Testing for misspecification in generalized linear mixed models

Peer-reviewed author version

ALONSO ABAD, Ariel; LITIERE, Saskia & MOLENBERGHS, Geert (2010) Testing for misspecification in generalized linear mixed models. In: BIOSTATISTICS, 11(4). p. 771-786.

DOI: 10.1093/biostatistics/kxq019
Handle: http://hdl.handle.net/1942/11149

Testing for Misspecification in Generalized Linear Mixed Models

Ariel ALONSO[1], Saskia LITIÈRE[*1], and Geert MOLENBERGHS[1,2]

Interuniversity Institute for Biostatistics and statistical Bioinformatics

[1] Hasselt University, Agoralaan 1, B3590 Diepenbeek, Belgium

[2] Katholieke Universiteit Leuven, Belgium

[*] *email:* saskia.litiere@uhasselt.be

Abstract

Generalized linear mixed models have become a frequently used tool for the analysis of non-Gaussian longitudinal data. Estimation is often based on maximum likelihood theory, which assumes that the underlying probability model is correctly specified. Recent research shows that the results obtained from these models are not always robust against departures from the assumptions on which they are based. Therefore, diagnostic tools for the detection of model misspecifications are of the utmost importance. In this paper, we propose two diagnostic tests that are based on two equivalent representations of the model information matrix. We evaluate the power of both tests using theoretical considerations as well as via simulation. In the simulations the performance of the new tools is evaluated in many settings of practical relevance, focusing on misspecification of the random-effects structure. In all the scenarios the results were encouraging, however, the tests also exhibited inflated type I error rates when the sample size was small or moderate. Importantly, a parametric bootstrap version of the tests seems to overcome this problem, although more research in this direction may be needed. Finally, both tests were also applied to analyze a real case study in psychiatry.

KEY WORDS: Generalized linear mixed model; Information matrix test; Linear mixed model; Random-effects Misspecification; Sandwich Estimator.

1 Introduction

Over the last decades, hierarchical models have developed into an effective tool for statistical analysis across a variety of applications. The parameters of interest are often estimated using maximum likelihood (ML) and assuming that the model is correctly specified. One of the basic assumptions underlying hierarchical models concerns the distribution of the random effects. To improve mathematical tractability and numerical performance, random effects are commonly assumed to be normally distributed. Nonetheless, since random effects are not observed, the validity of this presumption is difficult to verify.

For linear mixed models (LMM), Verbeke and Lesaffre (1997) showed that the maximum likelihood estimators of the fixed effects and variance components, obtained under the assumption of normal random effects, are consistent and asymptotically normal, even when the random-effects distribution is misspecified. However, recent research suggests that this does not hold for generalized linear mixed models (GLMM). According to Agresti *et al.* (2004), the choice of the random-effects distribution seems to have, in most situations, little effect on the maximum likelihood estimators. In spite of that, when there is a severe polarization of subjects, e.g., by omitting an influential binary covariate, the predictive qualities of characteristics involving the random effects as well as the fixed effects can be affected. Similarly, Heagerty and Kurland (2001) found substantial bias while using a random-intercept model when the random-effect distribution depends on measured covariates. Litière *et al.* (2008) found that the estimates of the variance components are always subject to considerable bias when the random-effect distribution is misspecified. The bias induced in the estimates of the linear predictor parameters appears to depend on the magnitude of the variance component, whereby a large bias is associated with a large random-effect variance. Furthermore, Litière *et al.* (2007) and Woods (2008) established

that the Type I error rate related to the tests of the linear predictor parameters, can also be severely impacted. The situation worsens when more complicated random-effects structures are used.

Evidently, in these circumstances, the development of diagnostic tools for GLMMs is of great importance. However, up to now, this problem has received only moderate attention in the literature. For instance, Waagepetersen (2006) proposed a simulation-based test to evaluate the appropriateness of the choice of the random-effects distribution, by generating random effects while conditioning on the observations. Although simulations with Poisson responses showed a reasonable power, this test required very large cluster and sample sizes to produce similar results with binary outcomes. Tchetgen and Coull (2006) introduced a diagnostic test to evaluate the assumed random-effects distribution, by comparing marginal and conditional maximum likelihood estimators of a subset of fixed effects in the model. The applicability of their test is restricted to those settings where at least one within-cluster covariate is available. Another limitation of this test is that it cannot be applied when auto-regressive random effects are present. However, as will be illustrated in Section 4, misspecification in this scenario may seriously affect our inferences.

White (1982) proposed a general test for model misspecification. Notwithstanding, his Information Matrix Test (IMT) requires third-order partial derivatives of the likelihood function. Even though the calculation of higher-order derivatives might not be an issue in cases where the likelihood is available in a closed form, it can become an important problem when working with complicated likelihood functions, like in generalized linear mixed models. Consequently, one has to resort to numerical approximations, which can be burdensome and less than straightforward to carry out using conventional statistical packages. In the present work, we propose two alternative diagnostic tools along the ideas of the IMT, but without the need for third-order partial

derivatives of the likelihood.

The motivating case study will be introduced in Section 2. In Section 3, the new tests are presented and some of their properties are discussed at a theoretical level. In Section 4 we study the performance of the newly proposed tools via simulations. Next, in Section 5, the appropriateness of the model chosen for the case study is evaluated using the diagnostic tools introduced in Section 3. Finally, in Section 6 some concluding remarks are given.

2 Case Study

The 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 *et al.*, 2004).The response variable is the *Clinical Global Impression* (CGI), a 7-grade scale used to characterize a subject's mental condition. Often, clinicians are interested in a dichotomic version of the scale that equals 1 for patients classified as normal to mildly ill, and 0 for patients classified as moderately to severely ill. In total, $n = 453$ patients were included in the trial, from which 226 were randomly assigned to the experimental treatment ($z_i = 1$) and the rest (227) to the control group ($z_i = 0$). Treatment was administered for 8 weeks and the outcome measured at 6 fixed time points: 0, 1, 2, 4, 6, and 8 weeks. Figure 1 summarizes the probability of being classified as normal to mildly ill ($P(Y = 1)$) by time point and treatment group. Note that, due to the random treatment allocation, the two treatment groups have similar average response at the start of the study, however, a very rapid onset of the treatment is observed at week 1 in the experimental group.

2.1 Generalized Linear Mixed Models

In this section we will introduce some general notation. Let us start by denoting y_{ij} the jth response of subject i, with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Conditional on a vector of individual random effects $\boldsymbol{b}_i$, the outcome variables are assumed to be independent, with density functions belonging to the exponential family

$$f(y_{ij}|\theta_{ij}, \varphi) = \exp[\varphi^{-1}\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\} + c(y_{ij}, \varphi)], \tag{1}$$

where φ is a scale parameter, $c(.)$ is a function only depending on y_{ij} and φ, and $\psi(.)$ is a function satisfying $\mathsf{E}(y_{ij}|\boldsymbol{b}_i) = \psi'(\theta_{ij})$ and $\mathsf{Var}(y_{ij}|\boldsymbol{b}_i) = \varphi\psi''(\theta_{ij})$. Further, $\mathsf{E}(y_{ij}|\boldsymbol{b}_i) = v(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)$, where $v(.)$ denotes a known link function, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors of covariates, and $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients. The subject-specific effects $\boldsymbol{b}_i$ are commonly assumed to be normal distributed with mean zero and variance-covariance matrix $\boldsymbol{D}$. Fitting the model requires maximization of the marginal likelihood, which is obtained by integrating over the random effects. Let the contribution of subject i to the marginal likelihood be given by

$$f(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{D}, \varphi) = \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \varphi) f(\boldsymbol{b}_i|\boldsymbol{D})\, d\boldsymbol{b}_i, \tag{2}$$

then we can write the marginal likelihood as

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \varphi) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \varphi) f(\boldsymbol{b}_i|\boldsymbol{D})\, d\boldsymbol{b}_i. \tag{3}$$

The frequently used normal distribution for the random effects generally leads to intractable likelihood functions. In response, several numerical approximations to the likelihood have been implemented in the available software packages. For example, Gaussian quadrature, as implemented in the SAS procedure NLMIXED, approximates the integral using Gaussian-Hermite polynomials, thereby employing specific properties of the normal distribution. All the analyses and simulations

in this manuscript were carried out using this procedure, choosing adaptive Gaussian quadrature with 50 quadrature points to approximate the likelihood.

2.2 Analysis of the Case Study

We analyzed the data using a random-intercept model and considering different link functions and linear predictors. In the model building exercise, a total of twelve models were fitted. These models were constructed as combinations of three link functions, i.e., the logit, log-log and probit link, and four linear predictors. For these predictors, expressions which model the fast onset of treatment, as in Model 4 below, were compared with structures in which this aspect of the data was ignored. Further, it was studied whether the evolution of the subjects over time differed by treatment group. The Akaike's Information Criterion (AIC) was used to select the best fitting model. The following final model emerged from this analysis

$$\text{logit}\{P(y_{ij}=1|b_{0i})\} = \begin{cases} \gamma_0 + \gamma_1 t_j + \gamma_2 z_i t_j + b_{0i} & \text{when } t_j \leq 1 \\ (\gamma_0 + \gamma_1 - \beta_1 - \beta_2) + \gamma_2 z_i + \beta_1 t_j + \beta_2 t_j^2 + b_{0i} & \text{when } t_j > 1 \end{cases}, \quad (4)$$

where i and t_j denote the patient and the measurement occasion, respectively, and b_{0i} is a random effect, assumed to follow a zero-mean normal distribution with variance σ_b^2. This model captures the impact of randomization at the beginning of the study through the common intercept parameter γ_0. Furthermore, γ_2 reflects the difference in the time evolutions of both groups during the first week of treatment. Finally, after the first week, both groups exhibit parallel quadratic evolutions over time, at the logit scale, characterized by the slope β_1 and the quadratic effect β_2. Figure 1(a) displays the plot of the fitted values obtained from (4) against the observed probability of being classified as normal to mildly ill ($P(Y=1)$) by time point and treatment group. The fitted probabilities are calculated by numerically integrating out the random effect for each subject. Until week 4 there seems to be a reasonable agreement between the fitted and

the observed values. Nevertheless, some discrepancy is observed in the last two measurement occasions. This may be ascribed to the rather high proportion of dropout at the end of the study (up to 27% and 21% for the control and experimental group respectively). We do not envisage entering here a full discussion on the missing data problem; rather, we will assume that the missing data generating mechanism is missing at random (MAR) making our likelihood approach a valid option (Molenberghs and Kenward, 2007).

The maximum likelihood estimates for the parameters in (4) are displayed in the first part of Table 1. Even though some evidence of treatment effect in the first week was observed, the parameter characterizing this effect γ_2 was not significant at the 5% level ($p = 0.094$). Moreover, the relatively large value obtained for the variance of the random intercept could be explained by the high proportion of patients that have a response pattern of nothing but zeros (60% and 55% in the control and experimental group, respectively). This high intra-subject correlation is accommodated in the model through a large value of σ_b^2. Note that such a large variance could imply a serious bias in the estimation of the linear predictor parameters, including the treatment effect, if the random-effect distribution is misspecified (Litière *et al.*, 2008) and, at the same time, it may also hint on the inappropriateness of the normal distribution for the random intercept. Therefore, in this setting, one would like to test for possible misspecification of the random-effect distribution, in particular, or any other misspecification, in general. In the following sections, we will address this issue in some detail.

3 Testing for Misspecification

Let us start by considering a random variable $\boldsymbol{y}$ with probability density (mass) function h, and a parametric family of probability density (mass) functions $\mathfrak{F} = \{f(\boldsymbol{y}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Gamma\}$. In

this manuscript, f denotes the marginal model (2) associated with the hierarchical model defined in (1), and the random effects are assumed to follow a normal distribution. Moreover, $\boldsymbol{\xi}$ represents the vector of all parameters in (1), including the linear predictor parameters in $\boldsymbol{\beta}$ and the variance components in $\boldsymbol{D}$.

We say that the model is correctly specified if there exists a $\boldsymbol{\xi}_0 \in \Gamma$ such that $h(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\xi}_0)$. White (1982) showed that, under some regularity conditions, the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}_n$ will (strongly) converge to the value of $\boldsymbol{\xi}$, denoted by $\boldsymbol{\xi}^*$, which minimizes the so-called Kullback-Leibler Information Criterion (KLIC)

$$I(h : f, \boldsymbol{\xi}) = E\left(\log \frac{h(\boldsymbol{y})}{f(\boldsymbol{y}, \boldsymbol{\xi})}\right),$$

where the expectation is taken with respect to the true distribution. Note that if the model is correctly specified, then the information criterion attains its unique minimum at $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$. In such a case, $\widehat{\boldsymbol{\xi}}_n$ is a consistent estimator for $\boldsymbol{\xi}_0$ and the inverse of the Fisher information matrix can be used to obtain standard errors for $\widehat{\boldsymbol{\xi}}_n$. Nonetheless, this matrix does not yield valid results when the model is misspecified. Instead, appropriate standard errors are obtained by replacing the asymptotic covariance matrix by the so-called sandwich estimator. To this end, we introduce the following additional notation

$$\boldsymbol{A}(\boldsymbol{\xi}) = \mathsf{E}\left(\left\{\frac{\partial^2 \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell}\right\}\right), \qquad \boldsymbol{B}(\boldsymbol{\xi}) = \mathsf{E}\left(\left\{\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell}\right\}\right),$$

$$\boldsymbol{A}_n(\boldsymbol{\xi}) = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell}\right\}, \qquad \boldsymbol{B}_n(\boldsymbol{\xi}) = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell}\right\},$$

where $k, \ell = 1, \ldots, p$, and p denotes the number of parameters in the model. If we further define $\boldsymbol{V}(\boldsymbol{\xi}) = \boldsymbol{A}^{-1}(\boldsymbol{\xi})\boldsymbol{B}(\boldsymbol{\xi})\boldsymbol{A}^{-1}(\boldsymbol{\xi})$ then, asymptotically, $\sqrt{n}\left(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*\right) \sim \mathsf{N}\left(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\xi}^*)\right)$. Additionally, when the model is correctly specified, the so-called *information matrix equality* holds, i.e.,

$$\boldsymbol{B}(\boldsymbol{\xi}^* = \boldsymbol{\xi}_0) + \boldsymbol{A}(\boldsymbol{\xi}^* = \boldsymbol{\xi}_0) = \boldsymbol{0}, \tag{5}$$

and therefore

$$\boldsymbol{V}(\boldsymbol{\xi}^* = \boldsymbol{\xi}_0) = -\boldsymbol{A}^{-1}(\boldsymbol{\xi}^* = \boldsymbol{\xi}_0), \tag{6}$$

i.e., we recover the inverse of the Fisher Information-Matrix. Based on these results, the elements of $\boldsymbol{B}(\boldsymbol{\xi}^*) + \boldsymbol{A}(\boldsymbol{\xi}^*)$ or $\boldsymbol{V}(\boldsymbol{\xi}^*) + \boldsymbol{A}^{-1}(\boldsymbol{\xi}^*)$ can be used as potential indicators of misspecification. This is the basic idea underlying the diagnostic tools that will be introduced in the following sections where, to simplify the notation, the * will be omited.

3.1 The Sandwich Estimator Test

Let us recall the equivalent form of the information matrix equality given in (6) and valid under a correctly specified model. Our first proposal focuses on the difference between $\boldsymbol{V}(\boldsymbol{\xi})$ and $-\boldsymbol{A}^{-1}(\boldsymbol{\xi})$ as a possible indicator of misspecification.

The Information-Matrix Test (IMT), introduced by White (1982), is based on $\boldsymbol{B}(\boldsymbol{\xi}) + \boldsymbol{A}(\boldsymbol{\xi})$. However, as this author claims, it may be prohibitive to base the test on all the elements of this matrix. Indeed, as the number of parameters in the model increases, so will the number of elements to be tested jointly, as well as the degrees of freedom of the test. Along these ideas, in the present work we will focus only on the diagonal of $\boldsymbol{V}(\boldsymbol{\xi}) + \boldsymbol{A}^{-1}(\boldsymbol{\xi})$.

Let us now define $\tilde{\boldsymbol{V}}_n(\boldsymbol{\xi}) = \boldsymbol{A}^{-1}(\boldsymbol{\xi})\boldsymbol{B}_n(\boldsymbol{\xi})\boldsymbol{A}^{-1}(\boldsymbol{\xi})$ and $\boldsymbol{v}_n(\boldsymbol{\xi}) = \text{diag}\left(\tilde{\boldsymbol{V}}_n(\boldsymbol{\xi}) + \boldsymbol{A}^{-1}(\boldsymbol{\xi})\right)$. Observe that $\boldsymbol{v}_n(\boldsymbol{\xi})$ can also be written as $\Delta\text{vec}\left[\tilde{\boldsymbol{V}}_n(\boldsymbol{\xi}) + \boldsymbol{A}^{-1}(\boldsymbol{\xi})\right]$. In this expression, the operator vec$(\cdots)$ is the vector obtained by stacking the columns of the matrix one below the other and Δ is the $p \times p^2$ matrix with elements

$$\Delta_{k\ell} = \begin{cases} 1 & \text{for } k = 1, \ldots, p \text{ and } \ell = (k-1)p + k, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Further, let us consider

$$\boldsymbol{b}_i(\boldsymbol{\xi}) = \text{vec}\left[\left\{\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell}\right\}\right],$$

and let $\boldsymbol{\mu}_b(\boldsymbol{\xi})$ and $\boldsymbol{V}_b(\boldsymbol{\xi})$ denote the mean and the covariance matrix of $\boldsymbol{b}_i(\boldsymbol{\xi})$. An unbiased estimator of $\boldsymbol{\mu}_b(\boldsymbol{\xi})$ is given by $\widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi}) = (1/n)\sum_{i=1}^n \boldsymbol{b}_i(\boldsymbol{\xi}) = \text{vec}\,[\boldsymbol{B}_n(\boldsymbol{\xi})]$, whereas $\boldsymbol{V}_b(\boldsymbol{\xi})$ can be estimated through

$$\widehat{\boldsymbol{V}}_b(\boldsymbol{\xi}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})\right)\left(\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})\right)^T. \tag{8}$$

In addition, let us define

$$\boldsymbol{C}_v(\boldsymbol{\xi}) = \Delta\{\boldsymbol{A}^{-1}(\boldsymbol{\xi}) \otimes \boldsymbol{A}^{-1}(\boldsymbol{\xi})\}\boldsymbol{V}_b(\boldsymbol{\xi})\{\boldsymbol{A}^{-1}(\boldsymbol{\xi}) \otimes \boldsymbol{A}^{-1}(\boldsymbol{\xi})\}\Delta^T, \tag{9}$$

a consistent estimator $\widehat{\boldsymbol{C}}_v(\boldsymbol{\xi})$ can then be obtained by plugging (8) into (9). Using all of these elements, we can now establish the following result.

Theorem 1 (Sandwich Estimator Test) *Under general regularity conditions, if the model is correctly specified then, asymptotically,* $\sqrt{n}\boldsymbol{v}_n(\boldsymbol{\xi}_0) \sim N_p\left(\boldsymbol{0}, \boldsymbol{C}_v(\boldsymbol{\xi}_0)\right)$. *From the previous expression and the Cochran theorem (Sen and Singer, 1993) it follows*

$$\delta_s(n) = n\boldsymbol{v}_n^T(\boldsymbol{\xi}_0)\widehat{\boldsymbol{C}}_v^{-1}(\boldsymbol{\xi}_0)\boldsymbol{v}_n(\boldsymbol{\xi}_0) \sim \chi_p^2. \tag{10}$$

A proof of this result can be found on `http://www.ibiostat.be/software`, the software webpage of I-Biostat. Some important comments come into place here. Notice first that, in all the previous deductions, it has been implicitly assumed that the matrix $\boldsymbol{A}$ is known. However, when Theorem 1 is applied in practice, $\boldsymbol{A}^{-1}(\boldsymbol{\xi}_0)$ in (10) needs to be substituted by its consistent estimator under the null $\boldsymbol{A}_n^{-1}(\widehat{\boldsymbol{\xi}}_0)$. As a consequence, the test statistic introduced in Theorem 1

clearly omits two potentially important sources of variability, i.e., the variability introduced by estimating $\boldsymbol{A}^{-1}(\boldsymbol{\xi}_0)$ by $\boldsymbol{A}_n^{-1}(\boldsymbol{\xi}_0)$ and the variability introduced by estimating $\boldsymbol{\xi}_0$ by its consistent estimator under the null $\widehat{\boldsymbol{\xi}}_n$. In what follows we will introduce an alternative of the IMT that not only evades the previous assumption about the matrix $\boldsymbol{A}^{-1}(\boldsymbol{\xi}_0)$ but also avoids the calculation of higher order derivatives.

3.2 The Modified Information Matrix Test

In this section we directly consider the information matrix equality given in (5) and valid under a correctly specified model. Let us first define $\boldsymbol{d}_n(\boldsymbol{\xi}) = \Delta \text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi}) + \boldsymbol{B}_n(\boldsymbol{\xi})\right]$, where Δ is given by (7). While developing the Sandwich Estimator Test (SET), the variability of $\text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]$ was estimated using the empirical covariance estimator (8). We will now use the same idea to obtain an empirical estimate of the variability of $\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right]$. Notice that from the previous expression for $\boldsymbol{d}_n(\boldsymbol{\xi})$ follows

$$n^{-1}\boldsymbol{C}_D(\boldsymbol{\xi}) = \text{cov}\left(\boldsymbol{d}_n(\boldsymbol{\xi})\right) = \Delta \text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right] + \text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right)\Delta^T \tag{11}$$

where

$$\begin{aligned}\text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right] + \text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right) =& \text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right]\right) + \text{cov}\left(\text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right) \\ &+ \text{cov}\left(\text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right], \text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right]\right) \\ &+ \text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right], \text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right).\end{aligned}$$

Further, $\text{cov}\left(\text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right) = n^{-1}\boldsymbol{V}_b(\boldsymbol{\xi})$, and $\boldsymbol{V}_b(\boldsymbol{\xi})$ can be consistently estimated using (8). Similarly to what was done in Section 3.1, we can now define

$$\boldsymbol{a}_i(\boldsymbol{\xi}) = \text{vec}\left[\left\{\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell}\right\}\right].$$

Let $\boldsymbol{\mu}_a(\boldsymbol{\xi})$ and $\boldsymbol{V}_a(\boldsymbol{\xi})$ represent the mean and the covariance matrix of $\boldsymbol{a}_i(\boldsymbol{\xi})$. It is then easy to show that $\text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right]\right) = n^{-1}\boldsymbol{V}_a(\boldsymbol{\xi})$, and a consistent estimator of $\boldsymbol{V}_a(\boldsymbol{\xi})$ is given by

$$\widehat{\boldsymbol{V}}_a(\boldsymbol{\xi}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})\right)\left(\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})\right)^T, \tag{12}$$

where $\widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i(\boldsymbol{\xi})$. Finally, let $\boldsymbol{C}_{ab}(\boldsymbol{\xi})$ and $\boldsymbol{C}_{ba}(\boldsymbol{\xi})$ denote the covariance matrices associated with $\boldsymbol{a}_i(\boldsymbol{\xi})$ and $\boldsymbol{b}_i(\boldsymbol{\xi})$, and $\boldsymbol{b}_i(\boldsymbol{\xi})$ and $\boldsymbol{a}_i(\boldsymbol{\xi})$ respectively. It then follows that $\text{cov}\left(\text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right], \text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right]\right) = \frac{1}{n}\boldsymbol{C}_{ab}(\boldsymbol{\xi})$ and $\text{cov}\left(\text{vec}\left[\boldsymbol{B}_n(\boldsymbol{\xi})\right], \text{vec}\left[\boldsymbol{A}_n(\boldsymbol{\xi})\right]\right) = \frac{1}{n}\boldsymbol{C}_{ba}(\boldsymbol{\xi})$. Consistent estimators for these matrices are given by

$$\begin{aligned}
\widehat{\boldsymbol{C}}_{ab}(\boldsymbol{\xi}) =& \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})\right)\left(\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})\right)^T, \\
\widehat{\boldsymbol{C}}_{ba}(\boldsymbol{\xi}) =& \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})\right)\left(\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})\right)^T.
\end{aligned}$$

Plugging the previous estimators into (11) we can obtain a consistent estimator $\widehat{\boldsymbol{C}}_D(\boldsymbol{\xi})$ for $\boldsymbol{C}_D(\boldsymbol{\xi})$. Using all these elements, we can now formulate the following theorem.

Theorem 2 (Modified Information Matrix Test) *Under general regularity conditions, if the model is correctly specified then, asymptotically,* $\sqrt{n}\boldsymbol{d}_n(\boldsymbol{\xi}_0) \sim N_p\left(\boldsymbol{0}, \boldsymbol{C}_D(\boldsymbol{\xi}_0)\right)$. *From the previous expression and the Cochran theorem (Sen and Singer, 1993) it follows*

$$\Im_m(n) = n\boldsymbol{d}_n^T(\boldsymbol{\xi}_0)\widehat{\boldsymbol{C}}_D^{-1}(\boldsymbol{\xi}_0)\boldsymbol{d}_n(\boldsymbol{\xi}_0) \sim \chi_p^2. \tag{13}$$

A proof for this result is also available on the aforementioned web-page. Observe that, unlike the SET, the Modified Information Matrix Test (MIMT) does take into account the variability introduced by replacing $\boldsymbol{A}(\boldsymbol{\xi}_0)$ with its consistent estimator $\boldsymbol{A}_n(\boldsymbol{\xi}_0)$. However, unlike the IMT, the MIMT still does not take into account the extra variability coming from replacing $\boldsymbol{\xi}_0$ with

$\widehat{\boldsymbol{\xi}}_n$. This is the only source of variability ignored by the test and it can be seen as the price to pay for avoiding high-order derivatives and gaining simplicity.

It is important to point out that both the SET and the MIMT have been constructed based on the diagonal elements of the matrices $\boldsymbol{V}(\boldsymbol{\xi}^*) + \boldsymbol{A}^{-1}(\boldsymbol{\xi}^*)$ and $\boldsymbol{A}(\boldsymbol{\xi}^*) + \boldsymbol{B}(\boldsymbol{\xi}^*)$, respectively. However, this does not preclude the use of other elements of these matrices if necessary. Indeed, both tests can be easily adapted by doing the appropriate modifications in the matrix Δ and selecting the adequate $\boldsymbol{a}_i(\boldsymbol{\xi})$ and $\boldsymbol{b}_i(\boldsymbol{\xi})$ vectors.

In the following sections we will study the power of both tests using a general theoretical framework and via simulations. In the simulation studies, we will focus on exploring the power of both tools to detect misspecifications of the random-effects structure, that have been shown to distort the inferences and introduce bias in the point estimators.

3.3 Asymptotic Power of the Diagnostic Tools

The SET and the MIMT can be applied to detect different types of misspecifications. Therefore, and following the notation of Section 3, the hypotheses of interest can be written as: $H_0 : h \in \mathfrak{F}$ versus the alternative $H_1 : h \notin \mathfrak{F}$. In what follows we will give an expression for the asymptotic power of the SET, but the same ideas can be used to calculate the asymptotic power of the MIMT. If we denote the power of the test by $\beta(h)$ then, asymptotically

$$\beta(h) = P\left(\delta_s(n) > \chi^2_{\alpha,p} \mid h\right), \tag{14}$$

where $\chi^2_{\alpha,p}$ denotes the α percentile of a chi-square distribution with p degrees of freedom. Note that when $h \in \mathfrak{F}$ the model is correctly specified and (14) quantifies the Type I error rate with $\beta(h) \leq \alpha$. On the other hand, if $h \notin \mathfrak{F}$ then the model has been misspecified and (14) quantifies

the power to detect such a misspecification, when the real data generating mechanism is h. It can be shown that, in general, $\delta_s(n) \sim \chi^2_p(\lambda)$, where $\chi^2_p(\lambda)$ denotes a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter $\lambda = \boldsymbol{\mu}_v^T(\boldsymbol{\xi}^*)\boldsymbol{C}_v^{-1}(\boldsymbol{\xi}^*)\boldsymbol{\mu}_v(\boldsymbol{\xi}^*)$ with $\boldsymbol{\mu}_v(\boldsymbol{\xi}^*) = \Delta\text{vec}\left[\boldsymbol{V}(\boldsymbol{\xi}^*) + \boldsymbol{A}^{-1}(\boldsymbol{\xi}^*)\right]$. Using all the previous elements we can now rewrite (14) as

$$\beta(\lambda) = P\left(\chi^2_p(\lambda) > \chi^2_{\alpha,p} \mid \lambda\right). \tag{15}$$

From (15) it is easy to see that the power is an increasing function of λ and, as a consequence, increasing λ will redound in an increased power of the test to detect the misspecification. To enhance insight into this issue, let us assume that one element of $\boldsymbol{\mu}_v(\boldsymbol{\xi}^*)$ increases in absolute value while the others remain constant and let us denote the increasing component of $\boldsymbol{\mu}_v(\boldsymbol{\xi}^*)$ by μ_k^*. Note that under the null $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$ and $\boldsymbol{\mu}_v(\boldsymbol{\xi}_0) = \boldsymbol{0}$ and, therefore, increasing absolute values of μ_k^* will correspond with distributions in the alternative that are "moving" away from the null. Moreover, λ is a positive definite quadratic form in $\boldsymbol{\mu}_v(\boldsymbol{\xi}^*)$ and this implies that $\lim_{|\mu_k^*|\to+\infty} \lambda(\mu_k^*) = +\infty$. As a result, the power of the test will go to one as we move away from the null hypothesis. The same conclusion is valid if more than one component of $\boldsymbol{\mu}_v(\boldsymbol{\xi}^*)$ increase in absolute value.

It is important to point out that the test will fail to detect misspecifications which do not distort (6) and a similar statement also holds for the MIMT. However, this potential lack of power of the SET and the MIMT to detect certain type of misspecifications is not surprising. Indeed, it is well known that uniformly most powerful tests (UMP) are difficult to find, even in the simpler scenario where the null and the alternative hypotheses are just subsets of the real line. For instance, it has been proven that when a testing problem has nuisance parameters, the UMP tests do not generally exist (Nomakuchi, 1992). Additionally, it is also well known that UMP tests do not

generally exist when the alternative hypothesis is two-sided (Sen and Singer, 1993). Hence, it is not surprising that a given test can have low power to detect certain elements in the alternative, specially in complicated settings like the one considered in the present work.

3.4 Implementation

One major advantage of the diagnostic tools proposed in this manuscript is their easy implementation using standard software, like the SAS procedures NLMIXED and IML. Indeed, note first that the Hessian $\boldsymbol{A}_n(\widehat{\boldsymbol{\xi}}_n)$ follows directly from NLMIXED. The subject contributions to $\boldsymbol{B}_n(\widehat{\boldsymbol{\xi}}_n)$ can also be obtained in a relatively straightforward way, but they need some extra calculations. To compute these values, we need to fit the final model in NLMIXED to each subject separately, keeping the maximum likelihood estimates fixed by setting `maxiter=0`, and saving the corresponding first order derivatives. Macros that compute both tests can be obtained from the authors.

4 Simulation Study

Heagerty and Kurland (2001) studied the impact of random-effects misspecification in a number of different settings. They presented bias calculations for plausible violations of the random-effects assumptions, using a logistic-normal model when: (i) the random effect is generated from a non-normal distribution; (ii) the variance of the random effect depends on a covariate in the mean structure; (iii) the random structure includes both a random intercept and slope; and (iv) the random effects are auto-correlated. In the following, we will study the performance of the SET and the MIMT in detecting these model misspecifications.

4.1 Non-normal Random Effects

In this section, binary responses were generated using the logistic random-intercept model given by

$$\text{logit}\{P(y_{ij} = 1|b_{0i})\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_{0i}, \tag{16}$$

including an intercept, a binary covariate z_i taking values 0 and 1, a within-cluster covariate t_j taking values 0, 1, 2, 4, 6, and 8 and a random intercept b_{0i} sampled from 5 mean-zero distributions: the normal, power function, lognormal distribution, a discrete distribution with equal probability at two support points and an asymmetric mixture of two normal densities. For all the previous distributions variances $\sigma_b^2 = 4$ and 32 were considered. Note that $\sigma_b^2 = 32$ is of the same order of magnitude as the estimate obtained from the case study, whereas $\sigma_b^2 = 4$ is used to analyze the performance of the tests in less extreme scenarios.

The parameters in the mean structure were fixed at $\beta_0 = -8$, $\beta_1 = 2$ and $\beta_2 = 1$. Six different sample sizes were considered, namely 50, 100, 200, 350, 500, and 1000. For each setting, 500 data sets were generated and (16) was fitted to these data under the assumption of normally distributed random effects. We then determined the proportion of cases in which a significant result at the pre-specified significance level of 5% was detected, using the the MIMT $\Im_m(n)$ and the SET $\delta_s(n)$. When the random effects are generated from a normal distribution, this proportion corresponds to the Type I error rate; otherwise, it represents the power of the tests. The results of these simulations are shown in Table 2.

In this setting the SET encountered problems to detect the misspecification when the random intercept was generated from a power function or an asymmetric mixture of two normal distributions, especially with $\sigma_{0b}^2 = 4$. The test performed clearly better when the variance of the random

intercept was large. This is, however, a desirable behavior given the results obtained by Litière *et al.* (2008) and presented in Section 1.

The MIMT on the other hand clearly outperforms the SET. Values of power above 70% were obtained in most of the settings with sample sizes as of 350 subjects. In addition, the MIMT has a remarkable power for detecting random-effects misspecification when $\sigma^2_{0b} = 32$. Indeed, in this scenario, the test could always detect the misspecification for sample sizes greater or equal than 350.

In spite of these encouraging results, the Type I error rates shown in Table 2 are larger than the pre-specified 5% level for small and moderate sample sizes. This behavior has been well documented for the IMT (Taylor, 1987) and given the relationship between the IMT and the proposed SET and MIMT, the inflated Type I error rates observed for the latter two are not entirely unexpected.

We approached this problem following the idea introduced by Horowitz (1994) and used a parametric bootstrap to account for the small sample bias. To study the performance of this bootstrap version of the tests, a new simulation study was carried out. The details of these additional simulations were as follows: 250 data sets were generated using Model (16), considering the normal, the power function, the discrete, the asymmetric mixture and the lognormal distributions with variance $\sigma^2_b = 4$ for the random effect. The samples sizes were fixed at 50, 100 and 200 subjects. For each generated data set the maximum likelihood estimates $\widehat{\beta}_{0n}$, $\widehat{\beta}_{1n}$, $\widehat{\beta}_{2n}$ and $\widehat{\sigma}^2_{bn}$ were calculated. Further, based on these values the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ were computed. Afterwards, new responses y_{ij} were generated using Model (16) with the linear predictor $\widehat{\beta}_{0n} + \widehat{\beta}_{1n} z_i + \widehat{\beta}_{2n} t_j + b_i^B$. Note that z_i and t_j follow from the original (simulated) data set and

the random effects b_i^B were generated from a normal distribution with mean zero and variance $\widehat{\sigma}_{bn}^2$. This process was repeated 1000 times, such that in total for each simulated data set 1000 new bootstrap replicas were created. In a final step, these bootstrap data sets were analyzed using Model (16) and assuming normal random effects. Estimates for the SET and the MIMT were obtained from each of the bootstrap samples and used to estimate the empirical distribution function and corresponding bootstrap-based critical value for each test. Table 2 shows, between parenthesis, the Type I error rates and power obtained after comparison of the original SET and MIMT values with their respective bootstrap-based critical values.

Clearly, parametric bootstrap does importantly reduce the Type I error rates. This is a very promising result and this approach should certainly be further explored. Note that, even though these bootstrap versions of the tests are computationally demanding, the current power of personal computers make them affordable alternatives in practical situations.

4.2 Random-intercepts Variance Depending on a Binary Covariate

Following the approach by Heagerty and Kurland (2001), let binary responses be generated using the model

$$\text{logit}\{P(y_{ij}=1|b_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + \beta_3 z_i t_j + b_{ij}, \tag{17}$$

where z_i is a binary covariate defined as before, and t_j is a within-cluster covariate representing a linear trend, with $t_j = (j-1)/(n_i-1)$ and $n_i = 6$. The variance of the random intercept $b_{ij} = b_{i0}$ depends on the value of the binary covariate z_i, such that

$$b_{i0} \sim \begin{cases} \text{N}(0, \sigma_0^2) & \text{when} \quad z_i = 0, \\ \text{N}(0, \sigma_1^2) & \text{when} \quad z_i = 1. \end{cases} \tag{18}$$

The parameters in the linear predictor were fixed at $\beta_0^0 = -2$, $\beta_1^0 = 1$, $\beta_2^0 = 0.5$, and $\beta_3^0 = -0.25$. For each setting, 500 data sets were generated with $n = 500$ subjects, and analyzed using the

model given by (17), assuming that $b_{ij} = b_{i0} \sim \mathsf{N}(0, \sigma_b^2)$.

Heagerty and Kurland (2001) found that substantial bias can occur for all coefficients in the model, when σ_0 and σ_1 are very different. For example, they reported 38% and 31% of relative bias in the estimation of β_1 and β_3 respectively, when $\sigma_0 = 1$ and $\sigma_1 = 2$. Moreover, they observed that, as the discrepancy between the two parameters increases, so does the bias in the parameter estimates.

To study the performance of our proposals in this particular setting, we have applied the SET and the MIMT to the generated data sets and determined the proportion out of 500 replications in which the tests were able to detect the misspecification (at a 5% significance level). The corresponding powers are displayed in the first part of Table 3 as a function of σ_0 and σ_1.

First, observe that, when $\sigma_0 = \sigma_1$, this corresponds to the Type I error rate of the tests. With a sample size of 500 subjects, in these settings the Type I error rate is always maintained under the pre-specified 5% level. Additionally, both tests show a good power as of differences between the two variance parameters exceeding 1.0. Here again, the MIMT has a better performance than the SET in most of the settings. In general, we conclude that both tests are able to detect the misspecification, especially in those settings where Heagerty and Kurland (2001) reported that the maximum likelihood estimators of the linear predictors could be most affected.

4.3 Ignoring a Random Effect

Another type of misspecification in the random structure occurs when a random slope is incorrectly assumed to be fixed. To study the performance of our proposals in this setting, we have generated binary responses from the model given by (17), with $b_{ij} = b_{i0} + b_{i1}t_j$, and σ_0^2 and σ_1^2 representing

the variance of the random intercept b_{i0} and the random slope b_{i1}, respectively.

Simulations by Heagerty and Kurland (2001) showed that, when these data are analyzed wrongly assuming that $b_{ij} = b_{i0}$, moderate bias can appear in the estimation of the regression coefficients. For instance, they observed asymptotic relative biases as large as 30–50% in the estimates of β_2 and β_3 when σ_0 is small and σ_1 is large. On the other hand, the bias for the estimators of the intercept β_0 and the cluster-level covariate effect β_1 remained below 15% for all (σ_0, σ_1) pairs considered.

The second panel of Table 3 shows the power of the diagnostic tools to detect this type of misspecification, as a function of σ_0 and σ_1. The power of both tests was more moderate in this scenario. Indeed, as one would expect, both tests fail to detect the misspecification when σ_1 is small. However, the bias calculations by Heagerty and Kurland (2001) showed that little bias is present in this case. The SET and the MIMT increase their power when σ_1 is increased, relative to σ_0. Nevertheless, when $\sigma_1 = 1$ and $\sigma_0 = 0.5$, precisely the setting in which bias as large as 52% was obtained for β_2 and β_3, we only observed a power of 55% with the SET to detect the misspecification, and 61% with the MIMT. Even though, as we stated before, these results are milder than the ones observed for the other misspecification, these levels of power can still be relevant in many practical applications.

4.4 Auto-regressive Random Effects

In the analysis of longitudinal data, one often observes that the dependence between two repeated measurements within a cluster decays with their separation in time. This could be accounted for in a generalized linear mixed model by including autocorrelated random effects b_{ij} for which $\text{cov}(b_{ij}, b_{ik}) = \sigma^2 \rho^{|t_{ij} - t_{ik}|}$. Simulations by Heagerty and Kurland (2001) for this type of misspec-

ification have shown that substantial negative bias can arise in the estimated fixed effects, with increasing bias as σ increases, especially when ρ is small. Note that the random-intercept model follows as a special case of the auto-regressive model when $\rho = 1$. For models with $\rho < 1$, a potentially large negative bias can be observed in $\widehat{\sigma}_n$, given that it estimates the common variance and, therefore, it approximates the true covariances $\sigma^2\rho^{|t_{ij}-t_{ik}|}$. These authors observed that, as ρ decreases, the negative bias in $\widehat{\sigma}_n$ increases, ranging between -30% and -50% when $\rho = 0.7$, and between -47% and -70% when $\rho = 0.5$. As a result, substantial negative bias can also arise in the estimated regression coefficients, with increasing bias as σ increases. For instance, when $(\rho, \sigma) = (0.5, 3.0)$, negative bias as high as -45% occurred in each of the parameter estimates in the linear predictor.

The third panel in Table 3 shows the power of the diagnostic tools to detect this type of misspecification as a function of σ and ρ. From the table it follows that both tests have a high power to detect this type of misspecification when σ is sufficiently large. Given that bias in the estimation of the linear predictor parameters was seen to be more substantial as of $\sigma \geq 2$, this is a very desirable property. Note that, for smaller values of σ the power of both tests is considerably lower but in these settings the impact of the misspecification is also minor.

5 Revisiting the Case Study

In this section, we apply the SET and the MIMT to assess the suitability of Model (4) with normal random effects for the analysis of the case study. It follows that $\delta_s(n) = 8.5$ and compared to a χ^2 distribution with 6 degrees of freedom, this leads to $p = 0.205$. In contrast, $\Im_m(n) = 13.1$ with corresponding $p = 0.041$. According to the MIMT, the data at hand give evidence of some model misspecification. In comparison, using parametric bootstrap, the p-value associated with

the SET reduces to 0.097, whereas a $p = 0.040$ was observed for the MIMT.

Therefore, a more detailed investigation of the model assumptions is required. To study how sensitive are our results to the choice of the random-effects distribution, we performed a sensitivity analysis including an exponential, a chi-square, a uniform, and a lognormal distribution. Such analysis can be easily carried out using probability integral transformations in the SAS procedure NLMIXED. We further considered a mixture of two and three normal distributions (see Molenberghs and Verbeke, 2005, Chapter 23). The corresponding parameter estimates are shown in Table 1.

The uniform distribution produced the best results according to the AIC, with an AIC weight of $\omega_{AIC-uni} = 0.804$ (Burnham and Anderson 2002). It was followed by the models with the mixture of two and three normals, which have AIC weights $\omega_{AIC-two} = 0.147$ and $\omega_{AIC-three} = 0.033$ respectively. The GLMM with a normal random effect ranked fourth with an AIC weight $\omega_{AIC-one} = 0.016$. Careful exploration of these models showed also a better fit of the marginal evolutions when the random effect distribution was the uniform (see Figure 1(b)). Moreover, the diagnostic tools do not detect any further misspecification when this model is adopted ($\delta_s(n) = 2.1$ with $p = 0.911$, and $\Im_m(n) = 6.5$ with $p = 0.366$).

Interestingly, the inferential results for the treatment effect may differ depending on the choice of the random-effects distribution. Indeed, while γ_2 was found to be non-significant (at the prespecified 5% level) in the best fitting model, it changed to very significant ($p = 0.007$) when the distribution of the random effects was assumed to be a two-component mixture, i.e., the second best model. This example clearly illustrates that inferences for the linear predictor parameters can depend on the distributional assumptions for the random effects. Therefore, the evaluation

of these assumptions is of utmost importance.

Finally, observe that the variance of the random intercept was rather large in all scenarios considered, with a median around 24. Hence, all models consistently hint on a very strong within-subject association.

6 Discussion

In this manuscript, we proposed the so-called Sandwich Estimator Test and the Modified Information Matrix Test to detect misspecifications in generalized linear mixed models. In our simulations we mainly focussed on the detection of misspecifications in the random effect structure. In general both tests showed a good power, with values frequently above 70%, in most of the settings considered. However, in scenarios with small or moderate sample sizes, the Type I error rates were inflated, jeopardizing the interpretation of the results. Therefore, we recommend the use of these tools in their original form, only when the sample size is greater than 350 subjects. Importantly, parametric bootstrap versions of the tests were able to control the Type I error rates and they could be an attractive option when the sample size at hand is small or moderate ($n \leq 350$). A finite sample correction to solve this problem would also be worth having and future research in this direction will surely follow.

Interestingly, the SET was clearly outperformed by the MIMT in most of the settings. This could be partly attributed to the fact that, unlike the SET, the MIMT accounts for the extra variability introduced by replacing $\boldsymbol{A}(\boldsymbol{\xi}_0)$ with its consistent estimator $\boldsymbol{A}_n(\widehat{\boldsymbol{\xi}}_n)$. Additionally, note that whereas the MIMT depends on $\boldsymbol{A}_n(\widehat{\boldsymbol{\xi}}_n)$ and $\boldsymbol{B}_n(\widehat{\boldsymbol{\xi}}_n)$ directly, more rounding errors may have been introduced to obtain $\boldsymbol{V}_n(\widehat{\boldsymbol{\xi}}_n)$ and $\boldsymbol{A}_n^{-1}(\widehat{\boldsymbol{\xi}}_n)$. These may interfere with the power of the

SET to detect misspecifications and could help to explain the differences observed between the performance of the SET and the MIMT.

When constructing the SET and the MIMT we used only the diagonal elements of the matrices $\boldsymbol{V}(\boldsymbol{\xi}^*) + A^{-1}(\boldsymbol{\xi}^*)$ and $\boldsymbol{A}(\boldsymbol{\xi}^*) + B(\boldsymbol{\xi}^*)$. Some reasons justified this choice. First, using all the elements of the previous matrices frequently led to a singular variance covariance matrix for the test statistic. Second, including more elements outside the diagonal also provoked a huge upward bias in the Type I error rates, especially in small samples. It is important to point out, however, that our choice does not preclude the use of other elements outside the diagonal and both tests can be easily adapted to do that.

Finally, even though we focused on misspecification of the random-effects structure, a significant result from the SET and/or the MIMT does not necessarily imply that the random effects are misspecified. These diagnostic tools are, in principle, also suitable to detect other types of model misspecification, such as a misspecified link function or a misspecified mean structure. Nevertheless, given the good power observed in our simulations to detect misspecifications in the random effect structure, this would obviously be a first place to start.

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., 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.

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

Burnham, K. P., and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd Edition). New York: Springer-Verlag.

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

Horowitz, J.L. (1994) Bootstrap-based critical values for the information matrix test. *Journal of Econometrics*, **61**, 394–411.

Litière, S., Alonso, A., and Molenberghs, G. (2007). Type I and Type II error random-effects misspecification in generalized linear mixed models. *Biometrics* **63**, 1038–1044.

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

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

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*. Hoboken, NJ : John Wiley & Sons.

Nomakuchi, K. (1992). A note on the uniformly most powerful tests in the presence of nuisance parameters. *Annals of the Institute of Statistical Mathematics* **44**, 141–145.

Sen, P.K., and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction With Applications*. New York: Chapman & Hall.

Taylor, L.W. (1987) The size bias of White's information matrix test. *Economics Letters* **24**, 63–67.

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

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.

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.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*, **32**, 511–526.

Table 1: *Parameter estimates and standard errors using the logistic random-intercept Model (4) with the random effect (RE) assumed to follow a normal distribution (GLMM), a chi-square, an exponential, a uniform, a lognormal, and a mixture of two and three normals.*

Model	$\widehat{\gamma}_0$ (s.e.)	$\widehat{\gamma}_1$ (s.e.)	$\widehat{\gamma}_2$ (s.e.)	$\widehat{\beta}_1$ (s.e.)	$\widehat{\beta}_4$ (s.e.)	$\widehat{\sigma}_b^2$ (s.e.)	AIC
1 GLMM	-6.79 (0.52)	1.41 (0.39)	0.73^{+} (0.44)	1.03 (0.18)	-0.06 (0.02)	24.1 (4.0)	1649.7
2 RE, uni.	-6.87 (0.58)	1.09 (0.35)	$0.70^{\dagger}$ (0.44)	1.00 (0.18)	-0.05 (0.02)	22.5 (3.7)	1641.9
3 RE, χ^2	-6.37 (0.50)	1.57 (0.40)	$0.70^{\dagger}$ (0.44)	1.06 (0.18)	-0.06 (0.02)	20.8 (3.7)	1658.4
4 RE, exp.	-5.86 (0.45)	1.76 (0.41)	$0.55^{\dagger}$ (0.41)	1.08 (0.18)	-0.06 (0.02)	19.3 (3.4)	1673.4
5 RE, logn.	-4.36 (0.49)	2.07 (0.42)	$0.29^{\dagger}$ (0.31)	1.06 (0.18)	-0.07 (0.02)	213 (115)	1735.3
6 Mixture, $k = 2$	-7.18 (2.00)	1.01 (0.33)	0.99^{*} (0.37)	0.94 (0.20)	-0.05 (0.02)	26.4 (10.8)	1645.3
7 Mixture, $k = 3$	-7.80 (40.7)	1.17 (0.35)	0.71^{+} (0.40)	0.97 (0.20)	-0.05 (0.02)	34.9 (513)	1648.3

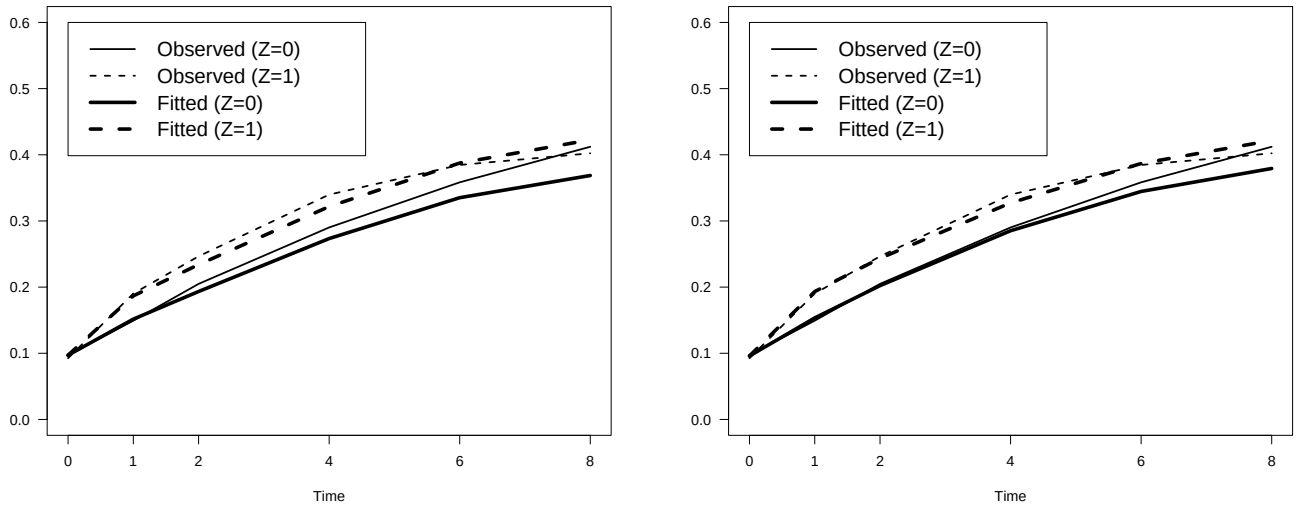
* $p = 0.007$, $^{+}$ $0.05 < p < 0.10$, † $p > 0.10$

Table 2: *Power and Type I error for detecting a misspecified random-effects distribution, at the 5% significance level, using the the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ in generalized linear mixed models: a normal random intercept is assumed, whereas the random effects are generated from a normal, a power function, a discrete, an asymmetric mixture of two normals or a lognormal distribution, with variance $\sigma^2_{0b} = 4$ or 32. For those settings for which a bootstrap correction was performed, corrected results are added inbetween parenthesis.*

	$\sigma^2_{0b} = 4$		$\sigma^2_{0b} = 32$		$\sigma^2_{0b} = 4$		$\sigma^2_{0b} = 32$	
n	$\delta_s(n)$	$\Im_m(n)$	$\delta_s(n)$	$\Im_m(n)$	$\delta_s(n)$	$\Im_m(n)$	$\delta_s(n)$	$\Im_m(n)$
	Normal				Power function			
50	0.278 (0.048)	0.259 (0.004)	0.234	0.154	0.242 (0.016)	0.589 (0.069)	0.264	0.762
100	0.192 (0.044)	0.242 (0.060)	0.122	0.102	0.172 (0.044)	0.515 (0.128)	0.288	0.922
200	0.098 (0.032)	0.180 (0.088)	0.072	0.072	0.092 (0.044)	0.620 (0.344)	0.406	0.996
350	0.046	0.108	0.038	0.056	0.086	0.710	0.650	1.000
500	0.044	0.080	0.040	0.048	0.096	0.820	0.812	1.000
1000	0.026	0.054	0.016	0.032	0.164	0.952	0.998	1.000
	Discrete				Asymmetric mixture			
50	0.304 (0.024)	0.782 (0.224)	0.710	0.982	0.230 (0.032)	0.471 (0.036)	0.242	0.788
100	0.306 (0.056)	0.914 (0.584)	0.830	0.998	0.146 (0.040)	0.374 (0.120)	0.172	0.926
200	0.430 (0.308)	0.968 (0.860)	0.948	1.000	0.092 (0.076)	0.458 (0.212)	0.114	0.996
350	0.688	0.988	0.996	1.000	0.070	0.514	0.136	1.000
500	0.848	1.000	1.000	1.000	0.064	0.588	0.098	1.000
1000	0.986	1.000	1.000	1.000	0.058	0.784	0.166	1.000
	Lognormal							
50	0.355 (0.048)	0.446 (0.024)	0.480	0.547				
100	0.251 (0.092)	0.545 (0.076)	0.338	0.772				
200	0.198 (0.084)	0.792 (0.288)	0.430	0.984				
350	0.224	0.964	0.724	1.000				
500	0.332	0.996	0.892	1.000				
1000	0.780	1.000	1.000	1.000				

Table 3: *Power of the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, at the 5% significance level, when a logistic-normal random-intercept model is assumed, but (i) the variance of the random intercept depends on a binary cluster-level covariate, $[b_{i0}|X_{i,1} = 0] \sim N(0, \sigma_0^2)$ and $[b_{i0}|X_{i,1} = 1] \sim N(0, \sigma_1^2)$; (ii) the data are generated using both a random intercept and slope $(b_{ij} = b_{i0} + b_{i1}x_j)$, with variance σ_0^2 and σ_1^2, respectively; and (iii) the data are generated using autocorrelated random effects b_{ij} such that cov$(b_{ij}, b_{ik}) = \sigma^2\rho^{|t_{ij}-t_{ik}|}$.*

(i)				(ii)				(iii)			
σ_1	σ_0	$\delta_s(n)$	$\Im_m(n)$	σ_1	σ_0	$\delta_s(n)$	$\Im_m(n)$	ρ	σ	$\delta_s(n)$	$\Im_m(n)$
0.5	0.5	0.040	0.064	0.2	0.5	0.030	0.066	0.5	0.5	0.013	0.054
	1.0	0.102	0.514		1.0	0.012	0.032		1.0	0.046	0.064
	2.0	0.984	1.000		2.0	0.008	0.018		2.0	0.696	0.264
	3.0	1.000	1.000		3.0	0.022	0.036		3.0	0.980	0.594
1.0	0.5	0.078	0.696	0.5	0.5	0.052	0.098	0.7	0.5	0.019	0.071
	1.0	0.018	0.038		1.0	0.026	0.056		1.0	0.044	0.056
	2.0	0.620	0.980		2.0	0.006	0.024		2.0	0.888	0.544
	3.0	0.994	1.000		3.0	0.013	0.032		3.0	1.000	0.958
2.0	0.5	0.770	1.000	0.8	0.5	0.230	0.282	0.9	0.5	0.024	0.064
	1.0	0.452	0.980		1.0	0.070	0.156		1.0	0.018	0.034
	2.0	0.020	0.020		2.0	0.022	0.076		2.0	0.368	0.306
	3.0	0.244	0.608		3.0	0.014	0.034		3.0	0.922	0.924
3.0	0.5	0.992	1.000	1.0	0.5	0.546	0.610				
	1.0	0.974	1.000		1.0	0.234	0.394				
	2.0	0.184	0.630		2.0	0.054	0.166				
	3.0	0.016	0.014		3.0	0.044	0.012				



(a) Normal random effects

(b) Uniform random effects

Figure 1: *Evolution of the observed and fitted (using Model (4) with normal and uniform random effects) probabilities to be classified as a normal to mildly ill patient by treatment group, denoted by* Z.