



**Manchester  
Metropolitan  
University**

---

Drikvandi, R (2017) Nonlinear mixed-effects models for pharmacokinetic data analysis: assessment of the random-effects distribution. *Journal of Pharmacokinetics and Pharmacodynamics*, 44 (3). pp. 223-232. ISSN 1567-567X

---

**Downloaded from:** <https://e-space.mmu.ac.uk/621854/>

**Publisher:** Springer Verlag

**DOI:** <https://doi.org/10.1007/s10928-017-9510-8>

Please cite the published version

<https://e-space.mmu.ac.uk>

# Nonlinear mixed-effects models for pharmacokinetic data analysis: assessment of the random-effects distribution

Reza Drikvandi

Received: date / Accepted: date

**Abstract** Nonlinear mixed-effects models are frequently used for pharmacokinetic data analysis, and they account for inter-subject variability in pharmacokinetic parameters by incorporating subject-specific random effects into the model. The random effects are often assumed to follow a (multivariate) normal distribution. However, many articles have shown that misspecifying the random-effects distribution can introduce bias in the estimates of parameters and affect inferences about the random effects themselves, such as estimation of the inter-subject variability. Because random effects are unobservable latent variables, it is difficult to assess their distribution. In a recent paper we developed a diagnostic tool based on the so-called gradient function to assess the random-effects distribution in mixed models. There we evaluated the gradient function for generalized linear mixed models only and in the presence of a single random effect. However, assessing the random-effects distribution in nonlinear mixed-effects models is more challenging, especially when multiple random effects are present, and therefore the results from linear and generalized linear mixed models may not be valid for such nonlinear models. In this paper, we further investigate the gradient function and evaluate its performance for such nonlinear mixed-effects models which are common in pharmacokinetics and pharmacodynamics. We use simulations as well as real data from an intensive pharmacokinetic study to illustrate the proposed diagnostic test.

**Keywords** Diagnostic test · Gradient function · Nonlinear mixed-effects models · Pharmacokinetic data · Random effects

## Introduction

Nonlinear mixed-effects models are being widely used in pharmacokinetics to study the pharmacological processes within the body after administration of a drug in order to characterize drug disposition (absorption, distribution, metabolism, and excretion). The term “Mixed-effects” refers to the presence of both fixed effects and random effects in the model. Fixed effects are regression parameters associated with covariates, while random effects are subject-specific random quantities incorporated to capture the inter-subject variability.

In pharmacokinetics, it is important to understand the inter-subject variation in pharmacokinetic parameters and its association with subject characteristics, which could provide useful information, for example, in developing dosing guidelines. Nonlinear mixed-effects models account for the inter-subject variability by adding subject-specific random effects to the model. The random effects are latent and hence unobservable variables that follow a distribution, which is unknown. Note that the random effects also capture the potential correlation between repeated measurements on the same subject, which must not be ignored in the analysis.

To fit a nonlinear mixed-effects model, one often needs to assume a distribution for the random effects. Inferences are then based on the marginal likelihood function after integrating out the random effects over their assumed distribution. It is common to assume that the random effects follow a (multivariate) normal distribution. The normality assumption makes computa-

---

Reza Drikvandi  
Statistics Section, Department of Mathematics, Imperial College London, London, UK  
E-mail: r.drikvandi@imperial.ac.uk

tion of the marginal likelihood more feasible. However, it is well understood that misspecifying the random-effects distribution can introduce bias in the estimates of fixed-effects parameters (see, for example, [11, 1, 16, 2, 9]) and it also affects the operating characteristics of hypothesis tests (see, for example, [10, 15]). More importantly, inferences about the random effects themselves, such as estimation of the inter-subject variability, 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 [23]. Therefore, to obtain reliable inference it is important to check the appropriateness of the assumed random-effects distribution.

In the literature, a number of diagnostic tools have been suggested for assessing the random-effects distribution in linear mixed models [14, 13, 5] and also in generalized linear mixed models [20, 18, 25, 22, 3, 12]. However, to the best of our knowledge, there is only one diagnostic tool available for checking the random-effects distribution in nonlinear mixed-effects models. We developed this diagnostic test recently [8], which is based on the so-called gradient function introduced in [24].

In our paper [8] we showed, via simulations, that the diagnostic test based on the gradient function performs well in detecting misspecification of the random-effects distribution. However, the performance of this test was evaluated in the case of generalized linear mixed models only and in the presence of a single random effect. It is not known how the diagnostic test performs in nonlinear mixed-effects models, especially with multiple random effects, which are common in pharmacokinetics and pharmacodynamics. Obviously, nonlinear mixed-effects models are more complex than linear and generalized linear mixed models, and moreover, the presence of multiple random effects makes assessment of the random-effects distribution much more difficult. Therefore, the results from linear and generalized linear mixed models may not be valid for such nonlinear models. In this paper, we aim to investigate and evaluate the performance of the diagnostic test based on the gradient function in such nonlinear mixed-effects models. We use simulations and real data from an intensive pharmacokinetic study.

## Methods

### Nonlinear mixed-effects models

In this subsection, we briefly explain the general form of nonlinear mixed-effects models for pharmacokinetic

data analysis. Consider a pharmacokinetic study in which  $N$  subjects are followed over time after administration of a drug. Let  $Y_{i1}, \dots, Y_{in_i}$  be  $n_i$  repeated measurements on the  $i$ th subject, where  $Y_{ij}$  is the outcome for subject  $i$  measured at time  $t_{ij}$ . For example,  $Y_{ij}$  could be the blood sample drawn after administration of the anti-asthmatic agent theophylline. Also, let  $U_i$  denote a vector of covariates representing conditions under which measurements on subject  $i$  are observed. For example, when each subject receives a single oral dose, say  $d_i$ , then  $U_i = d_i$ . Finally, assume that  $A_i$  is a vector of characteristics of subject  $i$  that do not change during the study. For example,  $A_i$  could include age, gender, height, ethnicity, and smoking status. Now, the nonlinear mixed-effects model can be expressed as a two-stage hierarchy as follows: (See [7])

Stage 1: Individual-Level Model

$$Y_{ij} = m(t_{ij}, U_i, \theta_i) + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad (1)$$

Stage 2: Population Model

$$\theta_i = d(A_i, \beta, b_i), \quad i = 1, \dots, N.$$

In Stage 1,  $m$  is a “known” nonlinear function of time which also depends on the subject conditions  $U_i$  and an  $r \times 1$  vector of pharmacokinetic parameters  $\theta_i$ , specific to subject  $i$ . For example,  $m$  could be a one-compartment model as in (5), where  $\theta_i = (k_{ai}, k_{ei}, Cl_i)'$  and  $U_i = d_i$ . Also,  $\varepsilon_{ij}$ 's are independent measurements errors, each of which has a normal distribution with mean 0 and variance  $\sigma^2$ . In Stage 2,  $d$  is an  $r$ -dimensional function that describes the relationship between the elements of  $\theta_i$  and between-subject covariates  $A_i$  in terms of a  $p \times 1$  fixed parameter  $\beta$  whose elements are referred to as fixed effects, and a  $q \times 1$  vector  $b_i$  of random effects representing the inter-subject variability. For instance,  $d$  could be the three-dimensional function in (6). Note that because the pharmacokinetic parameters can vary from subject to subject, the random effects  $b_i$  are incorporated into the model to capture the inter-subject variability through a hierarchical analysis. The random effects  $b_i$ , which are unobservable variables with an unknown distribution, are assumed to have a zero mean and a covariance matrix  $D$  whose elements are known as variance components.

Let  $\gamma = (\beta, \sigma^2, D)'$  represent all unknown parameters in model (1). The estimates of parameters can then be obtained using the maximum likelihood estimation method. The maximum likelihood approach requires some assumption on the distribution of random effects. For this, assume that the random effects  $b_i$  follow a specific distribution, which is denoted by  $G$ . A typical choice for  $G$  is the multivariate normal distribution, i.e.,  $b_i \sim N(0, D)$ . The log-normal distribution

is also used in pharmacokinetics. Denoting the conditional distribution of the outcome  $Y_{ij}$  given the random effects  $b_i$  by  $f_i(y_{ij}|b_i)$ , one can write the marginal log-likelihood function for model (1) as follows

$$l(\gamma) = \ln \prod_{i=1}^N \int_{R^q} \left[ \prod_{j=1}^{n_i} f_i(y_{ij}|b_i) \right] dG(b_i). \quad (2)$$

The integral in marginal log-likelihood (2) may not be solved analytically, and it requires numerical approximation. We use Gauss-Hermite quadrature (see, e.g., [19]), which can provide reliable approximations. The integral approximation and the maximisation of the log-likelihood function (2) with respect to  $\gamma$  can be done simultaneously in a standard software package like NONMEM, SAS or R. In this paper, we use PROC NLMIXED in SAS to obtain the maximum likelihood estimates of parameters.

Obviously, the random-effects distribution  $G$  is crucial in the calculation of the log-likelihood function (2) and, as discussed in the introduction, a misspecified random-effects distribution can lead to biased parameter estimates and invalid inferences. It is therefore important to check whether or not the assumed random-effects distribution is correctly specified. It should be pointed out that since random effects are unobservable variables, assessing their distribution is difficult.

#### A diagnostic tool for the random-effects distribution

As discussed in the previous subsection, to obtain the maximum likelihood estimates of parameters in the non-linear mixed-effects model (1), one needs to assume a distribution for the random effects  $b_i$ . In this subsection, we describe a diagnostic tool to verify whether or not an assumed random-effects distribution  $G$  is correctly specified.

Let  $Y_i = (Y_{i1}, \dots, Y_{in_i})'$  be the vector of repeated measurements on subject  $i$ ,  $i = 1, \dots, N$ . To check the adequacy of an assumed random-effects distribution  $G$ , [24] suggested to use the so-called gradient function given by

$$\Delta(G, b) = \frac{1}{N} \sum_{i=1}^N \frac{f_i(y_i|b)}{f_i(y_i|G)}, \quad b \in R^q, \quad (3)$$

where  $f_i(y_i|b)$  and  $f_i(y_i|G)$  are, respectively, the conditional (given random-effect point  $b$ ) and marginal distributions of  $Y_i$ . 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 subject  $i$  if the corresponding random effect  $b_i$  equals  $b$  rather than being sampled from the distribution  $G$ . Note that

calculation of the gradient function is easy because it only requires the calculation of the marginal and conditional distributions of all  $N$  subjects. [24] showed that if the random-effects distribution  $G$  is correctly specified, then  $\Delta(G, b) \leq 1$  for all  $b \in R^q$ , and moreover  $\Delta(G, b) = 1$  for all  $b$  in the support of  $G$ . Hence, deviations of the gradient function from 1 in the support points of  $G$  indicate inadequacy of  $G$  and 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  $G$ , and if the gradient plot is close to 1 then the adequacy of  $G$  is confirmed. Despite the simplicity of this graphical approach, it is not clear how misspecification can be distinguished from random variability by such a plot, especially when  $b_i$  is not a scalar which requires evaluating a three or higher dimensional plot of gradient function.

Based on the gradient function (3), we recently developed a formal diagnostic test for the random-effects distribution (see [8]). Let the null hypothesis  $H_0$  says the random-effects distribution  $G$  is correctly specified and the alternative hypothesis  $H_a$  says otherwise. Having considered all deviations of the gradient function from 1, we constructed a test statistic for testing  $H_0$  versus  $H_a$  as follows

$$T = \int_{R^q} (\hat{\Delta}(\hat{G}, b) - 1)^2 d\hat{G}(b), \quad (4)$$

where  $\hat{G}$  is the estimated random-effects distribution obtained by replacing the covariance matrix  $D$  by its maximum likelihood estimate, and  $\hat{\Delta}$  denotes the estimated gradient function based on  $\hat{G}$  obtained simply by replacing the unknown parameters in  $f_i(y_i|b)$  and  $f_i(y_i|G)$  by their maximum likelihood estimates. If  $T$  deviates much from 0, one can reject  $H_0$  indicating that the assumed random-effects distribution  $G$  is not adequate for random effects. The test statistic (4) considers a weight for each deviation of the gradient function from 1 for each random-effect point  $b$ . The weight is the estimated probability mass in point  $b$ . Note that  $T$  can be calculated using Monte Carlo integration.

The asymptotic distribution of  $T$  is given in Theorem 1 of [8], which is actually the distribution of weighted sum of independent chi-squared random variables each with one degree of freedom. The weights are eigenvalues of the square matrix  $A'QA$  (specified in Theorem 1 of [8]), which can be calculated using a software package.

The asymptotic distribution of  $T$  should be used when the sample size  $N$  is sufficiently large. For small-sample situations, we propose a parametric bootstrap procedure to obtain the finite-sample distribution of the test statistic  $T$  in (4). The key step in our bootstrap

procedure, in order to obtain a bootstrap sample, is to first generate random effects  $b_i^s$ ,  $i = 1, \dots, N$ , from  $\hat{G}$  and then generate a bootstrap sample  $Y_i^s$ ,  $i = 1, \dots, N$ , from  $\hat{f}_i(y_i|b_i^s)$ . We use 200 bootstrap samples to conduct the bootstrap test. Below we illustrate how the asymptotic and bootstrap tests based on (4) can be performed.

#### Implementation of the asymptotic test

The asymptotic test can be carried out by the following steps:

1. Generate  $K$ , say 1000, random-effect points  $b_k$  from  $\hat{G}$ .
2. Compute the gradient function (3) and its squared deviation from 1 for each  $b_k$ .
3. Calculate the test statistic  $T$  being the average of the  $K$  squared deviations obtained in step 2 (which is a Monte Carlo approximation of  $T$ ).
4. If  $T$  is greater than the 95% quantile of the asymptotic distribution (mentioned above), then reject  $H_0$ , indicating that the assumed random-effects distribution  $G$  is not appropriate for random effects. Otherwise,  $G$  is correctly specified.

#### Implementation of the bootstrap test

The bootstrap test can be carried out by the following steps:

1. Generate  $K$ , say 1000, random-effect points  $b_k$  from  $\hat{G}$ .
2. Compute the gradient function (3) and its squared deviation from 1 for each  $b_k$ .
3. Calculate the test statistic  $T$  being the average of the  $K$  squared deviations obtained in step 2 (which is a Monte Carlo approximation of  $T$ ), and denote it by  $T_{obs}$ .
4. For each bootstrap step  $s$ ,  $s = 1, \dots, 200$ , repeat the following two steps:
  - i. First generate random effects  $b_i^s$ ,  $i = 1, \dots, N$ , from  $\hat{G}$  and then generate a bootstrap sample  $Y_i^s$ ,  $i = 1, \dots, N$ , from  $\hat{f}_i(y_i|b_i^s)$ .
  - ii. Calculate the test statistic  $T$  for the bootstrap sample obtained in step i and denote it by  $T^s$ .
5. If the proportion of  $T^s$  exceeding  $T_{obs}$  is less than 0.05, then reject  $H_0$ , indicating that the assumed random-effects distribution  $G$  is not appropriate for random effects. Otherwise,  $G$  is correctly specified.

The above steps for the asymptotic and bootstrap tests can be coded in a standard software package. A SAS code is available from the author upon request.

In [8] we showed, via simulations, that both the asymptotic and bootstrap tests perform well in detecting misspecification of the random-effects distribution in the case of generalized linear mixed models with one random effect. However, it is not known how the tests perform in nonlinear mixed-effects models, especially in the presence of multiple random effects, as in (1). In those cases, as discussed in the introduction, assessing the random-effects distribution is more challenging. In the next section, we use simulations and real data to investigate the performance of both the asymptotic and bootstrap tests for assessing the random-effects distribution in the nonlinear mixed-effects model (1).

## Results

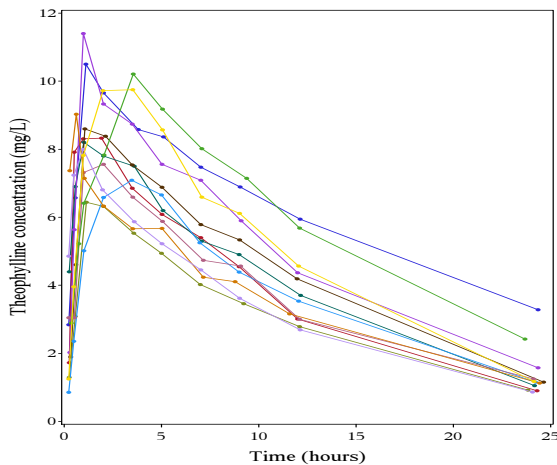
### Real data: theophylline data

Theophylline is a well-known anti-asthmatic agent, administered orally [4,6,7]. In an intensive pharmacokinetic study, 12 subjects were given the same oral dose (mg/kg) of theophylline, and blood samples were taken at several times following administration were assayed for theophylline concentration [7]. The individual profiles, presented in Figure 1, show that the theophylline concentrations have a similar shape for all subjects, but peak concentration achieved, rise, and decay vary significantly across the subjects. These differences are due to the inter-subject variation in the underlying pharmacokinetic processes, understanding of which is critical for developing dosing guidelines. To characterize these processes formally, we consider the following one-compartment model with first-order absorption and elimination: (see also [7])

$$C_i(t_{ij}) = \frac{d_i k_{ai} k_{ei}}{CL_i(k_{ai} - k_{ei})} \left[ \exp(-k_{ei} t_{ij}) - \exp(-k_{ai} t_{ij}) \right] + \varepsilon_{ij}, \quad (5)$$

where  $C_i(t_{ij})$  is the observed theophylline concentration on subject  $i$  at time  $t_{ij}$ ,  $d_i$  is the dose administered to subject  $i$ ,  $k_{ai}$  is the fractional absorption rate constant for subject  $i$ ,  $k_{ei}$  is the fractional elimination rate constant for subject  $i$ , and  $CL_i$  is the clearance for subject  $i$ . Note that the one-compartment model (5) can equivalently be written based on the volume of distribution ( $V$ ) by using the relationship  $CL = k_e V$  (see [17] p. 354).

From Figure 1, the pharmacokinetic parameters vary from subject to subject, therefore we include subject-specific random effects in the pharmacokinetic param-



**Fig. 1** Theophylline data: individual profiles for 12 subjects.

eters as follows

$$\begin{aligned} CL_i &= \exp(\beta_1 + b_{i1}), \\ k_{ai} &= \exp(\beta_2 + b_{i2}), \\ k_{ei} &= \exp(\beta_3 + b_{i3}), \end{aligned} \quad (6)$$

in which the vector of random effects  $b_i = (b_{i1}, b_{i2}, b_{i3})'$  is assumed to have mean 0 and covariance matrix  $D$  given by

$$\text{Cov}(b_i) = D = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{12} & d_{22} & d_{23} \\ d_{13} & d_{23} & d_{33} \end{bmatrix}.$$

Assuming that  $b_i$  follows a multivariate normal distribution, we obtained the maximum likelihood estimates of parameters (fixed-effect parameters, residual variance, and variance components of random effects) in the one-compartment model (5-6) along with their associated standard errors. The results are reported in Table 1. Using the estimates of variance components in this table, we computed the estimated correlation matrix of  $b_i$  as follows

$$\widehat{\text{Corr}}(b_i) = \begin{bmatrix} 1.000 & -0.098 & 0.672 \\ -0.098 & 1.000 & -0.261 \\ 0.672 & -0.261 & 1.000 \end{bmatrix},$$

which indicates that the clearance is highly correlated with the fractional elimination ( $\text{Corr}(b_{1i}, b_{i3}) = 0.672$ ) while it has a very small correlation with the fractional absorption ( $\text{Corr}(b_{1i}, b_{i2}) = -0.098$ ). Hence, it might be reasonable to assume the clearance and the fractional elimination are described by the same random effect, rather than two (highly) correlated ones. Considering the relationship  $CL = k_e V$ , we might let  $b_{i3} = \lambda b_{i1}$ ,

**Table 1** Theophylline data: the maximum likelihood estimates of parameters (fixed-effect parameters, residual variance, and variance components of random effects) in the one-compartment model (5) with the random effects  $b_i = (b_{i1}, b_{i2}, b_{i3})'$  given in (6), along with the associated standard errors.

Parameter	Estimate (s.e.)
Fixed effects:	
$\beta_1$	-3.277 (0.046)
$\beta_2$	0.537 (0.063)
$\beta_3$	-2.454 (0.064)
Residual variance:	
$\sigma^2$	0.624 (0.083)
Variance components of $b_i$ :	
$d_{11}$	0.057 (0.022)
$d_{12}$	-0.012 (0.018)
$d_{22}$	0.264 (0.054)
$d_{13}$	0.030 (0.020)
$d_{23}$	-0.025 (0.017)
$d_{33}$	0.035 (0.017)
-2 log-likelihood	341.7

where  $\lambda$  is a fixed variance inflation parameter to be estimated. Then, (6) can be simplified as

$$\begin{aligned} CL_i &= \exp(\beta_1 + b_{i1}), \\ k_{ai} &= \exp(\beta_2 + b_{i2}), \\ k_{ei} &= \exp(\beta_3 + \lambda b_{i1}). \end{aligned} \quad (7)$$

Now, we fit the one-compartment model (5) with the two random effects  $b_{i1}$  and  $b_{i2}$  given in (7) to the theophylline data. Again it is needed to assume a distribution for the random effects, and here a bivariate normal distribution is a handy choice. The parameter estimates under the bivariate normality assumption of  $b_{i1}$  and  $b_{i2}$  are calculated and reported in Table 2. The results suggest that (7) produces a better fit than (6) in terms of log-likelihood.

To verify the appropriateness of the bivariate normal distribution for the random effects  $b_{i1}$  and  $b_{i2}$ , we would apply the diagnostic test (4). Since the performance of this diagnostic test has not yet been examined for nonlinear mixed-effects models and subsequently for one-compartment models, we first conduct a simulation study to see whether the diagnostic test (4) has a good power in detecting misspecification of the random-effects distribution in such nonlinear models.

## Simulations

In this subsection, we conduct a simulation study in accordance with the one-compartment model (5) used for the theophylline data. For each sample size  $N =$

**Table 2** Theophylline data: the maximum likelihood estimates of parameters (fixed-effect parameters, residual variance, variance components of random effects, and variance inflation parameter) in one-compartment model (5) with the random effects  $b_i = (b_{i1}, b_{i2})'$  specified in (7), along with the associated standard errors.

Parameter	Estimate (s.e.)
Fixed effects:	
$\beta_1$	-3.216 (0.081)
$\beta_2$	0.464 (0.199)
$\beta_3$	-2.439 (0.064)
Residual variance:	
$\sigma^2$	0.516 (0.075)
Variance components of $b_i$ :	
$d_{11}$	0.063 (0.031)
$d_{12}$	-0.023 (0.053)
$d_{22}$	0.435 (0.202)
Variance inflation parameter:	
$\lambda$	0.515 (0.132)
-2 log-likelihood	333.3

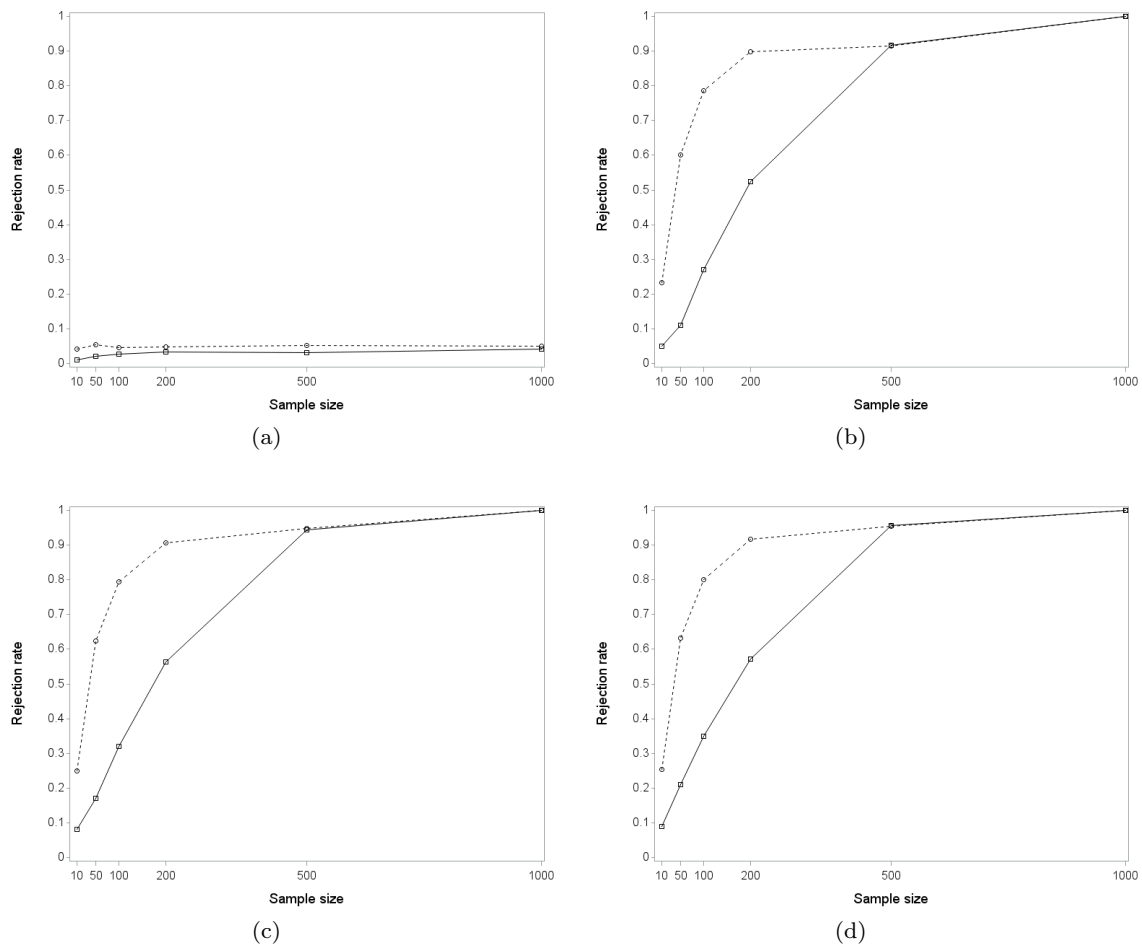
10, 50, 100, 200, 500, 1000 and with 10 repeated measurements per subject, we generated 1000 data sets from the one-compartment model (5) with the pharmacokinetic parameters given in (7). In simulations, the fixed-effect parameters were fixed at  $\beta_1 = -3.2$ ,  $\beta_2 = 0.5$ ,  $\beta_3 = -2.4$ , in accordance with the estimates in Table 2. Also, we set  $\lambda = 0.5$  and  $\sigma^2 = 0.5$ . For simplicity in the simulations, we here assumed that the two random effects  $b_{i1}$  and  $b_{i2}$  are independent (in Table 2 the estimate of  $\text{Cov}(b_{i1}, b_{i2})$  is very close to 0). Note that this assumption may not be realistic in practice and we would not make this assumption in our analysis of the theophylline data. We generated each of the random effects  $b_{i1}$  and  $b_{i2}$  from four distinct distributions:  $N(0, 1)$ , Chi-squared(2), Log-normal(3, 1), and  $F(1, 7)$ . All the generated random effects were shifted and rescaled such that  $b_{i1}$  and  $b_{i2}$  have both zero mean, but with different variances equal to 0.1 and 0.5 respectively, in accordance with the estimates of their variances in Table 2.

For each simulation setting, we first fitted the one-compartment model (5)-(7) to each of the generated data sets under a bivariate normality assumption of  $b_{i1}$  and  $b_{i2}$ , and then for each fitted model we carried out both the asymptotic and bootstrap tests using the algorithms on page 4. We computed the rejection rates of the asymptotic and bootstrap tests. When the true random-effects distribution was normal, the rejection rate would actually correspond to the Type I error rate, otherwise it represents the power of the test to detect misspecification.

The rejection rates (powers) of the asymptotic test and the bootstrap test are shown in Figure 2 with solid and dashed lines, respectively. Figure 2(a) indicates that the asymptotic test shows a Type I error rate smaller than the nominal level 0.05 while it gets closer to the nominal level when  $N$  increases. Thus, the asymptotic test is conservative in terms of Type I error when the sample size is not very large. The results in Figure 2(b,c,d) show that the power of the asymptotic test is not high when the sample size  $N$  is small, but it increases with the sample size. On the other hand, the results suggest that the bootstrap test has the correct Type I error rate with a high power and it is more powerful than the asymptotic test when the sample size  $N$  is small. Based on the simulation results, we conclude that a sample size of at least 200 is required to apply the asymptotic test to the one-compartment model (5)-(7) fitted to the theophylline data. Therefore, we should use the bootstrap test for the theophylline data.

In the above simulations, we used  $n_i = 10$  because the theophylline data contain 10 repeated measurements for each subject. However, in many other applications there might be a smaller number of repeated measurements per subject (i.e., sparse sampling). To examine the performance of the proposed diagnostic test in such situations, we repeated the previous simulation study but with  $n_i = 5$ . The simulation results, presented in Figure 3, suggest that both the asymptotic and bootstrap tests lose power when the number of repeated measurements per subject gets smaller. However, the power loss from 10 repeated measurements to 5 repeated measurements is not substantial especially when the number of subjects  $N$  is large.

We also ran simulations to see how the diagnostic test performs when the estimates of random effects are shrunk. For this, a good example is when random effects are generated from a mixture of normals, then the estimates of random effects obtained under the unimodal normality assumption are potentially shrunk toward the mean (see [21] and [23]). We repeated the previous simulation study but with random effects  $b_{i1}$  and  $b_{i2}$  each generated from  $\frac{1}{2}N(-2, 1) + \frac{1}{2}N(2, 1)$ . The simulation results for  $n_i = 10$  and  $n_i = 5$ , presented in Figure 4, show that the asymptotic and bootstrap tests perform reasonably well for the case when estimates of random effects are shrunk. It is probably because the proposed diagnostic method does not use the estimates of random effects, and instead it is based on the marginal likelihood after integrating out the random effects. Note that only the estimates of fixed-effects parameters and residual variance are used in our method just for calculation of the estimated gradient function.



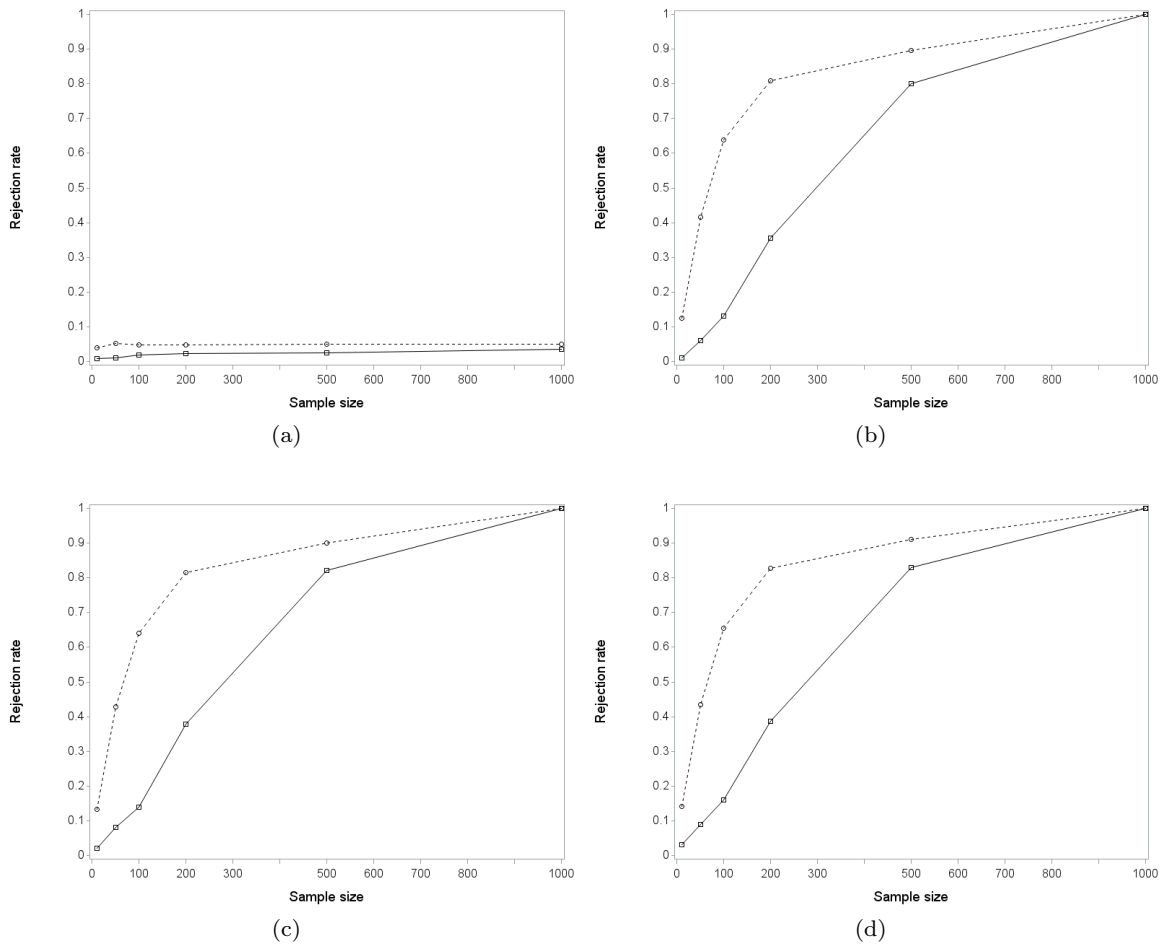
**Fig. 2** Rejection rates (powers) of the asymptotic test (solid line) and the bootstrap test (dashed line) at the significance level 0.05 with  $n_i = 10$  repeated measurements per subject and under four random-effects distributions: (a)  $N(0, 1)$ , (b) Chi-squared(2), (c) Log-normal(3, 1), and (d)  $F(1, 7)$ . Note that all the generated values of random effects were shifted and rescaled such that random effects  $b_{i1}$  and  $b_{i2}$  have both zero mean and variances equal to 0.1 and 0.5, respectively.

### Analysis of the theophylline data

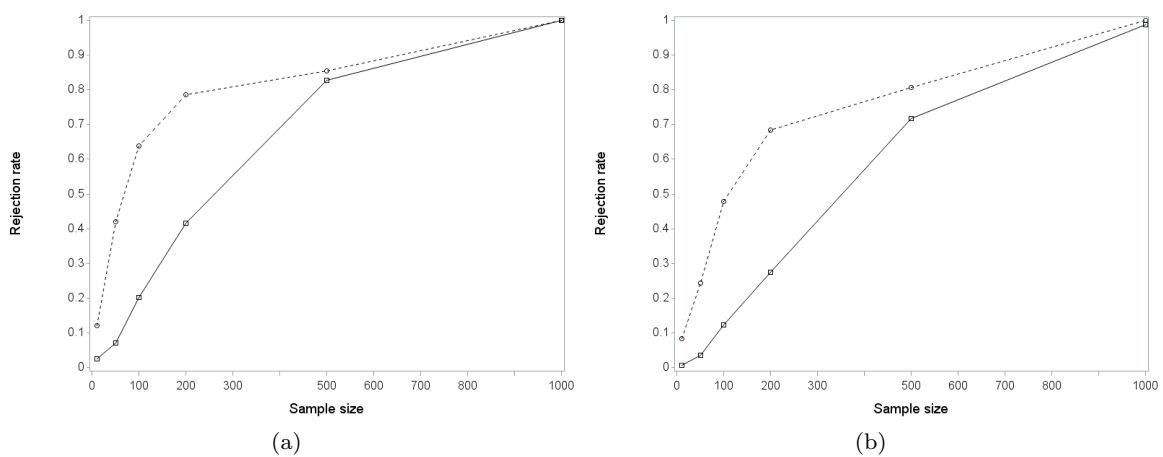
In our initial analysis of the theophylline data above, we found that the one-compartment model (5) with the pharmacokinetic parameters given in (7) produces a better fit than with the pharmacokinetic parameters in (6). There it was assumed that the random effects are normally distributed. Now, we want to check if the normality assumption is valid. For this purpose, we first look at the gradient function plot obtained from fitting the one-compartment model (5) with the pharmacokinetic parameters (7) to the theophylline data. The gradient plot, presented in Figure 5, shows some fluctuations from 1 and one may conclude that the bivariate normal distribution is not a proper choice. However, these fluctuations may be due to random variability in the estimates of parameters needed for calculation of the gradient function. Thus, we shall apply the diagnostic test based on (4) to formally check if a bi-

variate normal distribution is appropriate for  $b_{i1}$  and  $b_{i2}$ . According to the simulation results, we should apply the bootstrap test as there are 12 subjects in this dataset. The bootstrap test with 200 bootstrap samples and with 1000 Monte Carlo integration nodes provides a test statistic of 3.82, giving a  $p$ -value of 0.14. Therefore, it suggests that a bivariate normal distribution is appropriate for the random effects  $b_{i1}$  and  $b_{i2}$ . Although for the theophylline data the diagnostic test suggests no evidence for misspecification of the random-effects distribution, one might be cautious as the simulations showed that the diagnostic test does not have a high power to detect misspecification when the sample size is very small.

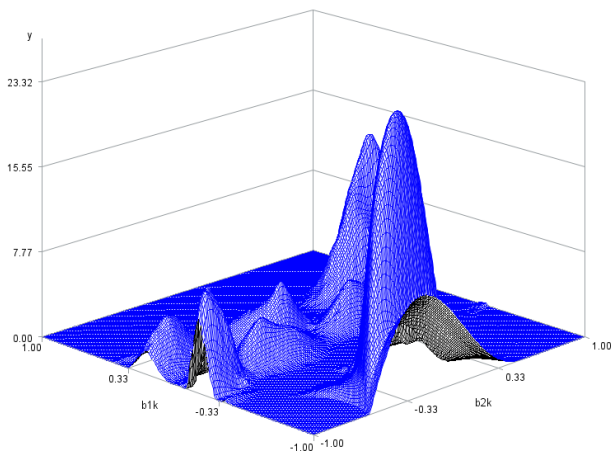




**Fig. 3** Rejection rates (powers) of the asymptotic test (solid line) and the bootstrap test (dashed line) at the significance level 0.05 with  $n_i = 5$  repeated measurements per subject and under four random-effects distributions: (a)  $N(0, 1)$ , (b) Chi-squared(2), (c) Log-normal(3, 1), and (d)  $F(1, 7)$ . Note that all the generated values of random effects were shifted and rescaled such that random effects  $b_{i1}$  and  $b_{i2}$  have both zero mean and variances equal to 0.1 and 0.5, respectively.



**Fig. 4** Rejection rates (powers) of the asymptotic test (solid line) and the bootstrap test (dashed line) at the significance level 0.05 when random effects are generated from  $\frac{1}{2}N(-2, 1) + \frac{1}{2}N(2, 1)$ , and with (a)  $n_i = 10$  repeated measurements per subject and (b)  $n_i = 5$  repeated measurements per subject. Note that all the generated values of random effects were shifted and rescaled such that random effects  $b_{i1}$  and  $b_{i2}$  have both zero mean and variances equal to 0.1 and 0.5, respectively.



**Fig. 5** Bivariate gradient function for the one-compartment model (5)-(7) fitted to the theophylline data under a bivariate normality assumption for  $b_{i1}$  and  $b_{i2}$ .

## Discussion

In this paper we have shown, via simulations and real data, that our diagnostic test based on the gradient function performs reasonably well in detecting misspecification of the random-effects distribution in nonlinear mixed-effects models. As expected, the asymptotic test does not show a high power when the sample size is small or moderate, while the bootstrap test performs better for small samples. Since the bootstrap test is time consuming when the sample size is very large, we suggest the asymptotic test for sufficiently large sample sizes, and the bootstrap test for smaller samples.

We have found that the power of our diagnostic test also depends on the number of repeated measurements per subject as well as the magnitude of variance components. In our real data example, the estimates of variance components of  $b_{i1}$  and  $b_{i2}$  were very small and consequently our simulation study conducted with those small variance components did not show high power values especially in small samples, while another simulation study (not reported in the paper) with the same settings but with larger variance components showed that the power of our diagnostic test increases considerably when the variance components are larger. Therefore, one must be cautious when applying the proposed diagnostic test to real data applications for which the estimates of variance components are very small, unless the sample size is very large.

When multiple random effects are present in the model, the gradient function plot is not much helpful because it needs evaluating a three or higher dimensional plot of gradient function which may not be easy to interpret and judge. Therefore, it is recommended to

apply the formal diagnostic test based on the gradient function when the model includes two or more random effects.

We previously developed a bootstrap test based on the gradient function (see [9]), however as shown in [8], that bootstrap test does not have a good power even for generalized linear mixed models, because it evaluates the gradient function for a grid of random effects only and moreover its bootstrap algorithm was developed under the normality of maximum likelihood estimates which is a large sample property. Instead, our bootstrap test in this paper uses a powerful test statistic as well as a simple bootstrap algorithm.

## References

1. Agresti, A., Caffo, B., Ohman-Strickland, P.: Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* **47**, 639–653 (2004)
2. Alonso, A., Litière, S., Laenen, A.: A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician* **64**, 318–324 (2010)
3. Alonso, A., Litière, S., Molenberghs, G.: A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. *Computational Statistics and Data Analysis* **52**, 4474–4486 (2008)
4. Boekmann, A., Sheiner, L., Beal, S.: *NONMEM user's guide, part V: introductory guide*. University of California, San Francisco (1992)
5. Claeskens, G., Hart, J.D.: Goodness-of-fit tests in mixed models. *Test* **18**, 213–239 (2009)
6. Davidian, M., Giltinan, D.M.: *Nonlinear models for repeated measurement data*. New York: Chapman and Hall (1995)
7. Davidian, M., Giltinan, D.M.: *Nonlinear models for repeated measurement data: an overview and update*. *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 387–419 (2003)
8. Drikvandi, R., Verbeke, G., Molenberghs, G.: Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics* (2016). DOI: 10.1111/biom.12551
9. Efendi, A., Drikvandi, R., Verbeke, G., Molenberghs, G.: A goodness-of-fit test for the random-effects distribution in mixed models. *Statistical Methods in Medical Research* (2014). DOI: 10.1177/0962280214564721
10. Hartford, A., Davidian, M.: Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational Statistics & Data Analysis* **34**, 139–164 (2000)
11. Heagerty, P.J., Kurland, B.F.: Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88**, 973–985 (2001)
12. Huang, X.: Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics* **65**, 361–368 (2009)
13. Jiang, J.: Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics* **29**, 1137–1164 (2001)
14. Lange, N., Ryan, L.: Assessing normality in random effects models. *The Annals of Statistics* **17**, 624–642 (1989)

15. Litière, S., Alonso, A., Molenberghs, G.: Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics* **63**, 1038–1044 (2007)
16. Litière, S., Alonso, A., Molenberghs, G.: The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine* **27**, 3125–3144 (2008)
17. Mentré, F., Escolano, S.: Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *Journal of Pharmacokinetics and Pharmacodynamics* **33**, 345–367 (2006)
18. Pan, Z., Lin, D.: Goodness-of-fit methods for generalized linear mixed models. *Biometrics* **61**, 1000–1009 (2005)
19. Pinheiro, J.C., Bates, D.M.: Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics* **4**, 12–35 (1995)
20. Ritz, C.: Goodness-of-fit tests for mixed models. *Scandinavian Journal of Statistics* **31**, 443–458 (2004)
21. Strenio, J.F., Weisberg, H.I., Bryk, A.S.: Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* **39**, 71–86 (1983)
22. Tchetgen, E.J., Coull, B.A.: A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika* **93**, 1003–1010 (2006)
23. Verbeke, G., Lesaffre, E.: A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221 (1996)
24. Verbeke, G., Molenberghs, G.: The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics* **14**, 477–490 (2013)
25. Waagepetersen, R.: A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scandinavian Journal of Statistics* **33**, 721–731 (2006)