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.

Anne Dubois, Sandro Gsteiger, Etienne Pigeolet, France Mentré

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

Dubois A.$^{(1)}$, Gsteiger S.$^{(2)}$, Pigeolet E.$^{(2)}$ and Mentré F.$^{(1)}$

$^{(1)}$ INSERM UMR738, Paris, France; University Paris Diderot, Paris, France

$^{(2)}$ Novartis Pharma AG, Basel, Switzerland

Correspondence and request for reprint to Anne Dubois

INSERM UMR 738, Université Paris Diderot

16 rue Henri Huchard, 75018 Paris,France

Tel: 33 (0)1 57 27 75 39; fax: 33(0)1 57 27 75 21

e-mail: anne.dubois@inserm.fr

Abstract

The main objective of this work is to compare the standard bioequivalence tests based on individual estimates of the area under the curve and the maximal concentration obtained by non compartmental analysis (NCA) to those based on individual empirical Bayes estimates (EBE) obtained by nonlinear mixed effects models. We evaluate by simulation the precision of sample means estimates and the type I error of bioequivalence tests for both approaches. Crossover trials are simulated under H_0 using different numbers of subjects (N) and of samples per subject (n). We simulate concentration-time profiles with different variability settings for the between-subject and within-subject variabilities and for the variance of the residual error. Bioequivalence tests based on NCA show satisfactory properties with low and high variabilities, except when the residual error is high which leads to a very poor type I error or when n is small which leads to biased estimates. Tests based on EBE lead to an increase of the type I error when the shrinkage is above 20% which occurs notably when NCA fails. In those cases, tests based on individual estimates cannot be used.

Keywords

pharmacokinetics; bioequivalence tests; non compartmental analysis; nonlinear mixed effects model; SAEM algorithm

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) (1) and the European Medicines Evaluation Agency (EMEA) (2). FDA and EMEA recommend to test bioequivalence from the log ratio of the geometric means of two parameters: the area under the curve (AUC) and the maximal concentration (C_{max}). These endpoints are usually estimated by non compartmental analysis (NCA) using the trapezoidal rule to evaluate AUC (3). NCA requires few hypotheses but a large number of samples per subject (usually between 10 and 20).

PK data can also be analyzed using nonlinear mixed effects models (NLMEM). This method is more complex than NCA but has several advantages: it takes benefit of the knowledge accumulated on the drug and can characterize the PK with few samples per subject. This allows to perform analyses in patients, the target population, and in whom pharmacokinetics can be different from healthy subjects. Non compartmental AUC is computed by trapezoidal rule which ignores assay error. NCA does not take into account non linear pharmacokinetics, which can bias the bioavailability estimation (4) and may amplify small bioavailability differences between drug products (5). The European guideline on similar biological medicinal products, which frequently exhibit non linear pharmacokinetics, recommends to estimate in the comparative PK studies, the elimination characteristics such as clearance (6). It is known that in these conditions, clearance is not accurately estimated by NCA. Models can also lead to better understanding of the biological system than a fully empirical approach and therefore help interpret ambiguous results.

However, the use of NLMEM is still rare in early phases of drug development or to analyze crossover studies. There are only seven published studies which use NLMEM to analyze bioequivalence trials (7, 8, 9, 10, 11, 12, 13) and except in Zhou et al (12), all analyze a dataset with many samples per subject. Five papers (7, 8, 9, 10, 13) compare tests based on individual NCA estimates to tests based on NLMEM and all conclude that the results are similar. Yet, they use different statistical approaches to test bioequivalence with NLMEM. Furthermore, none perform bioequivalence tests on individual estimates of AUC and C_{max} obtained from NLMEM. Pentikis et al (8) propose the estimation of AUC and C_{max} by standard nonlinear regression as an alternative to the NCA and Zhou et al (12) perform bioequivalence tests on the individual empirical Bayes estimates (EBE) of the volume of distribution and the steady-state through concentration. Otherwise, bioequivalence tests are performed on treatment effect parameters (7, 8, 9, 10, 11, 13). All authors agree that simulation studies are needed to evaluate bioequivalence tests based on NLMEM and to compare them to tests based on individual NCA estimates.

In this work, we compare the standard analysis of bioequivalence crossover trials based on NCA to the same usual analysis based on individual EBE obtained by NLMEM. We study the influence of the design for each approach. There is already one published simulation study of Panhard and Mentré which evaluates bioequivalence tests based on EBE estimated through NLMEM (14). Our present study relies on the work of Panhard and Mentré as starting point and adds several new features.

The major distinctness concerns the studied tests on the individual estimates (EBE or NCA). Panhard and Mentré perform the Student paired test and the Wilcoxon paired signed rank test whereas we use a linear mixed effects model (LMEM). As specified in the regulatory guidelines (1, 2), 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, LMEM including treatment, period, sequence and subject effects are usually used to analyze the log-transformed data (15).

Panhard and Mentré limit their comparison to bioequivalence tests on AUC and do not evaluate tests based on C_{max}. In the present study, both endpoints are analyzed; indeed we expect some issues for bioequivalence test performed on C_{max} as the estimation of C_{max} by NCA is sensitive to the design and the computation of C_{max} from EBE is more complex than for AUC. To simulate PK profiles and then to estimate individual parameters by NLMEM, Panhard and Mentré use a pharmacokinetic model parametrized using AUC as one of the PK parameters whereas we choose a more common parameterization, replacing AUC by the clearance of the drug.

For the estimation of NLMEM parameters, Panhard and Mentré use an algorithm based on a first order linearization with respect to the random effects, the first order conditional estimates (FOCE) algorithm (16) implemented in the R function `nlme` (17). The FOCE algorithm is the more widely used algorithm and corresponds to the industry standard for model-based PK analyses as it is implemented in NONMEM . Yet, this algorithm presents some convergence issues which could be avoided with the use of a stochastic algorithm using the exact maximum likelihood, such as the stochastic approximation expectation maximisation (SAEM) algorithm (18, 19, 20). The SAEM algorithm is implemented in the free software MONOLIX

(21) (first version February 2005) and is applied to several population PK analyses (22, 23, 24).

The main objective of this work is to compare standard bioequivalence tests based on individual estimates of AUC and C_{max} obtained by NCA or by NLMEM. The comparison is based on the precision of the sample means of $\log(AUC)$ and $\log(C_{max})$ and on the type I error of bioequivalence tests for both estimation methods. In section 2 of the article, we describe the model, the simulation study, both estimation methods (NCA and NLMEM), the evaluation of precsion of sample means, how bioequivalence tests are performed and how shrinkage on the tested parameters is estimated. The main results of the simulation are exposed in section 3. Finally, the study results and perspectives are discussed.

2 Methods

2.1 Simulation study

2.1.1 Simulation model

We analyze two-period two-sequence crossover PK trials where subjects are randomly allocated to one of two treatment sequences. In the first sequence ($Ref - Test$), subjects receive the reference treatment (Ref) and the test treatment ($Test$) in period one and two, respectively. In the second sequence ($Test - Ref$), subjects receive treatments in the reverse order ($Test$ then Ref). Designs are balanced, $i.e.$ there is the same number of subjects $N/2$ for each sequence.

In the following, we denote y_{ijk} the concentration for individual i ($i = 1, \cdots, N$) at sampling time j ($j = 1, \cdots, n_{ik}$) for period k ($k = 1, 2$). We also denote f the nonlinear pharmacokinetic function which links concentrations to sampling times. The nonlinear mixed effects model can be written as follows:

$$y_{ijk} = f(t_{ijk}, \theta_{ik}) + \epsilon_{ijk} \tag{1}$$

where $\theta_{ik} = (\theta_{ikl}; l = 1, \cdots, p)'$ is the p-vector of the PK parameters of subject i for period k. ϵ_{ijk} is the residual error assumed to be normally distributed with zero mean and variance σ^2_{ijk}, with:

$$\sigma^2_{ijk} = (a + b\ f(t_{ijk}, \theta_{ik}))^2 \tag{2}$$

This is a combined error model with two parameters: a for the additive and b for the proportional part. We assume a multivariate log-normal distribution for the individual parameters θ_{ik}. In absence of covariates, the l^{th} individual parameter can be decomposed as:

$$\theta_{ikl} = \mu_l\ e^{\ \eta_{il} + \kappa_{ikl}} \tag{3}$$

with $\mu = (\mu_l;\ l = 1, \cdots, p)'$ the p-vector of fixed effects, $\eta_i = (\eta_{il};\ l = 1, \cdots, p)'$ the vector of random effects of subject i and $\kappa_{ik} = (\kappa_{ikl};\ l = 1, \cdots, p)'$ the vector of random effects of subject i at period k. η_i represents the variability between individuals and it is named between-subject variability (BSV). κ_{ik} represents the variability between two periods of treatment for the same individual and it is called within-subject variability (WSV). η_i and κ_{ik} are assumed to be normally distributed with zero mean and with covariance matrices of size $p \times p$ denoted Ω and Γ, respectively. In this study, we assume that Ω and Γ are diagonal. η_i, κ_{ik} and ϵ_{ijk} are assumed to be independent.

We introduce three categorical covariates into the statistical model: the treatment T_{ik}, the period P_k and the sequence S_i. The reference classes for each covariate are defined as follows: T_{ik} is fixed to zero for the treatment Ref and is equal to 1 for the treatment $Test$; P_k is fixed to zero for the first period and is equal to 1 for the second one; S_i is fixed to zero for the first sequence $Ref-Test$ and is equal to 1 for the second one $Test-Ref$. $\beta_T = (\beta_{T,l}\,;\ l = 1, \cdots, p)'$, $\beta_P = (\beta_{P,l}\,;\ l = 1, \cdots, p)'$ and $\beta_S = (\beta_{S,l}\,;\ l = 1, \cdots, p)'$ correspond to vectors of the treatment, period and sequence effect. With these three covariates, μ_l of Eq.(3) is replaced by μ_{ikl} defined as:

$$\mu_{ikl} = \lambda_l \ e^{\ \beta_{T,l}\, T_{ik} \,+\, \beta_{P,l}\, P_k \,+\, \beta_{S,l}\, S_i} \tag{4}$$

with $\lambda = (\lambda_l\,;\ l = 1, \cdots, p)'$ the p-vector of the fixed effects for the reference classes.

2.1.2 Theophylline pharmacokinetics

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 ones in population pharmacokinetics (17) and are used in previous simulation studies done by Panhard et al. (14, 25). The theophylline data include twelve subjects receiving a single oral dose of theophylline depending on their body weight (from 3 to 6 mg). For each patient, ten blood samples were taken at 0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12 and 24 h after administration and serum concentrations were measured. A one compartment model with first order absorption and first order elimination adequally describes the data and can be written as follows:

$$f(t, \theta) = \frac{FDk_a}{CL - Vk_a} \left(e^{-k_a \ t} - e^{-CL/V \ t} \right) \tag{5}$$

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. As only data after oral administration are obtained, the bioavailability cannot be estimated and, consequently, the vector θ of PK parameters is equal to $(k_a, CL/F, V/F)$.

2.1.3 Simulation features

In this simulation study, we use rather similar settings as those of the simulation studies performed by Panhard et al. (14, 25). However we simulate two-period, two-sequence crossover pharmacokinetic trials whereas they simulate two-period, one-sequence crossover trials. For each trial, $N/2$ subjects are allocated to the sequence $Ref-Test$ and $N/2$ subjects are allocated to the sequence $Test - Ref$. We fix the dose for all subjects to 4 mg which corresponds to the rounded median dose of the theophylline study. The vector of population parameters λ is composed of $(\lambda_{k_a} = 1.48\ h^{-1}, \lambda_{CL/F} = 40.36\ mL/h, \lambda_{V/F} = 0.48\ L)$ for the reference treatment. In order to mimic a change in bioavailability, we add a treatment effect $\beta_T = (0, \beta_{T,CL/F}, \beta_{T,V/F})'$ on $\log(\lambda)$, *i.e.* we multiply $\lambda_{CL/F}$ by $e^{\beta_{T,CL/F}}$ and $\lambda_{V/F}$ by $e^{\beta_{T,V/F}}$ for the test treatment. The modification of bioavailability also affects AUC and C_{max}. Indeed, $AUC = FD/CL$ and C_{max} is defined as:

$$\begin{aligned} C_{max} &= f(t_{max}, \theta) = \frac{FD}{V}\ e^{-CL/V\ t_{max}} \\ \text{with } t_{max} &= \frac{\log(k_a) - \log(CL/V)}{k_a - CL/V} \end{aligned} \tag{6}$$

We do not simulate a period effect or a sequence effect. We simulate with two levels of variability for the between-subject and within-subject variability. In the following, BSV and

WSV are given as standard deviations of the log-transformed parameters multiply by 100 to be expressed in percent. The standard deviation on the log scale corresponds approximately to the coefficient of variation on the ordinary scale. For the low level, we fix BSV to 20% for k_a and CL/F and to 10% for V/F; WSV is fixed to half BSV for the three parameters. For the high level, we fix BSV to 50% and WSV to 15% for the three parameters. We also simulate with two levels of variability for the residual error: $a = 0.1\ mg/L,\ b = 10\%$ for the low level, and $a = 1\ mg/L,\ b = 25\%$ 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 and WSV and for the residual error, $S_{h,l}$, the variability setting with high variability for BSV and WSV and low for the residual error, and $S_{h,h}$, the variability setting with high variability for BSV and WSV and for the residual error. The three variability settings are summarized in Table I.

2.1.4 Simulation process

For each subject $i = 1, \cdots, N$ of each simulated trial $m = 1, \cdots, M$, we simulate a vector of random effects η_i in $\mathcal{N}(0, \Omega)$ and two vectors of random effects κ_{ik} in $\mathcal{N}(0, \Gamma)$, one for each period $k = 1, 2$. To get the logarithm of each individual parameters $\log(\theta_{ikl})$, we add the logarithm of the mean parameter $\log(\lambda_l)$, the treatment effect $\beta_{T,l}$ if needed (depending on the treatment group and the PK parameter considered), and both random effects η_{il} and κ_{ikl}. The concentrations $f(t_{ijk}, \theta_{ik})$ predicted by the PK model at time t_{ijk} $(j = 1, \cdots, n_{ik})$ are then computed using the individual parameters. In these simulations, the sampling times for all subjects and both periods are similar. So $j = 1, \cdots, n$, where n is a fixed number of sampling times for each simulated design. Finally, we add a residual error, generated from a normal distribution $\mathcal{N}(0,(a+b\ f(t_{ijk},\theta_{ik}))^2)$, to each predicted concentration to obtain the simulated concentrations y_{ijk}. We do not incorporate in the simulation a limit of quantification (LOQ) because NCA cannot handle such data, contrary to the SAEM algorithm, and we do not want to favour the later. In the rare cases where the simulated concentration is below zero, we fix it to the value 0.1 mg/L.

We expect more of these fixed concentrations when variability increases but their proportion could also differ from a design to another if the sampling times differ. Consequently, for each simulated design and each variability setting, we compute the proportion of the concentrations fixed to 0.1 mg/L and study the corresponding sampling times.

2.1.5 Simulation designs

We simulate trials with four different designs, which are also used by Panhard et al (14, 25). 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 each 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. For each design and each variability setting, we simulate 1000 trials under two different hypotheses: $H_{0;80\%}$ where $\beta_T = (0, \log(0.8), \log(0.8))'$ and $H_{0;125\%}$ where $\beta_T = (0, \log(1.25), \log(1.25))'$. For each simulated trial, each simulated design and each variability setting, the simulated concentrations for the reference treatment are equal in both simulated hypotheses. In the following, we call simulation setting the association of one design with one variability setting and one hypothesis ($H_{0;80\%}$ or $H_{0;125\%}$). Considering this, there are 18 different simulation settings (8 for $S_{l,l}$ and $S_{h,l}$, and 2 for $S_{h,h}$). All simulations are performed using the statistical software R 2.7.1. Figure 1 displays the individual data of one trial simulated under $H_{0;80\%}$ and $H_{0;125\%}$ with the intermediate design and three variability settings ($S_{l,l}$, $S_{h,l}$ and $S_{h,h}$).

2.2 Estimation of individual parameters

2.2.1 Notations

We perform bioequivalence tests on AUC and C_{max}. To estimate the individual parameters using NCA or NLMEM, we do not consider periods or sequences. Only the treatment group (Ref or $Test$) is taken into account. For each simulated trial $m = 1, \cdots, 1000$ of one simulation setting, there are $2N$ individual AUC and $2N$ individual C_{max}, one for each subject $i = 1, \cdots, N$ and each treatment group.

In the following, for one simulated trial, we call $AUC_i^{(Ref)}$ the true value of the individual AUC of subject i for the reference treatment and $AUC_i^{(Test)}$ the true value of the individual AUC of subject i for the test treatment; we also define $\widehat{AUC}_i^{(Ref)}$ the estimated value of individual AUC of subject i for the treatment Ref obtained from NCA or NLMEM and $\widehat{AUC}_i^{(Test)}$ the corresponding AUC for the treatment $Test$.

Same notations are applied to C_{max}. $C_{max\,i}^{(Ref)}$ and $C_{max\,i}^{(Test)}$ are the true value of the individual C_{max} of subject i for the treatment Ref and $Test$, respectively. $\widehat{C_{max}}_i^{(Ref)}$ and $\widehat{C_{max}}_i^{(Test)}$ are the corresponding estimated value of individual C_{max} obtained from NCA or NLMEM.

In some cases, we may refer to these different individual parameters without specifying the treatment group. For each simulated trial, $AUC_i^{(Ref)}$, $AUC_i^{(Test)}$, $C_{max\,i}^{(Ref)}$ and $C_{max\,i}^{(Test)}$ are computed from the corresponding individual parameters k_a, CL/F and V/F simulated as described in section 2.1.4.

2.2.2 Estimation based on non compartmental analysis

First, we estimate AUC and C_{max} by non compartmental analysis (3) using a R function named `mnca` which we develop. For each simulated trial, this function provides the estimation of different NCA parameters for each subject and each treatment group. Different options have to be specified in `mnca`. 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 infinity), we compute the terminal slope equal to CL/V using the logarithm of the last concentrations to perform a linear regression. To do so, we use a fixed number of concentrations which depends on the number of samples per subject in the design.

To avoid biased estimation of the terminal slope, the first point used for its computation should be on the descending side of the concentration curve and not too close to C_{max}. Using the mean value of PK parameters, t_{max}, the sampling time corresponding to C_{max}, is about 2.06 h for both treatment groups (contrary to C_{max}, t_{max} is not affected by the change of bioavailability). Consequently, 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. NCA is normally performed on PK profiles containing ten sampling times per subject or more. For intermediate and sparse designs where $n = 5$ and $n = 3$ respectively, the total AUC is estimated by NCA for completeness. For these two designs, 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.

Figure 2 displays the individual concentration curves of one simulated trial for the original, intermediate and sparse designs and the two variability settings $S_{l,l}$ and $S_{h,l}$. The bottom left graphic of the Figure 1 presents a similar graphic for the intermediate design and $S_{h,h}$, completing our illustration. For rich and intermediate designs, the number of concentrations used to compute the terminal slope seems reasonable. Same observation can be done for the rich design because the sampling times are similar to those of the original design, only the number of subjects differs. For sparse design, the number of concentration used to compute the terminal slope is chosen by default, first point being close to C_{max}.

Other assumptions are made to compute the terminal slope, to handle particular PK profiles, especially for the intermediate and sparse designs where only two points are used for the estimation. If the last two concentrations increase instead of decreasing or if they are similar up to the sixth digit, we consider the terminal slope be missing, *i.e.* there is no estimation of the total AUC for the subject and treatment concerned. The proportion of missing $\widehat{AUC}_i$ should increase with variability and could differ from a design to another due to different sampling times. Consequently, for each design and each variability setting, we compute the proportion of missing $\widehat{AUC}_i$.

For all designs, C_{max} is estimated as the maximal concentration observed. Contrary to AUC, there is no missing C_{max}.

2.2.3 Estimation based on nonlinear mixed effects model

We also estimate AUC and C_{max} from the individual empirical Bayes estimates of the PK parameters after population analyses. In this study, we use the SAEM algorithm implemented in MONOLIX 2.4 to estimate the NLMEM parameters (population and individual parameters). For each simulated trial, we analyze separately the concentrations of each treatment group using NLMEM without taking into account periods and sequences. As each subject receives both treatments, data of each treatment group contain observations from all subjects. In the following, we describe the statistical model used to fit the data of the reference treatment. We consider $y_{ij}^{(Ref)}$ the concentration for individual i ($i = 1, \cdots, N$) at time t_{ij} ($j = 1, \cdots, n$) and for the treatment Ref. Depending on the sequence of the subject i, $y_{ij}^{(Ref)}$ corresponds to concentration of the first or second period. The statistical model used has no covariate because no period or sequence effect are incorporated. Furthermore, since periods are not considered, WSV cannot be separated from BSV. Consequently, the l^{th} individual parameter is defined as:

$$\theta_{il}^{(Ref)} = \mu_l^{(Ref)} \ e^{\ \eta_{il}^{(Ref)}} \tag{7}$$

$\Omega^{(Ref)}$ is the covariance matrix of the vector of random effects $\eta_i^{(Ref)}$. A similar statistical model is applied to fit the data of the treatment $Test$.

Of note, given the BSV and WSV, the overall variability is equal for both treatment groups, *i.e.* $\Omega^{(Ref)} = \Omega^{(Test)}$. However, for each simulated trial, their estimates, $\widehat{\Omega}^{(Ref)}$ and $\widehat{\Omega}^{(Test)}$, are different. The overall simulated variability is 22.4% for k_a and CL/F and 11.2% for V/F under $S_{l,l}$, and 52.2% for the three PK parameters under $S_{h,l}$ and $S_{h,h}$.

After having estimated the population parameters for the data of one treatment group of one simulated trial, we estimate the conditional modes of the corresponding individual parameters which are defined as the individual empirical Bayes estimates. These EBE provide the individual estimates of PK parameters (k_a, CL/F and V/F). We then derive individual $\widehat{AUC}_i^{(Ref)}$ and $\widehat{C_{max}}_i^{(Ref)}$ or $\widehat{AUC}_i^{(Test)}$ and $\widehat{C_{max}}_i^{(Test)}$ depending on the treatment group considered. Contrary to NCA, there is no missing $\widehat{AUC_i}$ obtained by NLMEM using the SAEM algorithm.

2.2.4 Evaluation of estimates of sample means

In this study, we compute individual $\widehat{AUC_i}$ and $\widehat{C_{max\,i}}$ for 1000 replicates of different designs, different variabilities and different treatment groups using two types of estimation. To analyze and compare the accuracy and precision of the estimates of the sample means of $\log(AUC)$ and $\log(C_{max})$ using NCA or EBE, we compute estimation error for each treatment group (Ref or $Test$) of each simulated trial. To take into account sampling variability, for each dataset we compute the estimation error as the difference between the sample mean of the estimates (NCA or EBE) and the sample mean of the true simulated values. In the following, definitions are given for $\widehat{AUC}_i^{(Ref)}$. Same definitions apply to $\widehat{AUC}_i^{(Ref)}$, $\widehat{C_{max}}_i^{(Ref)}$ and $\widehat{C_{max}}_i^{(Test)}$. For each simulated trial, the estimation error for the sample mean of $\log(AUC)$ for the reference treatment is computed as:

$$ee_{AUC}^{(Ref)} = \frac{1}{N^*}\sum_{i=1}^{N^*}\log(\widehat{AUC}_i^{(Ref)}) - \frac{1}{N}\sum_{i=1}^{N}\log(AUC_i^{(Ref)}) \qquad (8)$$

with $\widehat{AUC}_i^{(Ref)}$ the AUC estimated by NCA or derived from EBE for subjects $i = 1, \cdots, N^*$, and $AUC_i^{(Ref)}$ the true simulatd parameter for subjects $i = 1, \cdots, N$. For the estimation of individual parameters by NCA, there may be missing $\widehat{AUC_i}$, so that $N^* \leq N$.

For one simulation setting, we call $ee^{(Ref)}_{AUC,m}$ the estimation error for the sample mean of $\log(AUC)$ computed for the reference treatment and the m^{th} simulated trial ($m = 1, \cdots, 1000$). We then define the bias and root mean square error (RMSE) computed from $ee^{(Ref)}_{AUC,m}$ over the 1000 replicates as:

$$\begin{aligned} bias^{(Ref)}_{AUC} &= \frac{1}{1000}\sum_{m=1}^{1000} ee^{(Ref)}_{AUC,m} \\ rmse^{(Ref)}_{AUC} &= \sqrt{\frac{1}{1000}\sum_{m=1}^{1000} (ee^{(Ref)}_{AUC,m})^2} \end{aligned} \tag{9}$$

As well as computing bias and RMSE, we compute the 95% confidence interval of $bias^{(Ref)}_{AUC}$ using the standard error of the mean and the 97.5% quantile of the Gaussian distribution. If zero does not belong to the 95% confidence interval of $bias^{(Ref)}_{AUC}$, we can conclude that bias is significantly different from zero with a type I error of 5%.

2.3 Bioequivalence test

2.3.1 Implementation of the two one-sided tests

We perform the standard bioequivalence analysis recommended by FDA and EMEA (1, 2). The individual parameters are log-transformed and analyzed using a linear mixed effects model written as follows:

$$\log(\theta_{ikl}) = \nu_l + \beta_{T,l}\ T_{ik} + \beta_{P,l}\ P_k + \beta_{S,l}\ S_i + \xi_{il} + \epsilon_{ikl} \tag{10}$$

where θ_{ikl} represents 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 ($k = 1, 2$). ν_l is the mean value for the studied log-transformed metric. The three covariates T_{ik}, P_k and S_i, for treatment, period and sequence are defined as before. It is assumed that the random subject effect ξ_{il} ($l = 1, 2$) and the residual error ϵ_{ikl} ($l = 1, 2$) are independently normally distributed with zero mean.

For each simulation setting, the individual estimates $\widehat{AUC}_i$ and $\widehat{C_{max\,i}}$ obtained from NCA and NLMEM are analyzed by the LMEM described above. To check the properties of the TOST, we also analyze the true simulated value AUC_i and $C_{max\,i}$. As specified before, for AUC estimated by NCA, they may be missing $\widehat{AUC}_i$. In that case, the LMEM is performed on less than $2N$ $\widehat{AUC}_i$.

After fitting the LMEM to individual metrics, a bioequivalence test is performed on the estimate of treatment effect $\widehat{\beta}_{T,l}$. The null hypothesis of the bioequivalence test recommended by the guidelines (1, 2) and performed on the l^{th} individual parameter is H_0: $\{\beta_{T,l} \leq \log(0.8)$ or $\beta_{T,l} \geq \log(1.25)\}$. H_0 is rejected if the 90% confidence interval (90% CI) of $\widehat{\beta}_{T,l}$ lies within $[\log(0.8); \log(1.25)]$. These limits of the bioequivalence test correspond to a ratio of the geometric mean falling within 80%-125%. This approach based on the 90% CI is equivalent to Schuirmann's two one-sided tests (TOST) procedure (26). H_0 is composed of two unilateral hypotheses $\{\beta_{T,l} \leq \log(0.8)\}$ and $\{\beta_{T,l} \geq \log(1.25)\}$. Both are tested separately by a one-sided test with a type I error of 5%. The p-value of the TOST is the maximum of both p-values of the one-sided tests and for each test the limit is the 95% quantile of the Student distribution with df degrees of freedom.

For balanced datasets, the $N/2$ subjects of each sequence are considered as two independent samples from normal populations with equal variances, and $df = N - 2$ (15, 27). For unbalanced datasets, *i.e.* when there is one or more missing $\widehat{AUC}_i$ in a dataset for NCA, the determination of the degrees of freedom is more complex. Different approximations are available as for example the containment method (28), the Kenward-Roger adjustment (29) or the Satterthwaite's procedure approximation (28, 29). In this study, we use the R function `lme` from the package `nlme` to perform the LMEM in which the degrees of freedom are estimated using the containment method (17). There, the degrees of freedom are calculated as: $df = n_{obs} - N - 2$ where n_{obs} is the total number of individual parameters. When there is no missing value, this approach coincides with the degrees of freedom computed in balanced datasets (because then $n_{obs} = 2N$).

2.3.2 Evaluation of the type I error

Bioequivalence tests are evaluated for $\widehat{AUC}_i$ and $\widehat{C_{max\,i}}$ estimated by NCA or NLMEM on trials simulated under the composite null hypothesis H_0. Bioequivalence tests are also performed on the true simulated values AUC_i and $C_{max\,i}$. The type I error of the TOST procedure is defined as the supremum of the type I errors over the null space (30). It corresponds to the supremum of the type I error of the two one-sided tests. As suggested by Liu and Weng (31), 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 for each design of each variability setting 1000 trials under each unilateral hypothesis $H_{0;80\%}$ and $H_{0;125\%}$ as specified before.

For each unilateral 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. If the bioequivalence tests were performed on the true parameters (AUC_i and $C_{max\,i}$), the results of both type I errors should be identical because $H_{0;80\%}$ and $H_{0;125\%}$ are symetric but we are working with estimates. As proposed by Panhard and Mentré (14), we define the global type I error as the maximum value of both type I errors estimated. Due to the 1000 replicates, the 95% prediction interval (95% PI) for a type I error of 5% is $[3.7\%; 6.4\%]$.

2.3.3 Shrinkage and tests based on empirical Bayes estimates

It is known in NLMEM that, with sparse individual information, the individual estimates of random effects shrink towards their mean value which is zero (32). For the reference treatment group of each simulated trial, the shrinkage on the l^{th} individual EBE (k_a, CL/F or V/F) can be defined as:

$$Sh_l^{(Ref)} = 1 - \frac{var(\widehat{\eta}_{il}^{(Ref)})}{\widehat{\omega}_l^{(Ref)\,2}} \tag{11}$$

where $var(\widehat{\eta}_{il}^{(Ref)})$ is the empirical variance of the l^{th} individual estimated random effects and $\widehat{\omega}_l^{(Ref)\,2}$ is the estimated variance of the corresponding random effects.

AUC and C_{max} are secondary parameters of the NLMEM because they are defined as functions of the PK parameters, k_a, CL/F and V/F. As the shrinkage on individual EBE, the shrinkage on $\log(AUC)$ and $\log(C_{max})$ can also be computed. Consequently, we can study the link between the type I error of bioequivalence tests based on EBE and the amount of shrinkage.

For $\log(AUC)$, Eq.(11) can be expressed as:

$$Sh_{AUC}^{(Ref)} = 1 - \frac{var(\log(\widehat{AUC}_i^{(Ref)}))}{\widehat{\omega}^2_{AUC^{(Ref)}}} \tag{12}$$

where $var(\log(\widehat{AUC}_i^{(Ref)}))$ is the empirical variance of the individual estimates $\log(\widehat{AUC}_i^{(Ref)})$ and $\widehat{\omega}^2_{AUC^{(Ref)}}$ is its estimated variance in the model. As $\log(AUC) = \log(D) - \log(CL/F)$, $\omega^2_{AUC^{(Ref)}} = \omega^2_{CL/F^{(Ref)}}$ and $\widehat{\omega}^2_{AUC^{(Ref)}}$ is the estimated value $\widehat{\omega}^2_{CL/F^{(Ref)}}$.

For one simulation setting, we call $Sh^{(Ref)}_{AUC,m}$ the shrinkage on $\log(AUC)$ computed for the reference treatment for the m^{th} simulated trial ($m = 1, \cdots, 1000$). To summarize the 1000 $Sh^{(Ref)}_{AUC,m}$ of each simulation setting, we compute the median shrinkage over these 1000 values.

Eq.(12) can be applied to $\log(C_{max})$; $var(\log(\widehat{C_{max}}_i^{(Ref)}))$ is computed from the individual estimates as for AUC. As the definition of C_{max} given in Eq.(6) is complex, the variance of $\log(C_{max})$ for the reference treatment, $\omega^2_{C_{max}^{(Ref)}}$ cannot be computed from $\omega^2_{k_a^{(Ref)}}$, $\omega^2_{CL/F^{(Ref)}}$ and $\omega^2_{V/F^{(Ref)}}$. It must be approximated for instance using the delta method (33). The expression and details are given in Appendix. As for AUC, the median shrinkage over the 1000 values of $Sh^{(Ref)}_{C_{max},m}$ is computed for each simulation setting.

3 Results

3.1 Simulated data and missing values

As explained in section 2.1.4, if the simulated concentration is below zero, it is fixed to 0.1 mg/L. As expected, the proportion of these fixed concentrations differs from one variability setting to another and from one design to another, except for the original and rich design where the sampling times are similar. The maximal proportion is rather small and is 0.03% for $S_{l,l}$, 1.6% for $S_{h,l}$ and 8.5% for $S_{h,h}$. For $S_{l,l}$, all fixed concentrations correspond to the last sampling time which is 24 h for all designs. For $S_{h,l}$, there are fixed concentrations corresponding to different sampling times but fixed concentrations at 24 h are majoritary, with a minimal proportion of 90%. For $S_{h,h}$, fixed concentrations corresponds mostly to 24 h (54%) and then mainly to 0.25 h (20%) and 12 h (19%).

Over all the simulations, some $\widehat{AUC}_i$ estimated by NCA are missing due to particular individual PK profiles (see section 2.2.2). The proportion of missing $\widehat{AUC}_i$ is similar in both hypotheses and remains rare for the four designs of $S_{l,l}$ and $S_{h,l}$. For both variability settings, the maximal proportion corresponds to the intermediate design ($N = 24$, $n = 5$) with 0.02% and 3.3% for $S_{l,l}$ and $S_{h,l}$, respectively. This proportion is 25% for $S_{h,h}$. Among missing $\widehat{AUC}_i$ of $S_{h,h}$, 12% are due to concentrations fixed to 0.1 mg/L, *i.e.* due to two similar last concentrations. Other missing $\widehat{AUC}_i$ are due to two last concentrations increasing instead of decreasing. As expected, there is no simulated trial where all $\widehat{AUC}_i$ for both treatment groups are missing. In other words, the estimation error for the sample mean of $\log(AUC)$ or $\log(C_{max})$ is computed on the 1000 simulated trial for each simulation setting, and the type I errors of bioequivalence test are estimated on 1000 replicates for AUC and C_{max} for both hypotheses $H_{0;80\%}$ and $H_{0;125\%}$.

3.2 Evaluation of estimates of sample means

Figure 3 displays the bias (top) and RMSE (bottom) on sample mean estimates for $\log(AUC)$ (left) and $\log(C_{max})$ (right) estimated for the reference treatment. Results are similar for both treatment groups (Ref and $Test$) and both unilateral hypotheses (results not shown). The 95% confidence interval of the bias is not shown in Figure 3 because this interval is tighter than the width of the displayed symbol and all biases are significantly different from zero. There is more bias and larger RMSE for NCA than for EBE for all designs and all variability settings. Note that biases and RMSE are computed on log scale, so that, for instance, a value of 0.038 corresponds approximatively to an error of 3.8% on the ordinary scale for the geometric mean. For NCA estimates, the bias and RMSE increase when the number of samples per subject decreases and are lower for $S_{l,l}$ compared to $S_{h,l}$. For the intermediate design ($N = 24, n = 5$), the bias on the sample mean of $\log(AUC)$ is 0.038, 0.094 and 0.15 for $S_{l,l}$, $S_{h,l}$ and $S_{h,h}$, respectively; RMSE is 0.044, 0.12 and 0.21, respectively.

For individual estimates based on EBE, the bias is small (less than 0.02) for both parameters ($\log(AUC)$ and $\log(C_{max})$), all designs and all variability settings whereas RMSE increase when the number of samples per subject decreases and is majoritary lower for $S_{l,l}$ compared to $S_{h,l}$. For instance, for the intermediate design, the bias on the sample mean of $\log(AUC)$ is -0.0096, -0.016 and -0.010 for $S_{l,l}$, $S_{h,l}$ and $S_{h,h}$ respectively; RMSE is 0.019, 0.031 and 0.10, respectively.

3.3 Bioequivalence test

Table II and Figure 4 provide the results of the type I error of bioequivalence tests performed on the treatment effect of $\log(AUC)$ and $\log(C_{max})$. Table II contains the estimated type I error for each unilateral hypothesis, each design of each variability setting, for the true simulated values and both types of estimates (NCA and EBE). Figure 4 represents the global type I error for $\log(AUC)$ (top) and $\log(C_{max})$ (bottom) versus the design for each variability setting and both types of estimates. The global type I error is defined as the supremum of both estimated type I errors.

For the bioequivalence test performed on the true simulated values, the type I error for all

designs, all variability settings and both null hypotheses lie in the 95% PI of the nominal level showing the good performance of the TOST. Mostly, for one type of estimates (NCA or EBE) and one design of one variability setting, the type I errors of both hypotheses are close.
For $\log(AUC)$, the global type I error of test based on NCA estimates lies between the 95% PI of the nominal level for the four designs of $S_{l,l}$ and $S_{h,l}$ and it is much too conservative for $S_{h,h}$. For instance, for the intermediate design, the global type I error is respectively 4.3%, 5.2% and 0.8% for $S_{l,l}$, $S_{h,l}$ and $S_{h,h}$. For C_{max}, test based on NCA estimates has a correct global type I error for the original and intermediate designs simulated with $S_{l,l}$ and $S_{h,l}$. The global type I error is above the 95% PI for the sparse design ($N = 40, n = 3$) simulated with $S_{l,l}$ and $S_{h,l}$ and the intermediate design simulated with $S_{h,h}$.
Surprisingly, tests based on EBE often lead to an increased type I error especially for the sparse design. For AUC, the global type I error remains at the nominal level for the rich design ($N = 40, n = 10$). For C_{max}, the global type I error lies between the 95% PI for the rich and the original designs simulated with $S_{l,l}$. The global type I error increases when the number of samples per subject decreases and is lower for $S_{h,l}$ compared to $S_{l,l}$ and $S_{h,h}$. Most of the type I errors are below 10% for $S_{l,l}$ and $S_{h,l}$. For AUC and the intermediate design, the global type I error is respectively 8.0%, 7.1% and 22.2% for $S_{l,l}$, $S_{h,l}$ and $S_{h,h}$.

Figure 5 represents the global type I errors of bioequivalence tests for the treatment effect on $\log(AUC)$ (top) and $\log(C_{max})$ (bottom) obtained from NLMEM versus the median shrinkage on the corresponding parameter for the reference treatment. The distribution of the shrinkage is similar for both treatment (Ref and $Test$) and both unilateral hypotheses (results not shown). For both parameters, the median shrinkage is lower for $S_{h,l}$ than for $S_{l,l}$. For $\log(AUC)$, the median shrinkage is also higher for $S_{h,h}$ than for $S_{h,l}$. There is a clear relationship between the inflation of the global type I error and the amount of shrinkage with type I error greater than 15% for shrinkage greater than 20%.

4 Discussion

In this study, we compare the standard bioequivalence analysis performed on individual estimates of AUC and C_{max} obtained by NCA to the same bioequivalence analysis performed on individual EBE obtained by NLMEM. To do so, we perform a simulation study with different designs and different levels of variability. The estimation of parameters and the type I error are evaluated for both types of estimates.

Compared with the simulation study of Panhard and Mentré (14), we use the bioequivalence analysis recommended in the guidelines (1, 2) and we study both parameters (AUC and C_{max}). Besides, the simulation study of Panhard and Mentré is performed using the FOCE algorithm implemented in R function `nlme`. The FOCE algorithm is widely used to perform population PK analyses but, in simulation studies which compared different algorithms available, stochastic EM algorithms (like the SAEM algorithm) obtained the best results for accuracy and precision of estimates (34, 35).

As Panhard and Mentré, we simulate under both null hypotheses assuming a modification in the bioavailability F, *i.e.* assuming the same modification for CL/F and V/F which also affects similarly both tested parameters AUC and C_{max}. Consequently, the number of simulations are reduced because the unilateral hypothesis $H_{0;80\%}$ ($H_{0;125\%}$ respectively) for AUC

corresponds to the unilateral hypothesis $H_{0;80\%}$ ($H_{0;125\%}$ respectively) for C_{max}; the same set of simulations is used for both parameters. However, other choices may be suitable as any PK parameter is likely to change between two formulations of the same drug. For instance, a change in the elimination rate CL/V due to interaction with excipient could be possible (36). Furthermore, we study only a one compartment model. We do not simulate multi-compartmental models. For both types of estimates (NCA and EBE), we perform bioequivalence test on AUC and C_{max}. Even with a multi-compartmental model, PK parameters would be summarized with these two endpoints even though the relationship between C_{max} and the PK parameters could be more complicated than for a one compartment model. As shown in Figure 5, the increase of the type I error of bioequivalence test based on EBE is linked to the shrinkage which already appears with one compartment model. We think this relationship should be similar for multi-compartmental models where more shrinkage is expected.

Conversely to the bias for estimates based on EBE, the bias for estimates based on NCA depends on the number of samples per subject and is large for sparse design ($N = 40$, $n = 3$) 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 and could be problematic for atypical concentration profiles. This later issue is perfectly illustrated by the simulation settings under $S_{h,h}$ where 77% of the missing $\widehat{AUC}_i$ are due to the two last concentrations increasing instead of decreasing. Contrary to NCA estimates, there is no missing $\widehat{AUC}_i$ estimated by NLMEM due to this kind of PK profiles because all subjects are analyzed together and information given by classical PK profiles offset information given by particular ones. NCA does not take into account all the knowledge accumulated on the PK of the studied drug as each new analysis by NCA erases the past contrary to NLMEM. Finally, although we do not simulate such data, NCA applied to nonlinear pharmacokinetics provides meaningless parameters and it cannot handle data below the limit of quantification. In this study, we choose to not introduce LOQ in the simulation because we do not want to favour the SAEM algorithm which can fit such data. We are aware that fixing some concentrations to 0.1 mg/L could introduce some bias. To avoid such arbitrary fixing, another common procedure is to resample until a valid value is obtained; however, resampling can also introduce a bias. Anyhow, the proportion of fixing value remains very low for $S_{l,l}$ and $S_{h,l}$. It is more important for $S_{h,h}$ but it is responsible for only 12% of the missing $\widehat{AUC}_i$ estimated by NCA.

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 sample mean estimates, the type I error is maintained at the nominal level of 5%. Though, for simulation with $S_{h,h}$, the type I error of AUC is very conservative (0.8%) which shows the limits of NCA for data with high residual error.

Tests based on individual EBE have higher type I error than tests based on NCA estimates. Our results on the type I error for $S_{l,l}$ are consistent with the results obtained by Panhard and Mentré with the same variability setting. For the sparse design, the type I error of tests based on EBE is surprisingly high. In that case, EBE shrink towards their mean value and they are more similar in both treatment groups. Therefore, the discrimination of the AUC or C_{max} between both treatment groups is more difficult which leads to an increase of the type I error (bioequivalence is obtained more easily). These results are consistent with the results of the simulation study performed by Bertrand et al (37). In that work, they evaluate by simulation the analysis of variance (ANOVA) performed on individual EBE to test the influence of a single nucleotide polymorphism on a pharmacokinetic parameter of a drug. They show the impact of the shrinkage on the power of ANOVA. The power is reduced when the shrinkage increases. In other words, it is more difficult to discriminate between the genotypes with high shrinkage even when data are simulated with a difference.

As discussed by Schuirmann (26), the TOST procedure can be very conservative for highly variable drugs. Consequently, several improvements of this procedure have been proposed as in Berger et al (30), Brown et al (38) or Cao et al (39) to mention only a few. We are aware that there is still a great arguing on which bioequivalence test should be performed. However, we study only the classical TOST in this paper because our main objective is to compare the same standard bioequivalence analysis recommended in the guidelines (1, 2) and performed on individual estimates obtained by two estimation methods (NCA and EBE). Nevertheless, in this simulation study, the type I error of bioequivalence test performed on the true individual simulated values is always at the nominal level of 5%, even for $S_{h,h}$ where the variability is particularly high. Therefore, we can conclude that, in this study, there is no issue about the TOST procedure. Consequently, liberal or conservative type I errors of bioequivalence tests performed on estimates cannot be imputed to the TOST but rather to the individual parameters estimation.

Tests based on individual estimates, NCA estimates or EBE, cannot be used for data with high residual error or when the number of samples per subject is small. In those cases, the type I error for tests based on NCA estimates is very poor or NCA estimates are biased and the shrinkage of EBE induces an increase of the type I error. In these situations, other tests based on a global analysis of all data should be considered. Panhard et al. already developed a global bioequivalence Wald test based on NLMEM (14, 25). This test is directly performed on the treatment effect parameter after fitting together the data of both treatment groups with the estimation of within-subject variability. In this study, they also used the FOCE algorithm implemented in `nlme`. Recently, Panhard and Samson developed an extension of the SAEM algorithm for NLMEM including the estimation of the within-subject variability (40). However, the likelihood ratio test for bioequivalence has not been developed, due to the composite null hypothesis. Additional methodological developments and simulations are needed to study bioequivalence tests after global analysis of all PK data. This will be especially useful for drugs with non linear pharmacokinetics and conditions where rich sampling is difficult to achieve, *i.e.* in pediatric studies or for drugs which cannot be administered in healthy subjects for safety reasons, such as oncology drugs.

Acknowledgments

We would like to thank the Modeling and Simulations group at Novartis Pharma AG, Basel, which supports by a grant Anne Dubois during this work.

References

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

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

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

4. W. J. Jusko, J. R. Koup, and G. Alván. Nonlinear assessment of phenytoin bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics.* **4**:327–336 (1976).

5. N. Hayashi, H. Aso, M. Higashida, H. Kinoshita, S. Ohdo, E. Yukawa, and S. Hiquchi. Estimation of rhG-CSF absorption kinetics after subcutaneous administration using a modified Wagner-Nelson method with a nonlinear elimination model. *European Journal of Pharmaceutical Sciences.* **13**:151–158 (2001).

6. EMEA. *Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues.* Technical report, EMEA. (2006).

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

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

9. M. Combrink, M.-L. McFadyen, and R. Miller. 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 (1997).

10. G. A. Maier, G. F. Lockwood, J. A. Oppermann, G. Wei, P. Bauer, J. Fedler-Kelly, and T. Grasela. 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 (1999).

11. C. Hu, K. Moore, Y. Kim, and M. Sale. 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 (2003).

12. H. Zhou, P. Mayer, J. Wajdula, and S. Fatenejad. Unaltered etanercept pharmacokinetics

with concurrent methotrexate in patients with rheumatoid arthritis. *Journal of Clinical Pharmacology.* **44**:1235–1243 (2004).

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

14. X. Panhard and F. Mentré. 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 (2005).

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

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

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

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

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

20. A. Samson, M. Lavielle, and F. Mentré. The SAEM algorithm for group comparison

tests in longitudinal data analysis based on non-linear mixed-effects model. *Statistics in Medicine.* **26**:4860–4875 (2007).

21. The MONOLIX software http://software.monolix.org/ (accessed 05/07/09).

22. M. Lavielle and F. Mentré. Estimation of population pharmacokinetic of saquinavir in HIV patients and covariate analysis with the SAEM algorithm. *Journal of Pharmacokinetics and Pharmacodynamics.* **34**:229–249 (2007).

23. E. Comets, C. Verstuyft, M. Lavielle, P. Jaillon, L. Becquemont, and F. Mentré. Modelling the influence of MDR1 polymorphism on digoxin pharmacokinetic parameters. *European Journal of Clinical Pharmacology.* **63**:437–449 (2007).

24. J. Bertrand, J.-M. Treluyer, X. Panhard, A. Tran, S. Auleley, E. Rey, D. Salmon-Céron, X. Duval, F. Mentré, and the COPHAR2-ANRS 111 study group. Influence of pharmacogenetics on indinavir disposition and short-term response in HIV patients initiating HAART. *European Journal of Clinical Pharmacology.* **65**:667–678 (2009).

25. X. Panhard, A. M. Taburet, C. Piketti, and F. Mentré. 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 (2007).

26. D. J. Schuirmann. 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 (1987).

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

28. G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data.* Springer, New-York, 2001.

29. H. Brown and R. Prescott. *Applied mixed models in medicine - second edition.* John Wiley & sons, Chichester, 2006.

30. R. Berger and J. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science.* **11**:283–319 (1996).

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

32. R. Savić and M. Karlsson. Shrinkage in empirical Bayes estimates for diagnostics and estimation, 2007. PAGE 16 Abstr 1087 available at http://www.page-meeting.org/pdf_assets/9436-EBE_PAGE07_1_web.pdf (accessed 05/07/09).

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

34. P. Girard and F. Mentré. A comparison of estimation methods in nonlinear mixed effects models using a blind analysis, 2005. PAGE 14 Abstr 834 available at http://www.page-meeting.org/page/page2005/PAGE2005O08.pdf (accessed 05/07/09).

35. R. Bauer, S. Guzy, and C. Ng. Survey of population analysis methods and software for complex pharmacokinetic and pharmacodynamic models with examples. *The* AAPS *Journal.* **9**:60–83 (2007).

36. A. Rescigno, J. Powers, and E. E. Herderick. Bioequivalent or nonbioequivalent ? *Pharmalogical Research*. **43**:543–546 (2001).

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

38. L. D. Brown, J. T. G. Hwang, and A. Munk. An unbiased test for the bioequivalence problem. *The Annals of Statistics*. **25**:2345–2367 (1997).

39. L. Cao and T. Mathew. A simple numerical approach toward improving the two-one sided test for average bioequivalence. *Biometrical Journal*. **50**:205–211 (2008).

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

Appendix

Approximation of the variance of $\log(C_{max})$ by the delta method

For a one compartment model with first order absorption and first order elimination, C_{max} is defined in Eq.(6) as a function of the three PK parameters, k_a, CL/F and V/F. The variance of $log(C_{max})$, $\omega^2_{C_{max}}$, is approximated by the delta method (33) as:

$$\omega^2_{C_{max}} \approx \left(\frac{\partial \log(C_{max})}{\partial \log(k_a)}\right)^2_{\log(\mu)} \omega^2_{k_a} + \left(\frac{\partial \log(C_{max})}{\partial \log(CL/F)}\right)^2_{\log(\mu)} \omega^2_{CL/F} + \left(\frac{\partial \log(C_{max})}{\partial \log(V/F)}\right)^2_{\log(\mu)} \omega^2_{V/F} \quad (13)$$

where $\log(\mu) = (\log(\mu_{k_a}), \log(\mu_{CL/F}), \log(\mu_{V/F}))'$. After computing the derivatives, $\omega^2_{C_{max}}$ can be approximated by:

$$\begin{aligned} &\omega^2_{C_{max}} \approx \Delta^2\ (\omega^2_{k_a}\ +\ \omega^2_{CL/F})\ +\ (\Delta-1)^2\ \omega^2_{V/F} \\ &\text{with } \Delta = \frac{\mu_{CL/F}\ (\mu_{CL/F} - \mu_{k_a}\,\mu_{V/F}) + \mu_{k_a}\,\mu_{CL/F}\,\mu_{V/F}\ \log\left(\dfrac{\mu_{k_a}\,\mu_{V/F}}{\mu_{CL/F}}\right)}{\left(\mu_{k_a}\,\mu_{V/F} - \mu_{CL/F}\right)^2} \end{aligned} \quad (14)$$

In this simulation study, the general formula above is applied to approximate the variance of $log(C_{max})$ for both treatment groups (Ref and $Test$). Given the treatment effect we simulate for the treatment $Test$, both approximations, $\omega^2_{C_{max}^{(Ref)}}$ and $\omega^2_{C_{max}^{(Test)}}$, are equal.

To approximate the variance of $log(C_{max})$ by the delta method, we use the true simulated values of $\mu^{(Ref)}$ and $\Omega^{(Ref)}$ described in section 2.2.3. To evaluate the delta method, we also estimate the variance of $log(C_{max})$, using the simulated parameter values of the rich design ($N = 40$, $n = 10$) for the reference treatment, under $S_{l,l}$ and $S_{h,l}$. For both variability settings, $\omega^2_{C_{max}^{(Ref)}}$ is estimated as the empirical variance of the 40000 true simulated values of $\log({C_{max}}_i^{(Ref)})$. For

$S_{l,l}$, the standard deviation of $log(C_{max})$ for the reference treatment expressed in percent is 10.5% both by simulation and the delta method. For $S_{h,l}$, it is 46.3% and 46.7% by simulation and the delta method, respectively.

These results on the true simulated values validate the approximation of the variance of $\log(C_{max})$ by the delta method. Consequently, we apply it to the data of each treatment group for each simulated trial of the simulation study to approximate $\widehat{\omega}^2_{C_{max}^{(Ref)}}$ ($\widehat{\omega}^2_{C_{max}^{(Test)}}$ respectively) using $\widehat{\mu}^{(Ref)}$ ($\widehat{\mu}^{(Test)}$ respectively) and $\widehat{\Omega}^{(Ref)}$ ($\widehat{\Omega}^{(Test)}$ respectively).

Table I: Summary of the three variability settings used in the simulation study. The between-subject (BSV) and within-subject (WSV) variability are given as standard deviations of the log-parameters multiply by 100 and expressed in percent.

Variability	$S_{l,l}$	$S_{h,l}$	$S_{h,h}$
BSV	20% for k_a and CL/F 10% for V/F	50%	50%
WSV	10% for k_a and CL/F 5% for V/F	15%	15%
Residual error	$a = 0.1\ mg/L$ $b = 10\%$	$a = 0.1\ mg/L$ $b = 10\%$	$a = 1\ mg/L$ $b = 25\%$

Table II: Type I error of the bioequivalence tests performed on the treatment effect of $log(AUC)$ and $log(C_{max})$ for each unilateral hypothesis, $H_{0;80\%}$ and $H_{0;125\%}$. The type I error is estimated from 1000 bioequivalence trials simulated under $H_{0;80\%}$ or $H_{0;125\%}$ for different designs (N: number of subjects, n: number of samples per subject), different variability settings $S_{l,l}$, $S_{h,l}$ and $S_{h,h}$, for the true simulated values (SIM) and both types of estimates (NCA and EBE). Due to the 1000 replicates, the 95% PI for a type I error of 5% is $[3.7\%; 6.4\%]$.

			$N = 40,\ n = 10$			$N = 12,\ n = 10$			$N = 24,\ n = 5$			$N = 40,\ n = 3$		
			SIM	NCA	EBE	SIM	NCA	EBE	SIM	NCA	EBE	SIM	NCA	EBE
$S_{l,l}$	AUC	$H_{0;80\%}$	3.9	4.0	5.5	5.4	5.2	7.7	4.3	4.3	8.0	3.9	5.9	14.8
		$H_{0;125\%}$	4.6	5.1	5.8	5.4	5.2	7.4	4.4	3.8	7.5	4.6	5.1	16.2
	C_{max}	$H_{0;80\%}$	4.5	6.6	10.0	5.7	5.1	9.0	5.8	5.3	14.6	4.5	6.8	30.6
		$H_{0;125\%}$	4.9	6.3	9.1	5.2	5.6	10.9	5.3	5.2	16.2	4.9	5.5	29.1
$S_{h,l}$	AUC	$H_{0;80\%}$	3.9	5.4	4.7	5.4	4.4	6.8	4.3	5.2	7.1	3.9	4.5	8.5
		$H_{0;125\%}$	4.6	6.1	5.2	5.4	4.7	6.1	4.4	3.9	5.8	4.6	5.1	11.5
	C_{max}	$H_{0;80\%}$	4.5	5.1	4.0	5.3	5.3	5.3	5.5	6.0	6.5	4.5	7.2	9.2
		$H_{0;125\%}$	5.0	5.4	5.0	5.2	5.1	5.8	5.7	6.1	7.1	5.0	6.2	7.8
$S_{h,h}$	AUC	$H_{0;80\%}$							4.3	0.8	20.6			
		$H_{0;125\%}$							4.4	0.4	22.2			
	C_{max}	$H_{0;80\%}$							5.5	7.0	13.8			
		$H_{0;125\%}$							5.7	9.3	17.0			

Legend to figures

Figure 1. Concentrations (mg/L) simulated for the intermediate design ($N = 24,\ n = 5$) for the reference treatment (left) and for the test treatment under $H_{0;80\%}$ (middle) and $H_{0;125\%}$ (right) using the variability settings $S_{l,l}$ (top), $S_{h,l}$ (middle) and $S_{h,h}$ (bottom).

Figure 2. Concentrations (mg/L) simulated for the original ($N = 12,\ n = 10$, left), intermediate ($N = 24,\ n = 5$, middle) and sparse ($N = 40,\ n = 3$, right) designs for the reference treatment using the variability settings $S_{l,l}$ (top) and $S_{h,l}$ (bottom).

Figure 3. Bias (top) and root mean square error (RMSE, bottom) of estimates of the sample mean for $log(AUC)$ (left) and $log(C_{max})$ (right) for the reference treatment from 1000 trials for different designs (N: number of subjects, n: number of samples per subject) and different variability settings $S_{l,l}$ ($\circ$), $S_{h,l}$ ($\square$) and $S_{h,h}$ ($\triangle$). The white symbols represent the individual estimates obtained from NCA and the grey ones the individual estimates obtained from EBE.

Figure 4. Global type I error of the bioequivalence tests performed on the treatment effect of $log(AUC)$ (top) and $log(C_{max})$ (bottom). The global type I error is estimated from 1000 bioequivalence trials simulated under $H_{0;80\%}$ and $H_{0;125\%}$ for different designs (N: number of subjects, n: number of samples per subject) and different variability settings $S_{l,l}$ ($\circ$), $S_{h,l}$ ($\square$) and $S_{h,h}$ ($\triangle$). The white symbols represent the individual estimates obtained from NCA and the grey ones the individual estimates obtained from EBE. The dashed lines represent the nominal level at 5% and its 95% prediction interval ($[3.7\%; 6.4\%]$).

Figure 5. Global Type I error of the bioequivalence tests performed on the treatment effect of $log(AUC)$ (top) and $log(C_{max})$ (bottom) versus the median shrinkage on the parameter of interest for the reference treatment and different simulation settings $S_{l,l}$ ($\circ$), $S_{h,l}$ ($\square$) and $S_{h,h}$

(△). The rich design design ($N = 40, n = 10$) is represented by white symbols, the original design ($N = 12, n = 10$) by light grey symbols, the intermediate design ($N = 24, n = 5$) by dark grey symbols and the sparse design ($N = 40, n = 3$) by black symbols. The dashed lines represent the nominal level at 5% and its 95% prediction interval ([3.7%; 6.4%]).

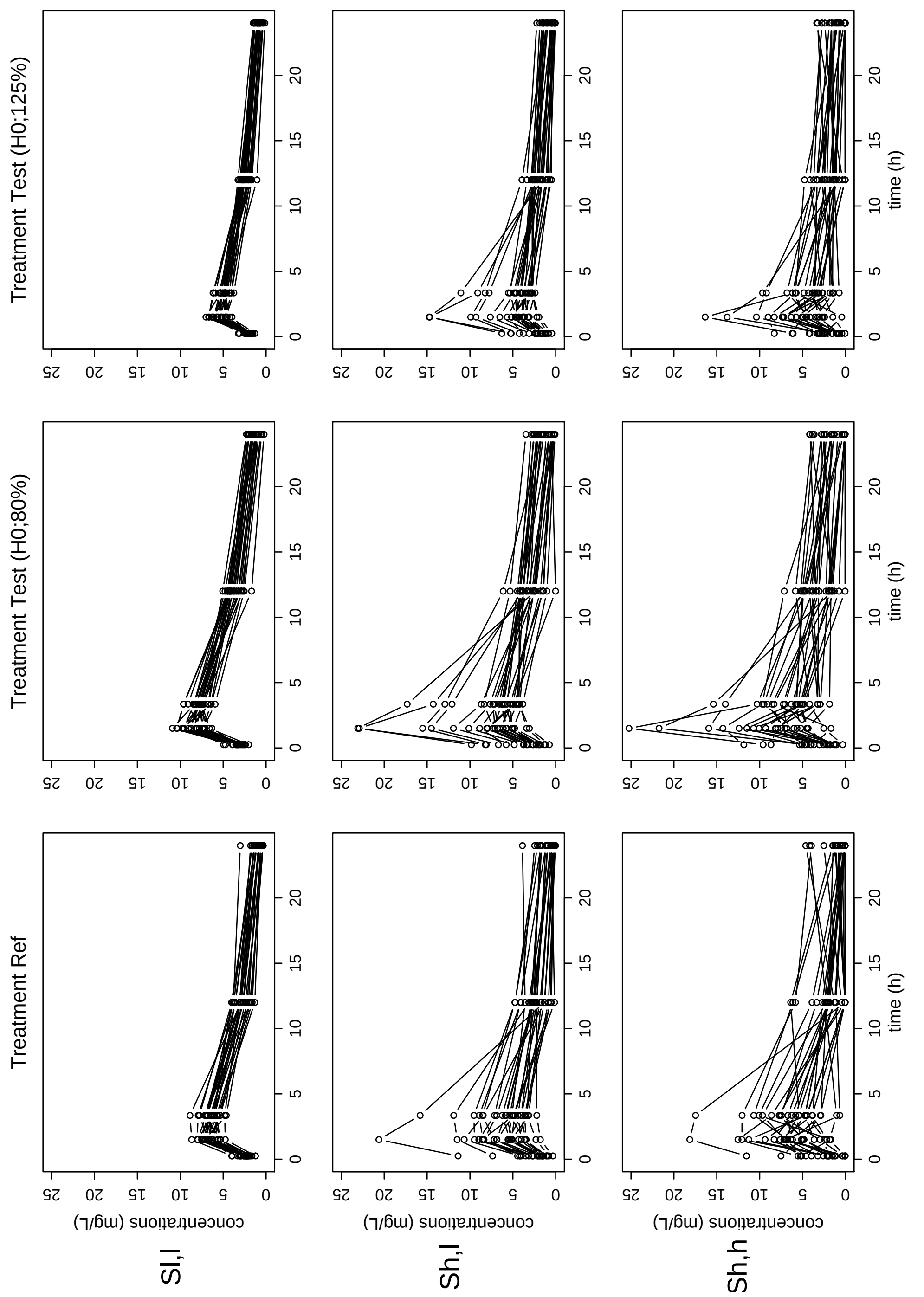
Treatment Ref
Treatment Test (H0;80%)
Treatment Test (H0;125%)
Sl,l
Sh,l
Sh,h
concentrations (mg/L)
0
5
10
15
20
25
time (h)
0
5
10
15
20

Figure 1

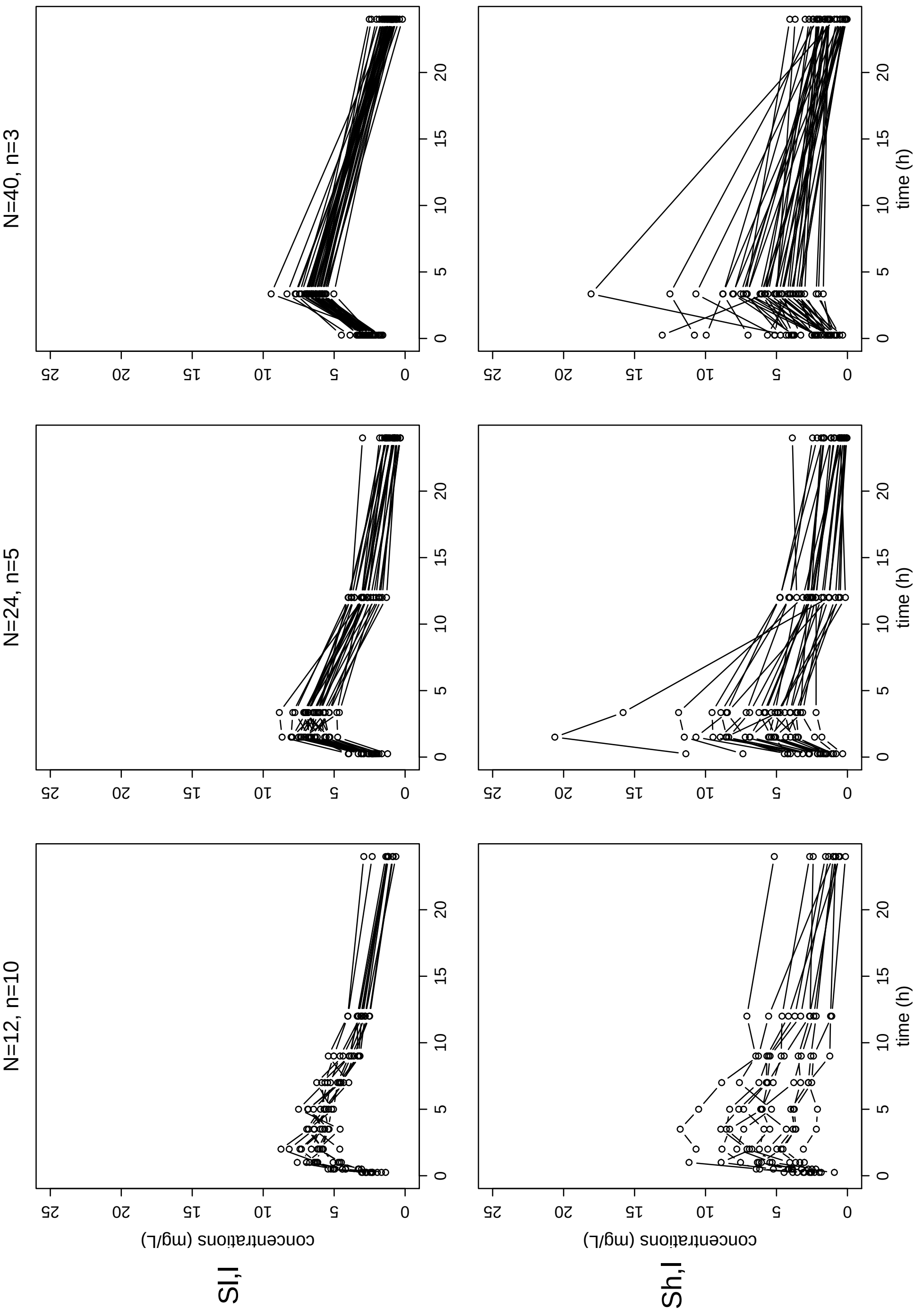
N=12, n=10
N=24, n=5
N=40, n=3
Sl,I
Sh,I
concentrations (mg/L)
concentrations (mg/L)
time (h)
time (h)
time (h)
0
5
10
15
20
25
0
5
10
15
20

Figure 2

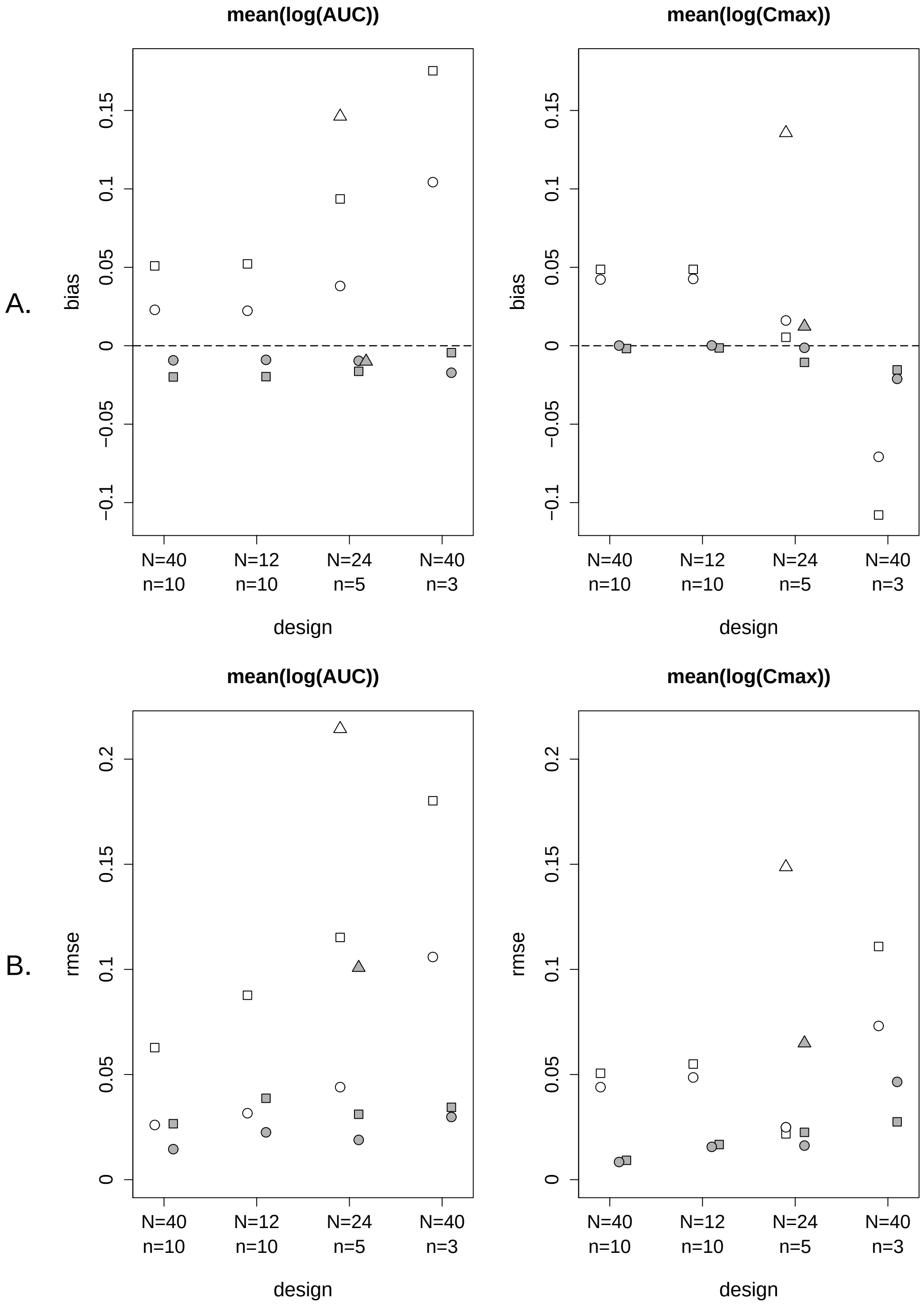
A.
mean(log(AUC))
mean(log(Cmax))
bias
0.15
0.1
0.05
0
−0.05
−0.1
N=40 n=10
N=12 n=10
N=24 n=5
N=40 n=3
design
B.
rmse
0.2
0.15
0.1
0.05
0

Figure 3

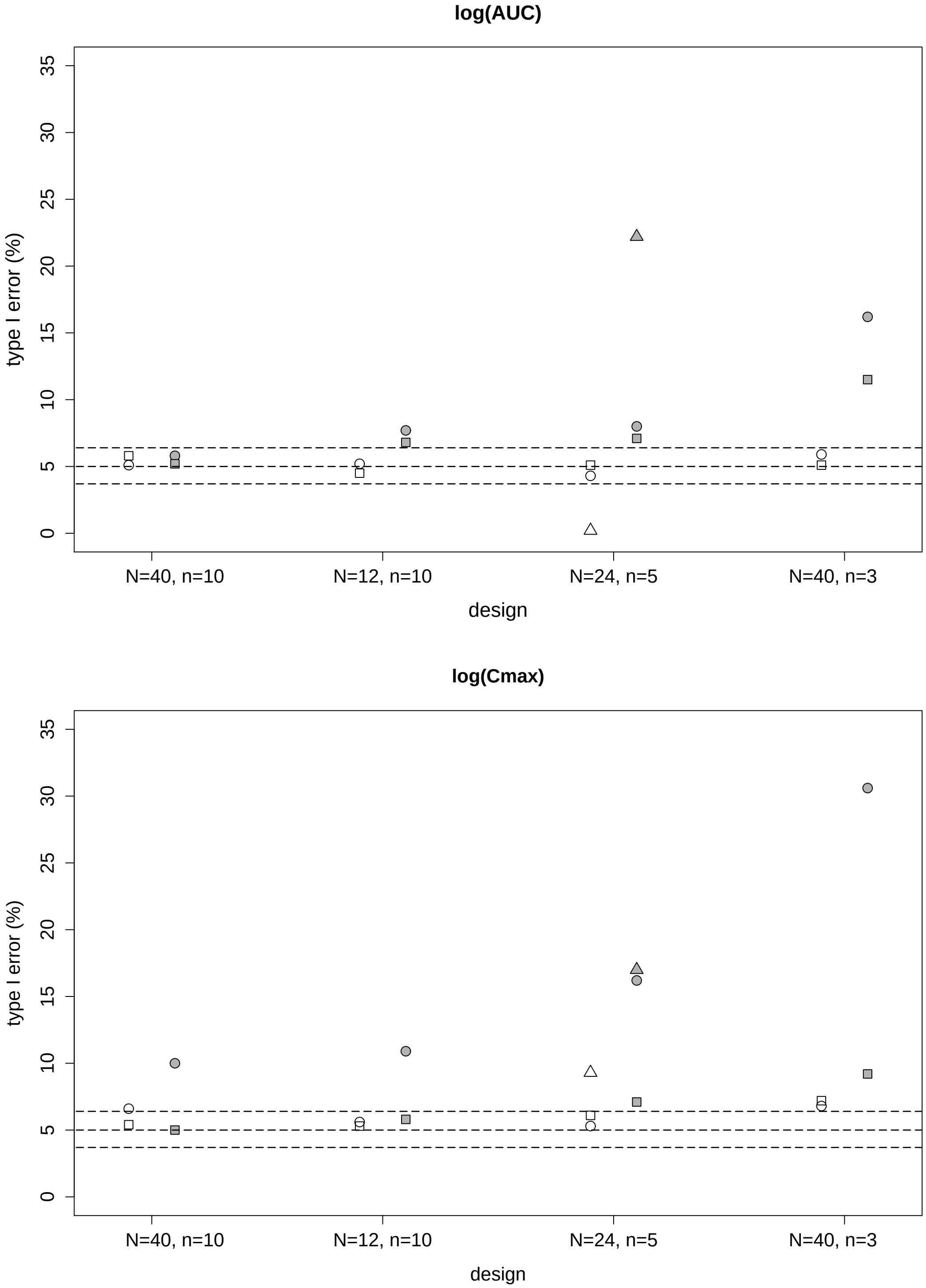
log(AUC)
type I error (%)
0
5
10
15
20
25
30
35
N=40, n=10
N=12, n=10
N=24, n=5
N=40, n=3
design
log(Cmax)
type I error (%)
0
5
10
15
20
25
30
35
N=40, n=10
N=12, n=10
N=24, n=5
N=40, n=3
design

Figure 4

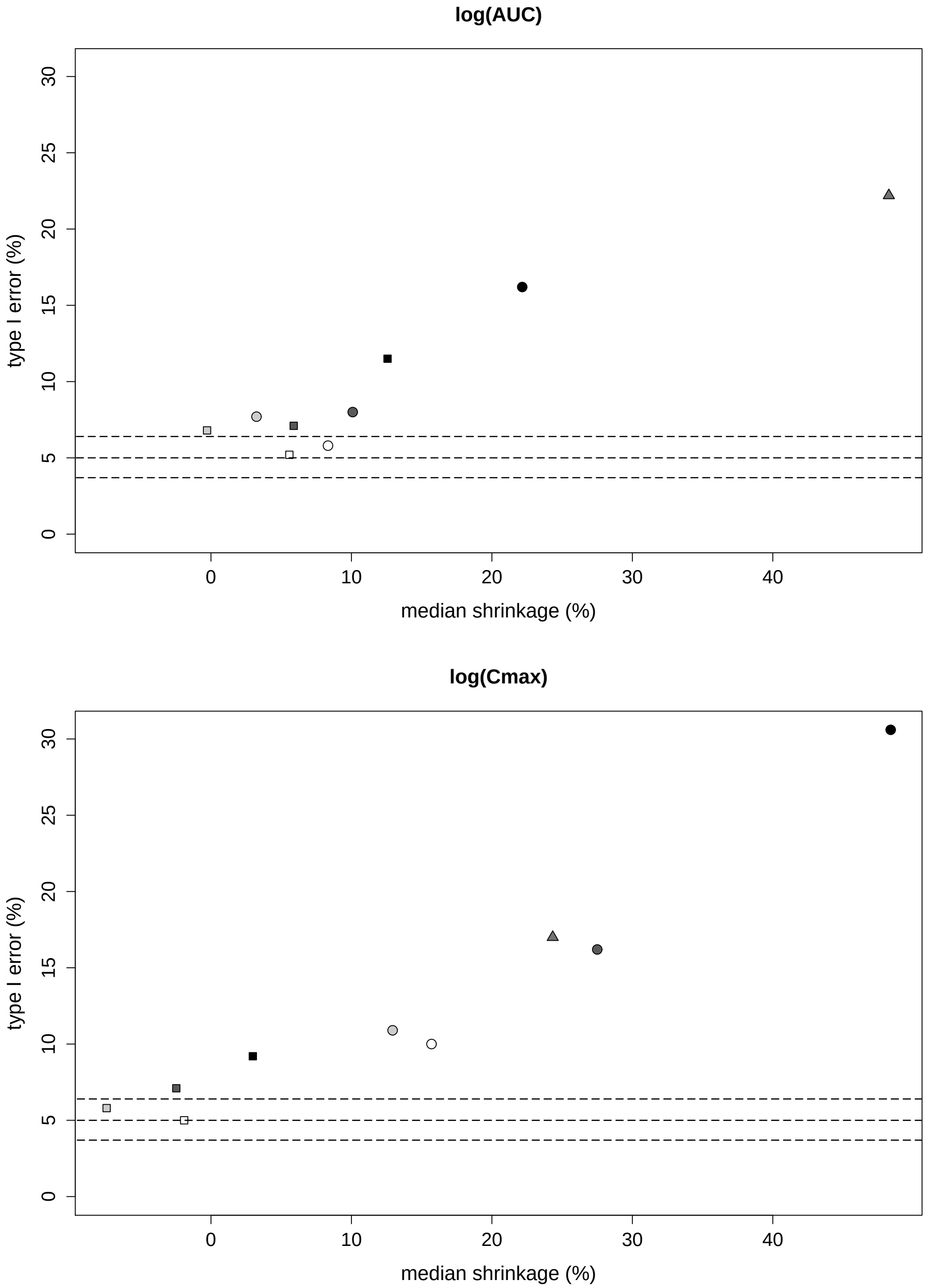
log(AUC)
type I error (%)
0
5
10
15
20
25
30
0
10
20
30
40
median shrinkage (%)
log(Cmax)
type I error (%)
0
5
10
15
20
25
30
0
10
20
30
40
median shrinkage (%)

Figure 5