MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1561

Published online 4 March 2013 in Wiley Online Library

A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models†

Hoai-Thu Thai,[a]* France Mentré,[a] Nicholas H. G. Holford,[b] Christine Veyrat-Follet,[c] and Emmanuelle Comets[a]

A version of the nonparametric bootstrap, which resamples the entire subjects from original data, called the case bootstrap, has been increasingly used for estimating uncertainty of parameters in mixed-effects models. It is usually applied to obtain more robust estimates of the parameters and more realistic confidence intervals (CIs). Alternative bootstrap methods, such as residual bootstrap and parametric bootstrap that resample both random effects and residuals, have been proposed to better take into account the hierarchical structure of multi-level and longitudinal data. However, few studies have been performed to compare these different approaches. In this study, we used simulation to evaluate bootstrap methods proposed for linear mixed-effect models. We also compared the results obtained by maximum likelihood (ML) and restricted maximum likelihood (REML). Our simulation studies evidenced the good performance of the case bootstrap as well as the bootstraps of both random effects and residuals. On the other hand, the bootstrap methods that resample only the residuals and the bootstraps combining case and residuals performed poorly. REML and ML provided similar bootstrap estimates of uncertainty, but there was slightly more bias and poorer coverage rate for variance parameters with ML in the sparse design. We applied the proposed methods to a real dataset from a study investigating the natural evolution of Parkinson's disease and were able to confirm that the methods provide plausible estimates of uncertainty. Given that most real-life datasets tend to exhibit heterogeneity in sampling schedules, the residual bootstraps would be expected to perform better than the case bootstrap.

Keywords: bootstrap; longitudinal data; Parkinson's disease; linear mixed-effects models; R

1. INTRODUCTION

Mixed-effects models are commonly used to analyze longitudinal data that consist of repeated measures from individuals through time [1]. They play an important role in medical research, particularly in clinical trials. These models incorporate the fixed effects, which are parameters representing effects in the entire population, and random effects, which are associated with individuals sampled from a population [2]. The parameters of a model are estimated by maximum likelihood (ML) or restricted maximum likelihood (REML) method. In linear mixed-effects models (LMEMs), REML is often preferred to ML estimation because it takes into account the loss of the degrees of freedom involved in estimating the fixed effects, resulting in unbiased estimates of variance components in many situations [2, 3].

The standard errors (SEs) of parameter estimates are obtained asymptotically from the inverse of the Fisher information matrix [2, 3]. The aforementioned estimates of SE might be biased when the asymptotic approximation is incorrect, for example, when the sample size is small. Sometimes, they cannot be obtained when the model is complex or the design is too sparse. Bootstrap methods represent an alternative approach for estimating the SE of parameters, as well as to provide a CI without assuming it is symmetrical. It was first introduced by Efron (1979) for independent and identically distributed observations. The principal idea of bootstrap is to resample the observed data repeatedly to create datasets similar to the original dataset and then fit them to construct the distribution of an estimator or a statistic of interest [4, 5]. Four main bootstrap approaches have been proposed for simple linear regression: case bootstrap, residual bootstrap, parametric bootstrap, and wild bootstrap [6–9]. The case bootstrap is the most simple and intuitive form that consists in resampling the entire vector of observations with replacement. The residual bootstrap resamples the residuals after model fitting. The parametric bootstrap adopts the principle of residual bootstrap, but instead of directly resampling observed residuals, we simulate the residuals from the estimated distribution, for example, the normal distribution, whose parameters are estimated using the original data. The wild bootstrap consists in resampling the residuals from an external distribution satisfying certain specifications.

† *Supporting information may be found in the online version of this article.*

[a] *Univ Paris Diderot, Sorbonne Paris Cité, UMR 738, F-75018 Paris, France; INSERM, UMR 738, F-75018 Paris, France*

[b] *Department of Pharmacology and Clinical Pharmacology, University of Auckland, Auckland, New Zealand*

[c] *Drug Disposition Department, Sanofi, Paris, France*

**Correspondence to: Hoai-Thu Thai, UMR738 INSERM, University Paris Diderot, 75018 Paris, France.*
E-mail: hoai-thu.thai@inserm.fr

The main concern when bootstrapping is how to generate a bootstrap distribution close to the true distribution of the original sample. To do that, the bootstrap resampling should appropriately mimic the 'true' data generating process that produced the 'true' dataset [10–12]. In the context of repeated measurement data and mixed-effects modeling, the bootstrap should therefore respect the true data generating process with the repeated measures within a subject and handle two levels of variability: between-subject variability (BSV) and residual variability (RUV). The classical bootstrap methods developed in simple linear regression should be modified to take into account the characteristics of mixed-effects models [13]. Resampling random effects may be coupled with resampling residuals [10, 13–15]. The case bootstrap can be combined with the residual bootstrap [8]. The performance of these approaches are, however, not well studied.

In this paper, we extend different bootstrap approaches that can be applied to LMEMs settings. The detail of bootstrap methods is described in Section 2. The simulation settings are described in Section 3. The results of the simulation studies and the application of the bootstraps to real data collected in a study in Parkinson's disease are described in Section 4. The discussion of the study is given in Section 5.

2. METHODS

2.1. Statistical models

Let $\mathbf{Y}$ be the response variable. N denotes the number of subjects. Let y_{ij} denote the observation j of $\mathbf{Y}$ in subject i, whereas $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ regroups the $(n_i \times 1)$ vector of measurements in this subject. Let $n_{\text{tot}} = \sum_{i=1}^{N} n_i$ denote the total number of observations. We use the following LMEM [16]:

$$\begin{cases} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\eta}_i \sim N(0, \Omega) \\ \boldsymbol{\epsilon}_i \sim N\left(0, \sigma^2\mathbf{I}_{n_i}\right) \end{cases} \tag{1}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ design matrices, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector containing the fixed effects, $\boldsymbol{\eta}_i$ is the $(q \times 1)$ vector containing the random effects, and $\boldsymbol{\epsilon}_i$ is the $(n_i \times 1)$ vector of residual components. Ω is the general $(q \times q)$ covariance matrix with (i,j) element $\omega_{ij} = \omega_{ji}$, and $\sigma^2\mathbf{I}_{n_i}$ is the $(n_i \times n_i)$ covariance matrix for residual errors in subject i where σ^2 is the error variance and $\mathbf{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix. The random effects $\boldsymbol{\eta}_i$ are assumed to be normally distributed with mean 0 and covariance matrix Ω, and the residual errors $\boldsymbol{\epsilon}_i$ are assumed to be normally distributed with mean 0 and variance $\sigma^2\mathbf{I}_{n_i}$. The random effects $\boldsymbol{\eta}_i$ and the residual errors $\boldsymbol{\epsilon}_i$ are assumed to be independent for different subjects and to be independent of each other for the same subject.

Conditional on the random effects $\boldsymbol{\eta}_i$, the response $\mathbf{Y}_i$ in subject i is normally distributed with mean vector $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\eta}_i$ and with covariance matrix $\sigma^2\mathbf{I}_{n_i}$.

2.2. Estimation methods

The parameters of LMEMs can be estimated in the framework of ML by two general methods: ML or REML. Let α denote the vector of all variance components of $\mathbf{V}_i = \mathbf{Z}_i\Omega(\mathbf{Z}_i)' + \Sigma_i$; it means α consists of the $q(q+1)/2$ different elements in Ω (or q elements if Ω is diagonal) and of all parameters in Σ_i. Let $\theta = (\beta', \alpha')'$ be the $(s \times 1)$ vector of all parameters in the marginal model for $\mathbf{Y}_i$. The parameter estimates are obtained by maximizing the marginal likelihood function with respect to θ. The ML likelihood function is defined as follows:

$$L_{\text{ML}}(\theta) = \prod_{i=1}^{N}\left\{(2\pi)^{-n_i/2}|\mathbf{V}_i|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right)\right\} \tag{2}$$

The REML likelihood function is derived from $L_{\text{ML}}(\theta)$ to correct the loss of the degrees of freedom involved in estimating the fixed effects:

$$L_{\text{REML}}(\theta) = \left|\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right|^{-\frac{1}{2}} L_{\text{ML}}(\theta) \tag{3}$$

In this study, we used REML as the estimation method. However, we also compared the results of REML with those of ML and presented them in the Appendix (available online as supplementary material).

When α is known, the maximum likelihood estimator of β, obtained from maximizing $L_{\text{ML}}(\theta)$, conditional on α, is given by

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i \tag{4}$$

When α is unknown, but an estimate $\hat{\alpha}$ is available, we can set $\mathbf{V}_i = \hat{\mathbf{V}}_i$ and estimate β by using Equation (4).

Estimates of the $\boldsymbol{\eta}_i$ can be obtained as the mean of the posterior distribution of $\boldsymbol{\eta}_i$ (empirical Bayes estimates):

$$\hat{\boldsymbol{\eta}}_i(\theta) = E[\boldsymbol{\eta}_i \mid \mathbf{y}_i] = \int \boldsymbol{\eta}_i f(\boldsymbol{\eta}_i \mid \mathbf{y}_i)\mathrm{d}\boldsymbol{\eta}_i = \Omega\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{5}$$

The maximum likelihood estimator $\hat{\theta}$ of θ is asymptotically normally distributed with mean θ and asymptotic covariance matrix given by the inverse of the Fisher information matrix M_F. The asymptotic SEs of parameters are then estimated as the square root of the diagonal element of the estimated covariance matrix.

2.3. Bootstrap methods

The principle of the bootstrap is to repeatedly generate pseudo-samples distributed according to the same distribution as the original sample. The unknown original distribution may be replaced by the empirical distribution of the sample, which is known as the nonparametric bootstrap [17]. The bootstrap exists in another version called the parametric bootstrap [7, 9]. In this version, the underlying distribution F is estimated from the data by a parametric model, for instance, normal distribution and bootstrap samples are generated by simulating within this distribution rather than from the empirical distribution as performed in the nonparametric version. The resampling can be carried out with an independent distribution; this procedure is called external bootstrap or wild bootstrap. This approach was proposed by Wu to deal with heteroscedasticity [18]. We have chosen to deal with the simpler case of homoscedasticity and therefore have not investigated the wild bootstrap.

Let B be the number of bootstrap samples to be drawn from the original dataset; a general bootstrap algorithm is as follows:

(1) Generate a bootstrap sample by resampling from the data and/or from the estimated model
(2) Obtain the estimates for all parameters of the model for the bootstrap sample

(3) Repeat steps 1 and 2 B times to obtain the bootstrap distribution of parameter estimates and then compute the mean, standard deviation, and 95% CI of this distribution.

Let $\hat{\theta}_b^*$ be the parameter estimated for the bth bootstrap sample. Given a data set, the expected value of the bootstrap estimator over the bootstrap distribution is calculated as the average of the parameter estimates from the B bootstrap samples:

$$\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^*\right) \tag{6}$$

The bootstrap SE is obtained as the sample standard deviation of the $\hat{\theta}_b^*$:

$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_B\right)^2} \tag{7}$$

A 95% bootstrap CI can be constructed by calculating the 2.5th and 97.5th percentile of bootstrap distribution:

$$\hat{\theta}^*_{(\alpha\cdot B)} \le \theta \le \hat{\theta}^*_{((1-\alpha)\cdot B)} \tag{8}$$

where $\alpha = 0.025$. An alternative approach is to use a normal approximation to construct a bootstrap CI, using the estimated $\widehat{\mathrm{SE}}_B$:

$$\hat{\theta}_B - \widehat{\mathrm{SE}}_B \cdot z_{1-\alpha/2} \le \theta \le \hat{\theta}_B + \widehat{\mathrm{SE}}_B \cdot z_{1-\alpha/2} \tag{9}$$

$z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, equal to 1.96 with $\alpha = 0.05$. However, it is preferable to use the bootstrap percentile CI when bootstrapping [7, 19].

The detailed algorithms of bootstrap methods to obtain a bootstrap sample (bootstrap generating process) are presented next in two separated groups: nonparametric and parametric bootstrap methods.

2.3.1. Nonparametric bootstrap. Nonparametric case bootstrap ($B_{\text{case,none}}$). This method consists of resampling with replacement the entire subjects, that is, the joint vector of design variables and corresponding responses ($\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{y}_i$) from the original data before modeling. It is also called the *paired bootstrap*. This procedure omits the second step of resampling the observations inside each subject. However, it is the most obvious way to do bootstrapping and makes no assumptions on the model.

Nonparametric case bootstrap coupled with global/individual residual bootstrap ($B_{\text{case,GR}}$ or $B_{\text{case,IR}}$). This method resamples first the entire subjects with replacement. The individual residuals are then resampled with replacement globally from the residual distribution of the original simulated dataset or individually from the residual distribution of new subjects obtained after bootstrapping in the first step. The bootstrap sample is obtained as follows:

(1) Fit the model to the data and then calculate the residuals $\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\boldsymbol{\eta}}_i$
(2) Draw N entire subjects $\left\{\left(\mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{y}_i^*\right)\right\}$ with replacement from $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{y}_i)\}$ in the original data and keep their predictions from model fitting $\mathbf{X}_i^*\hat{\boldsymbol{\beta}} + \mathbf{Z}_i^*\hat{\boldsymbol{\eta}}_i^*$ and their corresponding residuals $\hat{\boldsymbol{\epsilon}}_i^* = \mathbf{y}_i^* - \mathbf{X}_i^*\hat{\boldsymbol{\beta}} - \mathbf{Z}_i^*\hat{\boldsymbol{\eta}}_i^*$. The new subject has n_i^* observations.
(3) Draw the residuals with replacement globally from all residuals of the original data or individually from each new subject
  (a) Global residual resampling: draw a sample $\{\boldsymbol{\epsilon}^*\} = \left\{\hat{\epsilon}_{i^*j^*}\right\}$ of size $n_{\text{tot}}^* = \sum_{i=1}^{N} n_i^*$ with replacement globally from $\{\hat{\boldsymbol{\epsilon}}\} = \{\hat{\epsilon}_{ij}\}_{i=1,..,N;j=1,..,n_i}$ by assigning an equal probability $\frac{1}{n_{\text{tot}}}$ to each value of the n_{tot} residuals (note that, n_{tot}^* may be different with n_{tot})
  (b) Individual residual resampling: draw individually N samples $\left\{\boldsymbol{\epsilon}_i^*\right\} = \left\{\hat{\epsilon}_{ij^*}^*\right\}$ of size n_i^* with replacement from samples $\hat{\boldsymbol{\epsilon}}_i^* = \left\{\hat{\epsilon}_{ij}^*\right\}_{j=1,..,n_i^*}$ by assigning an equal probability $\frac{1}{n_i^*}$ to each residual of new subject i
(4) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i^*\hat{\boldsymbol{\beta}} + \mathbf{Z}_i^*\boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i^*$

Here, and in the following, we note the vector of estimated residuals for all subjects as $\hat{\boldsymbol{\epsilon}} = \{\hat{\epsilon}_{ij}\}_{i=1,..,N;j=1,..,n_i}$ and the vector of estimated residuals in subject i as $\hat{\boldsymbol{\epsilon}}_i = \{\hat{\epsilon}_{ij}\}_{j=1,..,n_i}$.

Nonparametric random effects bootstrap coupled with global/individual residual bootstrap ($B_{\eta,\text{GR}}$ or $B_{\eta,\text{IR}}$). This method consists of resampling with replacement the random effects obtained after model fitting, as well as the residuals globally or individually. The bootstrap sample is obtained as follows:

(1) Fit the model to the data and then estimate the random effects $\hat{\boldsymbol{\eta}}_i$ and the residuals $\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\boldsymbol{\eta}}_i$
(2) Draw a sample $\left\{\boldsymbol{\eta}_i^*\right\}$ of size N with replacement from $\{\hat{\boldsymbol{\eta}}_i\}$ by assigning an equal probability $\frac{1}{N}$ to each value
(3) Draw a sample $\{\boldsymbol{\epsilon}^*\} = \left\{\hat{\epsilon}_{i^*j^*}\right\}$ of size n_{tot} with replacement globally from $\{\hat{\boldsymbol{\epsilon}}\}$ or draw individually N samples $\left\{\boldsymbol{\epsilon}_i^*\right\} = \left\{\hat{\epsilon}_{ij^*}\right\}$ of size n_i with replacement from $\{\hat{\boldsymbol{\epsilon}}_i\}$
(4) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i^*$

Nonparametric global/individual residual bootstrap ($B_{\text{none,GR}}$/$B_{\text{none,IR}}$). For the sake of completeness, we also implemented a bootstrap where only RUV is resampled. These procedures do not resample the BSV (which remains in the model through the estimated random effects $\hat{\boldsymbol{\eta}}_i$). The bootstrap sample is obtained as follows:

(1) Fit the model to the data and then calculate the residuals $\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\boldsymbol{\eta}}_i$
(2) Draw a sample $\{\boldsymbol{\epsilon}^*\} = \left\{\hat{\epsilon}_{i^*j^*}\right\}$ of size n_{tot} with replacement globally from $\{\hat{\boldsymbol{\epsilon}}\}$ or draw individually N samples $\left\{\boldsymbol{\epsilon}_i^*\right\} = \left\{\hat{\epsilon}_{ij^*}\right\}$ of size n_i with replacement from $\{\hat{\boldsymbol{\epsilon}}_i\}$
(3) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\boldsymbol{\eta}}_i + \boldsymbol{\epsilon}_i^*$

The nonparametric bootstrap methods that resample the random effects or the residuals depend on the structural model to calculate the raw random effects or residuals. However, they do not require particular assumptions on their distributions.

2.3.2. Parametric bootstrap. The parametric bootstrap requires the strongest assumptions as it depends both on the model and the distributions of parameters and errors.

Case bootstrap coupled with parametric residual bootstrap ($B_{\text{case,PR}}$). Similar to $B_{\text{case,GR}}$, this methods resamples firstly the subjects and then resamples the residuals by simulating from the estimated distribution. This method combines elements of both the nonparametric bootstrap (case bootstrap in the first step) and the parametric bootstrap (residual bootstrap in the second step).

However, to simplify the classification, we keep it in the group of parametric bootstraps. The bootstrap sample is obtained as follows:

(1) Fit the model to the data
(2) Draw N entire subjects $\{(\mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{y}_i^*)\}$ with replacement from $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{y}_i)\}$ in the original data and keep their predictions from model fitting $\mathbf{X}_i^*\hat{\boldsymbol{\beta}} + \mathbf{Z}_i^*\hat{\boldsymbol{\eta}}_i^*$
(3) Draw N samples $\{\boldsymbol{\epsilon}_i^*\}$ of size n_i^* from a normal distribution with mean 0 and covariance matrix $\hat{\sigma}^2\mathbf{I}_{n_i}$
(4) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i^*\hat{\boldsymbol{\beta}} + \mathbf{Z}_i^*\hat{\boldsymbol{\eta}}_i + \boldsymbol{\epsilon}_i^*$

Parametric random effects bootstrap coupled with residual bootstrap ($B_{\mathrm{P}\eta,\mathrm{PR}}$). This methods resamples both random effects and residuals by simulating from estimated distribution after model fitting. The bootstrap sample is obtained as follows:

(1) Fit the model to the data
(2) Draw a sample $\{\boldsymbol{\eta}_i^*\}$ of size N from a multivariate normal distribution with mean 0 and covariance matrix $\hat{\Omega}$
(3) Draw N samples $\{\boldsymbol{\epsilon}_i^*\}$ of size n_i from a normal distribution with mean zero and covariance matrix $\hat{\sigma}^2\mathbf{I}_{n_i}$
(4) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i^*$

Parametric residual bootstrap ($B_{\mathrm{none},\mathrm{PR}}$). Again for exhaustiveness, we implemented a bootstrap that resamples only the residuals by simulating from the estimated distribution after model fitting. Similar to $B_{\mathrm{none},\mathrm{GR}}$ or $B_{\mathrm{none},\mathrm{IR}}$, this procedure omits the first step of resampling the BSV. The bootstrap sample is obtained as follows:

(1) Fit the model to the data
(2) Draw N samples $\{\boldsymbol{\epsilon}_i^*\}$ of size n_i from a normal distribution with mean 0 and covariance matrix $\hat{\sigma}^2\mathbf{I}_{n_i}$
(3) Generate the bootstrap responses $\mathbf{y}_i^* = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\boldsymbol{\eta}}_i + \boldsymbol{\epsilon}_i^*$

2.3.3. Transformation of random effects and residuals. Previous results in the literature show that the nonparametric bootstrap of the raw residuals obtained by model fitting usually yield downwardly biased variance parameter estimates [7, 8, 20]. In ordinary linear models, this underestimation is due to the difference between estimated and empirical variance of residuals [21]. That is why it is advisable to rescale the residuals so that they have the correct variance. Efron suggested to multiply centered raw residuals with the factor $\sqrt{(n-p)/n}$ where p is the number of parameters and n is the number of observations [22]. For the same reason, Davison *et al.* proposed to use the factor $\sqrt{1/(1-h_i)}$ where h_i is the ith diagonal element of the hat matrix that maps the vectors of observed values to the vector of fitted values [7]. In the mixed-effects models, the raw variance–covariance matrix is different from the ML estimate, as the raw random effects or residuals are 'shrunk' towards 0 [20]. Carpenter *et al.* proposed to take this into account by centering the random effects and residuals to resample from a distribution with mean 0 and multiplying them by the ratio between their corresponding estimated and empirical variance–covariance matrices to account for the variance underestimation [20, 23]. These corrections were used in our study.

Transforming random effects. The transformation of random effects was carried out in the following steps:

(1) Center the raw estimated random effects: $\tilde{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\eta}}_i - \bar{\boldsymbol{\eta}}_i$
(2) Calculate the ratio between the estimated and empirical variance–covariance matrix (A_η). Let $\hat{\Omega}$ be the model estimated variance–covariance matrix of random effects and Ω_{emp} denote the empirical variance–covariance matrix of the centered random effects $\tilde{\boldsymbol{\eta}}_i$. The ratio matrix A_η is formed by using the Cholesky factors L_{est} and L_{emp}, which are the lower triangular matrix of the Cholesky decomposition of $\hat{\Omega}$ and Ω_{emp}, respectively: $A_\eta = \left(L_{\mathrm{est}} \cdot L_{\mathrm{emp}}^{-1}\right)'$
(3) Transform the centered random effects using the ratio A_η: $\hat{\boldsymbol{\eta}}_i' = \tilde{\boldsymbol{\eta}}_i \times A_\eta$

Transforming residuals. The transformation of residuals was carried out globally for all residuals in the following steps:

(1) Center the raw estimated residuals: $\tilde{\epsilon}_{ij} = \hat{\epsilon}_{ij} - \bar{\epsilon}_{ij}$
(2) Calculate the ratio between the estimated and empirical variance–covariance matrix (A_σ). Let $\hat{\Sigma}$ be the model estimated variance–covariance matrix for residuals, which is assumed to be equal to $\hat{\sigma}^2\mathbf{I}_{n_i}$ and Σ_{emp} denote the empirical variance–covariance matrix of centered residual $\tilde{\epsilon}_{ij}$. Because the residuals are assumed to be uncorrelated and to have equal variance, the ratio matrix A_σ is then simply the ratio between the square root of the model estimated residual variance $\hat{\sigma}^2$ and the empirical standard deviation of the centered residuals, respectively: $A_\sigma = \hat{\sigma}/sd(\tilde{\epsilon}_{ij})$.
(3) Transform the centered residuals using the ratio A_σ: $\hat{\epsilon}_{ij}' = \tilde{\epsilon}_{ij} \times A_\sigma$

3. SIMULATION STUDIES

3.1. Motivating example

The motivating example for bootstrap evaluation was a disease progression model inspired from the model of Parkinson's disease developed by Holford *et al.* [24]. In that study, the subjects were initially randomized to treatment with placebo, deprenyl, tocopherol, or with both and, when clinical disability required, received one or more dopaminergic agents (levodopa, bromocriptine, or pergolide). The aim was to study the influence of various drugs on the changes in Unified Parkinson's Disease Rating Scale (UPDRS) over time. Several components describing the disease progression and the effect of treatment were developed. However, in this study, we are only interested in the linear part of the model, describing the natural evolution of Parkinson's disease using a random intercept and slope model as

$$y_{ij} = S_0 + \alpha \cdot t_{ij} + \eta_{S_0 i} + \eta_{\alpha i} \cdot t_{ij} + \epsilon_{ij} \tag{10}$$

where y_{ij} is the UPDRS score (u) representing the state of disease at time t_{ij} that was considered to be continuous, S_0 (u) is the expected score at randomization, α (u/year) is the progression rate, $\eta_{S_0 i}$ (u/year) and $\eta_{\alpha i}$ (u/year) are the random effects of S_0 and α, respectively, and ϵ_{ij} is the residual errors. In the formulation of Equation (1), the design matrix $\mathbf{X}_i = \mathbf{Z}_i$ for subject i is a two-column matrix with a first column of 1s and a second column containing the n_i times t_{ij} for subject i.

For our simulations, we used the subset of patients who remained in the placebo group over the first 2 years. The UPDRS scores were measured at randomization and at regular (unscheduled) visits up to 2 years after entry to the study. This subset contains 109 subjects with an average of six observations per subject and a total of 822 observations. Figure 1 describes the evolution of UPDRS score over a two-year period.

In the paper of Holford *et al.*, the baseline S_0 was assumed to be log-normally distributed, and the progression rate α was

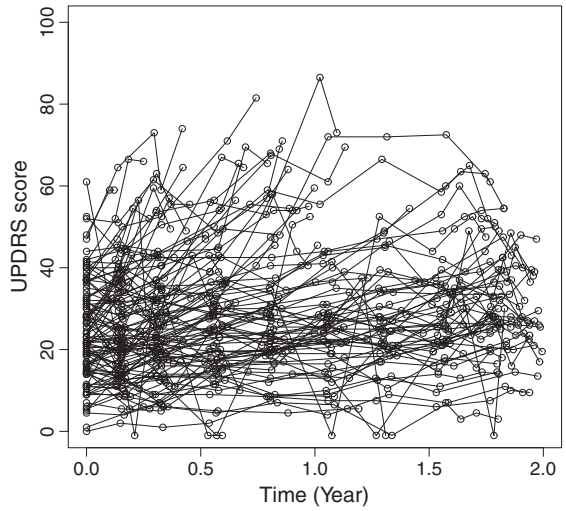

Figure 1. The evolution of Unified Parkinson's Disease Rating Scale (UPDRS) score over time in the real dataset used for the simulations, including 109 patients who remained in the placebo group over the first 2 years.

assumed to be normally distributed [24]. The variance of the random effects and the correlation ρ between S_0 and α was estimated. The residual unexplained variability was described by an additive model error with constant variance σ^2. In the present study, both S_0 and α were assumed to be normally distributed. We estimated the parameters of the real dataset by ML, the same estimation method as used in the original publication [24], using the `lme` function in R. The detail of parameter estimates is given in Section 4.2.

3.2. Simulation settings

Our simulations were inspired from the original design described earlier using estimates from the real dataset.

Three designs were planned to evaluate the performance of bootstrap. For each design, the sampling times were similar for all subjects.

Rich design. We simulated $N = 100$ subjects with $n = 7$ observations per subject at 0, 0.17, 0.33, 0.5, 1, 1.5, and 2 years after being entered in the study.

Sparse design. We simulated $N = 30$ subjects with $n = 3$ observations per subject at 0, 0.17, and 2 years after being entered in the study. This design is sparse with respect to estimation of variance parameters, including only 90 observations in total.

Large error design. We simulated $N = 100$ subjects with $n = 7$ observations per subject at 0, 0.17, 0.33, 0.5, 1, 1.5, and 2 years after being entered in the study. In this design, we modified the level of variability. The variability for random effects η_{S_0} and η_α was changed to $\omega_{S_0} = 11.09$ and $\omega_\alpha = 6.98$, respectively (equivalent to 50% of the corresponding fixed effects), and the standard deviation for the residual error was changed to $\sigma = 17.5$. We also removed the correlation ρ between S_0 and α in this design because convergence was obtained for only 78.3% simulated datasets with the presence of this correlation.

For each design, we simulated $K = 1000$ replications.

3.3. Software

The `lme` function in the `nlme` library in R was used to fit the data using REML as the estimation method. ML was also used to compare with the results of REML. For both methods, we fitted datasets with the initial values of variance parameters generated by the optimization procedure implemented in the `lme` function [25]. All the analysis and figures were carried out with R.

3.4. Evaluation of bootstrap methods

Table I presents all the bootstrap methods that we implemented and evaluated. The resampling of the BSV can be carried out by resampling subjects or random effects. The resampling of the RUV can be carried out by resampling of residuals obtained from all subjects, called global residuals, or resampling of residuals within each subject, called individual residuals. To compare the performance of these bootstraps to that of classical bootstraps, the residual bootstrap, which resamples only the RUV, and the case bootstrap, where the whole vector of observations, including both the BSV and RUV, is resampled, were also evaluated in our study. The bootstrap methods were classified as nonparametric and parametric methods. For the parametric approach, there is no difference between the global and the individual residual resampling because it is performed by simulation from the estimated distributions. In total, we had seven nonparametric methods and three parametric methods.

We drew $B = 1000$ bootstrap samples for each replication of simulated data and for each bootstrap method. The parameters were estimated by REML method using the `lme` function in R. For each method, we therefore performed 1 million fits (1000 simulated datasets $\times$ 1000 bootstrap datasets). If the convergence was

Table I. Bootstrap methods that can be applied in mixed-effects models.

				Variability related to subject		
				None	Resample individuals	Resample random effects
Variability related to observation	Resample residuals	None		Original data	$B_{\text{case,none}}$	
		Globally	NP	$B_{\text{none,GR}}$	$B_{\text{case,GR}}$	$B_{\eta,\text{GR}}$
			P	$B_{\text{none,PR}}$	$B_{\text{case,PR}}$	$B_{\text{P}\eta,\text{PR}}$
		Individually	NP	$B_{\text{none,IR}}$	$B_{\text{case,IR}}$	$B_{\eta,\text{IR}}$

NP, nonparametric; P, parametric.

not obtained, the NA(not applicable) was recorded in the table of parameter estimates and excluded for further calculation.

For the kth simulated dataset and for a given bootstrap method, we computed the bootstrap parameter estimate $\hat{\theta}_B^k$ as in Equation (5), using 1000 bootstrap samples, as well as the boostrap SE and CI, for each parameter θ. The relative bias (RBias) of bootstrap estimate was obtained by comparing the bootstrap estimate $\hat{\theta}_B^k$ and the asymptotic estimate $\hat{\theta}_k$ as follows:

$$\text{RBias}\left(\hat{\theta}_B\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{\theta}_B^k - \hat{\theta}_k}{\hat{\theta}_k} \times 100\right) \tag{11}$$

The average bootstrap SE was obtained by averaging the SE from Equation (6) over the $K = 1000$ datasets. The true SE is unknown, but we can have an empirical estimate as the standard deviation of the differences between the estimate of the parameter in the K datasets and the true value (θ_0):

$$\text{SE}_{\text{empirical}}\left(\hat{\theta}\right) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\theta}_k - \theta_0\right)^2} \tag{12}$$

The relative bias on bootstrap SE was then obtained by comparing the average bootstrap SE to this empirical SE:

$$\text{RBias}\left(\text{SE}\left(\hat{\theta}_B\right)\right) = \frac{\frac{1}{K}\sum_{k=1}^{K}\text{SE}_B\left(\hat{\theta}_B^k\right) - \text{SE}_{\text{empirical}}\left(\hat{\theta}\right)}{\text{SE}_{\text{empirical}}\left(\hat{\theta}\right)} \times 100 \tag{13}$$

The coverage rate of the 95% bootstrap CI was defined as the percentage of the $K = 1000$ datasets in which the bootstrap CI contains the true value of the parameter.

The bootstrap approaches were compared in terms of the Rbias on the bootstrap parameter estimates, the Rbias on SE, and the coverage rate of the 95% CI of all parameter estimates from one million bootstrap samples.

The performance of the bootstrap methods were also compared with the performance of the asymptotic method. It is worth noting that in the simulation studies, random variables were simulated according to normal distributions, which may have contributed to the good performance of the asymptotic method. The relative bias of asymptotic estimate was obtained by comparing the asymptotic estimate $\hat{\theta}_k$ and the true value θ_0 as follows:

$$\text{RBias}\left(\hat{\theta}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{\theta}_k - \theta_0}{\theta_0} \times 100\right) \tag{14}$$

The relative bias of asymptotic SEs given by the software (obtained as the inverse of the Fisher information matrix) was defined in the same way as Equation (13) but with respect to $\hat{\theta}_k$ instead of $\hat{\theta}_B^k$. The coverage rate of the 95% asymptotic CI was defined as the percentage of datasets in which the asymptotic CI contains the true value of the parameter.

The asymptotic and bootstrap parameters estimates and their SE were defined as unbiased when relative bias was within $\pm 5\%$, moderately biased (relative bias from $\pm 5\%$ to $\pm 10\%$) or strongly biased (relative bias $> \pm 10\%$). The coverage rate of the 95% CI was considered to be good (from 90% to 100%), low (from 80% to 90%), or poor($< 80\%$). A good bootstrap was defined as a method providing unbiased estimates for the parameters and their corresponding SE and ensuring a good coverage rate of the 95% CI.

3.5. Application to real data

All the bootstrap methods with good performance evaluated in the simulations studies were applied to the real data by drawing $B = 1000$ bootstrap samples for each method. The bootstrap parameter estimates and bootstrap SE were compared with each other and compared with the parameter estimates and their SEs obtained by the asymptotic approach.

4. RESULTS

4.1. Simulation studies

Examples of simulated data for each given design are illustrated in Figure 2. Our simulations gave some negative values for the observations because of the homoscedastic error model. These values were kept as is, because the purpose of the simulations was not to provide a realistic simulation of the trial but only to

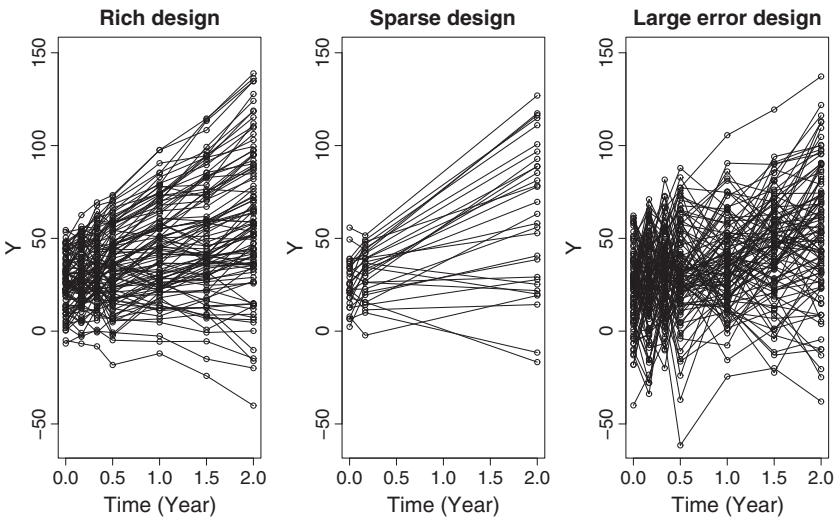

Figure 2. Examples of simulated data for each design.

Table II. Relative bias of parameter estimates by REML and their standard errors (SE) for the asymptotic method and the bootstrap methods in the three studied designs.

		Relative bias of parameters (%)						Relative bias of SE (%)					
Design	Method	S_0	α	ω_{S_0}	ω_α	ρ	σ	S_0	α	ω_{S_0}	ω_α	ρ	σ
Rich	Asymptotic	**0.05**	**0.19**	**−0.13**	**−0.03**	**−0.27**	**0.03**	**2.79**	**1.62**	**−2.72**	**−1.02**	**4.99**	**−0.92**
	$B_{\text{case,none}}$	**0.00**	**0.01**	**−0.85**	**−0.81**	**0.15**	**−0.05**	**2.31**	**1.17**	−5.47	**−2.97**	5.04	**−1.99**
	$B_{\text{none,GR}}$	**0.00**	**0.00**	**−3.63**	**−2.53**	13.48	**−0.12**	−70.16	−75.26	−60.48	−66.13	−41.84	**0.14**
	$B_{\text{none,IR}}$	**0.00**	**0.00**	**0.51**	**−2.53**	9.73	**−0.61**	−70.28	−75.37	−60.26	−65.89	−39.25	**−2.66**
	$B_{\text{none,PR}}$	**0.04**	**0.19**	**−3.72**	**−2.53**	12.97	**−0.02**	−70.15	−75.22	−60.45	−66.06	−41.80	**−1.04**
	$B_{\text{case,GR}}$	**0.00**	**0.01**	**−4.34**	**−3.25**	13.23	**−0.13**	**−1.05**	**−1.07**	−7.75	**−4.89**	−7.49	**0.24**
	$B_{\text{case,IR}}$	**0.05**	**0.21**	**−0.46**	**−3.32**	9.43	**−0.67**	6.97	**1.87**	**3.30**	**0.92**	14.82	65.57
	$B_{\text{case,PR}}$	**0.04**	**0.18**	**−4.43**	**−3.26**	12.70	**−0.01**	**−1.19**	**−1.28**	−7.54	**−4.87**	−7.34	**−1.01**
	$B_{\eta,\text{GR}}$	**0.00**	**−0.01**	**−0.80**	**−0.78**	**−0.03**	**−0.12**	**2.34**	**1.10**	**−4.88**	**−2.65**	5.59	**0.07**
	$B_{\eta,\text{IR}}$	**0.00**	**0.00**	**−0.65**	**−0.77**	**−0.17**	**−0.62**	**2.26**	**1.00**	**−4.68**	**−2.68**	6.28	**−2.68**
	$B_{\text{P}\eta,\text{PR}}$	**0.04**	**0.19**	**−0.44**	**−0.32**	**−0.30**	**−0.02**	**2.75**	**1.57**	**−2.29**	**−0.64**	5.94	**−0.93**
Sparse	Asymptotic	**0.08**	**0.48**	**−0.92**	**−0.79**	**0.60**	**−0.99**	**2.96**	**0.04**	**−1.30**	**−2.51**	**−3.86**	**0.74**
	$B_{\text{case,none}}$	**−0.01**	**−0.02**	**−2.67**	**−2.63**	**1.24**	**−1.29**	**1.09**	**−1.85**	−8.96	−9.77	−6.88	−8.12
	$B_{\text{none,GR}}$	**0.00**	**0.00**	−5.09	**−3.10**	19.24	**−2.82**	−62.90	−72.23	−50.93	−62.00	−40.45	**4.04**
	$B_{\text{none,IR}}$	**0.00**	**0.00**	6.83	**−3.23**	10.52	**−6.46**	−64.31	−73.39	−50.54	−63.16	−32.37	11.24
	$B_{\text{none,PR}}$	**0.00**	**−0.01**	**−5.08**	**−3.11**	19.25	**−2.04**	−62.72	−72.11	−50.49	−61.97	−40.03	**−4.61**
	$B_{\text{case,GR}}$	**0.02**	**0.03**	**−7.33**	−5.44	17.86	**−2.98**	**−3.14**	**−4.43**	−10.84	−10.95	−16.09	**3.38**
	$B_{\text{case,IR}}$	**0.09**	**0.49**	**3.05**	−6.19	5.37	−10.61	12.39	**−1.15**	6.84	−5.09	10.56	59.45
	$B_{\text{case,PR}}$	**0.00**	**0.02**	−7.27	−5.44	17.89	**−2.24**	**−3.10**	**−4.32**	−10.62	−10.83	−15.62	**−5.00**
	$B_{\eta,GR}$	**0.02**	**0.02**	**−2.41**	**−2.46**	**0.37**	**−2.16**	**1.21**	**−1.60**	−6.94	−8.51	**−4.38**	6.33
	$B_{\eta,\text{IR}}$	**−0.01**	**0.01**	**−0.94**	**−2.45**	**−1.26**	−6.50	**0.82**	**−1.95**	−5.42	−8.94	**0.29**	10.81
	$B_{\text{P}\eta,\text{PR}}$	**0.00**	**0.03**	**−0.84**	**−0.88**	**−0.10**	**−1.34**	**2.97**	**0.12**	**−1.86**	**−2.92**	**−4.03**	**−2.11**
Large error	Asymptotic	**0.02**	**−0.06**	**−0.93**	**−1.24**		**0.02**	**−2.85**	**−2.05**	**−2.21**	**−3.57**		**1.12**
	$B_{\text{case,none}}$	**0.00**	**−0.01**	**−1.20**	**−3.58**		**−0.08**	**−3.39**	**−2.55**	**−3.53**	**1.06**		**−0.42**
	$B_{\text{none,GR}}$	**−0.01**	**0.01**	−10.64	−24.75		**−0.84**	−37.47	−20.85	−29.94	5.66		**−1.46**
	$B_{\text{none,IR}}$	**−0.01**	**0.01**	18.63	−24.96		**−2.70**	−38.53	−22.32	−34.90	15.70		−5.98
	$B_{\text{none,PR}}$	**0.00**	**0.00**	−10.64	−24.74		**−0.77**	−37.44	−20.88	−29.85	5.73		**−1.09**
	$B_{\text{case,GR}}$	**−0.01**	**−0.01**	−11.37	−25.40		**−0.83**	−12.20	−12.45	−7.29	11.61		**−1.29**
	$B_{\text{case,IR}}$	**0.01**	**0.00**	16.85	−30.12		**−2.96**	19.51	15.75	19.38	61.88		60.31
	$B_{\text{case,PR}}$	**0.00**	**0.00**	−11.35	−25.43		**−0.75**	−12.11	−12.30	−7.22	11.80		**−1.13**
	$B_{\eta,\text{GR}}$	**0.00**	**0.00**	**−1.07**	**−2.97**		**−0.12**	**−3.27**	**−2.13**	**−2.90**	**1.69**		**0.87**
	$B_{\eta,\text{IR}}$	**0.01**	**0.00**	**3.20**	**−3.02**		**−1.97**	**−3.92**	**−3.51**	**−1.81**	**3.35**		**−4.05**
	$B_{\text{P}\eta,\text{PR}}$	**0.00**	**0.01**	**−0.59**	**−2.46**		**−0.05**	**−2.87**	**−2.05**	**−1.78**	**2.13**		**1.01**

Relative bias within ±5% is typeset in bold font.
REML, restricted maximum likelihood.

evaluate the bootstrap methods. For the same reason, we did not take into account other real-life factors such as drop-outs or missingness.

We present firstly the results of REML. Complete results for all bootstrap methods as well as the asymptotic method in the three evaluated designs are reported in Table II, for relative bias of the bootstrap estimates of parameters and their corresponding SEs, and Table III, for the coverage rate of the 95% CI. The correlation between S_0 and α was not estimated in the large error design for both simulated and bootstrap datasets to have better convergence rate and to keep the same model that we simulated. The transformation of random effects and residuals were carried out before resampling for the methods in which they were resampled nonparametrically. In the rich design, we found that the bootstrap methods that resample only the residuals ($B_{\text{none,GR}}$, $B_{\text{none,IR}}$, $B_{\text{none,PR}}$) yielded higher bias for the correlation term (>9.73%). The same was observed for the case bootstrap coupled with the residual bootstraps ($B_{\text{case,GR}}$, $B_{\text{case,IR}}$, $B_{\text{case,PR}}$). The case bootstrap and the bootstraps of both random effects and residuals ($B_{\eta,\text{GR}}$, $B_{\eta,\text{IR}}$, $B_{\text{P}\eta,\text{PR}}$) showed essentially no bias for all parameters. In terms of SE estimation, these four bootstrap methods estimated correctly all the SEs, as the estimates were very close to empirical SEs, whereas the residual-alone bootstraps greatly underestimated the SEs of all parameters, except for that of σ. On the contrary, the SE of σ was overestimated by the $B_{\text{case,IR}}$. The case bootstrap, coupled with global residual bootstrap ($B_{\text{case,GR}}$) or parametric bootstrap ($B_{\text{case,PR}}$), gave better estimates for SE of parameters than the residual-alone bootstraps, but they did not work as well as the case bootstrap and the bootstrap of random effects and residuals. In terms of coverage rate, the

Table III. Coverage rate of the 95% CI of parameters by REML obtained by the asymptotic method and the bootstrap methods in the three studied designs.

Design	Method	Coverage rate of the 95% CI of parameters					
		S_0	α	ω_{S_0}	ω_α	ρ	σ
Rich	Asymptotic	**0.95**	**0.96**	**0.94**	**0.96**	**0.96**	**0.94**
	$B_{\text{case,none}}$	**0.94**	**0.96**	**0.92**	**0.94**	**0.95**	**0.94**
	$B_{\text{none,GR}}$	0.44	0.36	0.50	0.44	0.46	**0.95**
	$B_{\text{none,IR}}$	0.43	0.36	0.56	0.44	0.60	**0.93**
	$B_{\text{none,PR}}$	0.43	0.36	0.50	0.44	0.45	**0.94**
	$B_{\text{case,GR}}$	**0.94**	**0.95**	0.85	0.89	0.72	**0.94**
	$B_{\text{case,IR}}$	**0.96**	**0.96**	**0.95**	**0.91**	**0.91**	**1.00**
	$B_{\text{case,PR}}$	**0.94**	**0.95**	0.85	0.88	0.72	**0.94**
	$B_{\eta,\text{GR}}$	**0.94**	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
	$B_{\eta,\text{IR}}$	**0.95**	**0.95**	**0.92**	**0.94**	**0.96**	**0.94**
	$B_{\text{P}\eta,\text{PR}}$	**0.95**	**0.96**	**0.93**	**0.96**	**0.96**	**0.94**
Sparse	Asymptotic	**0.94**	**0.94**	**0.94**	**0.94**	**0.91**	**0.94**
	$B_{\text{case,none}}$	**0.94**	**0.93**	0.89	0.89	**0.92**	**0.90**
	$B_{\text{none,GR}}$	0.53	0.39	0.60	0.51	0.66	**0.95**
	$B_{\text{none,IR}}$	0.51	0.38	0.65	0.48	0.84	**0.92**
	$B_{\text{none,PR}}$	0.53	0.40	0.61	0.50	0.66	**0.93**
	$B_{\text{case,GR}}$	**0.93**	**0.93**	0.83	0.86	0.84	**0.95**
	$B_{\text{case,IR}}$	**0.96**	**0.94**	**0.98**	0.88	**0.99**	**0.99**
	$B_{\text{case,PR}}$	**0.93**	**0.93**	0.83	0.86	0.85	**0.93**
	$B_{\eta,\text{GR}}$	**0.94**	**0.93**	**0.91**	0.89	**0.94**	**0.95**
	$B_{\eta,\text{IR}}$	**0.94**	**0.94**	**0.93**	0.89	**0.96**	**0.92**
	$B_{\text{P}\eta,\text{PR}}$	**0.94**	**0.94**	**0.93**	**0.94**	**0.94**	**0.94**
Large error	Asymptotic	**0.93**	**0.95**	**0.95**	**0.98**		**0.95**
	$B_{\text{case,none}}$	**0.93**	**0.94**	**0.92**	**0.93**		**0.95**
	$B_{\text{none,GR}}$	0.78	0.88	0.58	0.74		**0.93**
	$B_{\text{none,IR}}$	0.78	0.87	0.38	0.79		0.82
	$B_{\text{none,PR}}$	0.78	0.88	0.58	0.74		**0.93**
	$B_{\text{case,GR}}$	**0.90**	**0.91**	0.71	0.75		**0.93**
	$B_{\text{case,IR}}$	**0.98**	**0.98**	0.84	**0.94**		**0.99**
	$B_{\text{case,PR}}$	**0.90**	**0.92**	0.71	0.75		**0.94**
	$B_{\eta,\text{GR}}$	**0.93**	**0.95**	**0.94**	**0.93**		**0.95**
	$B_{\eta,\text{IR}}$	**0.94**	**0.95**	**0.97**	**0.94**		0.88
	$B_{\text{P}\eta,\text{PR}}$	**0.94**	**0.95**	**0.94**	**0.94**		**0.95**

Coverage rate from 90% to 100% is typeset in bold font.
CI, confidence interval; REML, restricted maximum likelihood.

residual-alone bootstraps had very poor coverage rates for all parameters (<0.6) except for the SE of σ, whereas the case bootstrap and the bootstraps of both random effects and residuals provided good coverage rates (close to the nominal value of 0.95). According to our predefined criteria, four methods showed good performance: $B_{\text{case,none}}$, $B_{\eta,\text{GR}}$, $B_{\eta,\text{IR}}$, and $B_{\text{P}\eta,\text{PR}}$. These methods remained the best methods for the sparse design and the large error design, with smaller relative bias for the parameter estimates and their SEs and better coverage rates. The simulation results also showed that the asymptotic approach provided good estimates for parameters and SEs as well as good coverage rates as do the bootstrap candidates. In this simple setting, the convergence rate was high: nearly all runs on the simulated datasets converged (respectively, 100%, 99.8%, and 99.9% for the rich, sparse, and large error designs). Convergence was also close to 100% when applied to the bootstrap samples in the rich (100%) and large error (99.9%) design, whereas for the sparse design, convergence rates were slightly lower and more dependent on the bootstrap method, going from around 90.2% for the bootstraps combining case and residuals, to 96.7% for the case bootstrap or the bootstraps of both random effects and residuals.

In this study, we also evaluated the influence of transformation of random effects and residuals using the ratio between the estimated and empirical variance–covariance matrices. Nontransformed and transformed resampling were compared for all nonparametric bootstrap methods except for the $B_{\text{case,none}}$ where no transformation was needed. The results of this transformation in the rich design are presented in Figure 3. We found that these corrections improved significantly the estimate of σ, its SE as well as the coverage rate for all applied methods. However, they only improved the estimates and the coverage rates of other variance parameters for $B_{\eta,\text{GR}}$ and $B_{\eta,\text{IR}}$.

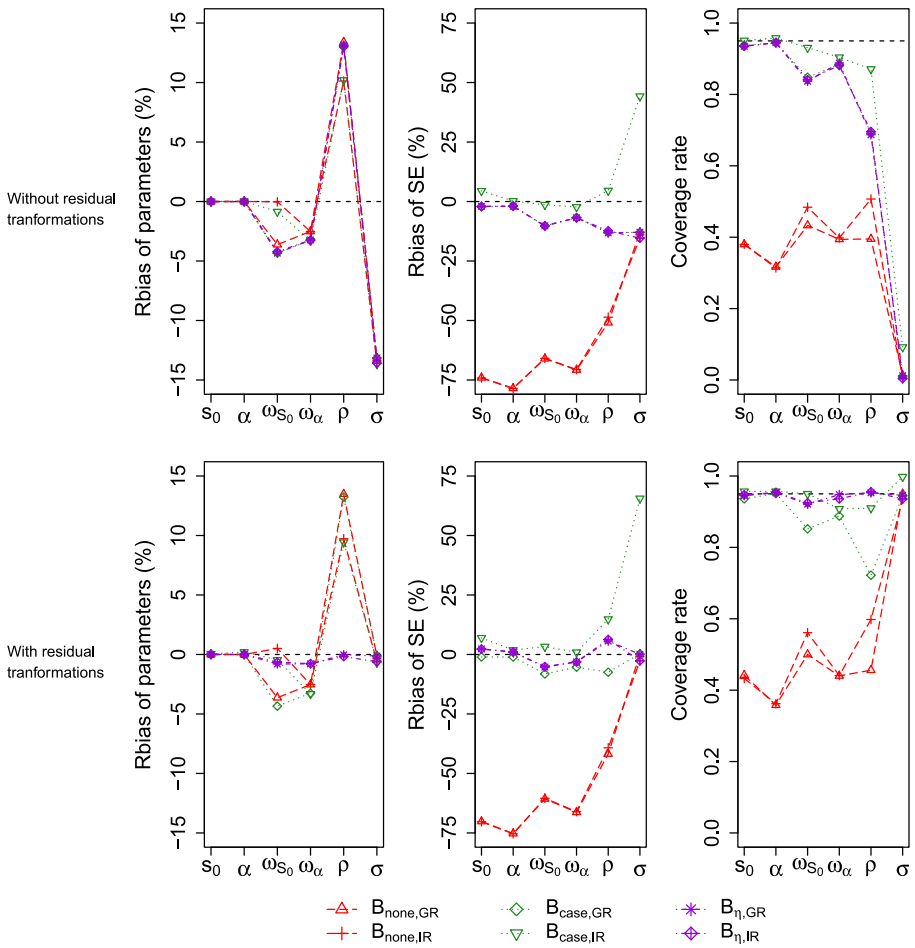

Figure 3. Relative bias of parameter estimates by restricted maximum likelihood (left), relative bias of standard errors (middle), and coverage rate of 95% confidence interval (right), for the nonparametric bootstrap methods without (top) and with (bottom) transformations of random effects and residuals in the rich design ($N = 100$, $n = 7$, $\sigma = 5.86$).

The results for the bootstrap candidates in the different evaluated designs, providing a more contrasted evaluation of their performance, were shown in Figure S1 in the Appendix. All the bootstrap candidates provided good parameter estimates in all evaluated designs with the relative bias within ± 5%, except for a higher bias on σ (−6.50%) observed for the $B_{\eta,\text{IR}}$ in the sparse design. The relative bias for the variance estimates were higher for the sparse and the large error designs compared with those of the rich design, whereas the relative bias of the fixed effects remained very small (<0.1%). All the bootstrap candidates estimated correctly the SE of parameters, with relative biases ranging from −5% to 5%, observed in both rich and large error designs. They estimated less correctly the SE of variance parameters in the sparse design with relative biases ranging from −10% to 10%, especially for the case bootstrap. The boxplots of the SEs of all parameter estimates obtained by the bootstrap candidates in all evaluated designs are shown in Figure S2 in the Appendix. The range of bootstrap SEs across the $K = 1000$ replications did not show any practically relevant differences across the bootstrap methods, with the exception of SE of σ in the sparse design. A good coverage rate was obtained for all parameters in the rich design setting, as well as the design with large error; the one exception is a low coverage rate for σ observed for $B_{\eta,\text{IR}}$ in the design with large error. In the sparse design, the bootstrap candidates provided lower coverage rate for variance parameters. Across the different designs, the $B_{\text{P}\eta,\text{PR}}$ worked slightly better than the $B_{\text{case,none}}$ and the $B_{\eta,\text{GR}}$ performed better than the $B_{\eta,\text{IR}}$.

To understand whether the estimation method influences the performance of the bootstrap methods, we compared the results of ML with those of REML. All the results of ML are given in the Appendix with Table S1 and Table S2. The difference between ML and REML in relative bias of parameter estimates and their SEs and coverage rate for all bootstrap methods and the asymptotic method in three evaluated designs are shown in Figure S3 in the Appendix. There was no difference in estimation of fixed effects and σ between two methods. However, the variance parameters (ω_{S_0} and ω_α) given by ML were slightly less well estimated compared with those of REML with increase of 0.5% to 2% in relative bias and had lower coverage rate in all evaluated designs. The difference between ML and REML in estimation of variance parameters was more apparent in the sparse design, resulting in bigger difference in their coverage rates (2% to 5%) observed for all bootstrap methods. In terms of SE estimation, ML and REML provided similar estimates, except for a small difference in relative bias (<2%) in the sparse design. On the basis of our criteria about relative bias, REML was better than ML by providing unbiased estimates for almost parameters obtained by four bootstrap candidates. REML also improved the coverage rate of variance parameters in the sparse design while it provided similar estimates of SEs compared with ML.

Table IV. Parameter estimates by REML and their standard errors (SE) obtained by the asymptotic method, the bootstrap candidates, and the stratified bootstrap in the real dataset.

	Parameter estimates						SE					
Method	S_0	α	ω_{S_0}	ω_α	ρ	σ	S_0	α	ω_{S_0}	ω_α	ρ	σ
Asymptotic (ML)	23.99	13.97	11.08	12.80	0.63	5.86	1.12	1.40	0.84	1.44	0.08	0.17
Asymptotic (REML)	23.98	14.01	11.13	12.93	0.63	5.86	1.12	1.41	0.85	1.46	0.08	0.17
$B_{\text{case,none}}$	24.02	14.10	11.10	12.99	0.63	5.81	1.11	1.82	0.78	2.04	0.10	0.54
$B^{\text{strat}}_{\text{case,none}}$	24.01	14.11	11.07	13.03	0.63	5.84	1.02	1.58	0.77	1.80	0.10	0.54
$B_{\eta,\text{GR}}$	23.97	14.02	10.64	11.68	0.78	5.84	1.08	1.27	0.74	0.94	0.08	0.37
$B_{\eta,\text{IR}}$	24.07	14.01	10.70	11.87	0.74	5.77	1.10	1.33	0.78	1.10	0.11	0.35
$B_{\text{P}\eta,\text{PR}}$	23.97	14.01	11.07	12.83	0.63	5.86	1.13	1.42	0.82	1.10	0.09	0.16

REML, restricted maximum likelihood; ML, maximum likelihood.

4.2. Application to real data

The asymptotic parameter estimates by ML and their SEs for the real dataset were given in the first row of Table IV; these were taken as the true parameter values used in the simulation study reported earlier. The parameter estimates by REML and the SEs of all parameters obtained by the asymptotic method and the four bootstrap candidates for the real dataset are also shown in Table IV. They provided similar values for both fixed-effect and variance parameters that were also close to the asymptotic estimates. However, there were some differences for the estimation of SE of α, ω_α, and σ between the bootstrap methods. The $B_{\text{case,none}}$ yield the highest SE for these parameters, whereas the $B_{\text{P}\eta,\text{PR}}$ and the asymptotic estimates were very similar, which were both different from the remaining bootstrap methods. This finding was not observed in the simulation study, where balanced designs with the same number of observations per subject were used and the residuals were normally distributed. The distribution of the empirical residuals in the real dataset was investigated and a nonnormal distribution was confirmed by the Shapiro–Wilk normality test. The similar results of ML for the real data are also presented in Table S3 in the Appendix.

To investigate the effect of the unbalance in the design, we stratified the case bootstrap based on the number of observations per subject (nobs). The real dataset was divided into three groups before bootstrapping: group 1 ($N = 40$ subjects with nobs $\leq$5), group 2 ($N = 38$ subjects with nobs $>$5 and $\leq$10), and group 3 ($N = 31$ subjects with nobs $>$10). The bootstrap samples were then built by sampling subjects within each group, keeping the same number of subjects from each group. The results obtained from this analysis are also given in Table IV. The case bootstrap with stratification $B^{\text{strat}}_{\text{case,none}}$ gave similar values for parameters as the case bootstrap and other bootstrap methods. In terms of SE of parameters, this method provided smaller values for SEs of α and ω_α and reduced the difference on SE estimation of these parameters with other bootstrap methods, which may indicate that the case bootstrap is more sensitive to unbalanced designs; on the other hand, stratification hardly affected the SE of σ.

5. DISCUSSION

In this paper, we evaluated different bootstrap approaches for estimating SEs and CIs of parameters in LMEMs with homoscedastic error. The proposed bootstraps take into account two levels of variability in the longitudinal data: BSV and RUV. They were also compared with the residual bootstrap, which resamples only one level of variability, and to the case bootstrap where the whole vector of observations is resampled.

Our simulations showed that bootstrapping only residuals underestimated greatly the SEs of parameters, except for σ, and provided poor coverage rates. This finding is to be expected because the large BSV for two parameters in the evaluated designs were not taken into account. On the contrary, the case bootstrap performed well as it provided nonbiased parameter estimates and SEs as well as good coverage rates for all parameters. Moreover, according to Van der Leeden *et al.*, it makes sense to resample only the individuals and collect the related observations from the original dataset when bootstrapping cases for the repeated measures data [14, 26]. Our results support the implication that the RUV is somewhat taken into account in this method; resampling cases preserves both the BSV and the RUV. It is in agreement with the worse performance of combining residual bootstrap with case bootstrap, in which the RUV is considered already resampled.

Another important result of this study is the good performance of bootstrapping both random effects and residuals, either in a nonparametric or in a parametric way. They worked as well as the case bootstrap. The incorporation of random effects into the classical residual bootstrap plays therefore a very important role for bootstrapping in mixed-effects models context, especially when the BSV is much higher than the RUV. This approach was proposed in various studies, such as for the resampling of multi-level data [14, 23], times series data [10], and more recently for longitudinal data [15]. It can be an alternative method to the case bootstrap when the model is correct (for nonparametric version) or both model and assumptions on distributions of parameters are correct (for parametric version). To our knowledge, there has not been any simulation study in the literature that compares this approach to the most commonly used method, the case bootstrap for mixed-effects models with longitudinal data.

Bootstrapping the raw random effects and residuals does not take into account variance underestimation, leading to shrinkage in the individual parameter estimates. To account for this issue, we employed the correction using the ratio between estimated and empirical variance–covariance matrix for the random effects and the residuals. It was shown to be an appropriate method for LMEMs because of the improvement of estimation for variance components. These ratios account for the degree of two shrinkages: η-shrinkage and ϵ-shrinkage, which quantify the amount of information in the individual data about the parameters [27–29].

When the data is not informative, the random effects and residuals are shrunk toward 0, and high degree of η-shrinkage and ϵ-shrinkage will be obtained. Sampling in the raw distribution will therefore underestimate the actual level of variability in the data, whereas correcting both empirical random effects and residuals for shrinkage restores this level. This idea of accounting for the difference between the estimated and empirical variance of residuals through an estimate of the shrinkage was proposed in bootstrapping ordinary linear models [21] and was extended for the two levels of variability found in mixed models by Wang *et al.* [23].

The performance of different bootstrap methods was evaluated under different conditions: the rich design in which all parameters can be well estimated, the sparse design in which the variance parameters are less well estimated, and the large error design in which the RUV is as important as the BSV to see whether the residual bootstraps work better. The convergence was obtained for almost all bootstrap datasets (90% to 100%), which should not provide substantial bias for estimating the uncertainty of parameters. The case bootstrap and three bootstrap methods where both random effects and residuals were resampled remained the best methods and selected as bootstrap candidates for LMEMs. The purpose of this work was not to determine which was the best method overall, but to eliminate bootstrap methods that do not perform well even with LMEMs. We did note that the global residual bootstrap was slightly better than individual residual bootstrap in the sparse and large error designs, especially in estimating σ, which is consistent with the noncorrelated structure of residuals. In addition, the distribution of resampled residuals obtained by the global residual bootstrap was slightly closer to the original distribution of residuals. The parametric bootstrap performed best across three evaluated designs, but it requires the strongest assumptions (good prior knowledge about model structure and distributions of parameters). If the model is misspecified and the assumptions of normality of random effects and residuals are not met, this method may not be robust. In practice, one of the main reasons for using bootstrap is the uncertainty of distribution assumption, the nonparametric bootstrap may therefore preferable to the parametric bootstrap in most applications [9].

We also investigated using ML instead of REML when performing the estimation step. The difference in performance was small with less than 5%, but there was slightly more bias with ML especially for the estimation of the variance parameters, which was most apparent in the sparse design where variances were less well estimated. This means that we may expect this to be true also for nonlinear mixed effect models where ML is more often used than REML for parameter estimation, although this needs to be verified in the nonlinear setting. This finding should be also explored further in other settings, because the superiority of REML over ML becomes more apparent as the number of fixed effect increases, especially when the number of subjects is limited [3].

The number of bootstrap replicates (B) depends on the estimation that we want to obtain. In simple regression models, B is recommended to be at least 100 for SE estimation and at least 1000 in the case of CI estimation [30, 31]. In this simulation study, 1000 bootstrap replicates were thought to be large enough to obtain both bootstrap SE and bootstrap CI for all bootstrap methods with LMEMs. Note that, to estimate directly the quantiles for 95% CI without interpolation, $B = 999$ should be used instead of 1000 in the future work [7]. In addition, further evaluation on choosing the number of bootstrap will be studied, especially when bootstrapping on nonLMEMs is much time-consuming and less stable.

Although in the simulation studies the performance of the case bootstrap was similar to that of the bootstrap methods resampling both random effects and residuals, when these methods were applied to the real dataset, the case bootstrap estimated a much larger SE for both α and its variability, and there was also smaller differences in the estimates of σ between the different bootstraps. We found a good agreement between the $B_{\mathrm{P}\eta,\mathrm{PR}}$ and the asymptotic method, which were both different from the remaining methods. The difference between $B_{\mathrm{P}\eta,\mathrm{PR}}$ and $B_{\eta,\mathrm{GR}}$ could come from different assumptions on distributions, because for the former, we sample in a normal distribution, whereas the empirical residuals used for resampling in $B_{\eta,\mathrm{GR}}$ were in fact not normally distributed. The difference between $B_{\mathrm{case,none}}$ and $B_{\eta,\mathrm{GR}}$ or $B_{\eta,\mathrm{IR}}$ might be due to the unbalanced design of the real dataset in which patients have different number of observations, a structure which is preserved by the residual bootstraps but not by the case bootstrap. Unbalanced designs, therefore, may be more challenging for the case bootstrap than for other bootstrap methods [7, 32], and stratification has been proposed to handle such situations; for example, when a study includes rich and sparse data, it is recommended to resample from both groups to maintain a similar structure in the bootstrap samples [33]. In this study, we tried to apply the stratified case bootstrap to the real dataset. The stratification could explain a part of the difference between the case bootstrap and other bootstrap candidates by decreasing the SEs of α and ω_α to the values closer to those obtained by other methods.

In conclusion, in this study, we found that the case bootstrap performs as well as the nonparametric/parametric bootstrap of random effects and residuals in LMEMs with balanced designs and homoscedastic error. However, the residual bootstraps always generate datasets with the same design as the original data and would be expected to perform better in situations where the design is not similar for every individual. This could be the explanation for the discrepancy between the bootstraps seen in the application to the real data from the Parkinson study. We now plan to compare these methods for nonLMEMs, addressing the issues of heteroscedasticity, the nonlinearity of model, and exploring the influence of designs with stratified bootstrap.

Acknowledgements

The authors would like to thank the Parkinson Study Group for providing us the data from Parkinson's study as well as IFR02 and Hervé Le Nagard for the use of the 'centre de biomodélisation'. We also thank the reviewer for a careful review and insightful comments that helped us to improve this paper.

During this work, Hoai-Thu Thai was supported by a research grant from Drug Disposition Department, Sanofi, Paris.

REFERENCES

[1] Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. John Wiley & Sons: New York, 2004.
[2] Pinheiro JC, Bates DM. *Mixed-effects models for s and s-plus*. Springer: New York, 2000.
[3] Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. Springer: New York, 2000.
[4] Efron B. Bootstrap methods: Another look at the jacknife. *The Annals of Statistics* 1979; **7**(1):1–26.

[5] Efron B, Tibshirani RJ. *An introduction to the bootstrap*. Chapman & Hall: New York, 1994.
[6] Shao J, Tu D. *The jacknife and bootstrap*. Springer: New York, 1995.
[7] Davison AC, Hinkley DV. *Bootstrap methods and their application*. Cambridge University Press: Cambridge, 1997.
[8] MacKinnon JB. Bootstrap methods in econometrics. *The Economic Record* 2006; **82**(s1):S2–S18.
[9] Wehrens R, Putter H, Buydens Lutgarde M C. The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems* 2000; **54**(1):35–52.
[10] Ocana J, El Halimi R, Ruiz de Villa MC, Sanchez JA. Bootstrapping repeated measures data in a nonlinear mixed-models context. *Mathematics Preprint Series* 2005; **367**:1–31.
[11] Flachaire E. Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis* 2005; **49**(2):361–376.
[12] Halimi R. *Nonlinear mixed-effects models and bootstrap resampling: Analysis of non-normal repeated measures in biostatistical practice*. VDM Verlag Dr. MÃijller: Berlin, 2005.
[13] Das S, Krishen A. Some bootstrap methods in nonlinear mixed-effect models. *Journal of Statistical Planning and Inference* 1999; **75**(2):237–245.
[14] Van der Leeden R, Busing FMTA, Meijer E. Bootstrap methods for two-level models. *Technical Report PRM 97-04*, Leiden University, Department of Psychology, Leiden, 1997.
[15] Wu H, Zhang J-T. The study of long-term hiv dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine* 2002; **21**(23):3655–3675.
[16] Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**(4):963–974.
[17] Ette EI. Stability and performance of a population pharmacokinetic model. *Journal of Clinical Pharmacology* 1997; **37**(6):486–495.
[18] Wu CFJ. Jacknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* 1986; **14**(4):1261–1295.
[19] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine* 2000; **19**(9):1141–1164.
[20] Carpenter JR, Goldstein H, Rasbash J. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics* 2003; **52**:431–443.
[21] Moulton LH, Zeger SL. Bootstrapping generalized linear models. *Computational Statistics and Data Analysis* 1991; **11**(1):53–63.
[22] Efron B. *The jacknife, the bootstrap and other resampling plans*. Society of Industrial and Applied Mathematics: Philadelphia, 1982.
[23] Wang J, Carpenter JR, Kepler MA. Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine* 2006; **82**(2):130–143.
[24] Holford NHG, Chan PLS, Nutt JG, Kieburtz K, Shoulson I, Parkinson Study Group. Disease progression and pharmacodynamics in Parkinson disease–evidence for functional protection with levodopa and other treatments. *Journal of Pharmacokinetics and Pharmacodynamics* 2006; **33**(3):281–311.
[25] Bates DM, Pinheiro JC. Computational methods for multilevel models. *Technical Memorandum BL0112140-980226-01TM*, Bell Labs, Lucent Technologies, Murray Hill, NJ, 1998.
[26] De Leeuw J, Meijer E. *Handbook of multilevel analysis*, chap. Resampling multilevel models. Springer: New York, 2007.
[27] Ette EI, Williams PJ. *Pharmacometrics: the science of quantitative pharmacology*. Wiley-Blackwell: New Jersey, 2007.
[28] Karlsson MO, Savic RM. Diagnosing model diagnostics. *Clinical Pharmacology and Therapeutics* 2007; **82**(1):17–20.
[29] Savic RM, Karlsson MO. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *The AAPS Journal* 2009; **11**(3):558–569.
[30] Chernick MR. *Bootstrap methods: A guide for practitionners and researchers*, (2nd edn). John Wiley & Sons: New Jersey, 2007.
[31] Bonate PL. *Pharmacokinetic-pharmacodynamic modeling and simulation*, (2nd edn). Springer: New-York, 2011.
[32] Davison AC, Kuonen D. An introduction to the bootstrap with applications in R. *Statisticial Computing and Statistical Graphics Newsletter* 2002; **13**(1):6–11.
[33] Lindbom L, Pihlgren P, Jonsson EN. PsN-Toolkit–A collection of computer intensive statistical methods for non-linear mixed effect modeling using NONMEM. *Computer Methods and Programs in Biomedicine* 2005; **79**(3):241–257.