Efficient model-based bioequivalence testing

Kathrin Möllenhoff, Florence Loingeville, Julie Bertrand, Thu Thuy Nguyen, Satish Sharan, Guoying Sun, Stella Grosser, Liang Zhao, Lanyan Fang, France Mentré, Holger Dette

Nr. 21/2019

Efficient model-based Bioequivalence Testing

Kathrin Möllenhoff, Florence Loingeville, Julie Bertrand, Thu Thuy Nguyen,
Satish Sharan, Guoying Sun, Stella Grosser, Liang Zhao, Lanyan Fang,
France Mentré, Holger Dette

Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany,
Université de Paris, IAME INSERM, 75018 Paris, France,
Faculty of Pharmacy, Univ. of Lille, EA 2694: Public health:
Epidemiology and Healthcare quality, 59000 Lille, France,
Division of Quantitative Methods and Modeling, Office of Research Standards,
Office of Generic Drugs, Center for Drug Evaluation and Research,
Food and Drug Administration, 10903 New Hampshire Ave Silver Spring MD 20993, USA,
Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research,
Food and Drug Administration, 10903 New Hampshire Ave Silver Spring MD 20993, USA

September 5, 2019

Abstract

The classical approach to analyze pharmacokinetic (PK) data in bioequivalence studies aiming to compare two different formulations is to perform noncompartmental analysis (NCA) followed by two one-sided tests (TOST). In this regard the PK parameters AUC and C_{max} are obtained for both treatment groups and their geometric mean ratios are considered. According to current guidelines by the U.S. Food and Drug Administration and the European Medicines Agency the formulations are deemed to be similar if the 90%- confidence interval for these ratios falls between 0.8 and 1.25. As NCA is not a reliable approach in case of sparse designs, a model-based alternative has already been proposed for the estimation of AUC and C_{max} using non-linear mixed effects models. Here we propose another test than the TOST, called BOT, and evaluate it through a simulation study both for NCA and model-based approaches. For products with high variability on PK parameters, this method appears to have closer type I errors to the conventionally accepted significance level of 0.05, suggesting its potential use in situations where conventional bioequivalence analysis is not applicable.

Keywords and Phrases: bioequivalence, nonlinear mixed effects model, pharmacokinetics, noncompartmental bioequivalence analysis, two one-sided tests

1 Introduction

In drug development the comparison of two different formulations is a frequently addressed issue. In this regard bioequivalence studies investigating the difference between two treatment groups are performed. According to current guidelines by the U.S. Food and Drug Administration (2003) and the EMA (2014) this question is commonly addressed by comparing the ratios of the geometric means of the pharmacokinetic (PK) parameters area under the curve (AUC) and the maximal concentration ($C_{\max}$) to a prespecified threshold. More precisely, bioequivalence is established if the boundaries of the 90%-confidence intervals for these ratios fall between 0.8 and 1.25 which is equivalent to performing two one-sided tests (TOST) proposed by Schuirmann (1987). As the data are usually log-transformed, we consider the log-ratio (also defined as the treatment effect) and hence the commonly used threshold of equivalence is given by $\delta = \log(1.25)$.

When performing bioequivalence studies, the classical approach to analyze PK data is given by noncompartmental analysis (NCA), see for example Gabrielsson and Weiner (2001), followed by a linear mixed effect analysis of the AUC or Cmax. The advantage of this approach is that it is very simple and comes without any further assumptions or knowledge of the data. However, it requires a sufficiently large number of samples and subjects which cannot be provided in each trial. As pointed out by Dubois et al. (2011) and Hu et al. (2004) the estimates obtained by NCA are biased if these conditions are not fulfilled. Further, in numerous studies a sufficiently large number of samples cannot be guaranteed. For instance, in pediatric research, ethical considerations lead to difficulties in the planning of studies which are therefore typically very small in size (for an example see Mentré et al. (2001)). But also in other areas where patients are especially frail, as for example in cancer research, these requirements are often not met and therefore methods for sparse designs are required. In such situations the Nonlinear Mixed Effects Models (NLMEM) have become very popular for analyzing pharmacokinetic data (see Sheiner and Wakefield (1999)). NLMEM turned out to be a promising alternative to the classical approach as the estimation of individual effects allows for incorporating variabilities, as the Between-subject-variability (BSV) and the Within-subject-variability (WSV), for a detailed comparison see Pentikis et al. (1996); Combrink et al. (1997); Panhard and Mentré (2005). Consequently the main advantage of the NLMEM consists in the improved accuracy of the estimates in particular when dealing with sparse designs (see also Hu et al. (2004)).

In order to assess bioequivalence between two products typically the two one-sided tests (TOST) proposed by Schuirmann (1987) is performed, where two level α-tests are combined for testing two seperate sub-hypotheses. This method is based on the Intersection-Union Principle (see Berger (1982)) and one concludes bioequivalence if for both one-sided tests the null hypotheses can be rejected. Due to its simplicity, this approach which is still recommended in the FDA guidelines has become very popular and is common practice nowadays (see for example Bristol (1993), Brown et al. (1997) and Midha and McKay (2009) among many others). However, it was demonstrated by Phillips (1990) and Tsai et al. (2014) that for a small number of individuals, high variability in the data or only few samples per patient this method is rather conservative and suffers from a lack of power.

The present paper addresses this problem. Here we propose a new approach for the assessment of bioequivalence which turns out to have always more power than the corresponding TOST. The superiority of the new approach is particularly visible in situations with a large variability in the data in parallel designs. We develop a methodology which mimics the uniformly most powerful test for normally distributed data with known variance, which can be found in many text books on mathematical statistics (see for example, Lehmann and Romano (2006), or Wellek (2010)). We argue that the superiority of this methodology for NCA also carries over to model-based inference for reasonable large sample sizes and demonstrate this fact by means of a simulation study.

This paper is organized as follows. In Section 2 we present the classical problem of bioequivalence and the commonly used TOST for NCA-based inference. We also introduce the new method proposed in this paper, which will be called bioequivalence optimal testing (BOT), and demonstrate its superiority by means of a small simulation study. In Section 3 we introduce the NLMEM, then we present the model based TOST as first introduced by Panhard and Mentré (2005) and Dubois et al. (2011) and after that the model based BOT. Subsequently, these tests are compared by means of a simulation study in Section 4 with NCA-based tests both for parallel and cross-over designs varying BSV and WSV. In particular we demonstrate that the BOT (model and NCA-based) usually yields larger power than methodology based on the TOST, where the superiority of the BOT is particularly visible in situations with a large variability. Some theoretical arguments for these finding can be found in the Appendix, where we investigate the properties of both methods in the problem of comparing the means from two normal distributions with known variance. This scenario corresponds to some kind of asymptotic regime for the problems considered in practice, if the sample sizes are reasonably large.

Summarizing, the new BOT introduced in the present paper improves the commonly used TOST for bioequivalence testing based on NCA or model-based inference. It has never lower power than this test, but substantially larger power in scenarios with a large variability in parallel designs.

2 Bioequivalence Tests

In this section we will briefly review a commonly used approach for bioequivalence testing, which is based on the well-known two one-sided test (TOST) introduced by Schuirmann (1987). We introduce a new method for testing bioequivalence, which will turn out to be more powerful than the TOST as illustrated by means of a small simulation study. Some theoretical explanation for our findings is given in the Appendix. For the sake of simplicity, both methods are described here in the case of a two groups parallel design, but can be applied to crossover design, more standard in BE.

In a bioavailability/bioequivalence study a test (T) and a reference product (R) are administered and it is investigated whether the two formulations of the drug have similar properties with respect to average exposure in the population. Exposure, in this context, is usually characterized by blood concentration profile variables and summarized by the area under the time concentration curve (AUC) and the maximum concentration ($C_{\max}$). More precisely, let μ_T and μ_R denote the average means of the test and reference product for $\log AUC$ or $\log C_{\max}$, then the common testing problem in bioequivalence is defined by the hypotheses

$$H_0 : |\mu_T - \mu_R| \geq \delta \text{ vs. } H_1 : |\mu_T - \mu_R| < \delta, \tag{2.1}$$

where δ is a given threshold. For example, according to the 80/125-rule considered in the guidelines by EMA (2014) and U.S. Food and Drug Administration (2003) the threshold δ is given by $\delta = \log 1.25$.

For the problem of testing for PK bioequivalence the metrics of interest are given by AUC and $C_{\max}$, which means that we consider

$$\begin{aligned} \beta^T_{AUC} &:= \mu_T - \mu_R = \log AUC_T - \log AUC_R \\ \beta^T_{C_{\max}} &:= \mu_T - \mu_R = \log C_{\max,T} - \log C_{\max,R} \end{aligned} \tag{2.2}$$

in (2.1), where β^T_{AUC} and $\beta^T_{C_{\max}}$ are the treatment effects on AUC and $C_{\max}$ respectively.

2.1 The two one-sided Tests (TOST)

We consider the following sub-hypotheses of H_0 as described in (2.1) given by

$$H_{0,-\delta} : \mu_T - \mu_R \leq -\delta \text{ and } H_{0,\delta} : \mu_T - \mu_R \geq \delta. \tag{2.3}$$

The idea of the TOST consists in testing each of these hypotheses separately by a one-sided test. The global null hypothesis H_0 in (2.1) is rejected with a type I error α if both one-sided hypotheses are rejected with a type I error α. To be precise let $X_{T,1}, \dots X_{T,N_T}$ and $X_{R,1}, \dots X_{R,N_R}$ denote the samples from the test (T) and a reference product (R) respectively and denote by $\bar{X}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{k,i}$ $(k = R, T)$ the mean measured endpoints (over all individuals for the two treatments). Under the assumption that the random variables $\{X_{k,i} \ : \ i = 1,\dots N_k, \ k = R,T\}$ are independent and normally distributed with a common (but unknown) variance σ^2, that is $X_{R,i} \sim \mathcal{N}(\mu_R, \sigma^2)$; $i = 1\dots, N_R$, $X_{T,i} \sim \mathcal{N}(\mu_T, \sigma^2)$; $i = 1\dots, N_T$ we have for the corresponding means

$$\bar{X}_R \sim \mathcal{N}(\mu_R, \tfrac{\sigma^2}{N_R}) \text{ and } \bar{X}_T \sim \mathcal{N}(\mu_T, \tfrac{\sigma^2}{N_T}). \tag{2.4}$$

In applications $X_{R,i}$ and $X_{T,i}$ usually represent AUC_k and C_{max_k}, $k = R,T$, which are typically assumed to be lognormally distributed (see Lacey et al. (1997)). We denote by $\sigma_P^2 := \big(\frac{1}{N_R} + \frac{1}{N_T}\big)\sigma^2$ the pooled variance and by $d := \mu_T - \mu_R$ the difference between the expectations of the reference and the treatment group. This yields for the difference of the means

$$\bar{X}_T - \bar{X}_R \sim \mathcal{N}(d, \sigma_P^2). \tag{2.5}$$

The unknown variance σ_P^2 is estimated by

$$\hat{\sigma}_P^2 := \big(\tfrac{1}{N_T} + \tfrac{1}{N_R}\big)\hat{\sigma}^2, \tag{2.6}$$

where

$$\hat{\sigma}^2 = \frac{1}{N_T + N_R - 2} \sum_{k \in \{R,T\}} \sum_{i=1}^{N_k} \big(X_{k,i} - \bar{X}_k\big)^2 .$$

Consequently the null hypothesis in (2.1) is rejected if

$$\frac{\bar{X}_T - \bar{X}_R - (-\delta)}{\hat{\sigma}_P} \geq t_{N-2,1-\alpha} \text{ and } \frac{\bar{X}_T - \bar{X}_R - \delta}{\hat{\sigma}_P} \leq -t_{N-2,1-\alpha}, \tag{2.7}$$

where $t_{N,1-\alpha}$ is the $(1-\alpha)$-quantile of the t-distribution with $N - 2 = N_R + N_T - 2$ degrees of freedom (see for example Chow and Liu (1992)). This method is equivalent to constructing a $(1 - 2\alpha)$-confidence interval for $\mu_T - \mu_R$ and concluding bioequivalence if its completely contained in the equivalence interval $[-\delta, \delta]$ (see Schuirmann (1987)).

The approach presented above has been extended for model-based bioequivalence inference by Dubois et al. (2011) and will be explained in detail in Section 3.2.

2.2 An efficient alternative to TOST

In this section we will propose an alternative test which turns out to be the (asymptotically) most powerful test in this setting. The proof of this property is deferred to the Appendix A.2. This new approach motivates the model-based method which will be derived in Section 3.3. For the sake of brevity we call this test BOT (Bioequivalence Optimal Test) throughout this paper. Let $\mathcal{N}_F(d, \sigma_P^2)$ denote the folded normal distribution with parameters (d, σ_P^2), that is the distribution of the random variable $|Z|$, where $Z \sim \mathcal{N}(d, \sigma_P^2)$. Due to (2.5) we have for the absolute difference

$$\big|\bar{X}_T - \bar{X}_R\big| \sim \mathcal{N}_F(d, \sigma_P^2).$$

This result motivates the choice of the quantile determining the decision rule of the test, which is described in the following algorithm.

Algorithm 2.1. (The BOT)

1. Estimate the parameters of interest $\hat{\mu}_R$ and $\hat{\mu}_T$ by $\bar{X}_R$ and $\bar{X}_T$ (for instance by non-compartmental analysis) and estimate the variance of the difference $\bar{X}_T - \bar{X}_R$ by the statistic defined in (2.6).

2. Reject the null hypothesis, whenever

$$\left|\bar{X}_T - \bar{X}_R\right| \; < \hat{u}_\alpha, \tag{2.8}$$

where $\hat{u}_\alpha$ is the α-quantile of the folded normal distribution $\mathcal{N}_F(\delta, \hat{\sigma}_P^2)$.

The quantile $\hat{u}_\alpha$ can be calculated solving the equation

$$\alpha = \Phi\left(\tfrac{1}{\hat{\sigma}_P}(u-\delta)\right) - \Phi\left(\tfrac{1}{\hat{\sigma}_P}(-u-\delta)\right).$$

Alternatively, it can directly be obtained by using statistical software, as for example the *VGAM* package by Yee (2015) in R.
The approach presented in Algorithm 2.1 is extended in Algorithm 3.1 for model-based bioequivalence inference, where we will estimate the parameters of interest $\hat{\mu}_R$ and $\hat{\mu}_T$ by fitting a nonlinear mixed model to the data.

2.3 A finite sample comparison

In the following we will compare the TOST and the BOT introduced in Section 2.1 and 2.2 respectively by means of a small simulation study. For this purpose we generate observations $X_{R,i}$, $i = 1, \dots, N_R$ and $X_{T,i}$, $i = 1, \dots, N_T$, respectively, from a normal distribution, that is

$$X_{R,i} \sim \mathcal{N}(\log(1.25) + \beta^T, \sigma^2),\; i = 1, \dots, N_R \text{ and } X_{T,i} \sim \mathcal{N}(\log(1.25), \sigma^2),\; i = 1, \dots, N_T,$$

where $\beta^T \in \mathbb{R}$ is chosen from the set $\{\log(1), \log(1.1), \log(1.2), \log(1.25), \log(1.35), \log(1.5)\}$. Note that β^T directly corresponds to the true underlying treatment effect $\beta^T = d = \mu_T - \mu_R$. We fix the number of subjects to $N_R = N_T = 20$, resulting in a total sample size of $N = 40$ and consider four different variance settings, that is σ^2 chosen from the set $\{0.05, 0.1, 0.2, 0.25\}$. Further we choose a threshold of $\delta = \log(1.25)$ in the hypotheses (2.1) corresponding to current guidelines (as explained in Section 2.1) and set the level of the test to $\alpha = 0.05$.
In Table 1 we display the type I error rates of the TOST (2.7) and the BOT (2.8). The simulations are based on 1000 simulation runs. It turns out that for small variances (that is $\sigma^2 = 0.05$ and $\sigma^2 = 0.1$) we obtain similar results for both methods. Further, for increasing

	$\sigma^2 = 0.05$		$\sigma^2 = 0.1$		$\sigma^2 = 0.15$		$\sigma^2 = 0.2$		$\sigma^2 = 0.25$	
β^T	TOST	BOT	TOST	BOT	TOST	BOT	TOST	BOT	TOST	BOT
log(1.5)	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.002	0.000	0.004
log(1.35)	0.006	0.007	0.009	0.009	0.010	0.016	0.002	0.016	0.002	0.030
log(1.25)	0.040	0.044	0.047	0.050	0.027	0.059	0.009	0.054	0.004	0.050

Table 1: *Type I error of the TOST defined in* (2.7) *and the BOT defined in* (2.8) *for the hypothesis* (2.1) *with* $\delta = \log(1.25)$. *The 95% prediction is given by* $[0.0373; 0.0656]$ *centered at* 0.05.

variance the TOST becomes very conservative as the proportion of rejection tends to zero, even on the boundary of the null hypothesis, that is $\beta^T = d = \log(1.25)$. On the other hand, the BOT yields a very precise approximation of the nominal level, as rejection probabilities for $\beta^T = d = \log(1.25)$ are very close to $\alpha = 0.05$.

In Table 2 we display the power of both tests. Again, for the low variance setting we obtain very similar results for both tests. For $\sigma^2 = 0.1$ we can already observe a higher power for the BOT and for $\sigma^2 = 0.15$ this effect becomes even more visible. Moreover, it turns out that the TOST does not have any power when considering $\sigma^2 = 0.2$ and $\sigma^2 = 0.25$ (the proportions of rejection are 0.041 and 0.006 respectively).

	$\sigma^2 = 0.05$		$\sigma^2 = 0.1$		$\sigma^2 = 0.15$		$\sigma^2 = 0.2$		$\sigma^2 = 0.25$	
β^T	TOST	BOT	TOST	BOT	TOST	BOT	TOST	BOT	TOST	BOT
log(1.2)	0.130	0.144	0.083	0.105	0.043	0.093	0.022	0.076	0.003	0.056
log(1.1)	0.548	0.565	0.254	0.286	0.097	0.195	0.032	0.141	0.009	0.113
log(1)	0.852	0.868	0.421	0.490	0.134	0.272	0.041	0.198	0.006	0.143

Table 2: *Power of the TOST defined in* (2.7) *and the BOT defined in* (2.8) *under the alternative, for the hypothesis* (2.1) *with* $\delta = \log(1.25)$.

Summarizing the BOT exhibits a better performance than the TOST in all scenarios under consideration. The superiority is negligible in settings with low variances but for larger variances the BOT has substantially more power. In such cases the TOST is conservative but the approximation of the level by the BOT is very accurate. These empirical findings are investigated from a theoretical point of view in Appendix A. Here we show that the probability of rejection by the BOT is always close (in the case of a known variance in both treatment groups precisely equal) to α on the boundary of the null hypothesis, that is $|\mu_T - \mu_R| = \delta$. On the other hand the level is only smaller or equal than α for the TOST, and the difference may be very substantial in scenarios with large variability. As a consequence the power of the BOT is always larger than the power of the TOST, and BOT offers a very promising alternative to the classical TOST.

2.4 Noncompartmental analysis

If we are testing for PK bioequivalence considering the hypotheses in (2.2) we need to calculate estimates of AUC and $C_{\max}$ directly from the data. In this regard the classical approach is given by NCA, as described for example in Gabrielsson and Weiner (2001). More precisely, $C_{\max}$ is directly obtained from the data, whereas AUC is approximated by the linear trapezoidal rule. This means that the total area under the curve is obtained by separating it into several smaller trapezoids and summing up these areas. Of course the accuracy of this approach strongly depends on the number of measurements as this gives the number of trapezoids but it does not require a model assumption and is widely applicable. As these methods do not take the profile of the blood concentration-time curve into account, we call them NCA-based methods throughout this paper. More precisely, we will denote the tests proposed in Sections 2.2 and 2.3 by NCA-TOST and NCA-BOT respectively. Model-based-methods will be discussed in more detail in Section 3.

3 Model-based Bioequivalence Tests

Classical NCA-based tests are a useful tool to establish bioequivalence if the blood concentration profile variables AUC and $C_{\max}$ can be calculated with a reasonable precision without using information about the form of the concentration profiles. For this purpose one usually needs a relatively dense design to determine the area under the curve or the maximum of the profile. However, there are many situations, where only a sparse design is available (for some examples see Hu et al. (2004)) and the NCA-based calculation of AUC and $C_{\max}$ might be misleading as the estimates are biased in this case (see Dubois et al. (2011)). In such situations where NCA is not reliable a model-based approach as proposed for the TOST by Panhard and Mentré (2005) and Dubois et al. (2011) might have important advantages.

Roughly speaking they proposed to use non-linear mixed effects models (NLMEM) to describe the blood concentration profile and derive AUC and $C_{\max}$ estimates. These quantities are then further analyzed using the methodology introduced in Section 2. By this approach they were willing to increase the accuracy of bioequivalence tests in the case of sparse designs.

We will use the same methodology to extend the BOT derived in Section 2.1 to situations with sparse designs. This new test achieves more power and simultaneously control the type I error.

3.1 Nonlinear mixed effects models (NLMEM)

We first consider crossover trials with K periods and N subjects, investigating the difference between a test and a reference treatment. A classical situation is given by the (balanced) two-period, two-sequence crossover design ($K = 2$), where the $N/2$ patients receive treatment R in the first period and treatment T in the second one while the other $N/2$ patients receive the treatments in the reverse order.

For each subject concentrations of the drug are measured in all periods and at different sampling points. In order to represent the dependence of the concentration on time for one subject we follow Dubois et al. (2011) and use a non-linear function, say f in order to fit one global model to the data, that is

$$y_{i,j,k} = f(t_{i,j,k}, \psi_{i,k}) + g(t_{i,j,k}, \psi_{i,k})\varepsilon_{i,j,k}, \tag{4.1}$$

where $y_{i,j,k}$ denotes the concentration of the i-th subject ($i = 1, \dots N$) at sampling time $t_{i,j,k}$ ($j = 1, \dots, n_{i,k}$) of period k ($k = 1, \dots K$). In (4.1) the residual errors $\varepsilon_{i,j,k}$ are independent and standard-normally distributed random variables and the function g is used to model heteroscedasticity. In particular we consider a combined error model with

$$g(t_{i,j,k}, \psi_{i,k}) = a + b \cdot f(t_{i,j,k}, \psi_{i,k}), \tag{4.2}$$

where the parameters $a, b \in \mathbb{R}_{\geq 0}$ account for the additive and the proportional part of the error respectively. This gives for the variance of the errors in (4.1)

$$\mathrm{Var}(y_{i,j,k}) = (g(t_{i,j,k}, \psi_{i,k}))^2 = |a + b \cdot f(t_{i,j,k}, \psi_{i,k})|^2.$$

The individual parameters $\psi_{i,k} = (\psi_{i,k,1}, \dots, \psi_{i,k,p})^\top$ (of length p) are defined by

$$\log(\psi_{i,k,l}) = \log \lambda_l + \beta_l^T Tr_{i,k} + \beta_l^P P_k + \beta_l^S S_i + \eta_{i,l} + \kappa_{i,k,l},\ l = 1, \dots, p, \tag{4.3}$$

where $\lambda = (\lambda_1, \dots, \lambda_p)^\top$ denotes a vector of fixed effects, $Tr_{i,k}$, P_k and S_i the (known) vectors of treatment, period and sequence covariates respectively and β^T, β^P and β^S the vectors of coefficients of treatment, period and sequence effects. In order to account for the variability between individuals, denoted as between-subject-variability (BSV), and the variability of one subject between two periods respectively, that is the within-subject-variability (WSV), we introduce random effects $\eta_i = (\eta_{i,1}, \dots, \eta_{i,p})^\top$ and $\kappa_{i,k} = (\kappa_{i,k,1},, \dots, \kappa_{i,k,p})^\top$. More precisely, the random effect η_i represents the BSV of subject i and $\kappa_{i,k}$ the WSV of subject i at period k respectively. Throughout this section we assume that the random effects are normal distributed, that is

$$\eta_i \sim \mathcal{N}(0, \Omega),\ \kappa_{i,k} \sim \mathcal{N}(0, \Gamma),\ i = 1, \dots N,\ k = 1, \dots K, \tag{4.4}$$

with $p \times p$-dimensional covariance matrices Ω and Γ and denote the diagonal elements of these matrices by ω_l^2 and γ_ℓ^2 respectively. Finally, the vector of all parameters in model (4.1) is given by

$$\theta = (\lambda, \beta^T, \beta^S, \beta^P, \Omega, \Gamma, a, b). \tag{4.5}$$

For biologics with a long half-life, such as monoclonal antibodies, a parallel group design, that is each individual receives only the test or the reference treatment, may be necessary (Dubois et al. (2012)). In that case, we consider only one period and the WSV can be omitted and (4.3) simplifies to

$$\log(\psi_{i,l}) = \log \lambda_l + \beta_l^T Tr_i + \eta_{i,l},\ l = 1, \dots, p. \tag{4.6}$$

Note that in this case we do not assume any period or sequence effects and hence the vector in (4.5) simplifies to $\theta = (\lambda, \beta^T, \Omega, a, b)$. For the sake of simplicity we now introduce a vector β which is defined by $\beta := \beta^T$ in case of parallel designs and $\beta := (\beta^T, \beta^S, \beta^P)$ for crossover designs. Consequently we can write for the vector of all parameters in model (4.1) $\theta = (\lambda, \beta, \Omega, \Gamma, a, b)$, where Γ disappears in case of parallel design.
Considering now the hypotheses in (2.2) the treatment effects β^T_{AUC} and $\beta^T_{C_{\max}}$ on AUC and $C_{\max}$ respectively can be directly obtained from the parameters of the global NLMEM. In other words, there exist functions, h_{AUC}, $h_{C_{\max}}$, such that

$$\beta^T_{AUC} = h_{\mathrm{AUC}}(\lambda, \beta),\ \beta^T_{C_{\max}} = h_{C_{\max}}(\lambda, \beta). \tag{4.7}$$

By this we obtain an estimate for its variance using the delta method (Oehlert (1992)), which has been proposed by Panhard et al. (2007). With these notations the hypotheses in (2.1) can be rewritten as

$$H_0 : |\beta^T| \geq \delta \quad \text{versus} \quad H_1 : |\beta^T| < \delta \tag{4.8}$$

where we do the same for AUC and $C_{\max}$.

3.2 Model-based TOST

A model-based version introduced by Panhard and Mentré (2005); Panhard et al. (2007) and Dubois et al. (2011) of the TOST for bioequivalence can be obtained by fitting the NLMEM (4.1) to the data and calculate the estimate $\hat{\beta}^T_c$ of the treatment effect β^T_c, $c = AUC,\ C_{max}$. We can assume from the theory of mixed effects modeling (see for example Demidenko (2013)) that this estimate $\hat{\beta}^T_c$ is asymptotically normal distributed and following the discussion in Section 2.1 the null hypothesis in (4.8) is rejected whenever

$$\frac{\hat{\beta}^T_c - (-\delta)}{SE(\hat{\beta}^T_c)} \geq z_{1-\alpha} \quad \text{and} \quad \frac{\hat{\beta}^T_c - \delta}{SE(\hat{\beta}^T_c)} \leq -z_{1-\alpha},\ c = AUC, C_{max}, \tag{4.9}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution and $SE(\hat{\beta}^T_c)$ is an estimate of the standard error of the estimate $\hat{\beta}^T_c$.
We obtain $SE(\hat{\beta}^T_c)$ by using an asymptotic approximation based on the estimated covariance matrix of the fixed effects (given by a submatrix of the inverse of the Fisher information matrix) and the Delta-method (see Oehlert (1992) and Dubois et al. (2011) for the concrete

calculation). More precisely, considering (4.7) and denoting the estimated covariance matrix of the fixed effects by $\hat{V}$, we have

$$SE(\hat{\beta}_c^T) = \sqrt{\nabla h_c(\hat{\lambda}, \hat{\beta}) \cdot \hat{V} \cdot \nabla h_c(\hat{\lambda}, \hat{\beta})}, \quad c = AUC,\ C_{\max}, \tag{4.10}$$

where ∇h_c denotes the gradient of the function h_c, expressing β_c^T as a function of the model parameters ($c = AUC$ or $C_{\max}$). As the functions h_{AUC} and $h_{C_{\max}}$ are known, all quantities of the rejection rule given in (4.9) can be directly obtained from the estimates of the parameters in model (4.1).

3.3 Model-based optimal Bioequivalence Test

In this section we extend the bioequivalence test described in Section 2.2 to NLMEM. It will be shown in Section 4 that the new method significantly improves currently used tests for bioequivalence of concentration curves measured by the pharmacokinetic parameters AUC and $C_{\max}$ as it can also be applied in the case of sparse designs. Further this test turns out to be more powerful than the model-based TOST described in Section 3.2, in particular for small sample sizes or data with high variability. The adaption of Algorithm 2.1 to model-based bioequivalence is very straight forward and is summarized in the following algorithm:

Algorithm 3.1. (The model-based BOT on AUC and C_{max})

1. Estimate a NLMEM to the data, resulting in the parameter estimate $\hat{\theta} = (\hat{\lambda}, \hat{\beta}, \hat{\Omega}, \hat{\Gamma}, \hat{a}, \hat{b})$. This can be done for example for parallel designs using the *saemix* package by Comets et al. (2011). The test statistic can be directly calculated as secondary parameter of the model parameters (see (4.7)) and is given by

   $$|\hat{\beta}_c^T| = |h_c(\hat{\lambda}, \hat{\beta})|,\ c = AUC, C_{max}.$$

   Approximate the standard error of the estimate $SE(\hat{\beta}_c^T),\ c = AUC, C_{max}$, by using the Delta-Method as describred in (4.10).

2. Reject the null hypothesis, whenever

   $$|\hat{\beta}_c^T| < \hat{u}_\alpha, \tag{4.11}$$

   where $\hat{u}_\alpha$ is the α-quantile of the folded normal distribution $\mathcal{N}_F(\delta, (SE(\hat{\beta}_c^T))^2)$.

Finite sample properties of this method are given in Section 4.

4 Numerical comparison of NCA- and model-based- approaches

In this section we investigate the finite sample properties of the different methods by means of a simulation study. For this purpose we consider eight different scenarios for parallel designs and for two-periods-two-sequence-cross-over studies respectively. Note that the latter represent the standard design for bioequivalence trials. More precisely, we will use the models as described in Section 3.1 in order to simulate pharmacokinetic (PK) data using a population PK model with several scenarios varying the study design, the number of sampling times per subject n and the magnitude of BSV and WSV (for the cross-over designs). The threshold for bioequivalence in (2.1) is as explained in Section 2 chosen as $\delta = \log(1.25)$ in all cases under consideration.

4.1 Settings

We use the same PK model as described in Dubois et al. (2011), which describes concentrations (mg/l) of the anti-asthmatic drug theophylline, for both reference and test group. More precisely, we consider a one-compartment model with first-order absorption and first-order elimination and hence the pharmacokinetic function f in (4.1) is defined by

$$f(t, D, k_a, CL/F, V/F) = \frac{F \cdot D \cdot k_a}{V(\frac{CL}{V} - k_a)} \left(\exp(-k_a \cdot t) - \exp(-\frac{CL}{V} \cdot t) \right), \tag{5.1}$$

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 and hence ψ is composed of k_a, CL/F and V/F. The value for the residual error model in (4.2) were set to $a = 0.1$mg/l and $b = 10\%$. The dose is fixed to $D = 4$mg for all subjects, and the fixed effects for the reference treatment group are $\lambda_{k_a} = 1.5\, h^{-1}$, $\lambda_{CL/F} = 0.04\, l\, h^{-1}$, and $\lambda_{V/F} = 0.5\, l$. The variance-covariance matrices Ω and Γ were chosen to be diagonal and we investigate two different levels of variability for the parallel and crossover design as specified in Table 3. To evaluate the type I error of the approaches, we simulate a treatment effect on parameters V and CL given by $\beta^T_V = \beta^T_{CL} = \log(1.25)$, which affects the AUC and C_{max} similarly, that is $|\mu_T - \mu_R| = \beta^T_{AUC} = |\log(AUC_T) - \log(AUC_R)| = \beta^T_{C_{max}} = |\log(C_{max_T}) - \log(C_{max_R})| = \log(1.25)$. The power of the bioequivalence test will be evaluated for $\beta_{CL} = \beta_V = \log(1)$. We will study two sampling time designs

- Rich design: $N = 40$, $n = 10$ samples taken at times $t = (0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, 24)$ hours after dosing,
- Sparse design: $N = 40$, $n = 3$ samples taken at times $t = (0.25, 3.35, 24)$ hours after dosing

as described in Dubois et al. (2011), where all subjects have the same vector of sampling times. Note that in this situation the sparse design reflects the most critical case as three sampling

Design	Variability Scenario	ω_{k_a}	$\omega_{V/F}$	$\omega_{CL/F}$	γ_{k_a}	$\gamma_{V/F}$	$\gamma_{CL/F}$
Parallel	Low BSV	22	11	22	NA	NA	NA
	High BSV	52			NA	NA	NA
Crossover	Low	20	10	20	10	5	10
	High	50			15		

Table 3: *Simulated values for the parallel and crossover design, low and high variability settings. ω and γ are expressed as coefficient of variation in %. Entries "NA" correspond to "not applicable".*

points are the minimum required for estimating a model with three parameters given in (5.1). For each scenario, we simulate 500 data sets. For the estimation of the model parameters we use the SAEM algorithm (see Kuhn and Lavielle (2005)). More precisely, in case of parallel designs, we use the R package *saemix* developed by Comets et al. (2011) with 10 chains and $(300, 100)$ iterations. For crossover studies, we used Monolix 2018 R2 developed by Lixoft (2018) to fit the model to the data with the same number of chains and interations as for parallel designs. For the standard NCA analysis (see for example Gabrielsson and Weiner (2001)) we used the R package $MESS$ developed by Ekstrom (2019). As this technique is not appropriate for sparse samples we only report results for NCA-based methods based on rich design.
We start considering a parallel design. Two-treatments parallel trials are simulated, that is 20 subjects receive the reference treatment R and the other 20 subjects are allocated to the test treatment T. Illustrations of the simulated concentrations in groups R and T under H_0 and H_1 in (4.8) are presented in Figure 1. Secondly, we observe a two-periods two-sequences crossover design. For each trial, the 20 subjects allocated to the first sequence receive the reference treatment first and then the test treatment. The other 20 subjects allocated to the second sequence receive treatments in the reverse order. Table 3 displays all variabilities under consideration.

4.2 Results

4.2.1 Type I error

In Table 4 we show the results for all tests proposed in Sections 2 and 3. For parallel designs it becomes obvious that both the NCA-based and the model-based TOST are conservative in settings with a high variability, while the BOT yields a very accurate approximation of the level. This corresponds to the empirical findings in Section 2 and the theoretical arguments given in the Appendix. However, we observe a slightly increased type I error for the sparse design with low variability for both model-based methods, probably due to standard error

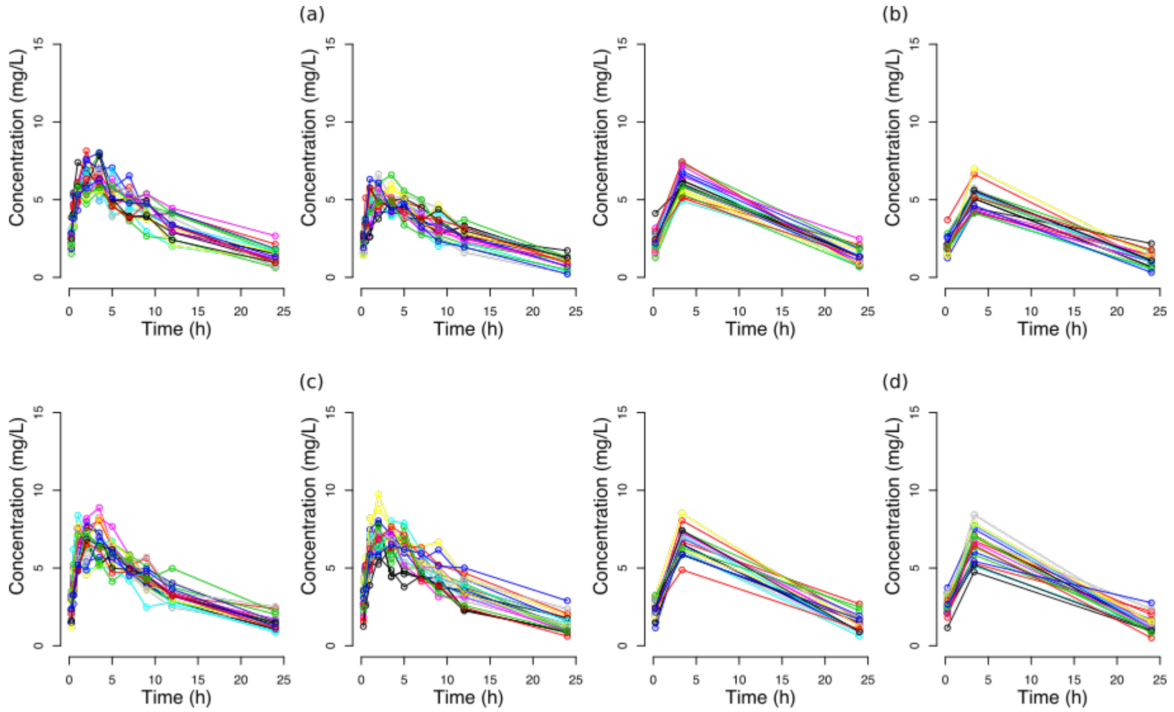

Figure 1: *Spaghetti plots of simulated concentrations for parallel design with* $N = 40/n = 10$ *((a) and (c))) and* $N = 40/n = 3$ *((b) and (d)), low variability under* H_0 *(top line), that is* $\beta^T = \log(1.25)$ *and* H_1 *(bottom line), that is* $\beta^T = \log(1)$*. On each plot, profiles on the left correspond to the reference group (R) and profiles on the right correspond to the treatment group (T).*

underestimation as mentioned by Dubois et al. (2011). For rich samples and low variability all four tests under consideration perform well and yield an accurate approximation of the nominal level at boundary of the hypotheses, that is $\delta = \log 1.25$.

In the case of crossover designs the approximation of the level is very precise for all four tests under consideration, even in the case of high variability. This can be explained by the fact that each individual receives a test and a reference treatment and hence we have twice as much data as for the parallel designs data. However, there is a slight type I error inflation (0.078) for the model-based TOST considering a sparse design with high variability. Concluding, the type I error rates are close to α in almost all scenarios under consideration. For increasing variances both versions of the TOST become very conservative whereas the BOT approximates the level still very precisely.

Study Design		Parallel				Crossover			
Sampling time		Rich		Sparse		Rich		Sparse	
Variability		Low	High	Low	High	Low	High	Low	High
NCA-TOST	AUC	0.052	**0.022**	-	-	0.046	0.042	-	-
	$C_{\max}$	0.062	**0.012**	-	-	0.062	0.070	-	-
NCA-BOT	AUC	0.052	0.054	-	-	0.046	0.042	-	-
	$C_{\max}$	0.062	0.052	-	-	0.062	0.070	-	-
MB-TOST	AUC	0.056	**0.004**	**0.076**	**0.006**	0.056	0.042	0.038	0.050
	$C_{\max}$	0.058	**0.008**	0.066	**0.002**	0.064	0.070	0.044	**0.078**
MB-BOT	AUC	0.056	0.064	**0.076**	0.034	0.056	0.044	0.038	0.056
	$C_{\max}$	0.070	0.060	0.070	0.058	0.064	0.054	0.044	0.056

Table 4: *Type I errors of the four tests under H_0. The numbers in boldface indicate that the type I error falls outside of the 95% prediction interval* $[0.0326; 0.0729]$ *centered at* 0.05.

4.2.2 Power

In order to investigate the power of the proposed methods we consider the scenarios summarized in Table 3 with a treatment effect of $\beta^T_{AUC} = 0$ and $\beta^T_{C_{max}} = 0$. In Table 5 we display the results for the four tests under consideration. In the case of parallel designs we observe that a sparse design does not affect the performance of the tests as much as the level of variability, which when high leads to a huge loss of power for all methods. Although in these settings the power is only close to 0.15 for the model-based BOT, a noticeable improvement compared to the model-based TOST is visible, as for this test the power is practically zero. For low variability the model-based tests perform very similarly, which confirms again the empirical findings in Section 2 and some theoretical explanation for these observations is given in the appendix. When considering rich designs the NCA-based methods achieve more power than the model-based ones but the difference turns out to be quite small. However, for sparse designs NCA-based methods are not applicable and in case of low variability we obtain a very high power for both the model-based BOT and the model-based TOST. For the cross-over designs all tests under consideration yield a power of one, irrespective of the sampling time, design and variability. This effect can again be explained by the larger sample size and each individual receiving both treatments.

Study Design		Parallel Design				Crossover Design			
Sampling Time		Rich		Sparse		Rich		Sparse	
Variability		Low	High	Low	High	Low	High	Low	High
NCA-TOST	AUC	0.998	0.132	-	-	1.000	1.000	-	-
	$C_{\max}$	0.998	0.056	-	-	1.000	1.000	-	-
NCA-BOT	AUC	0.998	0.228	-	-	1.000	1.000	-	-
	$C_{\max}$	0.998	0.154	-	-	1.000	1.000	-	-
MB-TOST	AUC	0.830	0.008	0.804	0.004	1.000	1.000	1.000	0.998
	$C_{\max}$	1.000	0.024	1.000	0.016	1.000	1.000	1.000	1.000
MB-BOT	AUC	0.838	0.140	0.808	0.132	1.000	1.000	1.000	1.000
	$C_{\max}$	1.000	0.138	1.000	0.116	1.000	1.000	1.000	1.000

Table 5: *Power of the four tests under H_1.*

5 Conclusions

In this paper we addressed the problem of sparse designs and high variability in bioequivalence studies. As described by Phillips (1990) and Tsai et al. (2014) we demonstrated that in general for data with high variability methods based on the TOST suffer from a lack of power. To address this problem we introduced a new method using quantiles of the folded normal distribution, which we called bioequivalence optimal testing (BOT) in this paper. In the case of known variances we proved in the Appendix that this test is uniformly most powerful in this setting and has consequently more power than the TOST. These arguments can be transferred to general bioequivalence testing using NCA or NLMEM if the sample variances can be estimated with reasonable accuracy.

By means of a simulation study we compared the TOST and the BOT based on NCA and NLMEM. We demonstrated that bioequivalence testing based on the BOT is a more powerful alternative to the commonly used TOST if the *AUC* and $C_{\max}$ are obtained by NCA. This superiority is also observed if these parameters are obtained by fitting an NLMEM, in particular for data with large variability.

Acknowledgements
This work has also been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt T1) of the German Research Foundation (DFG) and by the Food and Drug Administration (FDA) under contract 10110C. The authors would also like to thank Dr. Martin Fink and Dr. Frank Bretz for their helpful discussions and comments on an earlier version of this paper.

Disclaimer
This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

Bristol, D. R. (1993). Probabilities and sample sizes for the two one-sided tests procedure. *Communications in Statistics-Theory and Methods*, 22(7):1953–1961.

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

Chow, S.-C. and Liu, P.-J. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

Combrink, M., McFadyen, M., and Miller, R. (1997). A comparison of the standard approach and the nonmem approach in the estimation of bioavailability in man. *Journal of pharmacy and pharmacology*, 49(7):731–733.

Comets, E., Lavenu, A., and Lavielle, M. (2011). Saemix, an r version of the saem algorithm. In *20th meeting of the Population Approach Group in Europe, Athens, Greece. Abstr*, volume 2173.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Dubois, A., Gsteiger, S., Balser, S., Pigeolet, E., Steimer, J. L., Pillai, G., and Mentré, F. (2012). Pharmacokinetic similarity of biologics: analysis using nonlinear mixed-effects modeling. *Clin. Pharmacol. Ther.*, 91(2):234–242.

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

Ekstrom, C. (2019). Mess: Miscellaneous esoteric statistical scripts. R package version 0.5.5, available at `https://CRAN.R-project.org/package=MESS`.

EMA (2014). Guideline on the investigation of bioequivalence. available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf`.

Gabrielsson, J. and Weiner, D. (2001). *Pharmacokinetic and pharmacodynamic data analysis: concepts and applications*, volume 2. CRC Press.

Hu, C., Moore, K. H., Kim, Y. H., and Sale, M. E. (2004). Statistical issues in a modeling approach to assessing bioequivalence or pk similarity with presence of sparsely sampled subjects. *Journal of pharmacokinetics and pharmacodynamics*, 31(4):321–339.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038.

Lacey, L., Keene, O., Pritchard, J., and Bye, A. (1997). Common noncompartmental pharmacokinetic variables: are they normally or log-normally distributed? *Journal of biopharmaceutical statistics*, 7(1):171–178.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lixoft (2018). The monolix software, version r2.

Mentré, F., Dubruc, C., and Thénot, J.-P. (2001). Population pharmacokinetic analysis and optimization of the experimental design for mizolastine solution in children. *Journal of pharmacokinetics and pharmacodynamics*, 28(3):299–319.

Midha, K. K. and McKay, G. (2009). Bioequivalence; its history, practice, and future.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

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

Panhard, X., Taburet, A.-M., Piketti, C., and Mentré, F. (2007). Impact of modelling intra-subject variability on tests based on non-linear mixed-effects models in cross-over pharmacokinetic trials with application to the interaction of tenofovir on atazanavir in hiv patients. *Statistics in medicine*, 26(6):1268–1284.

Pentikis, H. S., Henderson, J. D., Tran, N. L., and Ludden, T. M. (1996). Bioequivalence: individual and population compartmental modeling compared to the noncompartmental approach. *Pharmaceutical research*, 13(7):1116–1121.

Phillips, K. F. (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of pharmacokinetics and biopharmaceutics*, 18(2):137–144.

Romano, J. P. et al. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, 33(3):1036–1047.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680.

Sheiner, L. and Wakefield, J. (1999). Population modelling in drug development. *Statistical methods in medical research*, 8(3):183–193.

Tsai, C.-A., Huang, C.-Y., and Liu, J.-p. (2014). An approximate approach to sample size determination in bioequivalence testing with multiple pharmacokinetic responses. *Statistics in medicine*, 33(19):3300–3317.

U.S. Food and Drug Administration (2003). Guidance for industry: bioavailability and bioequivalence studies for orally administered drug products-general considerations. *Food and Drug Administration, Washington, DC.* available at `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070124.pdf`.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority.* CRC Press.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R.* Springer, New York, USA.

A Theoretical comparison of tests for bioequivalence

In this section we provide some theoretical explanation, why the BOT proposed in Section 2.2 has more power than the TOST. For this purpose we now assume that variances in the reference and treatment group are known. In this case the quantiles of the t-distribution in (2.7) can be replaced by those of a normal distribution and the power functions of all tests can be calculated explicitly. We also note that this assumption is very well justified, if the sample sizes in both groups are sufficiently large. In other words: all arguments presented in this section can be applied to the NCA-based tests discussed in Section 2.1 and 2.2 provided that the sample sizes are sufficiently large. A similar comment applies to the model-based test for bioequivalence introduced in Section 3. We begin with a discussion of the TOST.

A.1 The two one-sided Test (TOST)

Consider the rejection rule of the TOST defined in (2.7), where we replace the estimate of the (pooled) variance $\sigma_P^2 = \frac{\sigma^2}{N_T} + \frac{\sigma^2}{N_R}$ by its true value and the quantile $t_{N-2,1-\alpha}$ by the $(1-\alpha)$ quantile of the standard normal distribution denoted by $z_{1-\alpha}$. If $z_{1-\alpha} > \delta/\sigma_P$ the probability of rejection is 0 (because the conditions in (2.7) are contradicting). On the other hand, and more importantly, if $z_{1-\alpha} \leq \delta/\sigma_P$ the probability of rejection for the test (2.7) is given by

$$\begin{aligned} \Psi_{\text{TOST}}(d) &:= \mathbb{P}_d\Big(\tfrac{\bar{X}_T-\bar{X}_R+\delta}{\sigma_P} \geq z_{1-\alpha}, \tfrac{\bar{X}_T-\bar{X}_R-\delta}{\sigma_P} \leq -z_{1-\alpha}\Big) \\ &= \mathbb{P}_d\Big(z_{1-\alpha} - \tfrac{\delta+d}{\sigma_P} \leq \tfrac{\bar{X}_T-\bar{X}_R-d}{\sigma_P} \leq -z_{1-\alpha} + \tfrac{\delta-d}{\sigma_P}\Big) \\ &= \Phi\Big(-z_{1-\alpha} + \tfrac{\delta-d}{\sigma_P}\Big) - \Phi\Big(z_{1-\alpha} - \tfrac{\delta+d}{\sigma_P}\Big), \end{aligned} \tag{3.1}$$

where Φ denotes the distribution function of the standard-normal distribution. From this formula we draw the following conclusions (if $z_{1-\alpha} \leq \delta/\sigma_P$):

(1) The test (2.7) controls its level. For example, if $d > \delta$ we have

$$\Psi_{\text{TOST}}(d) < \Phi(-z_{1-\alpha}) = \alpha$$

and with a similar argument the same inequality can be derived for $d < -\delta$.

(2) At the "boundary" of the null hypothesis (that is $d \in \{-\delta, \delta\}$) we have

$$\Psi_{\text{TOST}}(\pm\delta) = \alpha - \Phi\left(z_{1-\alpha} - \tfrac{2\delta}{\sigma_P}\right) \leq \alpha,$$

As $\Phi\left(z_{1-\alpha} - \frac{2\delta}{\sigma_P}\right)$ converges to 0 if $\frac{\delta}{\sigma_P}$ converges to infinity, we expect that the level of the test (2.7) is close to α at the "boundary" of the null hypothesis, if σ is small. This happens, for example, if the variance σ^2 (and hence the pooled variance σ_P^2) is small or, alternatively, if the sample sizes N_R and N_T in both groups are very large. On the other hand the test (2.7) is conservative if the variance σ^2 is large. In the extreme case $\frac{\delta}{\sigma_P} = z_{1-\alpha}$ we have

$$\Psi_{\text{TOST}}(\pm\delta) = \alpha - \Phi(-z_{1-\alpha}) = 0.$$

A.2 The BOT is uniformly most powerful

Similar to the TOST the test proposed in Section 2.2 simplifies under the additional assumption of a known variance. As the variance is assumed to be known, the null hypothesis is rejected, whenever,

$$\left|\bar{X}_T - \bar{X}_R\right| \; < u_\alpha, \tag{3.2}$$

where u_α denotes the α-quantile of the folded normal distribution $\mathcal{N}_F(\delta, \sigma_P^2)$. The following result shows that the test defined by (2.8) is the uniformly most powerful test for the hypotheses (2.1). It is well known in the mathematical statistics literature and we present a proof here for the sake of completeness (see also Lehmann and Romano (2006), Romano et al. (2005) or Wellek (2010)).

Theorem A.1. *The test defined by* (2.8) *is the uniformly most powerful (UMP) for the hypotheses* (2.1)*. Moreover, among all tests for the hypotheses* (2.1) *with power function* Ψ *satisfying* $\Psi(\delta) = \Psi(-\delta) = \alpha$ *the test defined by* (2.8) *has also minimal type I error.*

Proof: In order to prove optimality recall (2.5), that is $X = \bar{X}_T - \bar{X}_R \sim \mathcal{N}(d, \sigma_P^2)$, and note that the hypotheses in (2.1) can be rewritten as

$$H_0 : |d| \geq \delta \text{ vs. } H_1 : |d| < \delta \; . \tag{3.3}$$

The test (2.8) rejects the null hypothesis whenever $\left|\bar{X}_R - \bar{X}_T\right| \; < u_\alpha$, where u_α is the quantile of the folded normal distribution with parameters (δ, σ_P^2), which is defined by

$$\alpha \; = \; \mathbb{P}\left(\left|\mathcal{N}(\delta, \sigma_P^2)\right| \leq u_\alpha\right) = \Phi\left(\tfrac{1}{\sigma_P}(u_\alpha - \delta)\right) - \Phi\left(\tfrac{1}{\sigma_P}(-u_\alpha - \delta)\right) \; . \tag{3.4}$$

The probability of rejection is now given by

$$\begin{aligned}
\mathbb{P}_d(\left|\bar{X}_T - \bar{X}_R\right| \; < u_\alpha) \; &= \; \mathbb{P}_d(-u_\alpha < \bar{X}_T - \bar{X}_R < u_\alpha) \\
&= \; \mathbb{P}_d\left(\tfrac{1}{\sigma_P}(-u_\alpha - d) < \tfrac{\bar{X}_T - \bar{X}_R - d}{\sigma_P} < \tfrac{1}{\sigma_P}(u_\alpha - d)\right) \\
&= \; \Phi\left(\tfrac{1}{\sigma_P}(u_\alpha - d)\right) - \Phi\left(\tfrac{1}{\sigma_P}(-u_\alpha - d)\right), \qquad (3.5)
\end{aligned}$$

where Φ denotes the distribution function of the standard-normal distribution.
On the other hand the uniformly most powerful test for the problem (3.3) is well known, see for example Theorem 6 in Section 3.7 of Lehmann and Romano (2006) or Example 1.1 in Romano et al. (2005) This test reject the null hypothesis in (3.3), whenever

$$|\bar{X}_T - \bar{X}_R| < C$$

where the constant $C = C(\alpha, \delta, \sigma_P)$ is the unique solution of the equation

$$\alpha \; = \; \Phi\left(\tfrac{1}{\sigma_P}(C - \delta)\right) - \Phi\left(\tfrac{1}{\sigma_P}(-C - \delta)\right) \tag{3.6}$$

[see Example 1.1 in Romano et al. (2005)]. As the equations (3.4) and (3.6) coincide, it follows that $u_\alpha = C$ and the test (2.8) coincides with the UMP test for the hypotheses (2.1). $\square$

As a consequence of Theorem A.1 the BOT has always more power than the test defined by (2.7). This is indicated in Figure 2, where we display the power of both tests in different scenarios ($\alpha = 0.05$, $\delta = \log(1.25)$). The left panel shows the power curves for $\sigma_P^2 = 0.0049$. In this case the curves basically coincide (although the power of the BOT is slightly larger as stated in Theorem A.1). For increasing variance ($\sigma_P^2 = 0.0144$) it becomes obvious that the power of the BOT is much higher than that for the TOST. This effect becomes even clearer in the right panel ($\sigma_P^2 = \big(\frac{\log(1.25)}{z_{1-\alpha}}\big)^2 \approx 0.14^2$), where the power curve of the TOST is identical to zero. Note that these results are in line with the findings from the simulation study in Section 2.3.

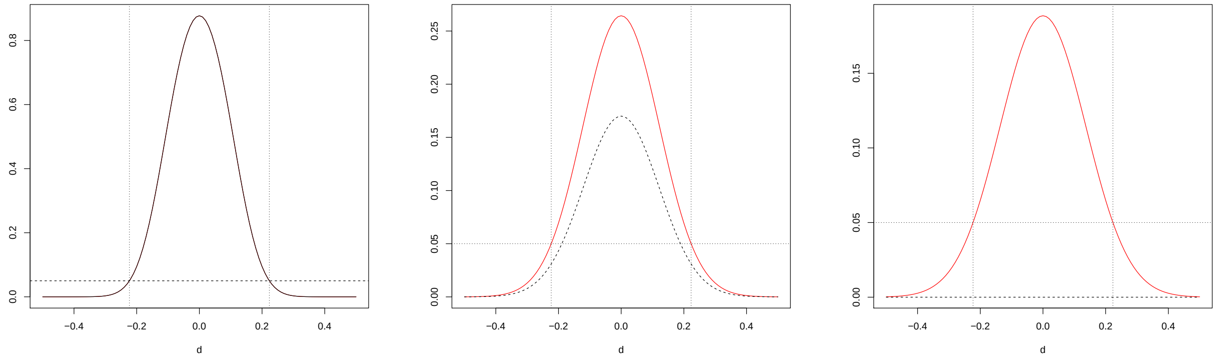

Figure 2: *Power curves of the tests* (2.8) *(solid red line) and the test* (2.7) *(dashed line) for different* $\sigma_P = 0.07$*,* $\sigma_P = 0.12$ *and* $\sigma_P = \frac{\log(1.25)}{z_{1-\alpha}} \approx 0.14$ *(from left to right). The horizontal line indicates the level* $\alpha = 0.05$ *and the vertical lines mark the threshold* $(\pm\delta = \pm log(1.25))$*.*

The proof of Theorem A.1 is based on the fact that for normal distributed data with known variance, that is $X = \bar{X}_T - \bar{X}_R \sim \mathcal{N}(d, \sigma_P^2)$, the form of the uniformly most powerful test is known, see for example Theorem 6 in Section 3.7 of Lehmann and Romano (2006) or Example 1.1 in Romano et al. (2005). Moreover, in the latter paper the author also derives the asymptotic optimal uniformly most powerful test for the hypothesis of bioequivalence in case of unknown variance. Using similar arguments as in this paper it can be shown that the test proposed in Section 2.2 coincides with the asymptotically uniformly most powerful for the hypotheses (2.1). Romano et al. (2005) also show that the TOST is asymptotically optimal. Note that this situation corresponds to the case $\sigma_P^2 = \sigma^2/N_R + \sigma^2/N_T \to 0$ in the present setup, where the power functions of the TOST and the BOT test basically coincide (see the left panel of Figure 2 and the results from the simulation study for the low variance setting in Section 2.3). However, for small or moderate sample sizes the results in this section indicate a superiority of the BOT proposed in Section 2.2, which is confirmed by the numerical results presented in Section 2.3.

We conclude noting that the same arguments also apply for model based methods introduced in Section 3, as the estimates of area under the curve and the maximal concentration follow approximately a normal distribution. Consequently, by the discussion given in the previous paragraphs, the model-based BOT is expected to have always more power than the model-based TOST, the superiority being more sensible in scenarios with larger variability. These theoretical arguments support the empirical results in the simulation study presented in Section 4.