# Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm.

Anne Dubois, Marc Lavielle, Sandro Gsteiger, Etienne Pigeolet, France Mentré

# Model-Based Analyses of Bioequivalence Crossover Trials Using the SAEM Algorithm

Anne Dubois, Marc Lavielle, Sandro Gsteiger, Etienne Pigeolet and France Mentré

INSERM UMR738,University Diderot Paris 7, Paris, France, 75018

INRIA Saclay, Orsay, France, 91400

Modeling and Simulation Department, Novartis Pharma AG, Basel, Switzerland, CH-4056

email: anne.dubois@inserm.fr

**Author's Footnote:**

Anne Dubois is a PhD student (anne.dubois@inserm.fr), INSERM UMR738, University Diderot Paris 7, Paris, France, 75018; Marc Lavielle is Professor of Mathematics (marc.lavielle@math.u-psud.fr), INRIA Saclay, Orsay, France, 91400; Sandro Gsteiger is a Senior Statistical Modeler (sandro.gsteiger@novartis.com), and Etienne Pigeolet is a Senior Expert Pharmacology Modeler (etienne.pigeolet@novartis.com), Modeling and Simulation Department, Novartis Pharma AG, Basel, Switzerland, CH-4056; France Mentré is Professor of Biostatistics (france.mentre@inserm.fr), INSERM UMR738, University Diderot Paris 7, Paris, France, 75018. The authors thank the Modeling and Simulation Department, Novartis Pharma AG, Basel, which supported by a grant Anne Dubois during this work, and Sigrid Balser from Sandoz Biopharmaceutical Development, Holzkirchen, Germany who provided the data set used in the example.

**Abstract**

In this work, we develop a bioequivalence analysis using nonlinear mixed effects models (NLMEM) that mimics the standard non-compartmental analysis (NCA). NLMEM parameters, including between (BSV) and within subject (WSV) variability, and treatment, period, and sequence effects are estimated. We explain how to perform a Wald test on a secondary parameter and we propose an extension of the likelihood ratio test (LRT) for bioequivalence. These NLMEM-based bioequivalence tests are compared to standard NCA-based tests. We evaluate by simulation the NCA and NLMEM estimates, and the type I error of the bioequivalence tests. For NLMEM, we use the SAEM algorithm implemented in MONOLIX. Crossover trials are simulated under $H_0$ using different numbers of subjects and of samples per subject. We simulate with different settings for BSV and WSV, and for the residual error variance. The simulation study illustrates the accuracy of NLMEM-based geometric means estimated with the SAEM algorithm, whereas the NCA estimates are biased for sparse design. NCA-based bioequivalence tests show good type I error except for high variability. For a rich design, type I errors of NLMEM-based bioequivalence tests (Wald test and LRT) do not differ from the nominal level of 5%. Type I errors are inflated for sparse design. We apply the bioequivalence Wald test based on NCA and NLMEM estimates to a three-way crossover trial, showing that Omnitrope® powder and solution are bioequivalent to Genotropin®. NLMEM-based bioequivalence tests are an alternative to standard NCA-based tests. However, caution is needed for small sample size and highly variable drug.

KEYWORDS: nonlinear mixed effects model; pharmacokinetics; non-compartmental bioequivalence analysis; two one-sided tests; Wald test; likelihood ratio test

# 1 INTRODUCTION

Pharmacokinetic (PK) bioequivalence studies are performed to compare different drug formulations. The most commonly used design for bioequivalence trials is the two-period, two-sequence, crossover design. This design is recommended by the Food and Drug Administration (FDA) (FDA 2001) and the European Medicines Evaluation Agency (EMEA) (EMEA 2001). FDA and EMEA recommend to test bioequivalence from the ratios of the geometric means of two parameters: the area under the curve ($AUC$) and the maximal concentration ($C_{max}$) estimated by non-compartmental analysis (NCA) (Gabrielson and Weiner 2006). As specified in the regulatory guidelines, the bioequivalence analysis should take into account sources of variation that can be reasonably assumed to have an effect on the endpoints $AUC$ and $C_{max}$. Therefore, linear mixed effects models (LMEM) including treatment, period, sequence, and subject effects are usually used to analyse the log-transformed individual parameters (Hauschke et al. 2007). Bioequivalence tests are then performed on the estimates of the treatment effect.

NCA requires few hypotheses but a large number of samples per subject (usually between 10 and 20). PK data can also be analysed using nonlinear mixed effects models (NLMEM). This method is more complex than NCA but has several advantages: it takes advantage of the knowledge accumulated on the drug and can characterize the PK with few samples per subject. This allows one to perform analyses in patients, the target population, in whom pharmacokinetics can be different from healthy subjects. In a previous work, Dubois et al (Dubois et al. 2010) compared the standard analysis of bioequivalence crossover trials based on NCA to the same usual analysis based on individual empirical Bayes estimates (EBE) obtained by NLMEM. PK data of each treatment group were analysed separately using NLMEM. Linear mixed effects models were then performed on individual $AUC$ and $C_{max}$ derived from the EBE. However, this methodology cannot be performed when the EBE shrinkage is above 20%. Panhard and Mentré (Panhard and Mentré 2005) developed different comparison and bioequivalence tests based on NLMEM for the analysis of PK crossover trials

comparing two treatments. For comparison tests, they proposed both the Wald test and the likelihood ratio test (LRT). For bioequivalence tests, they proposed the Wald test but the LRT was not developed, due to the composite null hypothesis. They applied these tests to two-period, one-sequence, crossover trials. In a later work, Panhard et al (Panhard et al. 2007) demonstrated the importance of modelling the between-subject (BSV) and within-subject (WSV) variability to control the inflation of the type I error using the same sets of simulations as previously. In both simulation studies, the NLMEM-based bioequivalence Wald test was performed on $AUC$ only because $C_{max}$ was a secondary parameter of the PK model, as often in PK modelling. The use of NLMEM is still rare to analyse bioequivalence crossover trials. Indeed, there are few published studies which use NLMEM to analyse bioequivalence trials (Kaniwa et al. 1990; Pentikis et al. 1996; Combrink et al. 1997; Maier et al. 1999; Hu et al. 2003; Zhou et al. 2004; Fradette et al. 2005; Zhu et al. 2008). Seven of these studies are previous to the different simulation studies (Kaniwa et al. 1990; Pentikis et al. 1996; Combrink et al. 1997; Maier et al. 1999; Hu et al. 2003; Zhou et al. 2004; Fradette et al. 2005). In six of these studies (Kaniwa et al. 1990; Pentikis et al. 1996; Combrink et al. 1997; Maier et al. 1999; Hu et al. 2003; Fradette et al. 2005), bioequivalence Wald tests were performed on treatment effects estimated by NLMEM, as Panhard et al. However, all of these applied works used different statistical approaches. The addition of the treatment effect on different PK parameters was not always justified. Only Hu et al (Hu et al. 2003) performed bioequivalence test on average $AUC$ and $C_{max}$ obtained from the fixed effect estimates using Monte Carlo simulation. Finally, none estimated the WSV or adds period or sequence effects as recommended in the guidelines. There is currently no published simulation study or applied work which takes into account treatment, period, sequence effects, BSV, and WSV for the bioequivalence analysis of crossover trials by NLMEM.

In the different NLMEM-based bioequivalence analysis, NLMEM parameters were estimated by maximum likelihood. Except for the simulation by Dubois et al (Dubois et al. 2010), an algorithm based on a first-order linearization was used, most often the

First-Order Conditional Estimation (FOCE) algorithm (Lindstrom and Bates 1990). The FOCE algorithm is a widely used algorithm and corresponds to the industry standard for model-based PK analyses. Yet, this linearization-based method cannot be considered as fully established theoretically. For instance, Vonesh (Vonesh 1996) and Ge et al (Ge et al. 2004) gave examples of specific designs resulting in inconsistent estimates, such as when the number of observations per subject does not increase faster than the number of subjects or when the variability of random effects is too large. Several estimation methods of maximum likelihood theory have been proposed as alternatives to linearization algorithms like the adaptative gaussian quadrature (AGQ) method (Pinheiro and Bates 1995) or methods derived from the Expectation-Maximisation (EM) algorithm (Dempster et al. 1977). The AGQ method requires a sufficiently large number of quadrature points implying an often slow convergence and is limited to a small number of random effects. Monte Carlo EM algorithms as proposed by Wei and Tanner (Wei and Tanner 1990), Walker (Walker 1996), or Wu (Wu 2004) are very time-consuming in computation since they require a huge amount of simulated data. Alternatively, Delyon et al (Delyon et al. 1999) introduced a stochastic approximation version of the EM algorithm (SAEM), which is more efficient in terms of computation. Later, Kuhn and Lavielle (Kuhn and Lavielle 2004) developed an algorithm which combined the SAEM algorithm with a Monte-Carlo procedure. They showed the good statistical convergence properties of this algorithm. Recently, Panhard and Samson (Panhard and Samson 2009) developed an extension of the SAEM algorithm for NLMEM including the estimation of the within-subject variability.

The main objective of this work is to develop a bioequivalence analysis based on NLMEM that mimics the standard bioequivalence analysis performed on NCA estimates. To do so, we use a NLMEM including treatment, period, sequence effects, BSV, and WSV. We also explain how to perform a Wald test on a secondary parameter of the model (like $C_{max}$), and we propose an extension of the LRT for bioequivalence. These NLMEM-based bioequivalence tests are compared to standard

NCA-based tests. We evaluate by simulation the NCA and NLMEM estimates, and the type I errors of the different bioequivalence tests. We use the same sets of simulations than the previous study by Dubois et al (Dubois et al. 2010) which allows to compare the results. As in Dubois et al, we use different sampling designs and levels of variability, investigating their influence on the results of the bioequivalence tests. To estimate NLMEM parameters, we use the SAEM algorithm implemented in MONOLIX. Then, we apply the Wald test based on NCA and NLMEM estimates to a three-way crossover trial comparing three formulations of somatropin. A somatropin is a biosynthetic version of human growth hormone (hGH) synthesised in bacteria modified by the addition of the gene for hGH. Replacement therapy with somatropin is a well accepted, effective treatment for hGH deficiency in children and adults (Fauci et al. 2008).

In Section 2 of this paper, we describe the LMEM for NCA estimates, the NLMEM on concentrations, and the bioequivalence tests based on both approaches. In Section 3, we present the simulation study, the estimation, and the evaluation of the estimates and of the type I errors. We present the example in Section 4, followed by a discussion in Section 5.

## 2 MODELS AND BIOEQUIVALENCE TESTS IN CROSSOVER TRIALS

In the following, we consider crossover pharmacokinetic trials with $C$ treatments, $K$ periods, and $Q$ sequences.

### 2.1 Models

**Linear Mixed Effects Model for NCA**

The standard bioequivalence analysis recommended by FDA and EMEA (FDA 2001; EMEA 2001) is based on NCA individual estimates of $AUC$ and $C_{max}$. We define $\theta_{ikl}$ the $l^{th}$ individual parameter ($AUC$ if $l = 1$ or $C_{max}$ if $l = 2$) for subject

$i$ ($i = 1, \cdots, N$) at period $k$ ($k = 1, \cdots, K$). The individual parameters are log-transformed and analysed using a linear mixed effects model written as follows:

$$\log(\theta_{ikl}) = \nu_l + \boldsymbol{\beta_l^{T\prime}}\,\boldsymbol{T_{ik}} + \boldsymbol{\beta_l^{P\prime}}\,\boldsymbol{P_k} + \boldsymbol{\beta_l^{S\prime}}\,\boldsymbol{S_i} + \eta_{il} + \epsilon_{ikl} \tag{1}$$

where $\nu_l$ is the expected value corresponding to the combination of covariate reference classes. $\boldsymbol{\beta_l^T}$, $\boldsymbol{\beta_l^P}$, and $\boldsymbol{\beta_l^S}$ are the vectors of coefficients of treatment, period, and sequence effects for the $l^{th}$ individual parameter ($AUC$ or $C_{max}$). $\boldsymbol{T_{ik}}$, $\boldsymbol{P_k}$, $\boldsymbol{S_i}$ are the known vectors of treatment, period, and sequence covariates of size $C$, $K$, and $Q$, respectively. $\boldsymbol{T_{ik}}$ is composed of zeros except for the $c^{th}$ element ($c = 1, \cdots, C$) which is one when treatment $c$ is given to patient $i$ at period $k$. Similarly, $P_k$ and $S_i$ are composed of zeros except for the $k^{th}$ and $q^{th}$ elements ($k = 1, \cdots, K$, $q = 1, \cdots, Q$) which are one. We consider that the first treatment, period, and sequence are the reference classes. The first elements of $\boldsymbol{\beta_l^T}$, $\boldsymbol{\beta_l^P}$, and $\boldsymbol{\beta_l^S}$ are fixed to zero, and other components are estimated. It is assumed that the random subject effect $\eta_{il}$ and the residual error $\epsilon_{ikl}$ are independently normally distributed with zero mean.

**Nonlinear Mixed Effects Modelling**

We denote by $y_{ijk}$ the concentration for subject $i$ ($i = 1, \cdots, N$) at sampling time $j$ ($j = 1, \cdots, n_{ik}$) for period $k$ ($k = 1, \cdots, K$). We define $f$ to be the nonlinear pharmacokinetic function which links concentrations to sampling times. The nonlinear mixed effects model can be written as:

$$y_{ijk} = f(t_{ijk}, \boldsymbol{\psi_{ik}}) + g(t_{ijk}, \boldsymbol{\psi_{ik}})\,\epsilon_{ijk} \tag{2}$$

with $\boldsymbol{\psi_{ik}}$ the $p$-vector of pharmacokinetic parameters of subject $i$ for period $k$. $g(t_{ijk}, \boldsymbol{\psi_{ik}})\,\epsilon_{ijk}$ is the residual error where $\epsilon_{ijk}$ is a Gaussian random variable with zero mean and variance one. All $\epsilon_{ijk}$ are independent and identically distributed. We consider a combined error model, additive plus proportional, with $g(t_{ijk}, \boldsymbol{\psi_{ik}}) =$

$a + bf(t_{ijk}, \boldsymbol{\psi_{ik}})$.

The statistical model used for the individual parameters $\boldsymbol{\psi_{ik}}$ is derived from the linear mixed effects model used to analyse the NCA individual estimates. So, the $l^{th}$ component of $\boldsymbol{\psi_{ik}}$ is defined as:

$$\log(\psi_{ikl}) = \log(\lambda_l) + \boldsymbol{\beta_l^{T\prime}}\boldsymbol{T_{ik}} + \boldsymbol{\beta_l^{P\prime}}\boldsymbol{P_k} + \boldsymbol{\beta_l^{S\prime}}\boldsymbol{S_i} + \eta_{il} + \kappa_{ikl} \tag{3}$$

with $\boldsymbol{\lambda} = (\lambda_l; l = 1, \cdots, p)$ the $p$-vector of fixed effects for the covariate reference classes. The known vectors of the treatment, period, and sequence covariates, $\boldsymbol{T_{ik}}$, $\boldsymbol{P_k}$, and $\boldsymbol{S_i}$, are defined as for NCA (section 2.1). $\boldsymbol{\beta_l^T}$, $\boldsymbol{\beta_l^P}$, and $\boldsymbol{\beta_l^S}$ are the vectors of coefficients of treatment, period, and sequence effects for the $l^{th}$ PK parameter. As previously mentioned, we consider that the first treatment, period, and sequence are the reference classes. The first elements of $\boldsymbol{\beta_l^T}$, $\boldsymbol{\beta_l^P}$, and $\boldsymbol{\beta_l^S}$ are fixed to zero, and other components are estimated. $\boldsymbol{\eta_i} = (\eta_{il}; l = 1, \cdots, p)$ is the vector of random effects of subject $i$ corresponding to the between-subject variability. $\boldsymbol{\kappa_{ik}} = (\kappa_{ikl}; l = 1, \cdots, p)$ is the vector of random effects of subject $i$ at period $k$ corresponding to the variability between periods of treatment for the same individual, or within-subject variability. These random effects are assumed to be normally distributed with zero mean and covariance matrix of size $p \times p$ named $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$, respectively. We define $\omega_l^2$ and $\gamma_l^2$ the variance for BSV and WSV of the $l^{th}$ parameter, corresponding to the $l^{th}$ element of the diagonal of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$. $\boldsymbol{\eta_i}$, $\boldsymbol{\kappa_{ik}}$, and $\epsilon_{ijk}$ are assumed to be mutually independent. Finally, the unknown population parameters of the statistical model are the fixed effects ("reference" and covariate effects) and the variance parameters ($\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}$, $a$, $b$).

### 2.2 Two-One Sided Tests

The bioequivalence test is performed on the $c^{th}$ treatment effect of the $l^{th}$ parameter, $\beta_{c,l}^T$ ($c = 2, \cdots, C$ and $l = 1, 2$ for NCA or $l = 1, \cdots, p$ for NLMEM). Its null hypothesis is $H_0$: $\{\beta_{c,l}^T \leq -\delta$ or $\beta_{c,l}^T \geq \delta\}$ which is decomposed in two one-sided hypotheses $H_{0,-\delta}$: $\{\beta_{c,l}^T \leq -\delta\}$ and $H_{0,\delta}$: $\{\beta_{c,l}^T \geq \delta\}$. The bioequivalence test is based

on Schuirmann's two one-sided tests (TOST) procedure (Schuirmann 1987). $H_{0,-\delta}$ and $H_{0,\delta}$ are tested separately by a one-sided test. The global null hypothesis $H_0$ is rejected with a type I error $\alpha$ if both one-sided hypotheses are rejected with a type I error $\alpha$. The p-value of the TOST is the maximum of both p-values of the one-sided tests. The major issue of a bioequivalence test is to define $\delta$. To assess pharmacokinetic bioequivalence, the guidelines (FDA 2001; EMEA 2001) recommend $\delta = \log(1.25) \approx 0.22$ (*i.e.* $-\delta = \log(0.8)$) for $\log(AUC)$ and $\log(C_{max})$. Due to the linear model on log-parameters, these bounds corresponds to 80%-125% on the parameter scale.

**Wald Tests Based on NCA Estimates**

In the following, we call $se(\beta_{c,l}^T)$ the standard error of the treatment effect estimate $\widehat{\beta}_{c,l}^T$. We also define $W_{-\delta} = (\widehat{\beta}_{c,l}^T + \delta)/se(\beta_{c,l}^T)$ and $W_{\delta} = (\widehat{\beta}_{c,l}^T - \delta)/se(\beta_{c,l}^T)$, the two Wald statistics for the one-sided hypotheses $H_{0,-\delta}$ and $H_{0,\delta}$, respectively. For the standard NCA-based bioequivalence analysis, we assume that $W_{-\delta}$ and $W_{\delta}$ follow a Student t-distribution with $df$ degrees of freedom under $H_{0,-\delta}$ and $H_{0,\delta}$, respectively. The global null hypothesis $H_0$ is rejected with a type I error $\alpha$ if $W_{-\delta} \geq t_{1-\alpha}(df)$ and $W_{\delta} \leq -t_{1-\alpha}(df)$, where $t_{1-\alpha}(df)$ is the $(1-\alpha)$ quantile of the Student t-distribution with $df$ degrees of freedom. For balanced datasets (*i.e.* with $N$ subjects for each period), $df = N - 2$ (Hauschke et al. 2007; Chow and Liu 2000). An alternative approach to perform a bioequivalence test is to compute the $(1-2\alpha)$ confidence interval (CI) of $\widehat{\beta}_{c,l}^T$. $H_0$ is rejected if this $(1-2\alpha)$ CI lies within $[-\delta; \delta]$.

**Wald Test Based on NLMEM Estimates**

For the bioequivalence Wald test using NLMEM estimates, we use a very similar approach to NCA-based bioequivalence Wald test. Same notations are used for NLMEM-based analyses as for NCA. For NLMEM, we assume that $W_{-\delta}$ and $W_{\delta}$ follow a Gaussian distribution under $H_{0,-\delta}$ and $H_{0,\delta}$, respectively. The global null hypothesis $H_0$ is rejected with a type I error $\alpha$ if $W_{-\delta} \geq z_{1-\alpha}$ and $W_{\delta} \leq -z_{1-\alpha}$,

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution. The rejection of $H_0$ can also be based on the $(1-2\alpha)$ CI as described previously (section 2.2).

To mimic the standard bioequivalence analysis, we would like to perform the NLMEM-based bioequivalence Wald test on $AUC$ and $C_{max}$ which are often secondary parameters of the PK model. So, we propose an approach to perform the NLMEM-based bioequivalence Wald on a secondary parameter. A secondary parameter is a function of the PK parameters of the structural model. Its $c^{th}$ treatment effect is $\beta^T_{c,SP} = h(\boldsymbol{\lambda}, \boldsymbol{\beta^T_c})$ with $h$ the function linking $\beta^T_{c,SP}$ to the PK parameters, $\boldsymbol{\lambda}$ the reference effects, and $\boldsymbol{\beta^T_c}$ the $c^{th}$ treatment effects. To perform a bioequivalence Wald test on the $c^{th}$ treatment effect of a secondary parameter, $\beta^T_{c,SP}$ and its standard error, $se(\beta^T_{c,SP})$, should be estimated. By definition, $\widehat{\beta}^T_{c,SP} = h(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}_{\boldsymbol{c}})$. However, $se(\beta^T_{c,SP})$ cannot be directly computed as $h$ is usually a nonlinear function. We propose to approximate it using the delta method (Oehlert 1992) or simulations. For the delta method, we use the partial derivatives of $h$, the fixed effect estimates $(\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}_{\boldsymbol{c}})$, and their estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$, which is a submatrix of the inverse of the Fisher information matrix estimate: $se(\beta^T_{c,SP}) = \sqrt{\nabla h(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}_{\boldsymbol{c}})' \; \widehat{\boldsymbol{\Sigma}} \; \nabla h(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}_{\boldsymbol{c}})}$. To estimate $se(\beta^T_{c,SP})$ by simulations, we simulate $\beta^T_{c,SP}$ $N_s$ times using a Gaussian distribution with mean the fixed effect estimates $(\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}_{\boldsymbol{c}})$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Then, $se(\beta^T_{c,SP})$ is estimated as the standard deviation of the $N_s$ simulated $\beta^T_{c,SP}$.

**Likelihood Ratio Test Based on NLMEM Estimates**

There is no simple extension of the likelihood ratio test for the composite null hypothesis of a bioequivalence test. Therefore, for a parameter of the PK model, we develop a methodology to perform a NLMEM-based bioequivalence LRT based on profile likelihood methods (Bates and Watts 1988; Meeker and Escobar 1995). Let us define $M_{all}$ to be the NLMEM where all fixed effects are estimateed and $M_{\delta,c,l}$ the NLMEM where $\beta^T_{c,l}$ is fixed to $\delta$ and all other parameters (including the other components of $\boldsymbol{\beta^T_l}$) are estimated. The proposed approach test whether the likelihood-based confidence interval of $\widehat{\beta}^T_{c,l}$ lies within $[-\delta;\delta]$. To do so, we perform two "one-sided"

LRT taking into account $\widehat{\beta}^T_{c,l}$ estimated with $M_{all}$, and the estimation of the log-likelihood of three models $M_{all}$, $M_{-\delta,c,l}$, and $M_{\delta,c,l}$. We define the statistic $\Lambda_{\delta,c,l}$ as follows: $\Lambda_{\delta,c,l} = -2 \times (L_{\delta,c,l} - L_{all})$ with $L_{all}$ and $L_{\delta,c,l}$ the estimated log-likelihoods for the models $M_{all}$ and $M_{\delta,c,l}$, respectively. The null hypothesis $H_{0,-\delta}$ is rejected with a type I error $\alpha$ if $\Lambda_{-\delta,c,l} \geq \chi^2_1(1-2\alpha)$ and $-\delta < \widehat{\beta}^T_{c,l}$, where $\chi^2_1(1-2\alpha)$ is the $(1-2\alpha)$ quantile of the Chi-squared distribution with one degree of freedom. $H_{0,\delta}$ is rejected if $\Lambda_{\delta,c,l} \geq \chi^2_1(1-2\alpha)$ and $\widehat{\beta}^T_{c,l} < \delta$. Consequently, the global null hypothesis $H_0$ is rejected with a type I error $\alpha$ if $-\delta < \widehat{\beta}^T_{c,l} < \delta$ and $\Lambda_{-\delta,c,l} \geq \chi^2_1(1-2\alpha)$ and $\Lambda_{\delta,c,l} \geq \chi^2_1(1-2\alpha)$.

# 3 SIMULATION STUDY

## 3.1 Simulation Settings

We use the concentration data of the anti-asthmatic drug theophylline to define the population PK model for the simulation study. These data are classical in population pharmacokinetics (Pinheiro and Bates 2000) and have been used in previous simulation studies (Panhard and Mentré 2005; Panhard et al. 2007; Dubois et al. 2010). We assume that concentrations can be described by a one-compartment model with first-order absorption and first-order elimination:

$$f(t, k_a, CL/F, V/F) = \frac{FDk_a}{CL - Vk_a} \left(\exp(-k_a\ t) - \exp(-CL/V\ t)\right) \tag{4}$$

where $D$ is the dose, $F$ the bioavailability, $k_a$ the absorption rate constant, $CL$ the clearance of the drug, and $V$ the volume of distribution.

We simulate two-treatment, two-sequence, crossover trials with two or four periods. For each two-period trial, the $N/2$ subjects of the first sequence receive the reference treatment (*Ref*) and the test treatment (*Test*) in period one and two, respectively. The other $N/2$ subjects allocated to the second sequence receive treaments in the reverse order. For each four-period trial, the $N/2$ subjects of the first sequence re-

ceive the treatment *Ref* in periods one and three, and the treatment *Test* in periods two and four. The $N/2$ subjects of the second sequence receive the treament *Test* in periods one and three, and the treatment *Ref* in periods two and four.

We consider that sampling times are similar for all subjects and all periods. So, $j = 1, \cdots, n$, where $n$ is a fixed number of sampling times for each simulated sampling design. We use four different sampling designs, which are also used by Dubois et al (Dubois et al. 2010). We simulate with the original design with $N = 12$ subjects and $n = 10$ samples per subject and per period, taken at the times of the initial study (0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, and 24 $h$ after dosing). We also simulate with an intermediate design with $N = 24$ subjects and $n = 5$ samples, taken at 0.25, 1.5, 3.35, 12, and 24 $h$ after dosing, a sparse design with $N = 40$ subjects and $n = 3$ samples, taken at 0.25, 3.35, and 24 $h$ after dosing, and a rich design $N = 40$ subjects and $n = 10$ samples, taken at the times of the initial study.

For the simulation study, we assume that $\alpha = 5\%$ and $\delta = \log(1.25)$. We fix the dose to 4 mg for all subjects. The vector of population parameters $\boldsymbol{\lambda}$ is composed of $\lambda_{k_a} = 1.48\ h^{-1}$, $\lambda_{CL/F} = 40.36\ mL/h$, and $\lambda_{V/F} = 0.48\ L$ for the reference treatment. We assume that the bioavailability changes between treatments, *i.e.*, we assume the same modification for $CL/F$ and $V/F$. It also similarly affects both secondary parameters $AUC$ and $C_{max}$ with $AUC = FD/CL$ and $C_{max}$ defined in Equation 5 of the Appendix. In the following, as we consider only two treatments in the simulation study, we omit the subscript $c$; we define $\beta^T_{CL/F}$ and $\beta^T_{V/F}$ the treatment effect on $CL/F$ and $V/F$ for the treatment *Test* ($\beta^T_{k_a} = 0$). As suggested by Liu and Weng (Liu and Weng 1995), the type I error of the bioequivalence test can be evaluated for each boundary of $H_0$ space, *i.e.*, $\log(0.8)$ and $\log(1.25)$. Consequently, we simulate under two different hypotheses: $\beta^T_{CL/F} = \beta^T_{V/F} = \log(0.8)$ and $\beta^T_{CL/F} = \beta^T_{V/F} = \log(1.25)$ which are the boundaries of $H_{0,\log(0.8)}$ and $H_{0,\log(1.25)}$, respectively. In the following, we call $H_{0;80\%}$ and $H_{0;125\%}$ these two simulated hypotheses. We assume no period effect or sequence effect, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ are diagonal.

We simulate with two levels of variability for the between-subject and within-subject

variability. For the low level of variability, we fix $\omega_{k_a}$ and $\omega_{CL/F}$ to 0.2, and $\omega_{V/F}$ to 0.1; $\gamma_{k_a}$, $\gamma_{CL/F}$ and $\gamma_{V/F}$ are fixed to half BSV for the three parameters. For the high level, we fix the three standard deviations to 0.5 for BSV, and 0.15 for WSV. We also simulate with two levels of variability for the residual error: $a = 0.1\ mg/L,\ b = 0.10$ for the low level, and $a = 1\ mg/L,\ b = 0.25$ for the high level. The high level of residual error is only used with the high level of BSV and WSV. We call $S_{l,l}$ the variability setting with low variability for BSV, WSV, and for the residual error. $S_{h,l}$ is the variability setting corresponding to high variability for BSV, WSV, and low for the residual error. Finally, $S_{h,h}$ is the variability setting with high variability for BSV, WSV, and for the residual error. In the following, we call a simulation setting the association of one design with one variability setting and one simulated hypothesis. We simulate crossover trials with 2 periods under $H_{0;80\%}$ and $H_{0;125\%}$. In that case, for each sampling design, we simulate using the variability settings $S_{l,l}$ and $S_{h,l}$. We simulate using $S_{h,h}$ only for the intermediate design. We simulate crossover trials with 4 periods under $H_{0;80\%}$ using rich and sparse sampling designs, and the two variability settings $S_{l,l}$ and $S_{h,l}$. All simulations are performed using the statistical software R 2.7.1. We use the function `rmvnorm` of the package `mvtnorm`, which is a pseudorandom number generator for the multivariate normal distribution. For each simulated trial, we specify the seed using the function `set.seed` in order to make simulations reproducible.

## 3.2 Estimation

### NCA

As in Dubois et al (Dubois et al. 2010), we estimate $AUC$ and $C_{max}$ by non compartmental analysis (Gabrielson and Weiner 2006) using a R function which we develop. For a crossover trial, this function provides the estimation of different NCA parameters for each subject and each treatment group. In this study, we use the linear trapezoidal rule to compute the $AUC_{0-last}$ between the time of dose (equal to 0) and the last sampling time. To obtain the total $AUC$ (between the time of dose and in-

finity), we estimate the terminal slope equal to $CL/V$ using the logarithm of the last concentrations to perform a log-linear regression. To do so, we use a fixed number of concentrations which depends on the number of samples per subject in the design. For the original and rich designs where $n = 10$, we use the last four concentrations which correspond to sampling times 7, 9, 12 and 24 $h$. For intermediate and sparse designs where $n = 5$ and $n = 3$ respectively, we use the last two concentrations which correspond to sampling times 12 and 24 $h$ for the intermediate design, and to 3.35 and 24 $h$ for the sparse design. For all designs, $C_{max}$ is estimated as the maximal observed concentration.

The analysis of log parameters by LMEM is then performed using the R function `lme` from the package `nlme`. For the estimation of LMEM parameters (including the treatment effect and its SE), the restricted maximum likelihood (REML) preocedure is used, as recommended in the guidelines (FDA 2001; EMEA 2001). For the original design where $N = 12$, $df = 10$. For the rich and sparse design where $N = 40$, $df = 38$. For the intermediate design where $N = 24$, $df = 22$.

All computations including the dataset simulation, the estimation by NCA, and the standard bioequivalence analysis by LMEM are made under R 2.7.1. The different R scripts are available upon request to the corresponding author.

**NLMEM**

We estimate the NLMEM parameters (included treatment, period, sequence effects, BSV and WSV) by maximum likelihood using the SAEM algorithm (Panhard and Samson 2009) implemented in MONOLIX (Lavielle et al. 2010). Estimation of standard errors (SE) and log-likelihood are also needed to perform Wald and likelihood ratio tests, respectively. SE can be evaluated as the square root of the diagonal elements of the inverse of the Fisher information matrix estimate which has no analytic form. In MONOLIX, one method to evaluate this matrix is to derive an approximate expression by the linearization of the function $f$ around the conditional mean of the individual parameters obtained with the SAEM algorithm. Although linearization-

based algorithms are not recommended to estimate NLMEM parameters, satisfactory results for SE estimation have been shown using this approach for computation of the FIM (Bazzoli et al. 2009). There is no analytical expression of the likelihood in NLMEM. In MONOLIX, it is proposed to estimate the log-likelihood of the observations without approximation using the Importance Sampling (IS) method (Kuhn and Lavielle 2005; Samson et al. 2007). The IS method is a Monte-Carlo procedure where individual parameters are simulated at each iteration using an instrumental distribution adequately chosen to reduce the variance of the estimator.

NLMEM parameters, standard errors and log-likelihoods are estimated with MONOLIX 2.4., supported by MATLAB R2007a. The use of MONOLIX software for the analysis of crossover bioequivalence trials is explained in a online Supplementary Material [1].

### 3.3 Evaluation Methods

**Estimates**

The SAEM algorithm used for the estimation of NLMEM parameters has not been evaluated for models including treatment, period, sequence effects, BSV, and WSV. Therefore, we evaluate the SAEM algorithm for two and four-period crossover trials with the rich and sparse design. We use the $H_{0;80\%}$ simulation settings of the two variability settings $S_{l,l}$ and $S_{h,l}$. There are 8 different simulation settings used (2 or 4 periods, rich or sparse design, $S_{l,l}$ or $S_{h,l}$ variability). We fit the statistical model $M_{all}$ for the 1000 trials of each simulation setting of both types of crossover trials (2 or 4 periods). Then, for the 1000 replicates, we compute the bias and the root mean square error (RMSE) for each estimated parameter.

Furthermore, in the standard bioequivalence analysis, the geometric means of $AUC$ and $C_{max}$ are reported for each treatment group. We evaluate those estimates for the reference treatment, and for NCA and NLMEM. We also evaluate the treatment effect estimates for $AUC$ and $C_{max}$, and for NCA and NLMEM. Lastly, good estima-

[1] Supporting information may be found in the online version of this article.

tion of the standard error is important when performing Wald tests. So, we evaluate the SE of the treatment effect estimates, for NCA and NLMEM. To evaluate the geometric means, the treatment effects, and their SE we use the two-period crossover trials simulated under the null hypothesis $H_{0;80\%}$ with the four different sampling designs (rich, original, intermediate and sparse), and the three variability settings ($S_{l,l}$, $S_{h,l}$ and $S_{h,h}$). There are 9 different used simulation settings, 4 for $S_{l,l}$ and $S_{h,l}$, and 1 for $S_{h,h}$ where only the intermediate design is simulated. For NCA, for each simulated trial, the geometric mean of $AUC$ ($C_{max}$) for the reference treatment is computed from the $N$ individual estimated $AUC$ ($C_{max}$). For NLMEM, due to the log-normal distribution of the random effects, the fixed effect estimates for the reference classes correspond to the geometric mean estimates for the reference treatment. So, the NLMEM-based geometric mean of $AUC$ for the reference treatment is directly obtained from the clearance estimate as $AUC = FD/CL$. For $C_{max}$, the geometric mean is computed from the fixed effect estimates using Equation 5 of the Appendix. For NCA and NLMEM estimates, the geometric means are compared to the $AUC$ or $C_{max}$ computed from the NLMEM simulated parameters. For NCA, the treatment effect on $AUC$ and $C_{max}$ are estimated by LMEM as explained in section 2.1. For NLMEM, due to the linear covariate model on log-parameters, $\widehat{\beta}^T_{AUC} = -\widehat{\beta}^T_{CL/F}$. The treatment effect $\widehat{\beta}^T_{C_{max}}$ is computed from $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$ using Equation 6. For NCA and NLMEM estimates, the treatment effect estimates are compared to the simulated value of the treatment effect. For NCA, standard errors of the treatment effect are estimated by LMEM. For NLMEM, as $\widehat{\beta}^T_{AUC} = -\widehat{\beta}^T_{CL/F}$, their standard error are equal. The SE of $\widehat{\beta}^T_{C_{max}}$ is estimated by the delta method and simulations (with $N_s = 10000$). For the delta method, the expression and details are given in the Appendix. The estimated standard errors of the treatment effect are compared to the corresponding empirical standard error for NCA and NLMEM estimates. For one simulation setting and one approach (NCA or NLMEM), the empirical standard error is computed as the standard deviation of the 1000 treatment effect estimates.

**Type I Error**

To evaluate the type I error of the bioequivalence tests, we use the two-period crossover trials simulated with the four different sampling designs, both hypotheses and the three variability settings. There are 18 different simulation settings used, 8 for $S_{l,l}$ and $S_{h,l}$, and 2 for $S_{h,h}$ where only the intermediate design is simulated. The simulation settings under $H_{0;80\%}$ are also used to evaluate the estimates of $AUC$ and $C_{max}$ (section 3.3).

We perform the bioequivalence Wald test based on NCA estimates on $AUC$ and $C_{max}$. For NLMEM, tests on $CL/F$ and $AUC$ are equivalent because $\widehat{\beta}^T_{AUC} = -\widehat{\beta}^T_{CL/F}$ and $se(\beta^T_{AUC}) = se(\beta^T_{CL/F})$. So, the NLMEM-based bioequivalence Wald test and LRT are performed on $AUC$. As $C_{max}$ is a secondary parameter of the NLMEM, only the NLMEM-based Wald test is performed on this parameter, and not the LRT. For NLMEM, the treatment effect $\widehat{\beta}^T_{C_{max}}$ is computed from $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$. Its standard error is estimated both by the delta method and by simulations (with $N_s = 10000$). For $AUC$ and $C_{max}$, the NLMEM-based bioequivalence Wald test is performed using estimated and empirical SE. For $C_{max}$, it is performed using estimated SE obtained from the delta method and simulations, for comparison. For each one-sided hypothesis $H_{0;80\%}$ and $H_{0;125\%}$, the type I error is estimated by the proportion of the simulated trials for which the null hypothesis $H_0$ is rejected. The global type I error is defined as the maximum value of both estimated type I errors (Dubois et al. 2010; Panhard and Mentré 2005). For 1000 replicates, the 95% prediction interval (95% PI) for a type I error of 5% is [3.7%; 6.4%].

## 3.4 Results

**Evaluation of the Estimates**

For the evaluated settings, all NLMEM parameters including treatment, period, sequence effects are estimated by the SAEM algorithm. Boxplots of the estimates of the clearance reference effect, the corresponding covariate effects and the standard deviations of BSV and WSV are displayed in Figure 1. For the six parameters and

both variability settings, the distribution is narrower when the number of samples or periods increases. For all simulation settings of both types of trials, the median of the fixed effects is close to the corresponding simulated value. For BSV and WSV, the median of the estimates is closer to the simulated value for four-period trials than for two-period trials. For the variability setting $S_{h,l}$, BSV and WSV are slightly underestimated especially for the sparse design. Similar results (not shown) are obtained for both PK parameters, $k_a$ and $V/F$. Table 1 provides the bias ($\times 100$) and RMSE ($\times 100$) of estimates of the reference effects and the standard deviations for BSV, WSV, and residual error. For all simulation settings and both types of crossover trials (2 or 4 periods), there is no bias and RMSE are small for the reference effects and the residual error. For BSV and WSV, bias decreases when the number of samples increases. For all parameters, RMSE decrease when the number of samples increases. Furthermore, RMSE are smaller for $S_{l,l}$ than for $S_{h,l}$ and smaller for four-period trials than for two-period trials. The same observations are made for covariate effects (results not shown).

For each simulation setting of two-period crossover trials of the hypothesis $H_{0;80\%}$ and for NCA and NLMEM, boxplots of the reference treatment geometric mean estimates of $AUC$ and $C_{max}$ are displayed in Figure 2. For $AUC$ and $C_{max}$, and for NCA and NLMEM estimates, the distribution is narrower when the variability is smaller. For NCA estimates, the median of the estimates is closer to the true simulated mean for the rich design, and there is a clear and very large bias of the geometric mean estimates for sparse design. For NLMEM estimates, the median of the estimates is close to the true simulated mean for all simulation settings. Figure 3 displays the boxplot of the treatment effect estimates on $AUC$ and $C_{max}$ and their standard errors for NCA and NLMEM estimates. The standard errors $se(\beta^T_{C_{max}})$ are estimated by the delta method, and very similar results are obtained by simulations. For NCA and NLMEM, for both parameters and all simulation settings, the median of the estimated treatment effects is close to the simulated value. Furthermore, the distribution is narrower when the variability decreases or when the number of subjects increases.

The distribution of the estimated standard errors is narrower and the empirical standard error is smaller when the variability decreases or when the number of subjects increases. For both parameters, the median of the estimated standard error is closer to the empirical one when the variability decreases. For the original design under $S_{h,l}$ and the intermediate design under $S_{h,h}$, standard errors of both parameters are underestimated for NCA and NLMEM estimates.

**Evaluation of the Type I Error**

Table 2 provides type I errors of bioequivalence tests performed on the treatment effects of $AUC$, and $C_{max}$ for each one-sided hypothesis and each sampling design of two-period crossover trials. Mostly, for all tests and both parameters, type I errors of both hypotheses are close. Only the type I errors for $C_{max}$ and the $S_{h,h}$ setting are somewhat different. For Wald tests based on NCA estimates, and for $S_{l,l}$ and $S_{h,l}$ settings, type I errors do not differ from the nominal level of 5%. For $S_{h,h}$ setting, the type I errors are much too conservative for $AUC$, and are inflated for $C_{max}$. For the NLMEM-based Wald test, type I errors for $C_{max}$ using SE obtained by the delta method or simulations are identical. For $AUC$, type I errors of the NLMEM-based Wald test are close to type I errors of the LRT. For the rich design ($N = 40$, $n = 10$), type I errors of both tests do not differ from the nominal level of 5%. However, for each simulation setting, there is an increase of the type I error of both tests when the number of subjects and/or the number of samples decreases.

The left hand side of Figure 4 displays the global type I error for $AUC$ (top) and $C_{max}$ (bottom) versus the design for each variability setting for the Wald test based on NCA estimates. For both parameters, the global type I error lies in the 95%PI of the nominal level for all the designs of $S_{l,l}$ and $S_{h,l}$ settings. For the $S_{h,h}$ setting, it is too conservative for $AUC$ and inflated for $C_{max}$. The right hand side of Figure 4 displays the global type I error of the NLMEM-based Wald test using the estimated or empirical standard error, and the NLMEM-based LRT. For the Wald tests using estimated SE and LRT, and for both parameters, the global type I error lies in the

95%PI of the nominal level for the rich design. It increases when the number of subjects or the number of samples decreases and is lower for $S_{l,l}$ than for $S_{h,l}$.For the NLMEM-based Wald test using the empirical SE, it can be seen that for both parameters the global type I errors almost never differ from the nominal level of 5% showing the influence of the underestimation of the standard errors on the properties of the NLMEM-based Wald test.

# 4 APPLICATION

In 2005, somatropins available in the United States (and their manufacturers) included Nutropin® (Genentech), Humatrope® (Lilly), Genotropin® (Pfizer), Norditropin® (Novo), and Saizen® (Merck Serono). In 2006, the FDA approved a new somatropin called Omnitropee® (Sandoz). For this approval, bioequivalence crossover trials were performed. We analyse one of them with the standard NCA-based approach and the proposed NLMEM-based approach. Then, we perfom the bioequivalence Wald test using NCA and NLMEM estimates.

## 4.1 Material and methods

A randomized, double-blind, single-dose, 3-way crossover study with three treatments, three periods, and six sequences was conducted to compare the pharmacokinetic parameters of Omnitrope® powder for solution for injection, Omnitrope® 3.3 mg/mL solution for injection, and Genotropin® powder for solution after a single subcutaneous dose of 5 mg. Thirty-six healthy caucasian adults were recruited and they received octreotide for endogenous hGH suppression before each treatment period. The three treatment periods were separated by a seven day wash-out period. Blood samples for pharmacokinetic assessments were collected after dose administration for each period at times 1, 2, 3, 4, 5, 6, 8, 10, 12, 16, 20, and 24 $h$. Concentrations were measured by chemiluminescent immunometric assay (Iranmanesh et al. 1994) with a limit of quantification (LOQ) of 0.2 $ng/mL$. Figure 5 (top) displays concentrations

versus time for the three formulations. There are very few concentrations below LOQ for the last sampling times.

We analyse the data with NCA and NLMEM using the SAEM algorithm implemented in MONOLIX 2.4. For NCA analysis, we use the linear trapezoid rule to estimate $AUC_{0-tlast}$. To obtain the total $AUC$, we compute the terminal slope by log-linear regression using 2 to 4 sampling times. As described in 2.1, the log-transformed individual $AUC$ and $C_{max}$ are then analysed using a LMEM including treatment, period, sequence, and subject effects. The reference classes are the Genotropin® treatment, the first period, and the sequence Genotropin® - Omnitrope® powder - Omnitrope® solution for the treatment, period, and sequence covariates, respectively.

For NLMEM analysis, we use a one-compartment model with first-order absorption with a lag time ($t_{lag}$) and first-order elimination to describe the data. With this model, for sampling times before $t_{lag}$, concentrations are null. For sampling times after $t_{lag}$, concentrations are described by Equation 4 replacing $t$ by $t - t_{lag}$. To determine the structure of the random effects matrices and the residual error model, we analyse the Genotropin® data. Models are compared using the Bayesian Information Criteria (BIC), the best statistical model corresponding to the smallest BIC (Bertrand et al. 2008). For the structure of the BSV matrix, we test diagonal, block diagonal, and complete matrices. Regarding the error model, we test a homoscedastic ($b = 0$) and a combined error model. For the analysis of all data, the structure of the WSV matrix is chosen to be identical to the structure of the BSV matrix determined during the analysis of the Genotropin® data. We add treatment, period, and sequence effects on the four PK parameters. The reference classes are the same as for NCA analysis. After fitting the data, the model is graphically evaluated using the individual weighted residuals (IWRES) and the 90% prediction interval for each formulation. For the model evaluation, from the final statistical model and its estimates, we simulate 200 datasets based on the structure of the original data (dose, covariates). For each formulation, we compute the 5% and 95% percentiles of the simulated time-course distribution to obtain the 90% prediction interval. The corre-

spoding graph is called a Visual Predictive Check.

We perform bioequivalence Wald tests on $AUC$ and $C_{max}$ using NCA and NLMEM estimates with a type I error of 5%. For NLMEM, we compute the treatment effect on $C_{max}$ using fixed effects estimates and its standard error by the delta method.

### 4.2 Results

For the analysis of the Genotropin® data, the best statistical model include BSV for all PK parameters with a correlation between the clearance and the volume of distribution, and a combined error model. Parameter estimates (except period and sequence effects) are displayed in Table 3 with their standard errors. Precision of estimation is judged satisfactory for all parameters. Concentrations of somatropin versus time with their 90% prediction interval and the IWRES versus time are displayed in Figure 5 for each treatment group. These model evaluation plots are judged satisfactory.

After estimating the parameters by NCA and NLMEM, we perform bioequivalence Wald tests on $AUC$ and $C_{max}$ for both formulations of Omnitrope®. The results of those tests are displayed in Table 4 with the ratios of $AUC$ and $C_{max}$, the corresponding 90% CI, and the p-values of the bioequivalence Wald tests. With a type I error of 5%, $AUC$ and $C_{max}$ of Omnitrope® powder and solution are bioequivalent to those of Genotropin® using NCA and NLMEM bioequivalence analysis.

## 5 DISCUSSION

In this study, we evaluate the type I error of NLMEM-based bioequivalence tests performed on the treatment effect estimates when treatment, period, and sequence effects but also within-subject variability are taken into account during the NLMEM estimation. This new approach is compared to the standard non-compartmental analysis where bioequivalence Wald tests are performed on the treatment effect estimated by linear mixed effects model taking into account the same three covariates,

BSV (corresponding to the random subject effect) and WSV (*i.e.* residual error). Concerning the NLMEM-based bioequivalence tests, we show how Wald tests can be performed on a secondary parameter such as $C_{max}$ which allows the extension of the standard bioequivalence analysis based on NCA estimates to the NLMEM context. Furthermore, for a parameter of the PK model, we extend the likelihood ratio test for bioequivalence.

As Panhard et al (Panhard et al. 2007), and Dubois et al (Dubois et al. 2010), we simulate under a one-compartment PK model and estimate the NLMEM parameters using the same model. So, we do not study the impact of having the incorrect model being used in the bioequivalence NLMEM-based tests, and how would it compare to the NCA approach in that case. Nevertheless, when bioequivalence analysis is performed, there is already accumulated information on the drug and the pharmacokinetic model is usually known. Furthermore, even if NCA is known as a "model-free" approach, it assumes linear terminal elimination and provides meaningless parameters when it is applied to nonlinear pharmacokinetics. So, the problem of estimating with a "wrong" model could exist for NCA and NLMEM.

The NLMEM-based bioequivalence analysis requires to estimate many parameters. So, a robust algorithm has to be used. The simulation study illustrates the accuracy of the SAEM algorithm, especially in the context of bioequivalence analysis. We show that biases and RMSE obtained by the SAEM algorithm are satisfactory for all parameters although BSV and WSV are slightly underestimated for large variability and low number of patients. These results are similar to those obtained by Panhard and Samson (Panhard and Samson 2009). As expected, biases and RMSE decrease when the amount of information increases (by the increase of the number of patients or periods). All fixed effects are correctly estimated with no bias, which is of great interest for testing treatment effect estimates. The good estimation of the fixed effects using the SAEM algorithm leads to a good estimation of the geometric means of $AUC$ and $C_{max}$, as illustrated by our evaluation. At the opposite, this evaluation also shows that geometric means estimated by NCA are biased for sparse design,

especially with high variability. Usually, NCA is used with rich designs where there are about ten to twenty samples per subject. This method is not well suited for trials performed in patients where the number of samples is often limited. In comparison to model-based approaches, the estimation of parameters through NCA has several drawbacks. It is giving equal weight to all concentrations without taking into account the measurement error. Furthermore, NCA is sensitive to missing data, especially for the determination of $C_{max}$ and the computation of the terminal slope. Even without missing data, the interpolation of the $AUC$ between the last sampling time and infinity is very sensitive to the number of samples used to compute the terminal slope. However, even with biased geometric mean, the treatment effect estimated by NCA are not biased which partly explains the good results for the type I error.

When the number of samples per subject is large and the variability is not too high, tests based on individual NCA estimates remain a good approach since they are simple and showed satisfactory properties for both tested parameters. For $C_{max}$ and the sparse design, we expected an increase of the type I error because there is no sampling time corresponding to the maximal concentration which is close to 2 $h$. But even with poor geometric mean estimates, the type I error is maintained at the nominal level of 5%. It could be explain by the good estimation of the treatment effect estimate despite the biased geometric mean. Though, for simulation with $S_{h,h}$, the global type I error of $AUC$ is very conservative (0.8%) which shows the limits of NCA for data with high residual error.

The type I error of the NLMEM-based bioequivalence Wald test and LRT are rather similar but Wald tests are easier to perform. Indeed, the bioequivalence LRT requires to estimate the parameters and log-likelihood of three statistical models. Furthermore, there is currently no methodology to perform a LRT on a secondary parameter if the model cannot be reparameterized using this parameter (e.g. $C_{max}$). For a Wald test on $C_{max}$, the delta method or simulations can be used to estimate its treatment effect standard error. Based on our simulation study, for a one-compartment PK model, the use of simulations is not more efficient than the delta method. Indeed,

for each simulation setting, standard errors estimated by delta method or simulations are really close and the results of the type I error are similar for both estimations. However, the use of the delta method can be tricky since the analytical expression of $C_{max}$ is not always available for complex or nonlinear PK models.

For NLMEM-based Wald tests and LRT, we found an inflation of the type I error when the conditions move away from asymptotic, *i.e.* for small sample size and/or data with high variability. The use of NLMEM-based bioequivalence analysis in its current proposed form would be questionable for regulatory agencies in these cases due to concerns about potential type I error inflation. For NLMEM-based Wald tests, the underestimation of the standard errors are responsible of the inflation of the type I error. Indeed, there is no inflation when the empirical standard error is used instead of the estimated. The empirical standard error can be used in practice but not easily because of the computing time. It requires first to estimate the parameters using the data of the clinical trial of interest, then to simulate trials with the same design as the original dataset and finally to re-estimate the parameters for each simulated trial. This approach also assumes that the underlying structural model is correct which is usually the case when bioequivalence analysis is performed, as previously mentionned. In our simulation, the number of subjects is more influential on the inflation of the type I error than the number of samples. Indeed, there is a slight inflation of the type I error for the sparse design ($N = 40,\ n = 3$) compared to the rich ($N = 40,\ n = 10$, same $N$) whereas the inflation is higher for the original design ($N = 12,\ n = 10$) also compared to the rich (same $n$). For NLMEM-based Wald test, this is explained by the slighter underestimation of the standard errors for the sparse design. The inflation of the type I error for NLMEM-based Wald tests and LRT is not specific to bioequivalence tests. It is due to the asymptotic properties of these tests and was also demonstrated for comparison tests by Panhard et al (Panhard and Mentré 2005) and Wälhby et al (Wählby et al. 2001). Similarly, the underestimation of the standard errors was also related to the inflation of the type I error for comparison NLMEM-based Wald tests (Bertrand et al. 2009). A good control of the type

I error for a bioequivalence test with sparse sampling should be therefore possible by increasing the number of patients. Furthermore, different approaches could be explored to correct the type I error inflation of NLMEM-based bioequivalence tests. For NLMEM-based Wald tests, the underestimation of BSV and WSV could explain the underestimation of the standard errors. Even though maximum likelihood estimation is the standard approach in NLMEM, the variance components are often underestimated for small sample size and high variability. In linear mixed effects models, the REML estimation is widely implemented, but in NLMEM it has been barely studied, although the REML procedure may improve the estimation of variance components in NLMEM. Meza et al (Meza et al. 2007) developped a REML estimation procedure for the standard SAEM algorithm. They showed that the SAEM-REML algorithm reduces bias and RMSE of the variance parameter estimates in a simulation study on a simple NLMEM. Further work is needed to propose the REML estimation procedure for the extended SAEM algorithm developed for crossover trial analysis.By improving the estimation of variance parameters, the REML estimation procedure should improve the bioequivalence Wald test. As explained in section 2.1 and 3.2, for bioequivalence Wald tests based on NCA estimates, the LMEM parameters are estimated by REML and both test statistics follow a Student t-distribution with degrees of freedom depending on the number of subjects. So, we perform the NLMEM-based bioequivalence Wald tests assuming a Student t-distribution under $H_0$ with the same number of degrees of freedom as the NCA-based bioequivalence Wald tests (unshown results). For all simulation settings, the type error decreases compared to the NLMEM-based Wald test with a Gaussian distribution but there is still a slight inflation of the type I error while the use of empirical SE corrects it. To our knowledge, there is no theoritical development or evaluation of the degrees of freedom in the context of NLMEM. The distribution we use is more or less empirical, and further work is needed.

Other approaches could be studied such as the correction of the nominal level using permutation tests or bootstrap methods to estimate the 90% CI. However, perform-

ing a permutation test may not be suitable for bioequivalence, and boostrap methods have not yet been properly studied in NLMEM. In NLMEM context, the paired bootstrap is usually used but without taking into account the different levels of variability of the NLMEM. Furthermore, there is no theoritical or simulation result to justify its application. To our knowledge, only two published studies adress the issue of bootstrap in NLMEM (Das and Krishen 1999; Ocaña et al. 2005). Ocaña et al (Das and Krishen 1999) proposed a bootstrap approach resampling the random effects and residual errors. They evaluated it by simulation but they performed it using two-stage fitting procedure (Steimer et al. 1984) where "population" mean parameters are estimated from individual parameters obtained after separate fitting of each subject data. Further simulations studies are needed to really understand bootstrap methods properties in NLMEM. So, we would favor a correction of the tests by degrees of freedom, which is also a less computer intensive method.

The analysis of the crossover trial of three somatropin formulations shows the ability to perform a NLMEM-based bioequivalence analysis using the SAEM algorithm on a real data set. Even with forty fixed effects and ten variance parameters in the statistical model, the SAEM algorithm converges. Furthermore, the SAEM algorithm can handle data below the limit of quantification contrary to NCA. The PK parameter estimates for Genotropin® are similar to those found by Stanhope et al (Stanhope et al. 2010). We perform NLMEM-based bioequivalence Wald tests and not LRT because results on Wald tests and LRT are similar in the simulation study, and we would like to perform tests on the treatment effects of $AUC$ and $C_{max}$, which is not possible by LRT. The results of the bioequivalence analysis based on NCA and NLMEM are similar. In both cases, we assess the bioequivalence of both Omnitropee® formulations. Bioequivalence tests based on NLMEM allow one to decrease the number of samples per subject, which is of great interest for trials performed in patients. However, caution is needed for small sample size and data with high variability. With sparse sampling, the choice of design is important notably to improve the properties of tests. For instance, Bertrand et al (Bertrand et al. 2009) showed that, for the same number

of samples, some designs have better power than others for detection of a pharmacogenetic effect in a one-period trial. Design optimisation algorithms for models with discrete covariates and different periods of treatment could be used for crossover studies. They are now available in the version 3.2 of PFIM software (Bazzoli et al. 2010).

## APPENDIX: DELTA METHOD FOR $C_{MAX}$

For a one-compartment model with first-order absorption and first-order elimation $C_{max}$ is a function of the three PK parameters:

$$C_{max} = \frac{FD}{V} \exp\left(\frac{CL\,\log(k_a) - \log(CL/V)}{V\,k_a - CL}\right) \tag{5}$$

So, $\beta^T_{C_{max}}$ is a function $h$ of $\lambda_{k_a}$, $\lambda_{V/F}$, $\lambda_{CL/F}$, $\beta^T_{k_a}$, $\beta^T_{V/F}$, and $\beta^T_{CL/F}$:

$$\begin{aligned}\beta^T_{C_{max}} =& h(\lambda_{k_a}, \lambda_{V/F}, \lambda_{CL/F}, \beta^T_{k_a}, \beta^T_{V/F}, \beta^T_{CL/F}) \\ =& -\beta^T_{V/F} - A_2\frac{\lambda_{CL/F}\exp(\beta^T_{CL/F})}{A_1} + \frac{\lambda_{CL/F}}{\sqrt{A_6}}\log\left(\frac{\lambda_{k_a}\lambda_{V/F}}{\lambda_{CL/F}}\right)\end{aligned} \tag{6}$$

The vector of partial derivatives of $h$ is:

$$\nabla h = \left(\tfrac{1}{\lambda_{k_a}}\left(\tfrac{A_4}{A_3} - \tfrac{A_5}{A_6}\right), \tfrac{1}{\lambda_{V/F}}\left(\tfrac{A_4}{A_3} - \tfrac{A_5}{A_6}\right), \tfrac{1}{\lambda_{CL/F}}\left(\tfrac{-A_4}{A_3} + \tfrac{A_5}{A_6}\right), \tfrac{A_4}{A_3}, \tfrac{A_4}{A_3} - 1, \tfrac{-A_4}{A_3}\right)' \tag{7}$$

where

$$\begin{aligned}A_1 =& \lambda_{k_a}\lambda_{V/F}\exp(\beta^T_{k_a} + \beta^T_{V/F}) - \lambda_{CL/F}\exp(\beta^T_{CL/F}) \\ A_2 =& \log\left(\frac{\lambda_{k_a}\lambda_{V/F}}{\lambda_{CL/F}}\right) + \beta^T_{k_a} + \beta^T_{V/F} - \beta^T_{CL/F} \\ A_3 =& \left(-A_1\lambda_{CL/F}\exp(\beta^T_{CL/F})\right)^2 \\ A_4 =& A_3 + A_2\,\lambda_{k_a}\,\lambda_{V/F}\,\lambda_{CL/F}\exp(\beta^T_{k_a} + \beta^T_{V/F} + \beta^T_{CL/F}) \\ A_5 =& \lambda_{k_a}\lambda_{V/F}\lambda_{CL/F}\log(\lambda_{k_a}\lambda_{V/F}/\lambda_{CL/F}) + \lambda_{CL/F}(\lambda_{CL/F} - \lambda_{k_a}\lambda_{V/F}) \\ A_6 =& \left(\lambda_{k_a}\lambda_{V/F} - \lambda_{CL/F}\right)^2\end{aligned} \tag{8}$$

## REFERENCES


Bates, D. M. and Watts, D. G. (1988), *Nonlinear regression analysis and its applications*, John Wiley & sons, Chichester.

Bazzoli, C., Retout, S., and Mentré, F. (2009), "Fisher information matrix for nonlinear mixed effects multiple response models: Evaluation of the appropriateness of the first order linearization using a pharmacokinetic/pharmacodynamic model," *Statistics in Medicine*, 28, 1940–1956.

Bazzoli, C., Nguyen, T. T., Dubois, A., Retout, S., Comets, E., and Mentré, F. (2010), "PFIM," url: http://www.pfim.biostat.fr/.

Bertrand, J., Comets, E., and Mentré, F. (2008), "Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters," *Journal of Biopharmaceutical Statistics*, 18, 1084–1102.

Bertrand, J., Comets, E., Laffont, C., Chenel, M., and Mentré, F. (2009), "Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm," *Journal of Pharmacokinetics and Pharmacodynamics*, 36, 317–339.

Chow, S. C. and Liu, J. P. (2000), *Design and analysis of bioavailability and bioequivalence studies*, Marcel Dekker.

Combrink, M., McFadyen, M.-L., and Miller, R. (1997), "A comparison of standard approach and the NONMEM approach in the estimation of bioavailability in man," *The Journal of Pharmacy and Pharmacology*, 49, 731–733.

Das, S. and Krishen, A. (1999), "Some bootstrap methods in nonlinear mixed-effect models," *Journal of Statistical Planning and Inference*, 75, 237–245.

Delyon, B., Lavielle, M., and Moulines, E. (1999), "Convergence of a stochastic approximation version of EM algorithm," *The Annals of Statistics*, 27, 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39, 1–38.

Dubois, A., Gsteiger, S., Pigeolet, E., and Mentré, F. (2010), "Bioequivalence tests based on individual estimates using non compartmental or model-based analyses: evaluation of estimates of sample means and type I error for different designs," *Pharmaceutical Research*, 27, 92–104.

EMEA (2001), "Note for guidance on the investigation of bioavailability and bioequivalence," Technical report, EMEA.

Fauci, A., Braunwald, E., Kasper, D., Hauser, S., Longo, D., Jameson, J., and Loscalzo, J. (2008), *Harrison's Principles of Internal Medicine - 17th Edition*, McGraw-Hill Professional, Columbus.

FDA (2001), "Guidance for Industry - Statistical Approaches to establishing bioequivalence," Technical report, FDA.

Fradette, C., Lavigne, J., Waters, D., and Ducharme, M. (2005), "The utility of the population approach applied to bioequivalence in patients," *Therapeutic Drug Monitoring*, 27, 592–600.

Gabrielson, J. and Weiner, D. (2006), *Pharmacokinetic and pharmacodynamic data analysis: concepts and applications*, Apotekarsocieteten, Stockholm.

Ge, Z., Bickel, P., and Rice, J. (2004), "An approximate likelihood approach to nonlinear mixed effects models via spline approximation," *Computational Statistics & Data Analysis*, 46, 747–776.

Hauschke, D., Steinijans, V., and Pigeot, I. (2007), *Bioequivalence studies in drug development*, John Wiley & sons, Chichester.

Hu, C., Moore, K., Kim, Y., and Sale, M. (2003), "Statistical issues in a modeling approach to assessing bioequivalence or PK similarity with presence of sparsely

sampled subjects," *Journal of Pharmacokinetics and Pharmacodynamics*, 31, 312–339.

Iranmanesh, A., Grisso, B., and Veldhuis, J. D. (1994), "Low basal and persistent pulsatile growth hormone secretion are revealed in normal and hyposomatotropic men studied with a new ultrasensitive chemiluminescence assay," *The Journal of Clinical Endocrinology and Metabolism*, 78, 526–535.

Kaniwa, N., Aoyagi, N., Ogata, H., and Ishii, M. (1990), "Application of the NONMEM method to evaluation of the bioavailability of drug products," *Journal of Pharmaceutical Sciences*, 79, 1116–1120.

Kuhn, E. and Lavielle, M. (2004), "Coupling a stochastic approximation version of EM with a MCMC procedure," *ESAIM Probability and Statistics*, 8, 115–131.

Kuhn, E. and Lavielle, M. (2005), "Maximum likelihood estimation in nonlinear mixed effects models," *Computational Statistics and Data Analysis*, 49, 1020–1038.

Lavielle, M., Meza, H., and Chatel, K. (2010), "The MONOLIX software," url: http://software.monolix.org.

Lindstrom, M. and Bates, D. (1990), "Nonlinear mixed effects models for repeated measures data," *Biometrics*, 46, 673–687.

Liu, J. P. and Weng, C. S. (1995), "Bias of two one-sided tests procedures in assessment of bioequivalence," *Statistics in Medicine*, 14, 853–861.

Maier, G. A., Lockwood, G. F., Oppermann, J. A., Wei, G., Bauer, P., Fedler-Kelly, J., and Grasela, T. (1999), "Characterization of the highly variable bioavailability of tiludronate in normal volunteers using population pharmacokinetic methodologies," *European Journal of Drug Metabolism and Pharmacokinetics*, 24, 249–254.

Meeker, W. Q. and Escobar, L. A. (1995), "Teaching about approximate confidence regions based on maximum likelihood estimation," *The American Statistician*, 49, 48–53.

Meza, C., Jaffrézic, F., and Foulley, J.-L. (2007), "REML estimation of variance parameters in nonlinear mixed effects models using SAEM algorithm," *Biometrical Journal*, 49, 876–888.

Ocaña, J., El Halimi, R., Ruiz de Villa, C., and Sánchez, J. (2005), "Bootstrapping repeated measures data in nonlinear mixed-models context," Universitat de Barcelona IMUB.

Oehlert, G. W. (1992), "A note on the delta method," *The American Statistician*, 46, 27–29.

Panhard, X. and Mentré, F. (2005), "Evaluation by simulation of tests based on non-linear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials," *Statistics in Medicine*, 24, 1509–1524.

Panhard, X. and Samson, A. (2009), "Extension of the SAEM algorithm for nonlinear mixed models with two levels of random effects," *Biostatistics*, 10, 121–135.

Panhard, X., Taburet, A. M., Piketti, C., and Mentré, F. (2007), "Impact of modelling intra-subject variability on tests based on non-linear mixed-effects models in cross-over pharmacokinetic trials with application to the interaction of tenofovir on atazanavir in HIV patients," *Statistics in Medicine*, 26, 1268–1284.

Pentikis, H., Henderson, J., Tran, N., and Ludden, T. (1996), "Bioequivalence: individual and population compartmental modeling compared to noncompartmental approach," *Pharmaceutical Research*, 13, 1116–1121.

Pinheiro, J. and Bates, D. (1995), "Approximations to the log-likelihood function in the non-linear mixed-effect models," *Journal of Computational and Graphical Statistic*, 4, 12–35.

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-effects models in S and Splus*, Springer, New-York.

Samson, A., Lavielle, M., and Mentré, F. (2007), "The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model," *Statistics in Medicine*, 26, 4860–4875.

Schuirmann, D. J. (1987), "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657–680.

Stanhope, R., Sörgel, F., Gravel, P., Schuetz, Y. B. P., Zabransky, M., and Muenzberg, M. (2010), "Bioequivalence studies of Omnitrope, the first biosimilar/rhGH follow-on protein: two comparative phase 1 randomized studies and population pharmacokinetic analysis," *Journal of Clinical Pharmacology*, 50, 1339–1348.

Steimer, J. L., Mallet, A., Golmard, J. L., and Boisvieux, J. F. (1984), "Alternative approaches to estimation of population pharmacokinetic parameters: comparison with the nonlinear mixed effect model." *Drug Metabolism Reviews*, 15, 265–269.

Vonesh, E. F. (1996), "A note on the use of laplace's approximation for nonlinear mixed-effects models," *Biometrika*, 83, 447–452.

Walker, S. (1996), "An EM algorithm for non-linear random effects models," *Biometrics*, 52, 934–944.

Wei, G. and Tanner, M. A. (1990), "Calculating the content and boundary of the highest posterior density region via data augmentation," *Biometrika*, 77, 649–652.

Wählby, U., Jonsson, E. N., and Karlsson, M. O. (2001), "Assessment of actual significance levels for covariate effects in NONMEM," *J Pharmacokinet Pharmacodyn*, 28, 231–252.

Wu, L. (2004), "Exact and approximate inferences for nonlinear mixed-effects models with missing covariates," *Journal of the American Statistical Association*, 99, 700–709.

Zhou, H., Mayer, P., Wajdula, J., and Fatenejad, S. (2004), "Unaltered etanercept pharmacokinetics with concurrent methotrexate in patients with rheumatoid arthritis," *Journal of Clinical Pharmacology*, 44, 1235–1243.

Zhu, M., Bifano, M., Wang, X. X. Y., LaCreta, F., Grasela, D., and Pfister, M. (2008), "Lack of an effect of human immunodeficiency virus coinfection on the pharmacokinetics of entecavir in hepatitis B virus-infected patients," *Antimicrobial agents and chemotherapy*, 52, 2836–2841.

Table 1: Bias (×100) and root mean square error (RMSE ×100) of estimates of reference effects and of standard deviations for BSV, WSV, and residual error.

| | Period | | $\lambda_{k_a}$ | $\lambda_{V/F}$ | $\lambda_{CL/F}$ | $\omega_{k_a}$ | $\omega_{V/F}$ | $\omega_{CL/F}$ | $\gamma_{k_a}$ | $\gamma_{V/F}$ | $\gamma_{CL/F}$ | $b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $N = 40,\ n = 10$ | | | | | | |
| $S_{l,l}$ | 2 | bias | 0 | 0 | 0 | -0.2 | -0.1 | -0.2 | -0.1 | 0 | -0.1 | 0 |
| | | RMSE | 9.2 | 1.3 | 0.2 | 1.3 | 0.3 | 1.1 | 0.6 | 0.1 | 0.3 | 0.4 |
| | 4 | bias | -1.1 | -0.2 | 0 | -0.2 | -0.1 | -0.2 | -0.1 | 0 | 0 | 0 |
| | | RMSE | 8.7 | 1.3 | 0.2 | 1.1 | 0.3 | 1 | 0.4 | 0.1 | 0.2 | 0.3 |
| $S_{h,l}$ | 2 | bias | -1.8 | -0.5 | 0 | 2.7 | 3.6 | -1.2 | -0.1 | -0.1 | -0.2 | 0.1 |
| | | RMSE | 18.5 | 5.8 | 0.5 | 8.6 | 8.9 | 6 | 1.3 | 0.7 | 0.7 | 0.4 |
| | 4 | bias | 0.2 | 0.2 | 0 | -0.4 | -0.4 | -1.2 | -0.1 | -0.1 | -0.1 | 0.1 |
| | | RMSE | 17.9 | 5.7 | 0.5 | 6.1 | 5.8 | 5.8 | 0.7 | 0.4 | 0.4 | 0.3 |
| | | | | | | $N = 40,\ n = 3$ | | | | | | |
| $S_{l,l}$ | 2 | bias | 0.3 | -0.1 | 0 | -0.2 | -0.1 | -0.2 | 0 | 0 | -0.1 | -0.4 |
| | | RMSE | 11 | 1.7 | 0.2 | 1.6 | 0.4 | 1.1 | 0.9 | 0.3 | 0.5 | 1.5 |
| | 4 | bias | -0.8 | 0 | 0 | -0.3 | -0.1 | -0.2 | -0.1 | 0 | 0 | -0.2 |
| | | RMSE | 9.5 | 1.6 | 0.2 | 1.2 | 0.3 | 1 | 0.6 | 0.2 | 0.3 | 0.9 |
| $S_{h,l}$ | 2 | bias | 6.9 | 1.6 | -0.1 | 2.1 | -3.2 | -4.2 | 0 | -0.3 | -0.3 | 0.3 |
| | | RMSE | 22.8 | 5.8 | 0.4 | 8.9 | 6.3 | 6.8 | 1.9 | 0.9 | 1 | 2.5 |
| | 4 | bias | 6.5 | 1.8 | -0.1 | 2.1 | -3.6 | -4.2 | -0.1 | -0.3 | -0.2 | 0.6 |
| | | RMSE | 21.4 | 5.9 | 0.4 | 8.6 | 6.2 | 6.5 | 1.3 | 0.6 | 0.6 | 1.7 |

NOTE: Bias and RMSE are estimated from 1000 crossover trials simulated under $H_{0;80\%}$ with two or four periods, for the rich ($N = 40$, $n = 10$) and sparse ($N = 40$, $n = 3$) designs, and two variability settings ($S_{l,l}$ and $S_{h,l}$).

Table 2: Type I error (×100) of bioequivalence tests performed on the treatment effect of $AUC$ and $C_{max}$ for each unilateral hypothesis, $H_{0;80\%}$ and $H_{0;125\%}$.

| | | | $N = 40,\ n = 10$ | | | $N = 12,\ n = 10$ | | | $N = 24,\ n = 5$ | | | $N = 40,\ n = 3$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | NCA | NLMEM | | NCA | NLMEM | | NCA | NLMEM | | NCA | NLMEM | |
| | | | | Wald | LRT | | Wald | LRT | | Wald | LRT | | Wald | LRT |
| $S_{ll}$ | $AUC$ | $H_{0;80\%}$ | 4.0 | 5.3 | 5.3 | 5.2 | 9.3 | 8.1 | 4.3 | 7.0 | 6.8 | 5.9 | 4.8 | 4.8 |
| | | $H_{0;125\%}$ | 5.1 | 5.2 | 5.2 | 5.2 | 9.3 | 7.6 | 3.8 | 5.8 | 5.6 | 5.1 | 5.6 | 5.2 |
| | $C_{max}$ | $H_{0;80\%}$ | 6.6 | 4.6 (4.7) | | 5.1 | 7.3 | | 5.3 | 5.2 | | 6.8 | 8.5 | |
| | | $H_{0;125\%}$ | 6.3 | 6.8 | | 5.6 | 8.0 | | 5.2 | 8.0 | | 5.5 | 6.9 (6.8) | |
| $S_{h,l}$ | $AUC$ | $H_{0;80\%}$ | 5.4 | 4.8 | 5.3 | 4.4 | 11.0 | 10.0 | 5.2 | 9 | 8.2 | 4.5 | 6.4 | 6.0 |
| | | $H_{0;125\%}$ | 6.1 | 6.6 | 6.0 | 4.7 | 10.7 | 8.9 | 3.9 | 6.7 | 6.8 | 5.1 | 8.6 | 7.2 |
| | $C_{max}$ | $H_{0;80\%}$ | 5.1 | 4.9 | | 5.3 | 9.1 (9.0) | | 6.0 | 6.3 (6.2) | | 7.2 | 6.9 | |
| | | $H_{0;125\%}$ | 5.4 | 5.3 | | 5.1 | 8.9 | | 6.1 | 7.0 | | 6.2 | 6.9 | |
| $S_{h,h}$ | $AUC$ | $H_{0;80\%}$ | | | | | | | 0.8 | 6.0 | 8.3 | | | |
| | | $H_{0;125\%}$ | | | | | | | 0.4 | 5.8 | 5.9 | | | |
| | $C_{max}$ | $H_{0;80\%}$ | | | | | | | 7.0 | 5.8 (5.3) | | | | |
| | | $H_{0;125\%}$ | | | | | | | 9.3 | 10.3 (9.9) | | | | |

NOTE: The Wald tests based on NCA and NLMEM estimates are performed on the treatment effect of $AUC$ and $C_{max}$. The NLMEM-based likelihood ratio test (LRT) is performed on $CL/F$ (*i.e.* $AUC$) only. The type I error is estimated from 1000 two-period crossover trials simulated under $H_{0;80\%}$ or $H_{0;125\%}$ for different sampling designs ($N$: number of subjects, $n$: number of samples per subject and period) and three variability settings ($S_{l,l}$, $S_{h,l}$, and $S_{h,h}$). For NLMEM-based bioequivalence Wald tests performed on the treatment effect of $C_{max}$, type I errors are estimated using the delta method or simulations. The values of both type I errors are reported only if they are not equal; in that case, the type I error of $C_{max}$ from simulations is in brackets.

Table 3: Pharmacokinetic parameter estimates of somatropin (standard errors) from the three-way crossover study on somatropin (period and sequence effects are not reported).

| | $t_{lag}$ $(h)$ | $k_a$ $(h^{-1})$ | $V/F$ $(L)$ | $CL/F$ $(L/h)$ | $corr_{CL/F,V/F}$ |
|---|---|---|---|---|---|
| $\lambda$ | 0.46 (0.08) | 0.32 (0.05) | 25.83 (6.24) | 8.66 (0.86) | |
| $\beta^T_{powder}$ | -0.25 (0.08) | -0.24 (0.1) | -0.14 (0.12) | 0.01 (0.03) | |
| $\beta^T_{solution}$ | -0.04 (0.06) | -0.11 (0.11) | 0.01 (0.13) | 0.05 (0.03) | |
| $\omega$ | 0.38 (0.06) | 0.15 (0.08) | 0.39 (0.04) | 0.23 (0.01) | 0.95 |
| $\gamma$ | 0.12 (0.06) | 0.27 (0.08) | 0.36 (0.04) | 0.10 (0.01) | 0.67 |
| $a$ $(ng/mL)$ | 0.12 (0.02) | | | | |
| $b$ | 0.14 (0.004) | | | | |

NOTE: The reference formulation is the Genotropin®. Treatment effects are estimated for Omnitrope® powder ($\beta^T_{powder}$) and Omnitrope® solution ($\beta^T_{solution}$)

Table 4: Bioequivalence Wald tests using NCA and NLMEM estimates for the three-way crossover study on somatropin.

| | | NCA | | | NLMEM | | |
|---|---|---|---|---|---|---|---|
| | Formulation | Ratio | 90% CI | p | Ratio | 90% CI | p |
| $AUC$ | powder | 0.99 | [0.94; 1.03] | $7\,10^{-11}$ | 0.99 | [0.95; 1.04] | $3\,10^{-17}$ |
| | solution | 0.95 | [0.90; 0.99] | $3\,10^{-8}$ | 0.95 | [0.92; 1.00] | $5\,10^{-12}$ |
| $C_{max}$ | powder | 0.95 | [0.88; 1.03] | $3\,10^{-4}$ | 0.94 | [0.84; 1.04] | 0.008 |
| | solution | 0.93 | [0.86; 1.01] | 0.001 | 0.92 | [0.83; 1.02] | 0.015 |

NOTE: p is the p-value of the bioequivalence Wald test. The reference formulation is the Genotropin®. The ratios correspond to Omnitrope® powder versus Genotropin® and to Omnitrope® solution versus Genotropin®.

**Legend to figures**

Figure 1. Boxplots of the estimates of the clearance reference effect ($\lambda_{CL/F}$), corresponding covariate effects ($\beta^T_{CL/F}$, $\beta^P_{CL/F}$ and $\beta^S_{CL/F}$), and standard deviation of the between-subject ($\omega_{CL/F}$) and within-subject ($\gamma_{CL/F}$) variability for the hypothesis $H_{0;80\%}$. Parameters are estimated from the 1000 crossover trials simulated under $H_{0;80\%}$ with two or four periods, for the rich ($N = 40$, $n = 10$) and sparse ($N = 40$, $n = 3$) designs, and two variability settings, $S_{l,l}$ (top) and $S_{h,l}$ (bottom). For four-period crossover trials, only the period effect estimates $\widehat{\beta}^P_{2,CL/F}$ are displayed. The horizontal lines correspond to the true simulated values.

Figure 2. Boxplots of the geometric mean estimates of $AUC$ (top) and $C_{max}$ (bottom) estimated by NCA (left) or NLMEM (right), for each simulation setting of two-period crossover trials, the hypothesis $H_{0;80\%}$, and the reference treatment. The horizontal lines correspond to the geometric means computed from the NLMEM simulated parameters.

Figure 3.Boxplots of the treatment effect on $AUC$ (first row) and $C_{max}$ (third row) and their standard errors (second and fourth rows) estimated by NCA (left) or NLMEM (right), for each simulation setting of two-period crossover trials and the hypothesis $H_{0;80\%}$. For NCA, $\beta^T_{AUC}$, $\widehat{\beta}^T_{C_{max}}$, $se(\beta^T_{AUC})$ and $se(\beta^T_{C_{max}})$ are obtained from LMEM analysis. For NLMEM, the estimates of $\beta^T_{AUC}$ and $se(\beta^T_{AUC})$ are directly obtained from $\widehat{\beta}^T_{CL/F}$ and $se(\beta^T_{CL/F})$. The treatment effect $\widehat{\beta}^T_{C_{max}}$ is computed from $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}$, and $se(\beta^T_{C_{max}})$ is estimated by the delta method. The horizontal lines correspond to the true simulated values of the treatment effects. The cross symbols correspond to the empirical standard errors of the treatment effect computed for each simulation setting.

Figure 4. Gobal type I error of the bioequivalence tests performed on the treatment effect of $AUC$ (top) and $C_{max}$ (bottom) from NCA (right) and NLMEM (left) estimates. The Wald tests based on NCA and NLMEM estimates are performed on both parameters, the likelihood ratio test (LRT) is performed only on $AUC$. For NLMEM-based bioequivalence Wald tests, $se(\beta^T_{C_{max}})$ are estimated by the delta

method. NLMEM-based bioequivalence Wald tests are performed with the estimated or empirical standard error. The type I error is estimated from 1000 bioequivalence trials simulated under $H_{0;80\%}$ and $H_{0;125\%}$ for different sampling designs ($N$: number of subjects, $n$: number of samples per subject) and different variability settings $S_{l,l}$, $S_{h,l}$, and $S_{h,h}$. The horizontal dashed lines represent the nominal level at 5% and its 95% prediction interval ($[3.7\%; 6.4\%]$).

Figure 5. Observed concentrations of somatropin versus time with their 90% prediction interval (top), and individual weighted residuals (IWRES) versus time (bottom) for each treatment, Genotropin® (left), Omnitrope® powder (middle), and Omnitrope® solution (right).

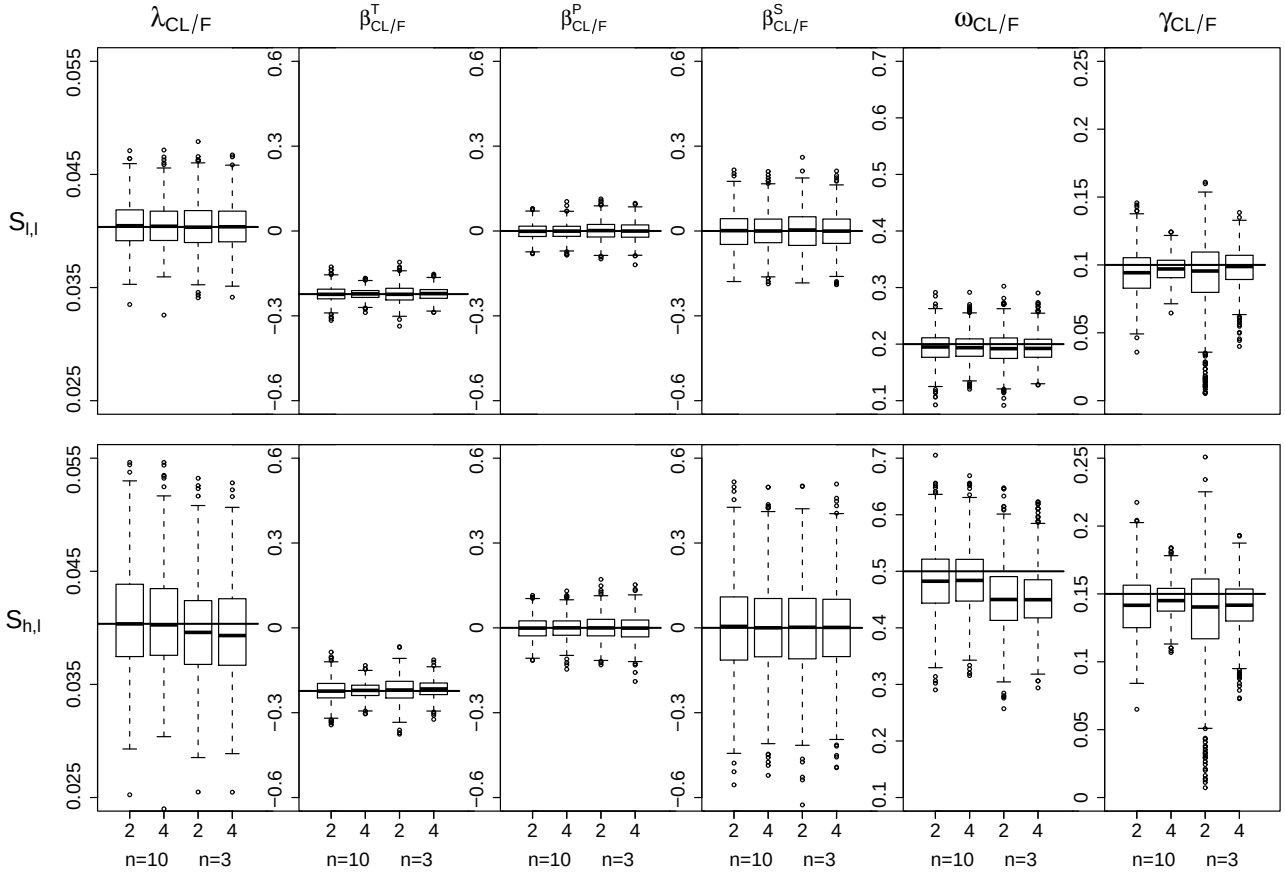


Figure 1

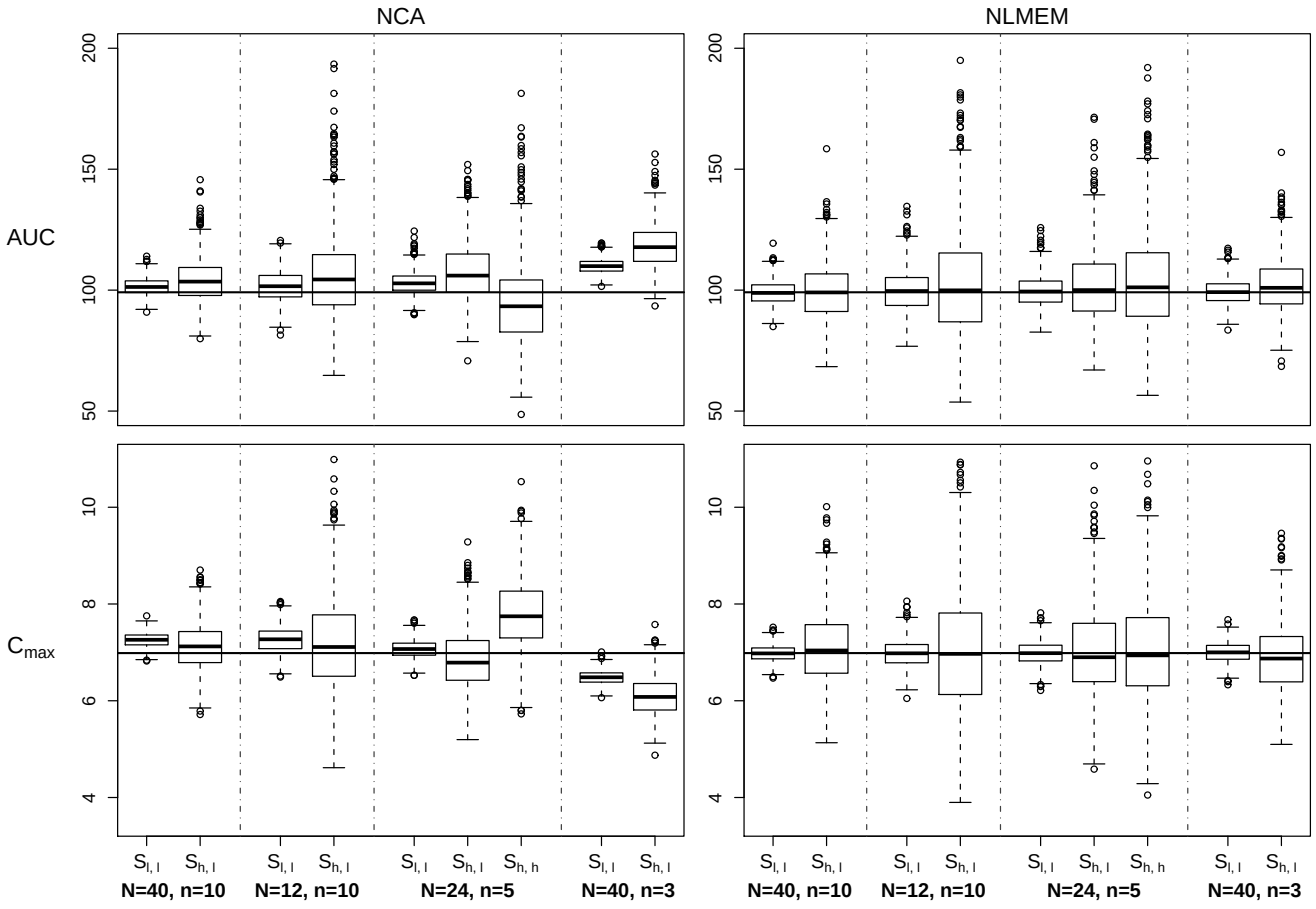
NCA
NLMEM
AUC
$C_{max}$
$S_{l,l}$ $S_{h,l}$ $S_{l,l}$ $S_{h,l}$ $S_{l,l}$ $S_{h,l}$ $S_{h,h}$ $S_{l,l}$ $S_{h,l}$
N=40, n=10
N=12, n=10
N=24, n=5
N=40, n=3

Figure 2

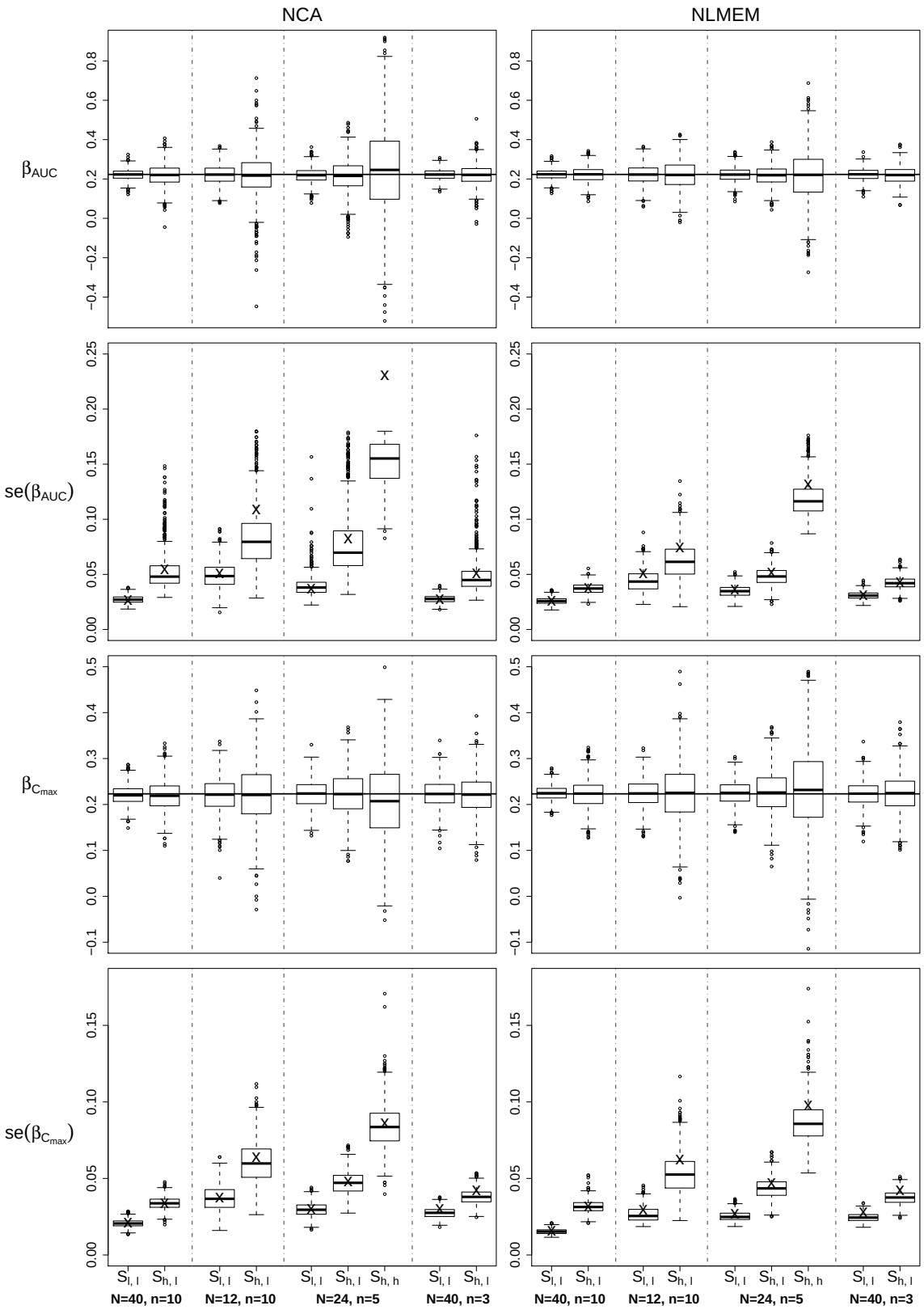
NCA
NLMEM
$\beta_{AUC}$
se($\beta_{AUC}$)
$\beta_{C_{max}}$
se($\beta_{C_{max}}$)
$S_{l,l}$
$S_{h,l}$
$S_{h,h}$
N=40, n=10
N=12, n=10
N=24, n=5
N=40, n=3

Figure 3

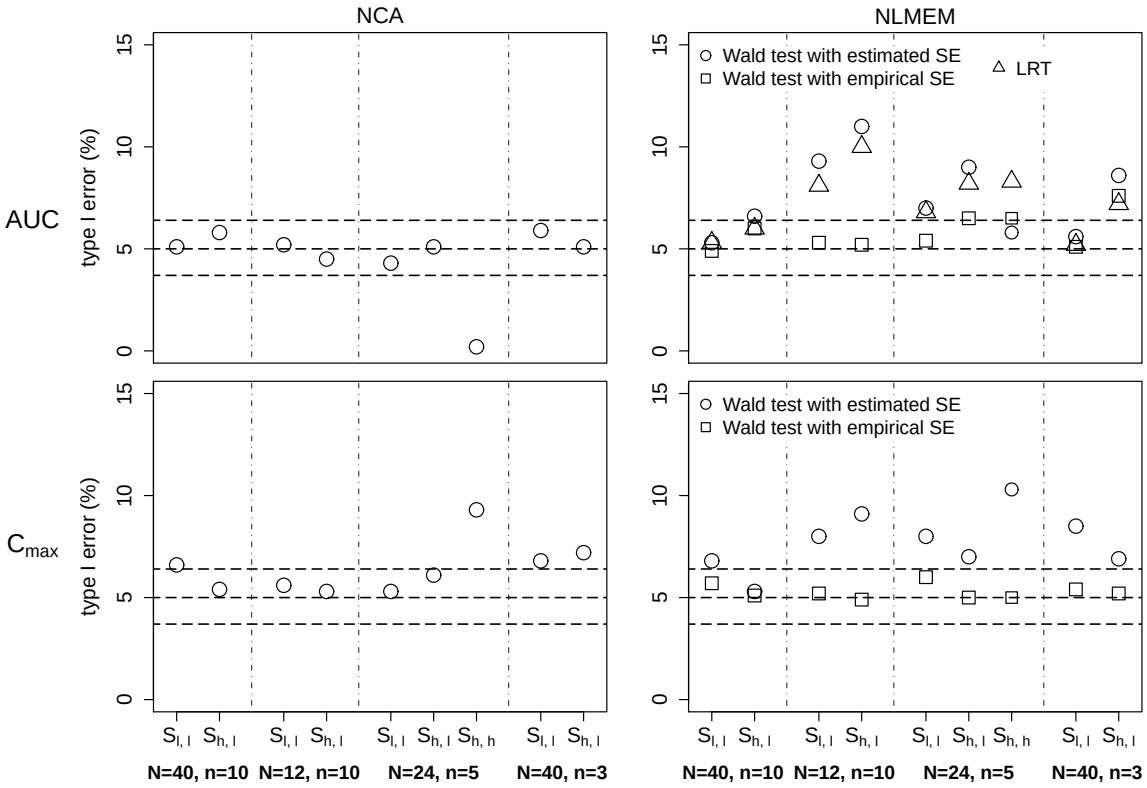
NCA
NLMEM
AUC
$C_{max}$
type I error (%)
Wald test with estimated SE
Wald test with empirical SE
LRT
$S_{l,l}$ $S_{h,l}$ $S_{l,l}$ $S_{h,l}$ $S_{l,l}$ $S_{h,l}$ $S_{h,h}$ $S_{l,l}$ $S_{h,l}$
N=40, n=10 N=12, n=10 N=24, n=5 N=40, n=3

Figure 4

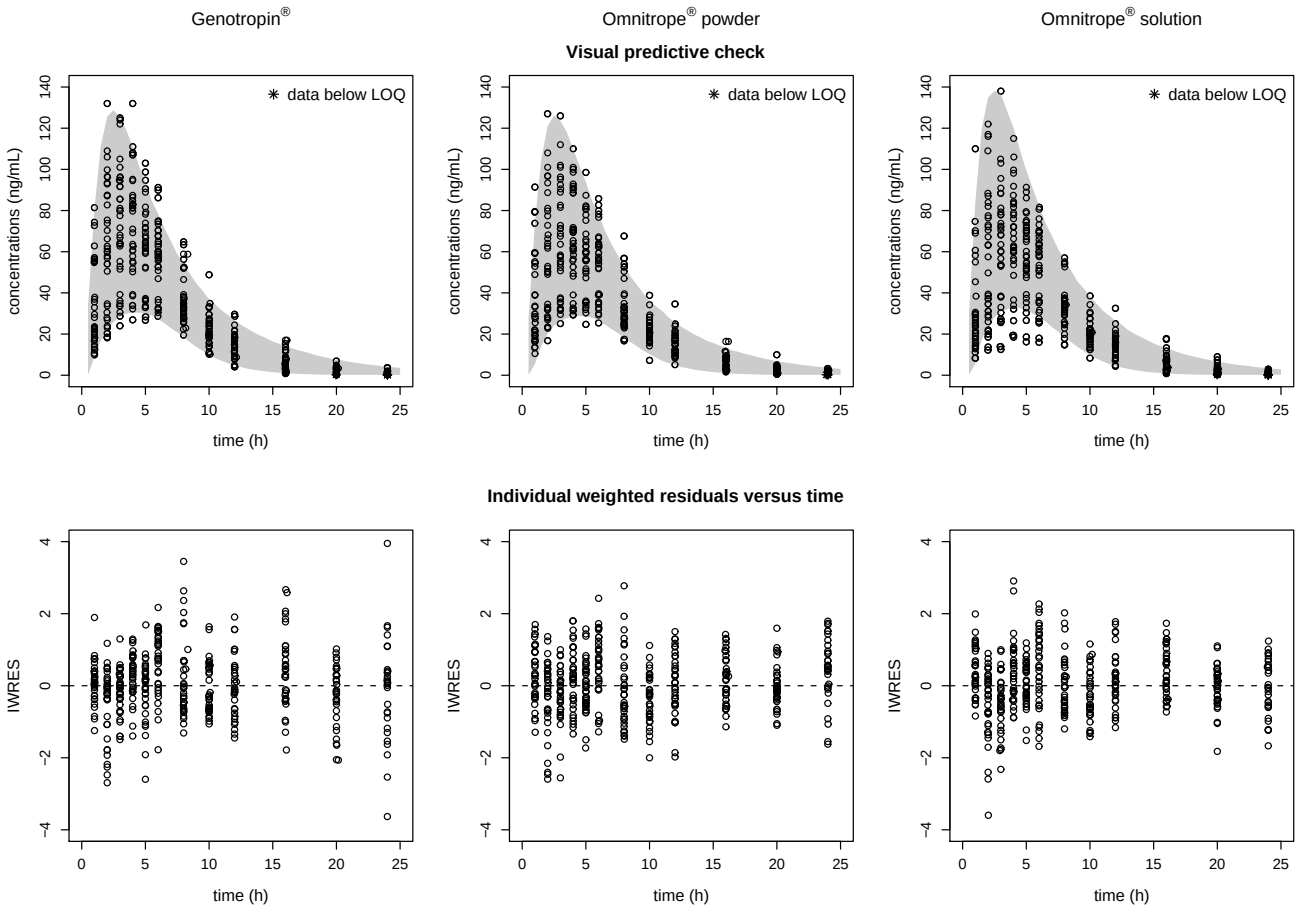
Genotropin®
Omnitrope® powder
Omnitrope® solution
Visual predictive check
∗ data below LOQ
concentrations (ng/mL)
time (h)
Individual weighted residuals versus time
IWRES

Figure 5