New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling

Florence Loingeville[1,2], Julie Bertrand[1], Thu Thuy Nguyen[1], Satish Sharan[3], Kairui Feng[3], Wanjie Sun[4], Jing Han[4], Stella Grosser[4], Liang Zhao[3], Lanyan Fang[3], Kathrin Möllenhoff[5,6], Holger Dette[5], France Mentré[1]

[1] University of Paris, IAME INSERM, UMR 1137, 75018 Paris, France

[2] Univ. Lille, CHU Lille, ULR 2694 - METRICS : Evaluation of Health Technologies and Medical Practices, F-59000 Lille, France

[3] Division of Quantitative Methods and Modeling, Office of Research Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring MD 20993, USA

[4] Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring MD 20993, USA

[5] Department of Mathematics, Ruhr-Universitat Bochum, Germany

[6] Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne, Cologne, Germany

[2] Corresponding author: florence.loingeville@univ-lille.fr; Laboratoire de biomathématiques, Faculté de Pharmacie, 3 Rue du Professeur Laguesse, 59 000 Lille, France

Abstract

Introduction: In traditional pharmacokinetic (PK) bioequivalence analysis, two one-sided tests (TOST) are conducted on the area under the concentration-time curve and the maximal concentration derived using a non-compartmental approach. When rich sampling is unfeasible, a model-based (MB) approach, using nonlinear mixed effect models (NLMEM) is possible. However, MB-TOST using asymptotic standard errors (SE) presents increased type I error when asymptotic conditions do not hold.
Methods : In this work, we propose three alternative calculations of the SE based on i) an adaptation to NLMEM of the correction proposed by Gallant, ii) the *a posteriori* distribution of the treatment coefficient using the Hamiltonian Monte Carlo algorithm, and iii) parametric random effects and residual errors bootstrap. We evaluate these approaches by simulations, for two-arms parallel and two-periods two-sequences cross-over design with rich (n=10) and sparse (n=3) sampling under the null and the alternative hypotheses, with MB-TOST.
Results: All new approaches correct for the inflation of MB-TOST type I error in PK studies with sparse designs. The approach based on the *a posteriori* distribution appears to be the best compromise between controlled type I errors and computing times.
Conclusion: MB-TOST using non-asymptotic SE controls type I error rate better than when using asymptotic SE estimates for bioequivalence on PK studies with sparse sampling.

Keywords and Phrases: pharmacokinetics, bioequivalence, nonlinear mixed effects model, two one-sided tests, non-asymptotic standard error

1 Introduction

Bioequivalence studies are routinely conducted for the development of generics or the adoption of new formulations of existing drug. According to current guidelines by regulation authorities both in the US and the EU [1, 2], bioequivalence between a reference (R) and a test (T) product is to be assessed based on the comparison of their respective area under the time-concentration curves (AUC) and maximal concentrations (C_{max}). The presently recommended statistical approach is to claim bioequivalence if the boundaries of the 90%-confidence intervals around the ratios of AUC and C_{max} geometric means of both groups do fall between 0.8 and 1.25. This is equivalent to performing a two one-sided tests (TOST) proposed by Schuirmann [3].

Traditionally, individual estimates of AUC are obtained using non-compartmental analysis (NCA-TOST). Based on few hypotheses, NCA requires dense pharmacokinetic (PK) sampling. In especially fragile populations (e.g., children or patients), or for specific indications (e.g., ophthalmic drugs), it may be challenging and/or unethical to collect such dense sampling. Therefore to assess the PK bioequivalence of two ophthalmic drugs on a study with only one-time point per subject, Shen et al. proposed a non parametric bootstrap NCA-based TOST[4]. A population PK model-based (MB) approach is another appealing alternative when dense PK samplings cannot be obtained, as it lowers the individual sampling burden by borrowing information across patients. .

In 2010, Dubois et al. compared the type I error and power of the NCA-TOST to a TOST based on individual empirical Bayes estimates (EBE) from a nonlinear mixed effect model (NLMEM)[5]. They found that, when the shrinkage is above 20%, using NCA TOST leads to a modestly increased type I error whereas using TOST on EBE leads to a more severe type I error inflation. They suggested to perform a TOST directly on the treatment effect estimate from the NLMEM (MB-TOST) using the asymptotic standard error (SE). In 2011, they evaluated the MB-TOST using Wald test and likelihood ratio test and found an inflation of the type I error when asymptotic conditions are not met, which is the underlying condition for applying Delta method, that is, for very sparse sample (number of samples per subject is limited), or small sample size (number of subjects is small), or high variability. Further, they associated this inflation to an under-estimation of the SE of the treatment effect coefficient, due to the use of an asymptotic approximation, i.e., the observed Fisher Information matrix (FIM). So the MB-TOST in its current form does not meet the standards of regulatory agencies for confirmatory tests.

Therefore, the primary objective of this work is to propose alternative approaches to calculate the SE, guarantying for the MB-TOST a nominal type I error on sparse sampling PK studies. First, we adapted the correction based on the work by Gallant [6] which Bertrand et al. extended to Wald tests in NLMEM, in case of small sample size studies [7]. Second, we proposed to sample in the *a posteriori* distribution of population parameters obtained by Hamiltonian Monte-Carlo using Stan [8], as proposed by Ueckert et al. [9]. Third, we used parametric bootstrap, which was shown to perform better than case bootstrap and non parametric residual bootstrap when the true model and variance distribution are used [10].

We evaluated MB-TOST using these approaches by clinical trial simulations with parallel and cross-over designs, with rich and sparse samplings.

Although the TOST is very efficient in most cases, it has proven to be too conservative on drugs with high variability [11]. Therefore, Möllenhoff et al. proposed a bioequivalence optimal test (BOT) as an alternative to the TOST for bioequivalence assessment in such situations [12]. They adapted this test to the MB approach (MB-BOT), and showed that this method appears to have closer type I errors to the conventionally accepted significance level of 0.05 than the MB-TOST for drugs with high variability. However, they also noticed an inflation of the type I errors on sparse designs, showing that the SE-computation method is also an issue with MB-BOT. In supplementary material 2, we evaluate MB-BOT along with the proposed SE-calculation approaches. Then, we further study the conjoint influence of design and variability on the SE of the treatment effect on AUC and C_{max}, and thus on type I error and power of MB tests. Thereafter, we determine a threshold on the SE above which MB-BOT should be recommended over MB-TOST.

In Section 2, we introduce the NLMEM, the MB-TOST as well as the different SE calculations. In Section 3, we present the clinical trial simulations performed to evaluate the approaches. In Section 4, we present the results, i.e., type I error and power of the different approaches and finally in Section 5, we discuss the conclusions and perspectives of this work.

2 Methods

2.1 Nonlinear mixed effects models

The concentration $y_{i,j,k}$ of subject i ($i = 1, \ldots, N$), at period k ($k = 1, 2$), at sampling time $t_{i,j,k}$ ($j = 1, \ldots, n_i$) is described by a nonlinear function f depending on the vector of individual parameters $\phi_{i,k}$ of subject i at period k

$$y_{i,j,k} = f(t_{i,j,k}, \phi_{i,k}) + g(t_{i,j,k}, \phi_{i,k})\varepsilon_{i,j,k}. \tag{1}$$

The l^{th} individual parameter $\phi_{i,k,l}$ ($l = (1, \ldots, p)$) is defined by the following equation, where p is the number of PK parameters

$$\log(\phi_{i,k,l}) = \log(\lambda_l) + \beta_l^{Tr} Tr_{i,k} + \beta_l^{P} P_k + \beta_l^{S} S_i + \eta_{i,l} + \kappa_{i,k,l}, \tag{2}$$

with λ_l the l^{th} element of the vector of fixed effects for the covariate reference class. $Tr_{i,k}$, P_k, S_i are known vectors of, respectively, the treatment, the period, and the sequence covariates. β_l^{Tr}, β_l^P, and β_l^S are the l^{th} elements of the vectors of coefficients of the treatment, the period, and the sequence effects for the individual parameter.
$\eta_{i,l}$ is the l^{th} element of the vector η_i of random effects of subject i capturing the between-subject variability (BSV). $\kappa_{i,k,l}$ is the l^{th} element of the vector $\kappa_{i,k}$ of random effects of subject i at period k capturing the within-subject variability (WSV). η_i and $\kappa_{i,k}$ are assumed independent and normally distributed with zero mean and covariance matrix, respectively Ω and Γ, both of size $p \times p$. We define ω_l^2 the between-subject variance of the l^{th} parameter, and γ_l^2 the within-subject variance of the l^{th} parameter.
The residual errors $\varepsilon_{i,j,k}$ are supposed independent and identically distributed according to a normal centered distribution with variance 1. The error model can be additive $g(t_{i,j,k}, \phi_{i,k}) = a$, proportionnal $g(t_{i,j,k}, \phi_{i,k}) = b \times f(t_{i,j,k}, \phi_{i,k})$, i.e., additive on log-concentrations, or combined $g(t_{i,j,k}, \phi_{i,k}) = a + b \times f(t_{i,j,k}, \phi_{i,k})$.

We denote by $\theta = (\lambda, \beta^{Tr}, \beta^P, \beta^S, \Omega, \Gamma, a, b)$ the vector of all parameters of the model, and by $\widehat{VAR}(\hat{\theta})$ the estimation variance-covariance matrix, derived as the inverse of the observed FIM.

Here, bioequivalence is assessed on β_{SP}^{Tr} the coefficient of the treatment on the secondary PK parameters of interest SP=$\{AUC$ or $C_{max}\}$. For each secondary parameter, β_{SP}^{Tr} is a function of λ and β^{Tr} and its SE is derived from $\widehat{VAR}(\hat{\theta})$.

2.2 Model-based TOST

The MB-TOST global null hypothesis is expressed as $H_0 : \beta_{SP}^{Tr} \leq -\delta$ or $\beta_{SP}^{Tr} \geq \delta$ and can be divided in two sub-hypotheses: $H_{0,-\delta} : \beta_{SP}^{Tr} \leq -\delta$ and $H_{0,\delta} : \beta_{SP}^{Tr} \geq \delta$.
Therefore, the MB-TOST consists in two Wald statistics: $W_{-\delta} = (\hat{\beta}_{SP}^{Tr} + \delta)/SE(\beta_{SP}^{Tr})$ and $W_\delta = (\hat{\beta}_{SP}^{Tr} - \delta)/SE(\beta_{SP}^{Tr})$, respectively testing $H_{0,-\delta}$ and $H_{0,\delta}$, with $SE(\beta_{SP}^{Tr})$ the standard error on the estimation of a secondary parameter $SP = AUC, C_{max}$. In an asymptotic setting, $W_{-\delta}$ and W_δ can be assumed to follow a Gaussian distribution under $H_{0,-\delta}$ and $H_{0,\delta}$, respectively.
So, the global null hypothesis H_0 is rejected with type I error α 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.
Alternatively, one can compute the $(1-2\alpha)$ confidence interval (CI) of β_{SP}^{Tr} and reject the global null hypothesis if it is included in the interval $[-\delta; \delta]$.

For MB-TOST, the asymptotic approach (Asympt.) consists in using $\widehat{VAR}(\hat{\theta})$, $\hat{\lambda}$, and $\hat{\beta^{Tr}}$ for deriving the SE of the secondary parameters (SE(β_{SP}^{Tr})) with the delta method [13]. The analytical formulas are shown in detail in Appendix A of Dubois et al. [14].

2.3 New approaches for standard error (SE) calculations

Gallant. This method consists in multiplying the asymptotic SE by a factor equal to $\sqrt{\frac{n_P \times N}{df_G}}$ where n_P is the number of periods, N is the number of subjects, and $df_G = n_P \times N - p$. For MB-TOST, the reference distribution is the Student distribution.

Sampling in the *a posteriori* distribution (Post). This method consists in sampling in the *a posteriori* distribution of β^{Tr}. The *a posteriori* distribution is obtained using Hamiltonian Monte-Carlo (HMC). We assigned default priors to the fixed effects $(\lambda, \beta^{Tr}, \beta^P, \beta^S)$ and non-informative half Cauchy priors to variance terms. The HMC chain was initialized at $\hat{\theta}$, $\hat{\eta}_i$, and $\hat{\kappa}_{i,k}$ from the NLMEM analysis using the SAEM algorithm. For each resulting sample of λ and β^{Tr}, we derive a corresponding β^{Tr}_{SP} and the standard deviation of this series is the *Post* $SE(\beta^{Tr}_{SP})$.

Parametric random effect and residual bootstrap (Boot). This method consists in simulating $b = 1, ..., B$ datasets with the original bioequivalence study design. The subject random effects and residuals are issued from distributions with means and variances equal to the estimated population parameters from the original bioequivalence study NLMEM analysis. Then, the B datasets are fitted with a NLMEM and B replicates of $\hat{\beta}^{Tr}_{SP_b}$ are calculated as functions of the $\hat{\lambda}_b$ and $\hat{\beta}^{Tr}_b$ estimates. The standard deviation of this series is the *Boot* $SE(\beta^{Tr}_{SP})$.

3 Simulation Study

3.1 Pharmacokinetic model

We used the PK model from Dubois et al. [14], which describes concentrations of the anti-asthmatic drug theophylline, for both reference and test group, with a one-compartment distribution (apparent volume, V/F) and first-order absorption (absorption rate, Ka) and elimination (apparent clearance, CL/F). We fixed the dose to $D = 4$ mg for all subjects.
For the reference treatment, we considered λ_{Ka}=1.50 /h, $\lambda_{CL/F}$=40.00 mL/h, and $\lambda_{V/F}$=0.50 L. We considered a combined error model with a=0.1 mg/L and b=10%, corresponding to a low residual variability.
The bioequivalence threshold δ was set at $log(1.25) \approx 0.22$ as recommended by the guidelines [15].

3.2 Treatment effect

We simulated under one null hypothesis $H_{0,\log(0.8)}$, i.e., $\log(AUC^T/AUC^R) = \log(C^T_{max}/C^R_{max}) = \log(0.8)$, where AUC^T/AUC^R and C^T_{max}/C^R_{max} are the ratios of geometric means of T to R formulations of AUC and C_{max} respectively [16]. The corresponding treatment effect coefficients modifying both CL/F and V/F are $\beta^{Tr}_V = \beta^{Tr}_{CL} = \log(1.25)$.
We also simulated under one alternative hypothesis by setting $\beta^{Tr}_{CL} = \beta^{Tr}_V = \log(1)$, which corresponds to $\beta^{Tr}_{AUC} = \beta^{Tr}_{C_{max}} = \log(1) = 0$.

3.3 Study Design

We simulated two-arms parallel and 2-periods 2-sequences cross-over designs (as in [14]). For both trials, we simulated a rich and a sparse design, both with N=40 subjects. For the rich design, there were n=10 samples per subject, taken at 0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, 24 hours after dosing. For the sparse design, we simulated n=3 samples per subject, taken at 0.25, 3.35 and 24 hours after dosing. We considered the same sparse design as in [14, 17]. The sampling times for this design were chosen by maximization of the determinant of the Fisher information matrix for an individual nonlinear model using the fixed effect values. This was done using the PFIM software [18] with a sampling window from 15 min to 24 hours.

We first simulated a parallel design (Figure 1), where $N/2$ subjects receive the reference treatment (R) whereas the other $N/2$ subjects are allocated to the test treatment (T). Such a design is often chosen to assess the bioequivalence of drugs with long half-lives preventing the use of each patient as his own control within the time constraints of drug development. We simulated BSV random effects with $\omega_{Ka} = \omega_{CL/F} = 22\%$ and $\omega_{V/F} = 11\%$, i.e., rather low BSV. For parallel trials, the period effects β^P_l, the sequence effects β^S_l, and the WSV $\kappa_{i,k,l}$ in the expression of the log of individual parameters (2) are null.

We also simulated a two-periods, two-sequences cross-over design (Figure 2), which is the gold-standard in bioequivalence trials. In these trials, the $N/2$ subjects of the first sequence (S_1) receive the reference (R) treatment at period 1 (P_1), and the test (T) treatment at period 2 (P_2), whereas the $N/2$ subjects of the second sequence (S_2) receive treatments in the reverse order. We simulated BSV and WSV random effects with $\omega_{Ka} = \omega_{CL/F} = \omega_{V/F} = 50\%$, and $\gamma_{Ka} = \gamma_{CL/F} = \gamma_{V/F} = 15\%$, i.e., rather high BSV and rather low WSV.

3.4 Implementation and evaluation

We evaluated the type I error and power of the MB-TOST on β^{Tr}_{AUC} and β^{Tr}_{Cmax} using the different approaches, at the nominal level $\alpha = 5\%$ on the different scenarios.

Five hundred data sets were simulated per scenario, using the R software.
The parameters of the NLMEM were estimated using the SAEM algorithm. For the parallel design, we used the R package saemix version 1.2 [19], and for the crossover design, we used the monolix software version 2018R2 [20].
We used the same parameterisation with both softwares; 10 Monte Carlo Markov chains, 300 iterations in the exploratory phase, and 100 iterations in the smoothing phase.

For the Asympt approach, the FIM was obtained by linearisation.
For the Post approach, we used the Rstan package with 1000 iterations and 100 burn-in, so that we obtained 900 samples.
For the Boot approach, we simulated B=250 data sets.

All calculations were run on an i7-5600 U CPU computer, with frequency 2.60 GHz, 4 cores, 8 GB of RAM.

4 Results

4.1 Two-arms parallel

For both the sparse and rich designs, the Relative Biais (RBiais) and the Relative Root Mean Square Errors (RRMSE) were below 10% for the fixed effects and close to 0 for the treatment effect coefficients (Table I in supplementary material 1). For the BSV and residual error standard deviations, there was a upward trend in the RBiais, below 10% for the rich design and up to 30% for the sparse design. The RRMSE also increased for the sparse design, up to 121% for the additive residual error standard deviation. Yet, there was no major concern with the estimation of the NLMEM parameters.

On the rich design, MB-TOST conserved a nominal type I error with all different SE calculations (Figure 3). On the sparse design with N=40, MB-TOST for β_{AUC}^{Tr} using the asymptotic SE led to an inflated type I error (Figure 3). However, using the alternative calculations of the SE, the inflation was corrected.
In supplementary material 2, we evaluate the proposed approaches to compute the SE along with both MB-TOST and MB-BOT on a sparser design (with N=12 subjects). Then, we further explore the relationship between $SE(\beta_{SP}^{Tr})$ and the MB-TOST type I error and derive a critical threshold when MB-TOST no longer controls its nominal level and MB-BOT should be used instead.

The simulated power of MB-TOST using the different SE calculations were of similar order and of reasonable size $> 70\%$ (Table I). We observe higher power to conclude to bioequivalence on C_{max} compared to AUC as we simulated $\omega_{Cmax} = 10\%$ and $\omega_{AUC} = 22\%$, respectively.
With regard to computational times, Asympt and Gallant approaches took a few seconds per data set, whereas running Post took a few minutes, and Boot close to 1 hour. In fact, we did not compute the bootstrap-based SE on the sparse design with N=20 (scenario simulated under the null only), given its computational burden and the good performances of the Gallant and Post alternatives.

4.2 Two-periods, two-sequences cross-over

The Rbias and RRMSE on all parameters for the scenario with a sparse design, which we expected to be the most challenging under the null, are listed in Supplementary Material 1, Table II. For the fixed effects and treatment effect coefficients, all Rbias were below 5% and the RRMSE were below 25%. For the BSV, WSV and residual error standard deviations, the Rbias showed a downward trend (but for γ_{Ka}), and the RRMSE were below 40%. Again, there was no major concern with the estimation of the NLMEM parameters.

MB-TOST for β_{AUC}^{Tr} controlled its nominal level using all SE calculation whatever the design (Figure 4). Whereas MB-TOST for β_{Cmax}^{Tr} using the asymptotic SE obtained an inflation of the type I error at 7.6% on the rich design and 7.8% on the sparse design. This inflation was corrected using all the alternative SE calculations. Given its computational burden and the good performance of other alternative SE calculations, we only evaluated the bootstrap-based SE on the most challenging design, i.e., the sparse design where it took about 5 hours per data set.
MB-TOST for β_{AUC}^{Tr} and β_{Cmax}^{Tr} obtained extremely high power (>95%) using all SE calculations whatever the design (Table II).

5 Discussion

In this work we proposed, and evaluated by simulations, three alternative SE calculations to correct for the type I error inflation of MB-TOST in PK bioequivalence studies with sparse sampling. MB-TOST using the three alternative SE calculations provided both a controlled nominal type I error and satisfactory power, on parallel and cross-over studies with rich and sparse sampling. Here, we used B=250 iterations for bootstrap. This relative small number nonetheless enabled the bootstrap approach to provide a controlled type I error. However, its computational burden proved particularly limiting, especially given the good performances of the SE calculations based on the work of Gallant [6] and the *a posteriori* distribution [9].

The latter calculation is particularly appealing given the SE calculation based on the work of Gallant will, by design, be of limited interest for high N. For now, it requires calling the Stan software and further work is needed to embed this calculation within the monolix software or saemix R package.

Besides sparse sampling, products with high PK variability present another methodological challenge in bioequivalence studies. Already, Haidar et al. proposed a scaling approach setting a constraint on the geometric mean ratio [21] and the US Food and Drug Administration's Office of Generic Drugs developed a reference-scaled average bioequivalence approach [22]. In [12], Möllenhoff et al. proposed a new test, the MB-BOT, more appropriate for drugs with high variability, when the sample size is not large enough. In Supplementary Material 2, we showed that MB-BOT can also benefit from alternative SE calculations in PK BE studies with sparse sampling. Further, we derived a threshold for the treatment effect SE above which MB-BOT should be prefered to MB-TOST.

One limitation of the present work is the number of simulated datasets for each scenario under consideration. We choose to simulate 500 data sets because of the computational burden and because we could effectively compare the approaches in term of type 1 error. Another non-negligible limitation is the use of the simulated model to perform the MB-TOST. Indeed, we did not investigate the robustness to model misspecification, or consider a model averaging approach [23]. However we reckon that when bioequivalence studies are performed, there exists some accumulated knowledge on the drug PK model (resulting from either a bottom-up or a top-down approach), at least in the reference treatment group.

Finally, statistical methods have recently been proposed to control the nominal type I error in bioequivalence studies using adaptive designs [24]. This methods rely on an adaptation of the NCA-TOST and we believe there is a case for exploring these methods using MB-TOST, and non-asymptotic SE for PK bioequivalence studies with sparse sampling.

6 Conclusion

We recommend to use non-asymptotic SE, based on the *a posteriori* distribution of the treatment effect coefficient, to test for bioequivalence on pharmacokinetic studies with sparse sampling with MB-TOST.

Acknowledgments

This work was supported by the Food and Drug Administration (FDA) under contract 75F40119C1011. The authors thank FDA for this funding. The views in this paper do not necessarily reflect the views or policies of the FDA.

References

[1] US Department of Health and Human Services and others. FDA Guidance for Industry, Statistical Approaches to Establishing Bioequivalence. https://wwwfdagov/media/70958/download. 2001;.

[2] EMEA C. Note for Guidance on the Investigation of Bioavailability and Bioequivalence. CPMP/EWP/QWP/1401/98, London; 2001.

[3] Schuirmann DJ. 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. 1987;15(6):657–680.

[4] Shen M, Machado SG. Bioequivalence evaluation of sparse sampling pharmacokinetics data using bootstrap resampling method. Journal of Biopharmaceutical Statistics. 2017;27(2):257–264.

[5] Dubois A, Gsteiger S, Pigeolet E, Mentré F. 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. 2010;27(1):92–104.

[6] Gallant AR. Seemingly unrelated nonlinear regressions. Journal of Econometrics. 1975;3(1):35–50.

[7] Bertrand J, Comets E, Chenel M, Mentré F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. Biometrics. 2012;68(1):146–155.

[8] Stan Development Team. RStan: the R interface to Stan, Version 2.12.0; 2016. `http://mc-stan.org/`.

[9] Ueckert S, Riviere MK, Mentre F. Improved Confidence Intervals and P-Values by Sampling from the Normalized Likelihood. In: Journal of Pharmacokinetics and Pharmacodynamics. vol. 42; 2015. p. S56–S57.

[10] Thai HT, Mentré F, Holford NH, Veyrat-Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. Pharmaceutical Statistics. 2013;12(3):129–140.

[11] Tsai CA, Huang CY, Liu Jp. An approximate approach to sample size determination in bioequivalence testing with multiple pharmacokinetic responses. Statistics in Medicine. 2014;33(19):3300–3317.

[12] Möllenhoff K, Loingeville F, Bertrand J, Nguyen TT, Sharan S, Sun G, et al.. Efficient model-based Bioequivalence Testing; arXiv:2002.09316[stat.ME]. 2020.

[13] Oehlert GW. A note on the delta method. The American Statistician. 1992;46(1):27–29.

[14] Dubois A, Lavielle M, Gsteiger S, Pigeolet E, Mentré F. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. Statistics in Medicine. 2011;30(21):2582–2600.

[15] FDA. Guidance for industry: statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland. 2001;.

[16] Liu JP, Weng CS. Bias of two one-sided tests procedures in assessment of bioequivalence. Statistics in Medicine. 1995;14(8):853–861.

[17] Panhard X, Mentré F. Evaluation by simulation of tests based on non-linear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials. Statistics in medicine. 2005;24(10):1509–1524.

[18] Dumont C, Lestini G, Le Nagard H, Mentré F, Comets E, Nguyen TT, et al. PFIM 4.0, an extended R program for design evaluation and optimization in nonlinear mixed-effect models. Computer methods and programs in biomedicine. 2018;156:217–229.

[19] Comets E, Lavenu A, Lavielle M. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. Journal of Statistical Software. 2017;80:1–42.

[20] Monolix version 2018R2 . Antony, France: Lixoft SAS; 2018. `http://lixoft.com/products/monolix/`.

[21] Haidar SH, Makhlouf F, Schuirmann DJ, Hyslop T, Davit B, Conner D, et al. Evaluation of a scaling approach for the bioequivalence of highly variable drugs. The American Association of Pharmaceutical Scientists Journal. 2008;10(3):450–454.

[22] Davit BM, Chen ML, Conner DP, Haidar SH, Kim S, Lee CH, et al. Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US Food and Drug Administration. The American Association of Pharmaceutical Scientists Journal. 2012;14(4):915–924.

[23] Buatois S, Ueckert S, Frey N, Retout S, Mentré F. Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. The AAPS journal. 2018;20(3):56.

[24] Maurer W, Jones B, Chen Y. Controlling the type I error rate in two-stage sequential adaptive designs when testing for average bioequivalence. Statistics in Medicine. 2018;37(10):1587–1607.

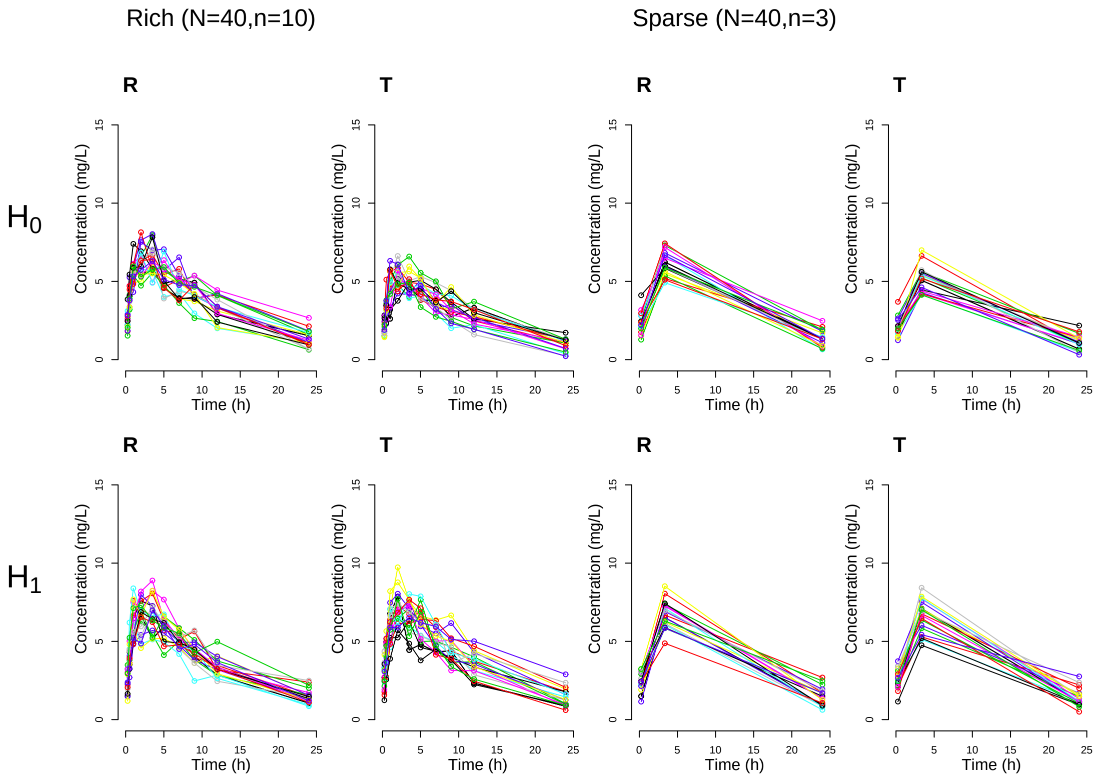


Figure 1: Spaghetti plots of simulated concentrations versus time for the two-arms parallel design under H_0 (top) and H_1 (bottom) for rich (columns 1 and 2) and sparse (columns 3 and 4) designs, in the reference (R, columns 1 and 3) and the test (T, columns 2 and 4) treatment groups.

Table I: Estimated power of MB-TOST on β^{Tr}_{AUC} and β^{Tr}_{Cmax}, using the different SE calculations, for the parallel rich and sparse designs on the 500 data sets.

	Rich (n=10)		Sparse (n=3)	
	β^{Tr}_{AUC}	β^{Tr}_{Cmax}	β^{Tr}_{AUC}	β^{Tr}_{Cmax}
Asympt	0.830	1.000	0.804	1.000
Gallant	0.782	1.000	0.762	0.998
Post	0.772	0.966	0.712	0.990
Boot	0.832	1.000	0.800	1.000

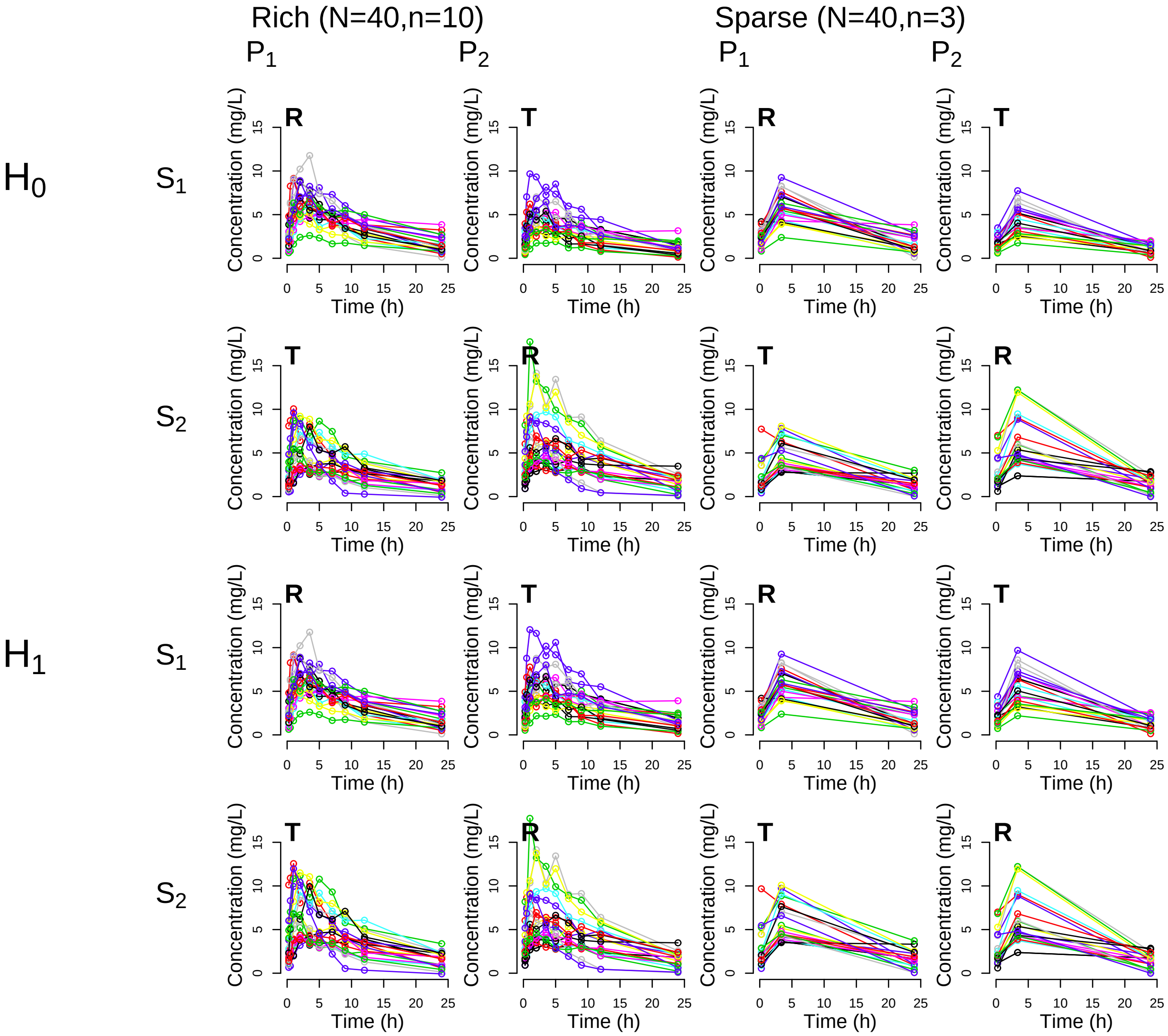


Figure 2: Spaghetti plots of simulated concentrations versus time for the 2-periods 2-sequences (S_1, S_2) cross-over design under H_0 (lines 1 and 2) and under H_1 (lines 3 and 4) for rich (columns 1 and 2) and sparse (columns 3 and 4) designs, in period 1 (P_1, columns 1 and 3) and period 2 (P_2, columns 2 and 4).

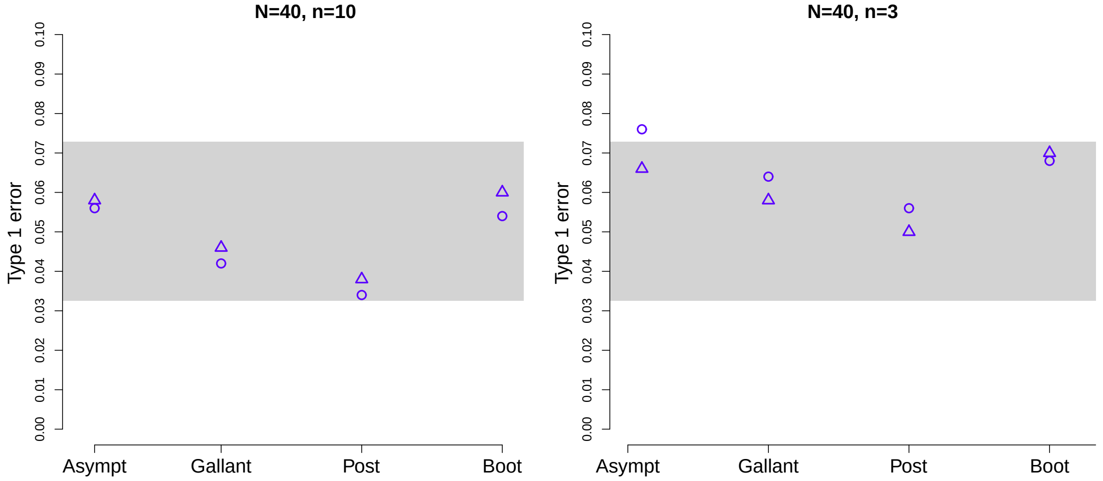


Figure 3: Type I errors of MB-TOST on β^{Tr}_{AUC} (o) and on β^{Tr}_{Cmax} ($\triangle$) using the different SE calculations on the parallel rich (left), and sparse (right) designs. The 95% prediction interval around 0.050 for 500 simulated data sets is indicated in grey ($PI_{95\%}(0.050) = [0.0326; 0.0729]$).

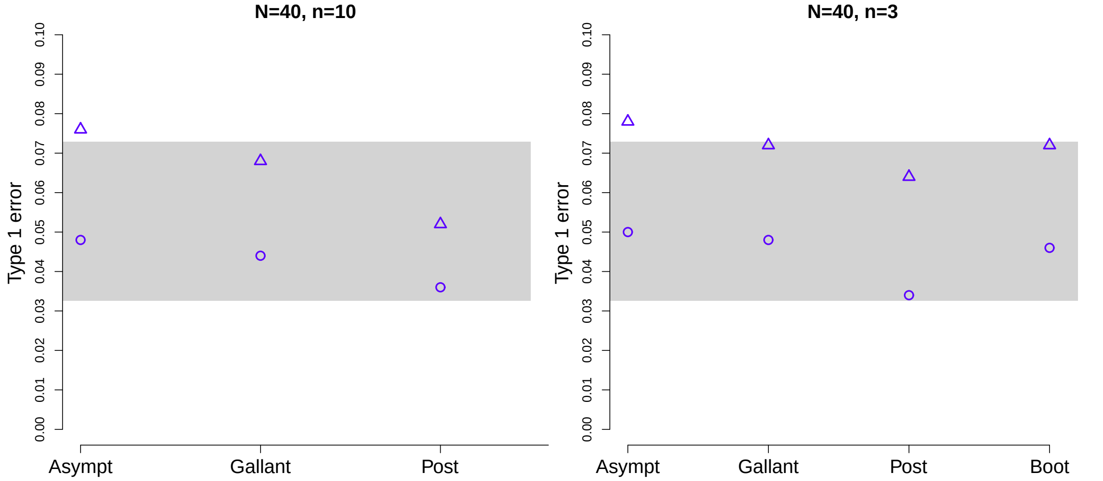


Figure 4: Type I errors of MB-TOST on β^{Tr}_{AUC} (o) and β^{Tr}_{Cmax} ($\triangle$) using the different SE calculations on the cross-over rich (left) and sparse (right) designs. The 95% prediction interval around 0.050 for 500 simulated data sets is indicated in grey ($PI_{95\%}(0.050) = [0.0326; 0.0729]$).

Table II: Estimated power of MB-TOST on β^{Tr}_{AUC} and β^{Tr}_{Cmax} using the different SE calculation, for the cross-over rich and sparse designs.

	Rich (n=10)		Sparse (n=3)	
	β^{Tr}_{AUC}	β^{Tr}_{Cmax}	β^{Tr}_{AUC}	β^{Tr}_{Cmax}
Asympt	1.000	1.000	0.998	1.000
Gallant	1.000	1.000	0.998	1.000
Post	0.988	0.998	0.996	0.996

Supplementary Material 1: Relative bias and Root Mean Square Errors

Suppl 1. Table I: Relative bias x 100 (RBiais x 100) and Relative Root Mean Square Errors x 100 (RRMSE x 100) of fixed effects in the reference group, treatment effect coefficients and standard deviations for BSV and the residual error using saemix, on the 500 data sets for the parallel rich and sparse design, with N=40, under the null hypothesis.

	Rich (n=10)		Sparse (n=3)	
	Rbiais × 100	RRMSE × 100	Rbiais × 100	RRMSE × 100
λ_{ka}	0.01	6.07	0.24	7.25
λ_{V}	0.12	3.16	0.23	4.13
λ_{CL}	0.01	5.33	-0.23	5.69
$\beta_{Ka}^{Tr\,*}$	0.15	0.09	0.76	10.90
β_{V}^{Tr}	-0.41	20.50	0.57	26.40
β_{CL}^{Tr}	-1.74	32.49	-1.28	35.0
β_{AUC}^{Tr}	-1.74	-32.49	-1.28	-35.0
β_{Cmax}^{Tr}	-0.66	-17.61	-0.05	-22.23
ω_{Ka}	-7.84	33.63	-38.44	63.91
ω_{V}	-6.06	33.02	-28.90	58.97
ω_{CL}	-6.96	25.42	-27.68	44.65
a	-2.35	39.83	75.43	121.13
b	0.32	11.67	-7.07	32.22

$*$: Root Mean Square Errors ×100 (RMSE ×100) and biais ×100 for β_{Ka}^{Tr}

Suppl 1. Table II: Relative bias x 100 (RBiais x 100) and Relative Root Mean Square Errors x 100 (RRMSE x 100) of fixed effects in the reference group, treatment effect coefficients and standard deviations for BSV, WSV and the residual error using monolix 2018 R2, on the 500 data sets simulated for the cross-over rich and sparse design, with N=40, under the null hypothesis.

	Rich (n=10)		Sparse (n=3)	
	Rbiais × 100	RRMSE × 100	Rbiais × 100	RRMSE × 100
λ_{ka}	0.37	12.32	0.04	12.98
λ_V	0.84	12.43	0.58	12.67
λ_{CL}	0.33	11.84	0.39	12.11
$\beta_{Ka}^{Tr\,*}$	-0.93	5.34	-0.67	7.13
β_V^{Tr}	-2.01	17.26	-2.31	22.32
β_{CL}^{Tr}	2.32	17.76	2.59	20.50
β_{AUC}^{Tr}	2.32	-17.76	2.59	-20.50
β_{Cmax}^{Tr}	-0.99	-14.52	-1.30	-18.34
ω_{Ka}	-2.69	13.48	-3.47	16.54
ω_V	-2.38	12.18	-2.79	12.99
ω_{CL}	-4.33	12.99	-4.54	13.43
γ_{Ka}	-3.51	26.25	8.52	36.01
γ_V	-4.75	14.15	-7.36	22.87
γ_{CL}	-5.35	17.47	-9.79	29.83
a	-0.15	15.19	-5.22	33.35
b	-0.08	6.01	-3.68	27.92

$*$: Root Mean Square Errors ×100 (RMSE ×100) and biais ×100 for β_{Ka}^{Tr}

Supplementary Material 2: MB-BOT - critical threshold for use and alternative $SE(\beta)$ calculations

1 Methods

1.1 MB-TOST

Assuming that $\beta \sim \mathcal{N}(\beta, SE(\beta))$, the type I error of MB-TOST on β can be written as:

$$\mathbb{P}(z_{1-\alpha} - \frac{\delta}{SE(\beta)} < \frac{\beta}{SE(\beta)} < -z_{1-\alpha} + \frac{\delta}{SE(\beta)}), \tag{1}$$

with α the nominal level of the test, so that we obtain:

$$\begin{cases} \Phi(-z_{1-\alpha}) - \Phi(z_{1-\alpha} - \frac{2\delta}{SE(\beta)}) & \text{if } SE(\beta) < \dfrac{\delta}{z_{1-\alpha}} \\ 0 & \text{if } SE(\beta) \geq \dfrac{\delta}{z_{1-\alpha}}, \end{cases} \tag{2}$$

where Φ is the cumulative probability function of the standard Normal distribution, and $z_{1-\alpha}$ the $1-\alpha$ quantile of this distribution.

Similarly for the power of MB-TOST we obtain:

$$\begin{cases} \Phi(-z_{1-\alpha} + \frac{\delta}{SE(\beta^{Tr}_{SP})}) - \Phi(z_{1-\alpha} - \frac{\delta}{SE(\beta^{Tr}_{SP})}) & \text{if } SE(\beta) < \dfrac{\delta}{z_{1-\alpha}} \\ 0 & \text{if } SE(\beta) \geq \dfrac{\delta}{z_{1-\alpha}}. \end{cases} \tag{3}$$

(2) and (3) highlight that when $SE(\beta) \geq \dfrac{\delta}{z_{1-\alpha}}$ both the type I error and the power of MB-TOST are null.

1.2 MB-BOT

The Bioequivalence Optimal Test (BOT), proposed by Möllenhoff et al., is an alternative to the TOST for bioequivalence assessment in situations of high variability [1] i.e. where $SE(\beta)$ is likely to be greater than $\dfrac{\delta}{z_{1-\alpha}}$.

The null hypothesis H_0 of MB-BOT is rejected with type I error α if $|\beta| < \hat{u}_\alpha$, where $\hat{u}_\alpha$ is the α-quantile of $\mathcal{N}_F(\delta, VAR(\beta))$, the folded normal distribution with parameters δ, the bioequivalence threshold, $VAR(\beta)$ the estimation variance of β.

In appendix of [1], Mollenhoff et al. derived the type I error and power of the BOT.

1.3 Critical threshold for the use of MB-BOT

We set $\delta = \log(1.25)$ and $\alpha = 5\%$, so that the threshold $\delta/z_{1-\alpha} = 0.136$.

We approximated $SE(\beta_{AUC})$ for various combinations of N, ω_{CL} and γ_{CL} with $SE(\beta_{AUC}) \approx \sqrt{\frac{4}{N}}\omega_{CL}$ for parallel studies, and $SE(\beta_{AUC}) \approx \sqrt{\frac{2}{N}}\gamma_{CL}$ for crossover studies and calculated the corresponding MB-TOST type I error and power using respectively (2) and (3).

Further, we calculated the MB-TOST and MB-BOT type I error and power for increasing $SE(\beta)$ using (2), (3) and the equations in appendix of [1].

1.4 Alternative $SE(\beta)$ calculation with MB-BOT

Using the settings described in sections 3.1 and 3.2, we evaluated the type I error of MB-BOT (and MB-TOST for comparison) for two parallel design scenarios i) with N=40 subjects / n=3 sampling times and ii) N=12 subjects (6 subjects in each treatment group) / n=3.

We used the same softwares and settings as described in section 3.4. However, for the Post approach, we increased the number of iterations to 2300, with 500 burn-in, so that we obtained 1800 samples and reached convergence.

2 Results

2.1 Critical threshold for the use of MB-BOT

For the cross-over design with γ_{CL}=10 or 15% (as simulated in [1]), MB-TOST type I error and power were appropriate whether N= 12, 40 or 150 (Suppl. 2 Table I).
Indeed, the corresponding SE are all far below the threshold.

For the parallel design with ω_{CL}=22%, MB-TOST type I errors were appropriate for N=40 or 20 subjects. However, for N=12, the type I error dropped dramatically as the $SE(\beta_{AUC})$ came close to the threshold established according to (2) and (3) ($\delta/z_{1-\alpha} = 0.136$). With ω_{CL}=52%, for N=40, $SE(\beta_{AUC})$ increased past this threshold and led to a null type I error (and power), whereas for N=150, $SE(\beta_{AUC})$ decreased and consequently the type I error and power came back to appropriate levels.

In Suppl. 2 Table II we compare the type I error and power of MB-TOST and MB-BOT for increasing $SE(\beta)$. We note that increasing $SE(\beta)$ led to decreasing MB-TOST type I error (down to 0 for $SE(\beta) = \delta/z_{1-\alpha} = 0.136$) while MB-BOT maintained the nominal level of 5%. We also note that the power of MB-BOT stayed consistently higher.

Suppl. 2 Table I: $SE(\beta_{AUC})$, type I error and power of MB-TOST on β_{AUC} for varying N, ω_{CL} for the parallel design and γ_{CL} for the cross-over design.

	ω_{CL} (%)	N	$SE(\beta_{AUC})$	Type 1 error	Power
Parallel design	22	**12**	**0.127**	**0.019**	**0.088**
		20	0.098	0.048	0.466
		40	0.070	0.050	0.088
	52	**40**	**0.164***	**0.000**	**0.000**
		150	0.085	0.050	0.673
	γ_{CL} (%)	N	$SE(\beta_{AUC})$	Type 1 error	Power
Cross-over design	10	12	0.041	0.050	1.000
		40	0.022	0.050	1.000
		150	0.011	0.050	1.000
	15	12	0.061	0.050	0.956
		40	0.033	0.050	1.000
		150	0.017	0.050	1.000

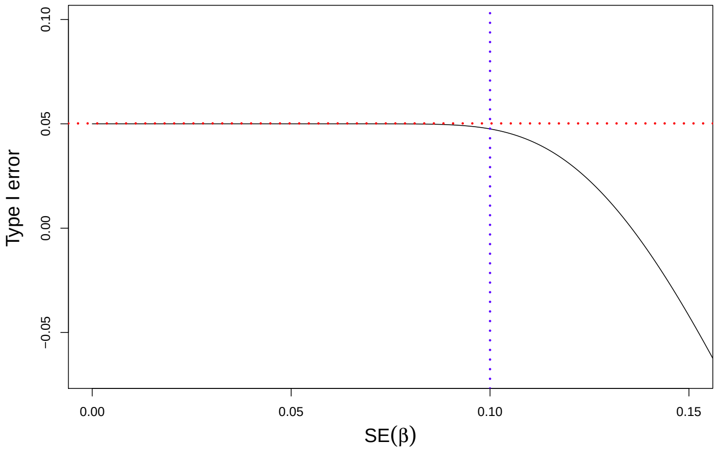
$*$ $SE(\beta_{AUC}) > 0.136$, leading to null expected type I error and power according to (2) and (3).

In **bold face** are the SE leading to a drop in expected type I error and power

Finally, Suppl. 2 Figure 1 further illustrates the relationship between $SE(\beta)$ and the MB-TOST type I error. We highlighted 0.1 as an inflection point for the drop in type I error when $\delta = \log(1.25)$ and $\alpha = 5\%$.

Suppl. 2 Table II: Expected type I errors and power of MB-TOST and MB-BOT on β for different values of $SE(\beta)$, and the corresponding quantiles u_α of the folded normal distribution of MB-BOT.

	MB-TOST		MB-BOT		
$SE(\beta)$	Type 1 error	Power	u_α	Type 1 error	Power
0.09	0.049	0.595	0.075	0.050	0.598
0.10	0.048	0.441	0.061	0.050	0.456
0.11	0.042	0.298	0.049	0.050	0.343
0.12	0.031	0.169	0.040	0.050	0.264
0.13	0.013	0.056	0.035	0.050	0.210
0.136	0.000	0.000	0.032	0.050	0.187



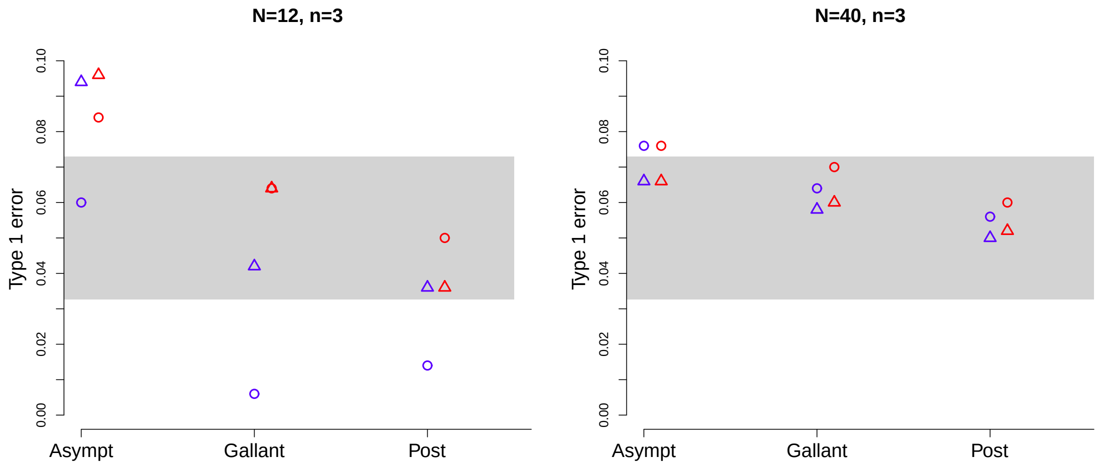
Suppl. 2 Figure 1: Expected type I error of MB-TOST on β as a function of $SE(\beta)$. The nominal level (=5%) is indicated by the horizontal dashed line and an inflection point at $SE(\beta)$ (=0.1) is indicated with the vertical dashed line.

2.2 Alternative $SE(\beta)$ calculation with MB-BOT

For β_{AUC}, the combination of $\omega_{AUC} = 22\%$ and N=12 pushed MB-TOST to its limit. $SE(\beta_{AUC})$ is expected to be approx 0.127 and consequently MB-TOST type I error to be around 0.019, according to Suppl. 2 Table I . However due to the design sparsity (n=3), the asymptotic SE underestimation inflated the type I error so that the estimate deceptively fell back within the nominal level 95% prediction interval. Indeed using the Post correction, the type I error of MB-TOST for β_{AUC} fell below the nominal level 95% prediction interval.

Conversely, MB-BOT type I error, always higher or equal to the MB-TOST estimate, was only impacted by the asymptotic SE underestimation. The alternative SE calculations enabled MB-BOT to keep a controlled type I error.

Computing times were similar to those indicated in the paper, except for Post approach on the sparse design with N=12. Indeed, for this design we increased the number of iterations in the HMC algorithm so that the approach reached about 35 minutes to run per data set.



Suppl. 2 Figure 2: Type I errors of MB-TOST (blue) and MB-BOT (red) on β_{AUC}^{Tr} (o) and on β_{Cmax}^{Tr} ($\triangle$) using the different SE calculations on the parallel sparse design with N=12 (left) and N=40 (right) subjects. The 95% prediction interval around 0.050 for 500 simulated data sets is indicated in grey ($PI_{95\%}(0.050) = [0.0326; 0.0729]$).

References

[1] Möllenhoff K, Loingeville F, Bertrand J, Nguyen TT, Sharan S, Sun G, et al.. Efficient model-based Bioequivalence Testing; arXiv:2002.09316[stat.ME]. 2020.