) and the test of Alonso et al. (2008) did not perform well when assessing autoregressive random effects. For more details, see Web Appendix C.

7.2. Evaluation of the bootstrap algorithm for small samples

To evaluate the behavior of our parametric bootstrap algorithm, we performed the same simulation study as in Section 7.1, but with smaller sample sizes of $N = 30, 50, 80, 100$. In each replication, 200 bootstrap samples were used to perform the bootstrap test as described in Section 5. We also compared our bootstrap test, say T, with the bootstrap test of Efendi et al. (2014), say E. The simulation results, displayed in Table 2, show that our parametric bootstrap test T has the correct Type I error rate with a considerable power. Our bootstrap test outperforms the bootstrap test of Efendi

et al. (2014) because we have used a more powerful test statistic that not only considers all deviations of the gradient function from 1 but also assigns an appropriate weight to each deviation. More importantly, their bootstrap algorithm was developed based on the normality of ML estimates which is a large sample property.

[Table 2 appears here]

Overall, the results of the simulations indicate that both the asymptotic and bootstrap tests perform reasonably well to detect misspecification of the random-effects distribution. As we expected, the power of our asymptotic test is not very large for small to moderate sample sizes. The proposed bootstrap algorithm performs better in small-sample situations and provides a powerful test. However, the bootstrap test is time consuming when the sample size is very large.

8. Real data example

We apply our method to real data from a randomized multicenter longitudinal study for the comparison of two oral treatments (coded as A and B) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer et al. (1996). In the study, 2×189 patients were randomly distributed over 36 centers, and 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 afterward, resulting in a maximum of seven measurements per subject. On the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. One of the responses of interest was the infection severity, coded as 0 (not severe) or 1 (severe). The main objectives were to investigate whether the percentage of severe infections decreased over time,

and to see whether the evolution was different for the two treatment groups. Similar to De Backer et al. (1996), we restrict our analysis to only those patients for whom the target nail was one of the two big toenails. This reduces the sample under consideration to 146 and 148 subjects in group A and group B, respectively. Figure 1(a) shows the observed percentage of severe infections at all time points, for both treatment groups separately.

[Figure 1 appears here]

Let Y_{ij} be the binary outcome indicating the severity of the toenail infection for patient i at occasion j. The mixed model we consider here is

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}),\ \text{logit}(\pi_{ij}) = \beta_0 + b_i + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}, \tag{8}$$

in which T_i is the treatment indicator for patient i, t_{ij} is the time point (in months) at which the jth measurement was taken for the ith patient, and b_i is a random subject-specific intercept with mean 0 and variance σ_b^2. Assuming a normal distribution for the random intercept b_i (the null hypothesis), the ML estimates of parameters and associated standard errors are calculated and reported in Table 3. The gradient function related to this model is shown in Figure 1(b), along with 95% pointwise confidence bands that have been obtained according to Verbeke and Molenberghs (2013). The gradient function plot reveals some departures from 1, suggesting that the random-effects distribution might be misspecified. To formally test the null hypothesis, our asymptotic test produces a test statistic of 0.0169, which results in a p-value of 0.094 using Web Appendix B. The asymptotic test does not provide evidence for misspecification at the significance level 0.05. In contrast, the adjusted test statistic T^*, which is equal to 11.4651 (with $T(\hat{\psi}) = 0.016911$ and $\bar{\lambda} = 0.001475$), yields a significant p-value

of 0.042, indicating that the normality assumption is inadequate for the random intercepts in this model. On the other hand, our bootstrap test declares severe violation of the normality assumption since it gives a significant p-value of 0.001 based on 500 bootstrap samples. We conjecture that $N = 294$ in this dataset is not sufficiently large for our asymptotic test to detect misspecification.

[Table 3 appears here]

As discussed by Verbeke and Molenberghs (2013), the shape of the gradient function gives some indication of how the random-effects distribution can be adapted to provide a better fit. In fact, a model with a gradient function exceeding 1 can be improved by moving probability mass from areas where the gradient function is small to areas where the gradient function is large. For the toenail data, the shape of the gradient function in Figure 1(b) suggests that we can replace the normality assumption of the random intercepts by a mixture of three normal distributions. Note also that since 163 patients never experienced a severe infection during the study and 16 patients experienced a severe infection at all visits, a 3-component mixture could capture the heterogeneity between the three different types of patient. We therefore refit model (8) by assuming the random intercepts to be distributed as $b_i \sim \pi_1 N(\mu_1, \sigma_b^2) + \pi_2 N(\mu_2, \sigma_b^2) + \pi_3 N(\mu_3, \sigma_b^2)$, with $\pi_1 + \pi_2 + \pi_3 = 1$, where we also include the restriction $\pi_1\mu_1 + \pi_2\mu_2 + \pi_3\mu_3 = 0$ to impose the assumption $E(b_i) = 0$. As shown by Liu and Yu (2008), the model with mixture of normals can still be fitted using the procedure NLMIXED in SAS. The parameter estimates are also presented in Table 3. The gradient function plot, shown in Figure 1(c), exhibits only small fluctuation around 1, suggesting that the mixture of normals seems appropriate for the random effects. To investigate this formally, we apply our asymptotic test which provides a test statistic of 0.0003, giving a p-value of 0.995 using Web Appendix B. Our asymptotic test concludes that the mixture of

normals is adequate for the random effects b_i. This is confirmed by the adjusted test statistic T^*, which is equal to 0.3309 (with $T(\hat{\psi}) = 0.000277$ and $\bar{\lambda} = 0.000837$), providing a p-value of 0.999. Our bootstrap test, with a p-value of 0.667 based on 500 bootstrap samples, also confirms the adequacy of the mixture of normals as the distribution of random effects.

Similar results are obtained using the determinant-trace test of Alonso et al. (2008). Their diagnostic test, with a p-value smaller than 0.001, suggests there is significant evidence against the normality assumption for the random intercepts in model (8), while it advocates that the model with finite mixture of normals provides an adequate fit to the data according to a non-significant p-value of 0.146.

9. Discussion

We presented a novel diagnostic test for assessing the random-effects distribution in mixed models. The proposed test statistic has been constructed based on the gradient function using the Cramér-von Mises measure. We established asymptotic properties of our test statistic and provided an explicit formula to compute critical values of our test using the asymptotic distribution. Moreover, as a simple approximation, the test statistic has been adjusted such that the asymptotic distribution of the modified statistic is approximately χ^2 with r degrees of freedom, where r is the number of eigenvalues. We also explored the finite-sample properties of our test statistic by developing a parametric bootstrap procedure.

The simulations showed that both the asymptotic and bootstrap tests perform reasonably well to detect misspecification of the random-effects distribution. While the bootstrap test has the correct Type I error rate, the asymptotic test is conservative in

terms of Type I error when the sample size is not sufficiently large. As we expected, the power of our asymptotic test is not very high for small or moderate samples, while the proposed bootstrap algorithm performs much better in small-sample situations. Since the bootstrap test is time consuming when the sample size is very large, we suggest our asymptotic test for sufficiently large samples, and our bootstrap test for smaller samples. We should point out that the required sample size for the asymptotic test may depend on several aspects of the data (e.g., cluster sizes) as well as of the outcome (e.g., binary versus continuous) and model used (e.g., presence of multiple random effects). The simulations indicated that our asymptotic test outperforms two recent tests proposed by Tchetgen and Coull (2006) and Alonso et al. (2008), respectively. Also, for small samples our bootstrap test is much more powerful than the bootstrap test of Efendi et al. (2014). Web Appendix C presents further simulations conducted with the objective of evaluating the performance of our diagnostic test in detecting some other forms of misspecification related to the random-effects part.

We employed a quasi-Monte Carlo integration method to facilitate calculations regarding our test statistic and the eigenvalues of the asymptotic distribution. For our work, we found that 1000 integration nodes are sufficient for a reliable integration approximation. However, increasing the number of integration nodes in quasi-Monte Carlo approach is not a major concern in the sense of computation time, even for high dimensional integrals (see Pan and Thompson, 2007).

While most emphasis in the literature has been placed on checking the normality assumption of random effects, our method can be used to verify the appropriateness of any parametric distribution for random effects. As illustrated in Section 8, we applied our method to check the adequacy of a finite mixture of normals as the distribution of random effects.

For the toenail data, one may argue that the model that assumes normally distributed random effects and the model that assumes a mixture of normal distributions provide approximately the same estimated regression coefficients. This apparent robustness does not hold in general, and examples of the severe impact of model misspecifications have been reported, e.g. by Litière et al. (2008), and can also be observed here if we focus attention on the treatment effect after 1 year, i.e. on inference for $\beta_1 + 12 * \beta_3$. Point estimates (and associated standard errors) are $-2.0498\,(0.8853)$ and $-1.4269\,(0.7378)$ for a normal and a mixture of normals, respectively. Under the normal random-effects distribution, a significant treatment effect ($p = 0.0213$) is obtained after 12 months, while under the mixture model, there is less evidence for such a treatment effect ($p = 0.0541$). This shows that different inferences could be obtained with different random-effects distributions.

Finally, in this paper we assumed the conditional distribution to be correctly specified. This assumption was made by Verbeke and Molenberghs (2013) to exploit the idea of the gradient function in the context of mixed models. In an attempt to relax this assumption, one can replace the marginal likelihood function with a quasi-likelihood function. The quasi-likelihood approach only needs the first and second moments of the conditional distribution to be specified. This is a subject of ongoing research.

10. Supplementary Materials

Web appendices referenced in Sections 4, 7, and 9 as well as the SAS code for the analysis of toenail data are available with this paper at the Biometrics website on Wiley Online Library.

References

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.

Aitkin, M. (1999), 'A general maximum likelihood analysis of variance components in generalized linear models', *Biometrics* **55**, 117–128.

Alonso, A., Litière, S. and Laenen, A. (2010), 'A note on the indeterminacy of the random-effects distribution in hierarchical models', *The American Statistician* **64**, 318–324.

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

Böhning, D. (1989), 'Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms', *Biometrika* **76**, 375–383.

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.

Claeskens, G. and Hart, J. D. (2009), 'Goodness-of-fit tests in mixed models', *Test* **18**, 213–239.

De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996), 'A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250 mg/day vs. itracona-

zole 200 mg/day–a double-blind comparative trial', *British Journal of Dermatology* **134**, 16–17.

Efendi, A., Drikvandi, R., Verbeke, G. and Molenberghs, G. (2014), 'A goodness-of-fit test for the random-effects distribution in mixed models', *Statistical Methods in Medical Research*. doi: 10.1177/0962280214564721.

Fang, K. T. and Wang, Y. (1994), *Number-theoretic methods in statistics*, Chapman and Hall, London.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008), *Longitudinal data analysis: a handbook of modern statistical methods*, Chapmann & Hall/CRC.

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

Huang, X. (2009), 'Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response', *Biometrics* **65**, 361–368.

Imhof, J. P. (1961), 'Computing the distribution of quadratic forms in normal variables', *Biometrika* **48**, 419–426.

Jiang, J. (2001), 'Goodness-of-fit tests for mixed model diagnostics', *The Annals of Statistics* **29**, 1137–1164.

Lange, N. and Ryan, L. (1989), 'Assessing normality in random effects models', *The Annals of Statistics* **17**, 624–642.

Lindsay, B. G. (1983*a*), 'The geometry of mixture likelihoods: a general theory', *The Annals of Statistics* **11**, 86–94.

Lindsay, B. G. (1983*b*), 'The geometry of mixture likelihoods, part II: the exponential family', *The Annals of Statistics* **11**, 783–792.

Litière, S., Alonso, A. and Molenberghs, G. (2008), 'The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models', *Statistics in Medicine* **27**, 3125–3144.

Liu, L. and Yu, Z. (2008), 'A likelihood reformulation method in non-normal random effects models', *Statistics in Medicine* **27**, 3105–3124.

McCulloch, C. E. and Neuhaus, J. M. (2011*a*), 'Misspecifying the shape of a random effects distribution: why getting it wrong may not matter', *Statistical Science* **26**, 388–402.

McCulloch, C. E. and Neuhaus, J. M. (2011*b*), 'Prediction of random effects in linear and generalized linear models under model misspecification', *Biometrics* **67**, 270–279.

Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M. and Strawderman, R. (2006), 'Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects', *Journal of Computational and Graphical Statistics* **15**, 39–57.

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.

Pan, J. and Thompson, R. (2007), 'Quasi-Monte Carlo estimation in generalized linear mixed models', *Computational Statistics and Data Analysis* **51**, 5765–5775.

Pan, Z. and Lin, D. (2005), 'Goodness-of-fit methods for generalized linear mixed models', *Biometrics* **61**, 1000–1009.

Rao, J. and Scott, A. (1981), 'The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables', *Journal of the American Statistical Association* **76**, 221–230.

Ritz, C. (2004), 'Goodness-of-fit tests for mixed models', *Scandinavian Journal of Statistics* **31**, 443–458.

Rotnitzky, A. and Jewell, N. P. (1990), 'Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data', *Biometrika* **77**, 485–497.

Tchetgen, E. J. and Coull, B. A. (2006), 'A diagnostic test for the mixing distribution in a generalised linear mixed model', *Biometrika* **93**, 1003–1010.

Verbeke, G. and Lesaffre, E. (1996), 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association* **91**, 217–221.

Verbeke, G. and Molenberghs, G. (2013), 'The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models', *Biostatistics* **14**, 477–490.

Waagepetersen, R. (2006), 'A simulation-based goodness-of-fit test for random effects in generalized linear mixed models', *Scandinavian Journal of Statistics* **33**, 721–731.

Zhang, D. and Davidian, M. (2001), 'Linear mixed models with flexible distributions of random effects for longitudinal data', *Biometrics* **57**, 795–802.

Table 1: Power and Type I error rate of our asymptotic diagnostic test, denoted by T, the test based on the adjusted test statistic, denoted by T^*, the determinant-trace test of Alonso et al. (2008), denoted by δ_{dt}, and the test of Tchetgen and Coull (2006), denoted by D, to detect misspecification of the random-effects distribution in the binary mixed model (7) at the nominal level 0.05. A normal distribution was assumed for the random effect b_i to fit the model, whereas the true random effect was generated from Normal$(0, 1)$, Chi-squared(2), Log-normal$(3, 1)$, and F$(1, 7)$. Note that all the generated values of b_i were shifted and rescaled such that each b_i has mean 0 and variance $\sigma_b^2 = 9$.

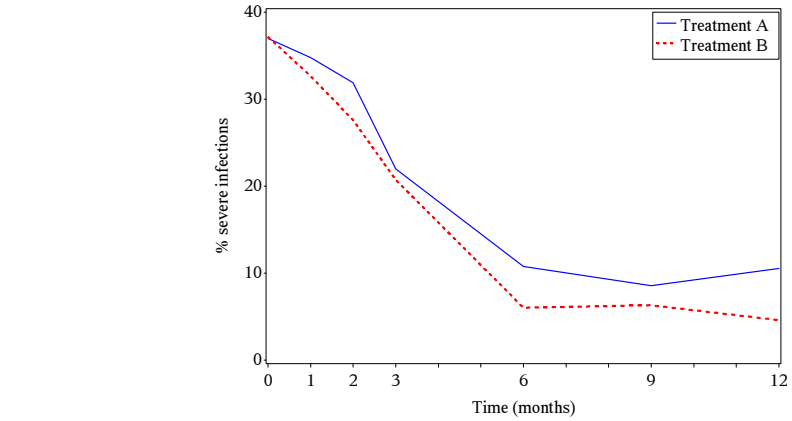
		$N = 100$		$N = 200$		$N = 300$		$N = 500$		$N = 1000$	
True RE distribution		$n = 10$	$n = 15$	$n = 10$	$n = 15$	$n = 10$	$n = 15$	$n = 10$	$n = 15$	$n = 10$	$n = 15$
Normal$(0, 1)$	T	0.001	0.002	0.002	0.013	0.003	0.017	0.011	0.019	0.018	0.029
	T^*	0.003	0.010	0.005	0.014	0.006	0.024	0.014	0.021	0.027	0.036
	δ_{dt}	0.080	0.050	0.053	0.033	0.040	0.050	0.056	0.040	0.047	0.055
	D	0.091	0.078	0.059	0.082	0.066	0.061	0.050	0.063	0.054	0.051
Chi-squared(2)	T	0.018	0.106	0.413	0.735	0.853	0.983	0.996	1.000	1.000	1.000
	T^*	0.069	0.290	0.684	0.960	0.971	0.996	1.000	1.000	1.000	1.000
	δ_{dt}	0.094	0.112	0.094	0.097	0.118	0.122	0.133	0.141	0.155	0.187
	D	0.061	0.069	0.070	0.201	0.120	0.383	0.277	0.806	0.775	1.000
Log-normal$(3, 1)$	T	0.011	0.087	0.479	0.740	0.912	0.994	1.000	1.000	1.000	1.000
	T^*	0.075	0.330	0.846	0.991	0.996	1.000	1.000	1.000	1.000	1.000
	δ_{dt}	0.148	0.122	0.243	0.246	0.261	0.267	0.338	0.350	0.405	0.550
	D	0.064	0.090	0.139	0.475	0.328	0.783	0.663	0.992	0.974	1.000
F$(1, 7)$	T	0.015	0.086	0.387	0.731	0.888	0.997	1.000	1.000	1.000	1.000
	T^*	0.113	0.282	0.839	0.994	1.000	1.000	1.000	1.000	1.000	1.000
	δ_{dt}	0.168	0.210	0.255	0.233	0.285	0.301	0.313	0.360	0.424	0.567
	D	0.051	0.128	0.179	0.574	0.322	0.919	0.786	0.997	0.986	1.000

Table 2: Power and Type I error rate of our parametric bootstrap test, denoted by T, and the bootstrap test of Efendi et al. (2014), denoted by E, to detect misspecification of the random-effects distribution in the binary mixed model (7) at the nominal level 0.05. A normal distribution was assumed for the random effect b_i to fit the model, whereas the true random effect was generated from Normal$(0, 1)$, Chi-squared(2), Log-normal$(3, 1)$, and F$(1, 7)$. Note that all the generated values of b_i were shifted and rescaled such that each b_i has mean 0 and variance $\sigma_b^2 = 9$.

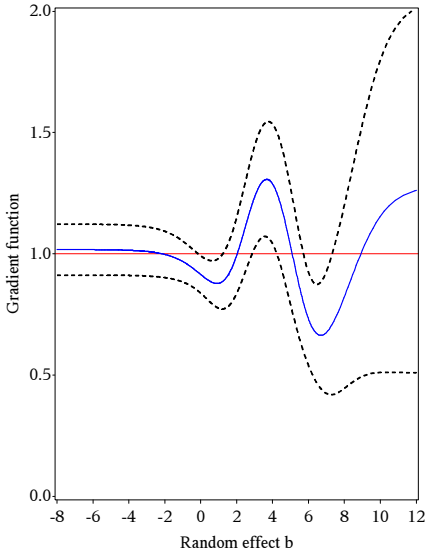
		$N = 30$		$N = 50$		$N = 80$		$N = 100$	
True RE distribution		$n = 10$	$n = 15$	$n = 10$	$n = 15$	$n = 10$	$n = 15$	$n = 10$	$n = 15$
Normal$(0, 1)$	T	0.068	0.044	0.036	0.045	0.038	0.053	0.047	0.051
	E	0.055	0.034	0.037	0.040	0.039	0.042	0.046	0.048
Chi-squared(2)	T	0.281	0.415	0.404	0.552	0.616	0.809	0.772	0.834
	E	0.198	0.286	0.295	0.384	0.372	0.453	0.515	0.661
Log-normal$(3, 1)$	T	0.642	0.651	0.790	0.803	0.941	0.986	0.973	0.995
	E	0.404	0.472	0.486	0.580	0.599	0.637	0.684	0.782
F$(1, 7)$	T	0.663	0.817	0.869	0.922	0.935	0.990	0.994	1.000
	E	0.429	0.537	0.570	0.631	0.624	0.703	0.718	0.813

Table 3: Toenail Data: the ML estimates of parameters and associated standard errors obtained from fitting model (8), once assuming the random effect b_i to be normal, once assuming the random effect b_i to follow a finite mixture of normals with 3 components.

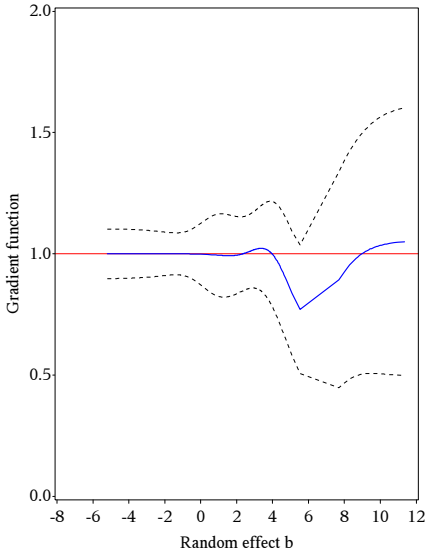
		Normal	Mixture
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	β_0	-1.6308 (0.4356)	-1.5644 (0.5311)
Treat	β_1	-0.1146 (0.5855)	0.4642 (0.4316)
Time	β_2	-0.4043 (0.0460)	-0.3970 (0.0468)
Treat×Time	β_3	-0.1614 (0.0719)	-0.1573 (0.0756)
Variance of b_i	σ_b^2	16.1318 (3.0643)	0.6925 (0.3327)
Prob-1	π_1		0.5759 (0.0435)
Prob-2	π_2		0.3788 (0.0439)
Prob-3	π_3		0.0453 (0.0129)
Mean-1	μ_1		-2.5912 (0.5309)
Mean-2	μ_2		2.8097 (0.3435)
Mean-3	μ_3		9.4479 (1.3290)
-2 log-likelihood		1247.8	1219.3



(a)



(b)



(c)

Figure 1: Toenail data: (a) Evolution of the percentage of severe toenail infections in the two treatment groups separately. (b) Gradient function (solid) and 95% pointwise confidence bands (dashed) for the logistic mixed model (8) with normal random effects. (c) Gradient function (solid) and 95% pointwise confidence bands (dashed) for the logistic mixed model (8) with a finite mixture of normals as the random-effects distribution. The confidence bands for the gradient function are obtained according to Verbeke and Molenberghs (2013).