# Finite Sample Adjustments for Average Equivalence Testing

Younes Boulaguiem[1], Julie Quartier[2,3], Maria Lapteva[2,3], Yogeshvar N. Kalia[2,3], Maria-Pia Victoria-Feser[1], Stéphane Guerrier[1,2,3] & Dominique-Laurent Couturier[4,5,*]

[1]Geneva School of Economics and Management, University of Geneva, Switzerland; [2]School of Pharmaceutical Sciences, University of Geneva, Switzerland; [3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Switzerland; [4]Medical Research Council Biostatistics Unit, University of Cambridge, England; [5]Cancer Research UK – Cambridge Institute, University of Cambridge, England; *Corresponding author: dominique.couturier@mrc-bsu.cam.ac.uk

Abstract: The objective of average (bio)equivalence tests is to determine whether a parameter, such as the mean variation in treatment response between two conditions, lies within a specified equivalence range, indicating that the means of the conditions are equivalent. The widely-used *Two One-Sided Tests* (TOST) procedure checks if the target parameter is significantly greater or lower than pre-defined upper and lower equivalence limits by examining whether its confidence interval falls within these limits. However, the TOST procedure can be overly conservative and may quickly lose power for highly variable responses, in many cases reaching a flat zero over the entire parameter space, resulting in its inability to conclude equivalence when it truly exists. To address this limitation, we propose a new procedure called the $\alpha$-TOST that incorporates a finite sample and variability correction by adjusting the size of the TOST to ensure a type I error rate of $\alpha$ in all situations. Our analysis shows that the $\alpha$-TOST is uniformly more powerful, simple to compute, and outperforms its competitors in terms of operating characteristics. We use a case study of econazole nitrate deposition in porcine skin to illustrate the advantages of our approach over other available procedures.

Keywords: similarity test, equivalence, bioequivalence, scaled average bioequivalence, two one-sided test, interval inclusion principle.

Acknowledgements: S. Guerrier is supported by the SNSF Professorships Grant #176843 and by the Innosuisse Grant #53622.1 IP-ENG. Y. Boulaguiem and M.-P. Victoria-Feser are partially supported by the SNSF Grant #182684. D.-L. Couturier is partially supported by the Cancer Research UK grant C9545/A29580. Y. N. Kalia thanks the University of Geneva for teaching assistantships for J. Quartier and M. Lapteva, for providing financial support and for the purchase of the Waters Xevo TQ-MS detector, and also thanks the Fondation Ernst and Lucie Schmidheiny and the Société Académique de Genève for providing equipment grants.

Conflict of Interest: None declared.

# 1. Introduction

Equivalence tests, also known as similarity or parity tests, have become a focal point of research interest over the last two decades. Originating from the field of pharmacokinetics (Metzler, 1974; Westlake, 1976), they are referred to as bioequivalence tests and have found wide-ranging applications in both research and production (Pallmann and Jaki, 2017). In the manufacturing of generic medicinal products, bioequivalence tests are frequently employed to speed up the approval process by demonstrating that the generic version has comparable bioavailability to its brand-name counterpart (for an overview, see e.g., Senn, 2021). Equivalence tests have also found applications in various domains beyond pharmacokinetics and for diverse purposes. In production, they are used to test changes in the mode of administration or production site (see e.g., Patterson and Jones, 2006). In social and behavioral sciences, equivalence tests are employed to assess replication results and to corroborate risky predictions (Lakens, 2017). Recent literature also reflects the ever-expanding use of equivalence tests in brand new domains such as assessing virtual reality imaging measurements by feature (Sureshkumar et al., 2022), cardiovascular responses to stimuli by gender (O'Brien and Kimmerly, 2022), children's neurodevelopment (Wehrle et al., 2022), chemotherapy efficacy and safety (Sansone et al., 2022), post-stroke functional connectivity patterns by patient group (Branscheidt et al., 2022), risk-taking choices by moral type (Feri et al., 2023), and even the turnout of 2020 US presidential election by political advertising condition (Aggarwal et al., 2023). Moreover, several review articles have been published, covering topics such as food sciences (Meyners, 2012), psychology (Lakens et al., 2018), sport sciences (Mazzolari et al., 2022) and pharmaceutical sciences (Wang et al., 2022).

The objective of equivalence testing is to to establish a range of values, known as the equivalence region, within which the parameter of interest, such as the difference in mean response between two treatments, must fall for the treatments to be regarded as equivalent. This ensures that deviations within this region would be considered insignificant and not affect the similarity of the therapeutic effects between the compared treatments. In contrast to standard hypothesis testing for equality of means, where the null hypothesis assumes that both means are equal (or that their difference is zero), and the alternative assumes they are not, equivalence testing reverses the roles of the hypothesis formulations and considers a region rather than a point. Specifically, it defines the alternative as the equivalence region within which the parameter of interest must lie for the treatments to be considered equivalent, and the null hypothesis as the opposite. This paradigm puts the burden of proof on equivalence, rather than non-equality, and emphasises the importance of assessing the similarity of treatments in addition to their differences. Formally, a canonical form for the average equivalence problem involves two independent random variables $\widehat{\theta}$ and $\widehat{\sigma}_\nu$, which are distributed as follows:

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma_\nu^2\right) \quad \text{and} \quad \frac{\nu \widehat{\sigma}_\nu^2}{\sigma_\nu^2} \sim \chi_\nu^2, \tag{1}$$

where $\theta$ and $\sigma_\nu^2$ respectively denote the target equivalence parameter and its variance. The value of $\sigma_\nu^2$ depends on the number of the degrees of freedom $\nu$, which is in turn determined by the sample size and total number of parameters. This general framework can be applied to a variety of situations, including cases where the equivalence parameter corresponds to the difference in means of responses between two experimental conditions with either independent or paired responses. It can also represent an element of the parameter vector of a (generalised) linear (mixed-effect) model, such as the shift in mean responses between two treatments in a longitudinal study.

The hypotheses of interest are defined as

$$H_0: \; \theta \notin \Theta_1, \quad vs. \quad H_1: \; \theta \in \Theta_1 := (\theta_L, \, \theta_U), \tag{2}$$

where $\Theta_1 = (\theta_L, \theta_U)$ denotes the range of equivalence limits with known constants $\theta_L$ and $\theta_U$. Is can be assumed, without loss of generality, that the equivalence limits are symmetrical around zero, and we define $c := \theta_U = -\theta_L$ so that $\Theta_1 = (-c, \, c)$. The investigation of equivalence is commonly carried out using the *Two One-Sided Tests* (TOST) procedure (Schuirmann, 1987), which assesses whether the target parameter is significantly greater than $-c$ and less than $c$, with the Type I Error Rate (TIER) controlled at level $\alpha$ (see Berger, 1982), and typically set at 5%. Equivalence is most commonly assessed through the *Interval Inclusion Principle* (IIP), which involves verifying whether the $100(1-2\alpha)\%$ Confidence Interval (CI) for the target parameter falls within the equivalence margins $(-c, \, c)$ (see, for example, Muñoz et al., 2016; Pallmann and Jaki, 2017). The IIP strategy yields the same test decision as the TOST procedure if the CI is equi-tailed (Hsu et al., 1994; Berger and Hsu, 1996)

However, it is widely acknowledged that the TOST procedure can be overly conservative, leading to a decrease in power and a concomitant reduction in the probability of detecting truly equivalent mean effects. This issue is particularly noticeable when the standard deviation $\sigma_\nu$ is relatively large, as in studies involving highly variable drugs or where the sample size has been underestimated based on a prior experiment. To address this challenge, the AH-test proposed by Anderson and Hauck (1983) has been shown to exhibit greater power than the TOST in cases where $\sigma_\nu$ is relatively large. However, this test is known to be liberal and does not control the TIER (see Berger and Hsu, 1996). Furthermore, it may lead to acceptance of equivalence even when the target parameter $\theta$ falls outside the equivalence interval (see Schuirmann, 1987). Brown et al. (1997) then proceeded by developing an unbiased test that is uniformly more powerful than the TOST. However, this test is computationally intensive and, in some cases, may exhibit rather irregular shapes in its rejection region. To mitigate these issues, Berger and Hsu (1996) proposed a smoothed version of this test at the cost of making it slightly biased. Notably, these three last tests do not require the calculation of confidence intervals and therefore do not rely on the IIP, and the latter two can be difficult to interpret due to the use of polar coordinates (see e.g., Liu and Chow, 1996).

The conservative nature of the TOST approach in determining bioequivalence between highly variable drugs, characterized by relatively large standard deviation $\sigma_\nu$, has prompted regulatory bodies such as the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) to recommend the adoption of an alternative strategy known as Scaled Average BioEquivalence (SABE). This method involves the linear adjustment of equivalence limits based on the value of $\sigma_\nu$ within the reference group, while still requiring that $\widehat{\theta}$ falls within the equivalence margins $(-c, \, c)$. Despite the authorities' constraints on the degree of expansion, recent studies have revealed that the TOST can exceed the nominal level $\alpha$ (see, for example, Wonnemann et al., 2015; Muñoz et al., 2016; Endrenyi and Tothfalusi, 2019; Molins et al., 2021; Schütz et al., 2022, and references therein), leading to proposals for correction methods that ensure a level-$\alpha$ test. Moreover, these corrections also result in more seamless changes to the acceptance regions as $\sigma_\nu$ varies. Additional details on the SABE can be found in Muñoz et al. (2016); Davit et al. (2012), while

Labes and Schütz (2016); Tothfalusi and Endrenyi (2016); Ocaña and Muñoz (2019); Deng and Zhou (2020) provide further insights into these correction methods.

In this paper, we present a novel finite sample correction method for the TOST as an alternative to existing approaches. Our proposed method involves a simple adjustment of the TOST's level, which ensures a TIER of exactly $\alpha$ when $\sigma_\nu$ is known. The *corrected level*, denoted as $\alpha^*$, can be easily computed and enables the construction of $100(1-2\alpha^*)\%$ CIs. As a result, the $\alpha$-TOST can be viewed as a finite sample continuous variability adjustment of the TOST, which enhances the probability of accepting equivalence when it is true, particularly for large values of $\sigma_\nu$ or small sample sizes.

Our study shows that the $\alpha$-TOST is uniformly more powerful than the TOST and, for small to moderate values of $\sigma_\nu$, is nearly equivalent to the TOST in terms of power, as $\alpha^* \approx \alpha$ in such cases. Furthermore, our proposed method provides a more powerful test than the level-corrected SABE.

It is worth noting that, in practice, $\sigma_\nu$ is typically estimated from the data. Thus, we also propose a straightforward estimator for $\alpha^*$ and demonstrate through simulations that this estimator does not significantly alter the TIER of the $\alpha$-TOST. Overall, this paper presents an efficient method for performing equivalence testing, which can be readily applied in various scientific domains.

The present paper is structured as follows. In Section 2, we showcase the effectiveness of our proposed approach by applying the TOST, the $\alpha$-TOST, and other existing methods to a case study. The $\alpha$-TOST is then introduced in Section 3, where we provide a suitable formulation of the TOST, describe its statistical properties, and present a straightforward algorithm to compute the corrected level $\alpha^*$. Additionally, we provide a power comparison with alternative methods. Next, in Section 4, we conduct a simulation study to compare the TIER and power of the $\alpha$-TOST to those of the TOST, the EMA implementation of the SABE, as well as to its corrected level version. Finally, in Section 5, we discuss potential extensions of our proposed method.

## 2. Evaluation of Bioequivalence for Econazole Nitrate Deposition in Porcine Skin

Quartier et al. (2019) study the cutaneous bioequivalence of two topical cream products: a reference medicinal product and an approved generic containing econazole nitrate (ECZ), an antifungal medication used to treat skin infections. The evaluation of the putative bioequivalence is based on the determination of the cutaneous biodistribution profile of ECZ observed after application of the reference medicinal product and the generic product. This involves quantification of the amounts of ECZ present as a function of depth within the skin descending from the surface to a depth of $\sim 800$ microns. It follows that the more similar the biodistribution profile, i.e., the more similar the amounts of drug present at each depth, the greater the likelihood of equivalent pharmacological effect.

In this section, we examine a dataset consisting of 17 pairs of comparable porcine skin samples, on which measurements of ECZ deposition were collected using the two creams. The data, depicted in Figure 1, were analyzed using the TOST and the $\alpha$-TOST procedures, both based on a paired $t$-test statistic. The bioequivalence limits of $c = \theta_U = -\theta_L = \log(1.25) \approx 0.223$, recommended by regulatory agencies such as the FDA (Food and Drugs Administration, 2001) and EMA (European Medicine Agency, 2010), were used for both procedures.

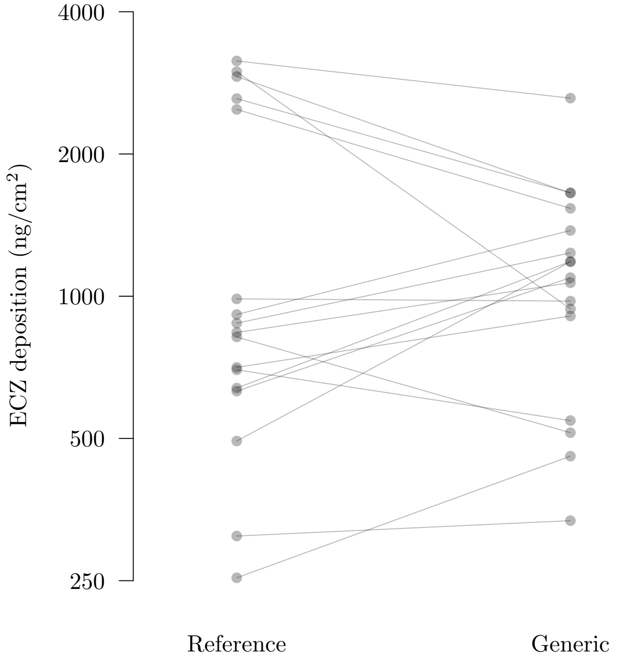

Figure 1: Econazole nitrate deposition levels (y-axis) measured for the reference and generic creams (x-axis) on 17 pairs of comparable porcine skin samples (lines).

The TOST confidence interval for the mean of the paired differences in ECZ levels was found to be $[-0.204, 0.250]$, with an estimated effect size of $\widehat{\theta} = 0.023$, an estimated standard error of $\widehat{\sigma}_\nu = 0.130$, a number of degrees of freedom of $\nu = 16$, and a significance level of $\alpha = 5\%$. As the upper bound of this CI exceeds the upper bioequivalence limit, we cannot conclude that the two topical products are equivalent using the standard TOST procedure. Remarkably, by computing a higher significant rate of $\widehat{\alpha}^* = 7.48\%$ to achieve an (empirical) TIER of 5%, the $\alpha$-TOST procedure produces a confidence interval of $[-0.166, 0.211]$ that is strictly embedded within the $(-c, c)$ bioequivalence limits. This indicates that the $\alpha$-TOST accepts bioequivalence, and yields an increase in power that is induced by the increased TIER. Note, on the other hand, that the empirical power of the TOST is equal to zero, regardless of the value $\widehat{\theta}$, because $t_{0.05,16}\ \widehat{\sigma}^2\nu > c$, with $t_{\alpha,\nu}$ denoting the upper quantile of a $t$-distribution evaluated at $\alpha$ with $\nu$ degrees of freedom (see Section 3.3 for a more detailed power analysis). Furthermore, it is important to highlight that the $\alpha$-TOST guarantees an adjusted TIER of $\alpha$ for all sample sizes, as demonstrated in Section 3.2. Thus, the conclusion drawn by the $\alpha$-TOST is more trustworthy than that of the standard TOST. The confidence intervals obtained from both methods are presented in Figure 2.

In order to gain a more comprehensive understanding of the advantages afforded by our proposed method, we have conducted a comparison of the characteristics and outcomes of the $\alpha$-TOST with other available approaches, as presented in Table 1. Additionally, we provided a sketch of their rejection regions in Figure 3. In addition to the TOST and $\alpha$-TOST, the analysed methods include the AH-test, the SABE implementation using the EMA margins (SABE), and the level-corrected version proposed by Labes and Schütz (2016) (SABE corrected). While the AH-test fails to meet the IIP, it serves as a decent proxy for the other tests that satisfy this property, while remaining relatively straightforward to implement. And even though the SABE test does not conform to the $\alpha$-level requirement, it is

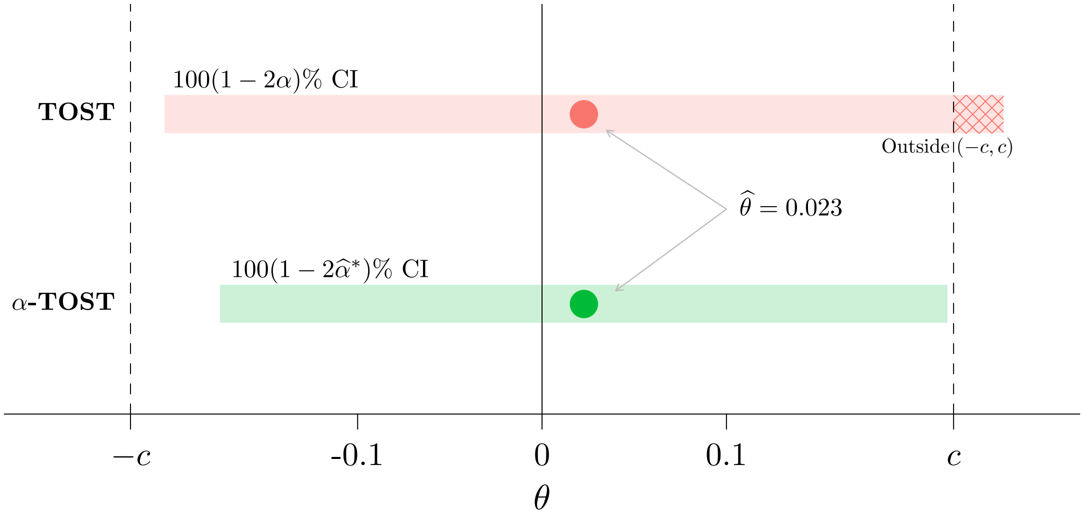

Figure 2: $100(1-2\alpha)$ and $100(1-2\widehat{\alpha}^*)$ confidence intervals resulting from the TOST and $\alpha$-TOST procedures, respectively, for the mean of the paired differences in ECZ levels of the reference and generic creams. The chosen and computed values for $\alpha$ and $\widehat{\alpha}^*$ are 5% and 7.48%, respectively. The dashed vertical lines depict the lower and upper bioequivalence limits, which are set at $c = \log(1.25)$. By comparing the confidence intervals of each approach to the bioequivalence limits, we find that the $\alpha-$TOST procedure indicates bioequivalence, while the standard TOST approach does not. Specifically, the upper limit of the TOST confidence interval exceeds $c$, as illustrated by the hatched area.

included in our comparison to contrast its performance with that of the level-corrected SABE. It is noteworthy that, out of the level-$\alpha$ tests, only the $\alpha$-TOST leads to an acceptance outcome.

Figure 3 depicts the regions in which bioequivalence is accepted for as a function of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$, using $c = \log(1.25)$ and $\nu = 16$ as represented in the case-study. The rejection regions of the different methods closely overlap for $\widehat{\sigma}\nu$ below 0.09, but their operating characteristics vary as this value increases. The TOST and the level-corrected SABE methods exhibit a conservative rejection region that fails to accept bioequivalence for values of $\widehat{\sigma}\nu$ above $\approx .15$, while the $\alpha$-TOST and AH-test maintain their stringent criteria for bioequivalence acceptance even for larger values of $\widehat{\sigma}\nu$, with the latter method being more liberal than the former.

Interestingly, the SABE method exhibits an anomalous rejection area that yields paradoxical outcomes. Specifically, for a range of values of $\widehat{\theta}$, non-equivalence is rejected for larger values of $\widehat{\sigma}_\nu$ and not rejected for smaller ones. This peculiar outcome is not unique to the SABE method but has also been reported for other methods, as noted in the literature (see, for example, Figures 1 of Berger and Hsu, 1996, Brown et al., 1997 and Cao and Mathew, 2008). The $\times$ symbol on Figure 3 represents the estimated values of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ obtained from the porcine skin dataset. Notably, these coordinates result in the acceptance of bioequivalence for the SABE, AH-test, and $\alpha$-TOST methods, among which only the latter maintains a size-$\alpha$ criterion (i.e., a TIER of $\alpha$). Overall, these findings underscore the superiority of the $\alpha$-TOST method for a reliable assessment of equivalence.

A more comprehensive power analysis of the selected methods is provided in Section 3.3, and a simulation study comparing their TIER and finite sample power is presented in Section 4.

| Method | IIP | Level-$\alpha$ | TIER $\alpha$ | Decision |
|---|---|---|---|---|
| AH-test | no | no | no | Accept |
| SABE | yes | no | no | Accept |
| SABE corrected | yes | yes* | no | Reject |
| TOST | yes | yes | no | Reject |
| $\alpha$-TOST | yes | yes* | yes* | Accept |

Table 1: Bioequivalence decision for the econazole nitrate deposition in the porcine skin dataset using the AH-test, the SABE implementation using the EMA margins (SABE), the level-corrected version proposed by Labes and Schütz (2016) (SABE corrected), the TOST and $\alpha$-TOST. The decision is based on the values of $\widehat{\sigma}_\nu = 0.130$, $\nu = 16$, $\widehat{\theta} = 0.0227$, and $\alpha = 5\%$. The columns indicate whether each method satisfies the Interval Inclusion Principle (IIP), is level-$\alpha$ (a TIER bounded by $\alpha$), and size-$\alpha$ (a TIER of exactly $\alpha$). The * symbol indicates that the property is valid only when the standard error $\sigma_\nu$ is known.

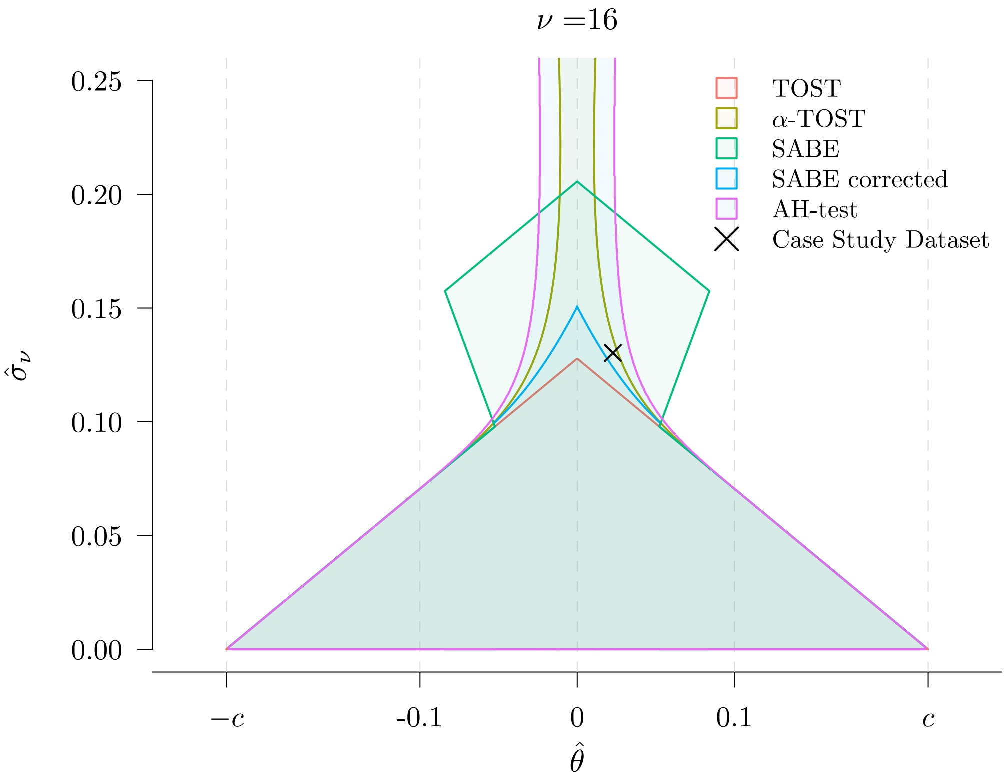

Figure 3: Bioequivalence test rejection regions as a function of the estimated effect size $\widehat{\theta}$ (x-axis) and estimated standard error $\widehat{\sigma}_\nu$ (y-axis) per method considered in Table 1. The coloured areas represent the rejection regions using $c = \log(1.25)$ and $\nu = 16$ for each method. The $\times$ symbol represents the estimated values of the effect size and standard error obtained from the porcine skin dataset, namely $\widehat{\theta} = 0.023$ and $\widehat{\sigma}_\nu = 0.130$.

## 3. Equivalence Testing

In this section, we delve into the methodology for the derivation of a size-adjusted statistical equivalence test. We begin by introducing the TOST and examine its properties. Next, we define the $\alpha$-TOST and detail its statistical properties. We then propose an algorithm with an exponential convergence rate to compute the corrected significance level $\alpha^*$. Finally, we show that the $\alpha$-TOST is uniformly more powerful than both the TOST and the level-adjusted variation of the SABE.

### 3.1. The TOST Procedure

The TOST employs two test statistics to evaluate the hypotheses in (2), namely

$$Z_L := \frac{\widehat{\theta} + c}{\widehat{\sigma}_\nu} \quad \text{and} \quad Z_U := \frac{\widehat{\theta} - c}{\widehat{\sigma}_\nu},$$

where $Z_L$ tests for $\mathrm{H}_{01} : \theta \leqslant -c$ versus $\mathrm{H}_{11} : \theta > -c$, while $Z_U$ tests for $\mathrm{H}_{02} : \theta \geqslant c$ versus $\mathrm{H}_{12} : \theta < c$. Rejecting $\mathrm{H0} := \mathrm{H01} \cup \mathrm{H02} = \theta \notin \Theta_1$ in favour of $\mathrm{H1} := \mathrm{H11} \cap \mathrm{H12} = \theta \in \Theta_1$ requires both tests to simultaneously reject their marginal null hypotheses at a significance level $\alpha$, that is, if

$$Z_L \geqslant t_{\alpha,\nu} \quad \text{and} \quad Z_U \leqslant -t_{\alpha,\nu},$$

where $t_{\alpha,\nu}$ denotes the upper $\alpha$ quantile of a $t$-distribution with $\nu$ degrees of freedom. The corresponding rejection region is given by

$$C_1\left(\widehat{\sigma}_\nu\right) := \left\{\widehat{\theta} \in \mathrm{I\!R},\, \widehat{\sigma}_\nu \in \mathrm{I\!R}_+ \,\middle|\, c \geqslant |\widehat{\theta}| + t_{\alpha,\nu}\widehat{\sigma}_\nu\right\}. \tag{3}$$

As illustrated in Figure 3, for all $\widehat{\sigma}_\nu > \widehat{\sigma}_{\max} := c/t_{\alpha,\nu}$, the TOST fails to accept equivalence.

More formally, given $\alpha$, $\theta$, $\sigma_\nu$, $\nu$ and $c$, the power function, i.e., the probability that equivalence is accepted, can be expressed as follows (see also Phillips, 1990):

$$p(\alpha, \theta, \sigma_\nu, \nu, c) := \Pr\left(Z_L \geqslant t_{\alpha,\nu} \ \text{ and } Z_U \leqslant -t_{\alpha,\nu}, \,\middle|\, \alpha, \theta, \sigma_\nu, \nu, c\right). \tag{4}$$

The vector $(T_L, T_U)$ has a bivariate non-central $t$-distribution, with non-centrality parameters $\frac{\theta - c}{\sigma_\nu}$ and $\frac{\theta + c}{\sigma_\nu}$ respectively. This particular case has been studied by Owen (1968), which developed a very convenient way to express the bivariate distribution in terms of the difference of two univariate distributions, namely

$$\begin{aligned} p(\alpha, \theta, \sigma_\nu, \nu, c) :=& \Pr\left(T_L \geqslant t_{1-\alpha,\nu} \ \text{ and } T_U \leqslant -t_{1-\alpha,\nu}, \,\middle|\, \alpha, \theta, \sigma_\nu, \nu, c\right) \\ =& Q_\nu\left(-t_{1-\alpha,\nu}, \frac{\theta - c}{\sigma_\nu}, \lambda\right) - Q_\nu\left(t_{1-\alpha,\nu}, \frac{\theta + c}{\sigma_\nu}, \lambda\right), \end{aligned} \tag{5}$$

where $\lambda := \frac{c\sqrt{\nu}}{\sigma_\nu t_{\alpha,\nu}}$, and where

$$Q_\nu(t,\, y,\, z) := \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^z G\left(\frac{tx}{\sqrt{\nu}} - y\right) x^{\nu-1} G'(x)dx,$$

with

$$G'(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{and} \quad G(x) := \int_{-\infty}^x G'(t)dt.$$

Subsequently, after fixing the values of $\alpha$, $\sigma_\nu$ and $\nu$, the size of the TOST is defined as the supremum of (5) (see e.g., Lehmann, 1986), which can be expressed as

$$\omega(\alpha, c, \sigma_\nu, \nu) := \sup_{\theta \notin \Theta_1} p(\alpha, \theta, \sigma_\nu, \nu, c) = p(\alpha, c, \sigma_\nu, \nu, c) = Q_\nu\left(-t_{\alpha,\nu},\, 0,\, \lambda\right) - Q_\nu\left(t_{\alpha,\nu},\, \frac{2c}{\sigma_\nu},\, \lambda\right). \tag{6}$$

We can deduce that the TOST is level-$\alpha$, by observing that when $\sigma_\nu > 0$, the following holds

$$\begin{aligned} \omega(\alpha, c, \sigma_\nu, \nu) < Q_\nu\left(-t_{\alpha,\nu},\, 0,\, \lambda\right) &= \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^\lambda G\left(-\frac{t_{\alpha,\nu}x}{\sqrt{\nu}}\right) x^{\nu-1} G'(x)dx \\ &< \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^\infty G\left(-\frac{t_{\alpha,\nu}x}{\sqrt{\nu}}\right) x^{\nu-1} G'(x)dx = \Pr\left(T_\nu \leqslant -t_{\alpha,\nu}\right) = \alpha, \end{aligned} \tag{7}$$

where $T_\nu$ denotes a random variable following a $t$-distribution with $\nu$ degrees of freedom. Conversely,

$$\lim_{\sigma_\nu \to 0} \omega(\alpha, c, \sigma_\nu, \nu) = \alpha.$$

This demonstrates that the TOST is a level-$\alpha$ test, and is only size-$\alpha$ in the theoretical case of $\sigma_\nu = 0$, a fact previously noted by several researchers (see, for example, Deng and Zhou, 2020, and the references therein). When $\widehat{\sigma}_\nu$ is small, the TOST effectively controls the TIER at values near $\alpha$. However, as $\widehat{\sigma}_\nu$ increases, the TOST becomes progressively more conservative and ultimately yields a TIER of zero when $\widehat{\sigma}_\nu > c/t_{\alpha,\nu}$. As a solution to this limitation, we propose a novel approach that we call the $\alpha$-TOST, which ensures a uniform control of the TIER at $\alpha$ over the pace of $\widehat{\sigma}_\nu$, and retains the ability to assess equivalence through the IIP principle, as illustrated in Figure 2.

## 3.2. Adjusted $\alpha$-TOST

A size-corrected version of the TOST can be constructed through the computation of new significance level $\alpha^*$ with the following paradigm

$$\alpha^* := \alpha^*(\sigma_\nu) = \underset{\gamma \in [\alpha, 0.5)}{\text{argzero}} \left[\omega(\gamma, c, \sigma_\nu, \nu) - \alpha\right], \tag{8}$$

with $\omega(\gamma, c, \sigma_\nu, \nu)$ defined in (6). We omit the dependence of $\alpha^*$ on $\alpha$ and $\nu$ as these are known. A similar type of correction was used to amend the level of the SABE procedure by Labes and Schütz (2016) and Ocaña and Muñoz (2019) (see also Palmes et al., 2022, for power adjustment), but in these cases, the aim of the correction is to decrease the level so that the TIER does not exceed the nominal level of $\alpha$. This is different from our objective, as our proposed method aims to ensure that the TIER of the $\alpha$-TOST is exactly $\alpha$ when $\sigma_\nu$ is known.

The condition under which $\alpha^*$ exists are elaborated in Appendix A. It relates to the maximal value the estimated standard error can take for the $\alpha$-TOST procedure to produce a unique solution, namely $\widehat{\sigma}_\nu < \frac{2c}{\Phi^{-1}(\alpha+0.5)}$. As outlined in Section 3.3, this upper limit permits the $\alpha$-TOST to achieve a strictly positive power for the larger values of $\sigma_\nu$, which sets it apart from other methods that satisfy the IIP.

Given that $\alpha^*(\sigma_\nu)$ is a population parameter that depends on the unknown quantity $\sigma_\nu$, an appropriate estimator for its sample value is given by

$$\widehat{\alpha}^* := \alpha^*(\widehat{\sigma}_\nu) = \underset{\gamma\in[\alpha,0.5)}{\text{argzero}} \left[\omega(\gamma, c, \widehat{\sigma}_\nu, \nu) - \alpha\right]. \tag{9}$$

Thus, in practice, the $\alpha$-TOST method rejects the null hypothesis of non-equivalence in favor of equivalence at level $\alpha$, based on the estimated corrected level $\widehat{\alpha}^*$, if both $Z_L > t_{1-\widehat{\alpha}^*,\nu}$ and $Z_U < -t_{1-\widehat{\alpha}^*,\nu}$ hold. We investigate the performance of various methods in a simulation study presented in Section 4 where $\sigma_\nu$ is estimated, and we compare their TIER and power. Our results demonstrate that the $\alpha$-TOST method maintains a relatively good TIER compared to other available methods.

When it comes to solving (8) (or indeed (9)), $\alpha^*$ can easily be computed using the following iterative approach. At iteration $k$, with $k \in \mathbb{N}$, we define

$$\alpha^{*(k+1)} = \alpha + \alpha^{*(k)} - \omega\left(\alpha^{*(k)}, c, \widehat{\sigma}_\nu, \nu\right). \tag{10}$$

Here, $\omega(\cdots)$ is the function defined in (6), and the procedure is initialised at $\alpha^{*(0)} = \alpha$. This simple iterative approach has exponentially fast convergence to $\alpha^*$ as it can be shown that

$$\left|\alpha^{*(k+1)} - \alpha^*\right| < \frac{1}{2}\exp(-bk),$$

for some positive constant $b$ (see Appendix B for more details).

## 3.3. Power Analysis

In comparison to the standard TOST method, the $\alpha$-TOST approach evaluates the acceptance of equivalence using an interval that is computed based on a smaller value of $t_{\alpha,\nu}$, since the corrected significance level $\alpha^*$ is greater than the nominal level $\alpha$. This results in a rejection region of the $\alpha$-TOST that is necessarily wider than its TOST counterpart, which means that the $\alpha$-TOST is uniformly more powerful than the TOST, and explains cases such as the one presented in the porcine skin case study presented in Section 2. Indeed, the $\alpha$-TOST approach accepts equivalence due to the shrinked CI, whereas the TOST fails to do so (yielding an empirical power of zero given the sample value of $\widehat{\sigma}_\nu = 0.13$).

Figure 4 presents a comparison of the rejection regions of the TOST, the $\alpha$-TOST, and the level-corrected EMA implementation of the SABE by Labes and Schütz (2016) (SABE corrected) for different values of $\nu$ ranging from $\nu = 12$ to $\nu = 36$. The level-corrected SABE approach was chosen as it is level-$\alpha$ and respects the IIP. It is worth mentioning that the implementation and correction of the FDA procedure would lead to the same conclusions (see e.g., Muñoz et al., 2016). The rejection regions of the TOST and level-corrected SABE approaches are embedded in the $\alpha$-TOST region for all values of $\nu$ and $\widehat{\sigma}_\nu$ considered in the study. This indicates that our approach is uniformly

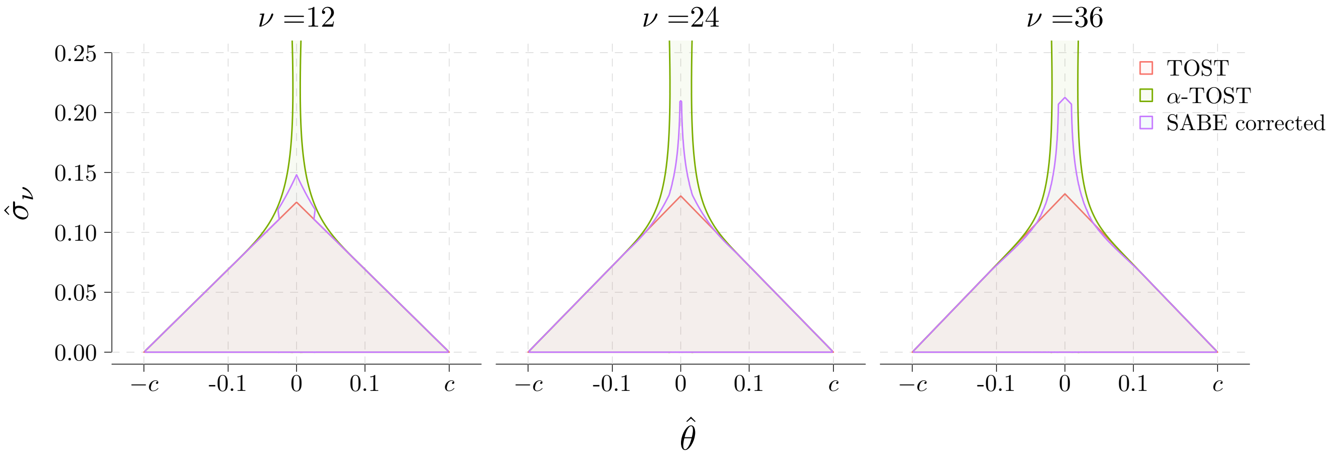

Figure 4: Rejection regions of three different methods for conducting equivalence tests: the TOST method (in red), the $\alpha$-TOST approach (in green), and the level-corrected EMA implementation of the SABE by Labes and Schütz (2016) (in purple). The plots show how the rejection regions vary as a function of $\widehat{\theta}$ (x-axis) and $\widehat{\sigma}_\nu$ (y-axis), for different values of $\nu$ ranging from $\nu = 12$ to $\nu = 36$ (columns). The values $c = \log(1.25)$ and $\alpha = 5\%$ are used throughout. The rejection regions of the TOST and the level-corrected SABE methods are contained within the rejection region of the $\alpha$-TOST, indicating that the later is more powerful.

more powerful than both the TOST and the level-corrected SABE. As the values of $\nu$ increase, the difference in power between the $\alpha$-TOST and the level-corrected SABE decreases. However, the $\alpha$-TOST still maintains a strictly positive power for larger values of $\widehat{\sigma}_\nu$, up to approximately $\frac{2c}{\Phi^{-1}(\alpha+0.5)} \approx 3.55$ (see Appendix A). It is important to note that allowing a testing procedure to accept for equivalence when the standard error is allowed to take very large values may be debatable. Ultimately, the determination of an acceptable range of standard errors goes beyond statistical considerations and requires the input of practitioners and regulatory authorities.

Finally, the concept of adjusting the critical values to correct the confidence interval length is not new. In fact, Cao and Mathew (2008) propose an adjustment that varies with the standard error and is obtained through linear interpolation, resulting in adjusted critical values that change depending on the values of $\widehat{\sigma}_\nu$ (as presented in Table 1 of the same paper). In the lower plot of Figure 7 of appendix C, we compare the critical values obtained with the method of Cao and Mathew (2008) to those obtained by the $\alpha$-TOST for varying values of $\widehat{\sigma}_\nu$ and $\nu$. It is worth noting that, for all values of $\nu$ examined in this study, and for $\widehat{\sigma}_\nu$ greater than 0.1, the critical values proposed by Cao and Mathew (2008) correspond to a piece-wise version of the critical values of the $\alpha$-TOST, which are derived under large $\nu$. Thus, their approach appears to be an approximation of the $\alpha$-TOST evaluated asymptotically, i.e., when $\nu$ approaches infinity.

## 4. Simulation Study: Comparing the TIER and Power of Different Methods

In Section 3.3, we analysed the TIER and power of different procedures assuming known values of $\sigma_\nu$ and $\theta$. In practice however, these must be estimated. To assess the operating characteristics of the procedures when the values are estimated, we conducted a Monte Carlo simulation study. We considered the same methods as in Section 3.3 and also included the non-corrected SABE method to demonstrate that its TIER can exceed the nominal level $\alpha$. For a given set of $\theta$, $\nu$, $\sigma_\nu$, and with $c = \log(1.25)$ and $\alpha = 5\%$, we evaluated the TIER and power of the $\alpha$-TOST using the following steps for each Monte Carlo sample $b = 1, \dots, B$:

1. simulate two values for $\widehat{\theta}$, namely $\widehat{\theta}_{b,0}$ and $\widehat{\theta}_{b,c}$, using (1), with $\theta = 0$ for the power and $\theta = c$ for the TIER,
2. simulate one value for $t = \frac{\nu\widehat{\sigma}_\nu^2}{\sigma_\nu^2}$ using (1) and set $\widehat{\sigma}_{\nu,b} = \sqrt{t \cdot \sigma_\nu^2/\nu}$,
3. compute $\widehat{\alpha}^*$ using (the algorithm associated to) (9),
4. compute $t_{\widehat{\alpha}*,\nu,b}$.

The finite sample TIER is then obtained as

$$\frac{1}{B}\sum_{b=1}^{B} \eta\left(c \geqslant |\widehat{\theta}_{b,c}| + t_{\widehat{\alpha}*,\nu,b}\widehat{\sigma}_{\nu,b}\right)$$

and the finite sample power as

$$\frac{1}{B}\sum_{b=1}^{B} \eta\left(c \geqslant |\widehat{\theta}_{b,0}| + t_{\widehat{\alpha}*,\nu,b}\widehat{\sigma}_{\nu,b}\right),$$

where $\eta(\cdot)$ is the indicator function with $\eta(A) = 1$ if $A$ is true and 0 otherwise. For the other methods, Steps 1 and 2 are the same, and the following ones are adapted to the specific methods. The parameters' values for the Monte Carlo simulation are presented below.

**Simulation 1:** Parameters' values for the simulation study.

- $c = \log(1.25)$,
- $\nu = 20, 30, 40$,
- $\sigma_\nu = 0.01, 0.012, \dots, 0.6$,
- $\alpha = 5\%$,
- $B = 10^4$, the number of Monte Carlo simulations per parameters combination.

Figure 5 shows the Monte Carlo finite sample TIER estimates as a function of $\sigma_\nu$ and $\nu$ for the different methods of interest. As expected, the SABE procedure can generate a TIER significantly exceeding the nominal level of $\alpha = 5\%$, while the other three methods ensure a TIER of at most 5%. Under all scenarios considered here, the $\alpha$-TOST maintains the highest non-liberal TIER, especially when the number of degrees of freedom is large. For both the TOST and the level-corrected SABE approaches, the TIER reaches zero for relatively small values of $\sigma_\nu$ (around 0.274).

The behaviour of the TIER and the power are closely linked as depicted by Figure 6. This figure displays the Monte Carlo finite sample power estimates as a function of $\sigma_\nu$ and $\nu$ for the different methods of interest. Indeed, for the TOST and level-corrected SABE, the power is zero when the TIER is zero, which occurs at values of $\sigma_\nu$ approximating 0.274. The seemingly large power of the SABE is a consequence of its TIER being well above the nominal level $\alpha$. As

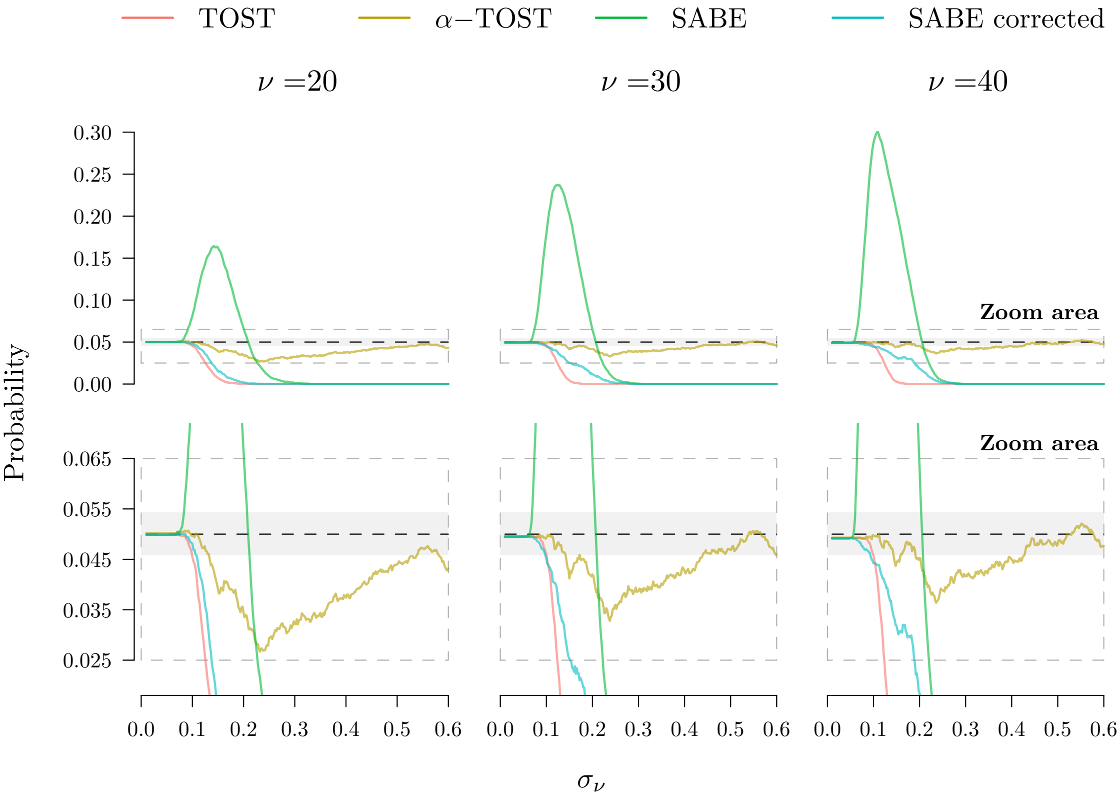

Figure 5: Finite sample TIER (y-axes) computed using the simulation setting in Simulation 1 as a function of $\sigma_\nu$ (x-axes), and $\nu$ (columns) for the different methods of interest. Top panels: complete results. Bottom panels: TIER truncated outside $(0.025; 0.065)$. The shaded gray areas correspond to the pointwise 95% Monte Carlo tolerance band. We can note that the TIER of the TOST and the level-corrected SABE approaches tend to be overly conservative for values of $\sigma_\nu > 0.1$, with a slight advantage of the level-corrected SABE approach compared to the TOST for values of $\sigma_\nu$ below 0.2, especially when $\nu$ is large. As expected, the original SABE approach is overly liberal for certain values of $\sigma_\nu$. Under all scenarios considered here, the $\alpha$-TOST maintains the highest non-liberal TIER.

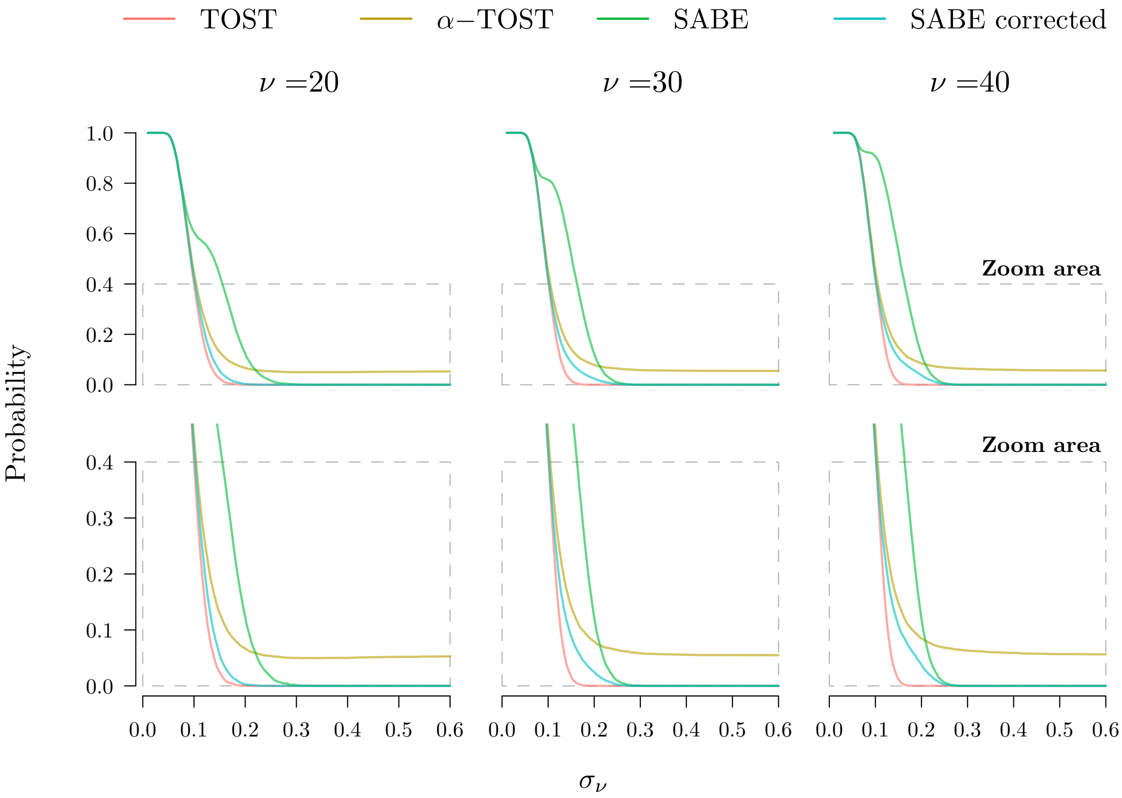

Figure 6: Finite sample power (y-axes) computed using the simulation setting in Simulation 1 as a function of $\sigma_\nu$ (x-axes), and $\nu$ (columns) for the different methods of interest. Top panels: complete results. Bottom panels: power truncated above 0.4. The power of the $\alpha$-TOST not only remains above the one of its competitors that exhibit a non-liberal TIER under all configurations of Simulation 1, but it is also greater than zero for all values of $\sigma_\nu$ considered here.

the latter decreases with larger values of $\sigma_\nu$, its power is dominated by that of the $\alpha$-TOST. The power of the $\alpha$-TOST not only remains higher than its competitors that exhibit a non-liberal TIER under all configurations of Simulation 1, but it is also greater than zero for all values of $\sigma_\nu$ considered here.

## 5. Discussion

The canonical framework considered in this paper, as described in (1), assumes normally distributed differences (in finite samples) with a known distribution for $\widehat{\sigma}_\nu$. This encompasses a broad range of data settings, including the common two-period crossover design (see, e.g., Jones and Kenward, 2014), and has the potential to be extended to incorporate covariates to potentially reduce residual variance. Extensions to non-linear cases such as binary responses (see, e.g., Dunnett and Gent, 1977; Tu, 1998; Schouten and Kester, 2010; Lui and Chang, 2011; Ostrovski, 2022, and the references therein) would follow the same logic, but would require specific treatment due to the nature of the responses and the associated function that is used (e.g., ratio of proportions, odds ratios). These extensions also deserve some attention but are left for further research.

Regarding sample size calculation, one could proceed with the $\alpha$-TOST, i.e., using the value of $\alpha^*$ obtained in (8) for given values of $c$, $\theta$ and $\sigma_\nu$. However, when considering high levels of power, as shown in Section 4, the correction becomes negligible as $\alpha^* \approx \alpha$. Thus, sample size calculations can be done using the standard TOST, as implemented in standard packages.

## 6. Bibliography

Aggarwal, M., J. Allen, A. Coppock, D. Frankowski, S. Messing, K. Zhang, J. Barnes, A. Beasley, H. Hantman, and S. Zheng (2023). A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. Nature Human Behaviour.

Anderson, S. and W. W. Hauck (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics A 12, 2663–2692.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics 24(4), 295–300.

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

Branscheidt, M., N. Ejaz, J. Xu, M. Widmer, M. D. Harran, J. C. Cortés, T. Kitago, P. Celnik, C. Hernandez-Castillo, J. Diedrichsen, A. Luft, and J. W. Krakauer (2022). No evidence for motor-recovery-related cortical connectivity changes after stroke using resting-state fmri. Journal of Neurophysiology 127, 637–650.

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

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

Davit, B. M., M. L. Chen, D. P. Conner, S. H. Haidar, S. Kim, C. H. Lee, R. A. Lionberger, F. T. Makhlouf, P. E. Nwakama, D. T. Patel, D. J. Schuirmann, and L. X. Yu (2012). Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US Food and Drug Administration. American Association of Pharmaceutical Scientists Journal 14, 915–924.

Deng, Y. and X.-H. Zhou (2020). Methods to control the empirical type I error rate in average bioequivalence tests for highly variable drugs. Statistical Methods in Medical Research 29, 1650–1667.

Dunnett, C. W. and M. Gent (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. Biometrics 33, 593–602.

Endrenyi, L. and L. Tothfalusi (2019). Bioequivalence for highly variable drugs: regulatory agreements, disagreements, and harmonization. Journal of Pharmacokinetics Pharmacodynamics 46, 117–126.

European Medicine Agency (2010). Guideline on the investigation of bioequivalence-cpmp. Technical report, EWP/QWP/1401/98 Rev. 1.

Federer, H. (2014). Geometric Measure Theory. Springer.

Feri, F., C. Giannetti, and P. Guarnieri (2023). A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. Journal of Behavioral and Experimental Finance 37.

Food and Drugs Administration (2001). Guidance for industry, statistical approaches to establishing bioequivalence. http://www.fda.gov/cder/guidance/index.htm.

Hsu, J. C., J. T. G. Hwang, H.-K. Liu, and S. J. Ruberg (1994). Confidence intervals associated with tests for bioequivalence. Biometrika 81, 103–114.

Jones, B. and M. G. Kenward (2014). Design and Analysis of Cross-over Trials. CRC press.

Labes, D. and H. Schütz (2016). Inflation of type I error in the evaluation of scaled average bioequivalence, and a method for its control. Pharmaceutical Research 33, 2805–2814.

Lakens, D. (2017). Equivalence tests: A practical primer for $t$-tests, correlations, and meta-analyses. Social Psychological and Personality Science 8, 355–362.

Lakens, D., A. M. Scheel, and P. M. Isager (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science 1, 259–269.

Lehmann, E. L. (1986). Testing Statistical Hypothesis, 2nd edition. New York: Wiley.

Liu, J. P. and S. C. Chow (1996). Bioequivalence trials, intersection-union tests and equivalence confidence set: Comment. Statistical Science 11, 306–312.

Lui, K.-J. and K.-C. Chang (2011). Test non-inferiority (and equivalence) based on the odds ratio under a simple crossover trial. Statistics in Medicine 30, 1230–1242.

Mazzolari, R., S. Porcelli, D. J. Bishop, and D. Lakens (2022). Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. Experimental Physiology 107, 201–212.

Metzler, C. (1974). Bioavailability – a problem in equivalence. Journal of Pharmaceutical Sciences 30, 309–317.

Meyners, M. (2012). Equivalence tests – a review. Food Quality and Preference 26, 231–245.

Molins, E., D. Labes, H. Schütz, E. Cobo, and J. Ocaña (2021). An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs. Biometrical Journal 63, 122–133.

Muñoz, J., D. Alcaide, and J. Ocaña (2016). Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. Statistics in Medicine 35, 1933–1943.

O'Brien, M. W. and D. S. Kimmerly (2022). Is "not different" enough to conclude similar cardiovascular responses across sexes? American Journal of Physiology-Heart and Circulatory Physiology 322, H355–H358.

Ocaña, J. and J. Muñoz (2019). Controlling type I error in the reference-scaled bioequivalence evaluation of highly variable drugs. Pharmaceutical Statistics 18, 583–599.

Ostrovski, V. (2022). Testing equivalence to binary generalized linear models with application to logistic regression. Statistics & Probability Letters 191, 109658.

Owen, D. B. (1968). A survey of properties and applications of the noncentral t-distribution. Technometrics 10, 445–478.

Pallmann, P. and T. Jaki (2017). Simultaneous confidence regions for multivariate bioequivalence. Statistics in Medicine 36(29), 4585–4603.

Palmes, C., T. Bluhmki, B. Funke, and E. Bluhmki (2022). Asymptotic properties of the two one-sided t-tests - new insights and the schuirmann-constant. The International Journal of Biostatistics 18, 19–38.

Patterson, S. and B. Jones (2006). Bioequivalence and Statistics in Clinical Pharmacology. Boca Raton, FL: Chapman & Hall/CRC.

Phillips, K. (1990). Power of the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics 18, 137–144.

Quartier, J., N. Capony, M. Lapteva, and Y. N. Kalia (2019). Cutaneous biodistribution: A high-resolution methodology to assess bioequivalence in topical skin delivery. Pharmaceutics 11, 484.

Sansone, P., L. G. Giaccari, C. Aurilio, F. Coppolino, M. B. Passavanti, V. Pota, and M. C. Pace (2022). Comparative efficacy of Tapentadol versus Tapentadol plus Duloxetine in patients with chemotherapy-induced peripheral neuropathy. Cancers 14, 4002.

Schouten, H. and A. Kester (2010). A simple analysis of a simple crossover trial with a dichotomous outcome measure. Statistics in Medicine 29, 193–198.

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.

Schütz, H., D. Labes, and M. J. Wolfsegger (2022). Critical remarks on reference-scaled average bioequivalence. Journal of Pharmacy & Pharmaceutical Sciences 25, 285–296.

Senn, S. (2021). Statistical Issues in Drug Development, 3rd edition. Wiley.

Sureshkumar, H., R. Xu, N. Erukulla, A. Wadhwa, and L. Zhao (2022). “snap on” or not? a validation on the measurement tool in a virtual reality application. Journal of Digital Imaging 35(3), 692–703.

Tothfalusi, L. and L. Endrenyi (2016). An exact procedure for the evaluation of reference-scaled average bioequivalence. American Association of Pharmaceutical Scientists Journal 18, 476–489.

Tu, D. (1998). On the use of the ratio or the odds ratio of cure rates in establishing therapeutic equivalence of non-systemic drugs with binary clinical endpoints. Journal of Biopharmaceutical Statistics 8, 263–282.

Wang, K., Y. Li, B. Chen, H. Chen, D. E. Smith, D. Sun, M. R. Feng, and G. L. Amidon (2022). In vitro predictive dissolution test should be developed and recommended as a bioequivalence standard for the immediate-release solid oral dosage forms of the highly variable mycophenolate mofetil. Molecular Pharmaceutics 19, 2048–2060.

Wehrle, F. M., T. Bartal, M. Adams, D. Bassler, C. F. Hagmann, O. Kretschmar, G. Natalucci, and B. Latal (2022). Similarities and differences in the neurodevelopmental outcome of children with congenital heart disease and children born very preterm at school entry. The Journal of Pediatrics 250, 29–37.e1.

Westlake, T. (1976). Symmetrical confidence intervals for bioequivalence trials. Biometrics 32, 741–744.

Wonnemann, M., C. Frömke, and A. Koch (2015). Inflation of the type I error: Investigations on regulatory recommendations for bioequivalence of highly variable drugs. Pharmaceutical Research 32, 135–143.

# Appendix

## A. Existence of $\alpha^*$

In this section, we state the conditions for $\alpha^*$, defined in (8), to be a singleton. Fixing $\alpha$, $c$, $\sigma_\nu$, and $\nu$, we simplify our notation so that $\omega(\gamma) := \omega(\gamma, c, \sigma_\nu, \nu)$ and let

$$\mathcal{A} := \left\{\gamma \in [\alpha,\, 0.5) \,\middle|\, \omega(\gamma) > 0\right\}.$$

The function $\omega(\gamma)$, defined in (6), is continuously differentiable and strictly increasing for $\gamma$ in $\mathcal{A}$. From (7), we have that $\alpha \geqslant \omega(\alpha)$. Thus, it is sufficient to show that $\alpha < \alpha_{\max} := \lim_{\alpha \to 0.5^-} \omega(\alpha)$, where $\alpha \to 0.5^-$ denotes the limit from below 0.5, to ensure that $\alpha^*$ is a singleton. Let $T_{\nu,\delta}$ denote a random variable following a non-central $t$-distribution with $\nu$ degrees of freedom and noncentrality parameter $\delta$, $\Phi(x)$ denote the cumulative distribution function of the standard normal distribution, and $\delta := 2c/\sigma_\nu$. Then, we have

$$\begin{aligned}\alpha_{\max} &= \lim_{\gamma \to 0.5^-} \omega(\gamma) = \lim_{\gamma \to 0.5^-} \left\{Q_\nu\left(-t_{\gamma,\nu},\, 0,\, \frac{c\sqrt{\nu}}{\sigma_\nu t_{\gamma,\nu}}\right) - Q_\nu\left(t_{\gamma,\nu},\, \delta,\, \frac{c\sqrt{\nu}}{\sigma_\nu t_{\gamma,\nu}}\right)\right\}\\ &= \Pr\left(T_{\nu,0} \leqslant 0\right) - \Pr\left(T_{\nu,\delta} \leqslant 0\right) = 0.5 - \Pr\left(T_{\nu,\delta} \leqslant 0\right) = 0.5 - \Phi(-\delta) = \Phi(\delta) - 0.5.\end{aligned}$$

Thus, the condition $\alpha < \alpha_{\max}$ can expressed as follows

$$\Phi(\delta) - 0.5 > \alpha \quad \Longleftrightarrow \quad \frac{2c}{\sigma_\nu} > \Phi^{-1}(\alpha + 0.5) \quad \Longleftrightarrow \quad \sigma_\nu < \frac{2c}{\Phi^{-1}(\alpha + 0.5)}.$$

Therefore, the condition $\sigma_\nu < \frac{2c}{\Phi^{-1}(\alpha+0.5)}$ implies that $\alpha < \alpha_{\max}$ and consequently that $\alpha^*$ is a singleton

## B. Convergence Rate of the Iterative Approach for $\alpha^*$

Using the notation of Appendix A and for $\gamma \in \mathcal{A}$, we have that $\omega(\gamma)$ is continuously differentiable and such that $0 < \dot{\omega}(\gamma) < 2$, where

$$\dot{\omega}(\gamma) := \left.\frac{\partial}{\partial x}\omega(x)\right|_{x=\gamma}.$$

Next, we define

$$T(\gamma) := \alpha + \gamma - \omega(\gamma).$$

For all $\alpha_1, \alpha_2 \in \mathcal{A}$, we have the mean value theorem stating that

$$T(\alpha_1) - T(\alpha_2) = \alpha_1 - \alpha_2 - \omega(\alpha_1) + \omega(\alpha_2) = \alpha_1 - \alpha_2 + \dot{\omega}(\alpha^*)(\alpha_2 - \alpha_1),$$

where $\alpha^* = \tau\alpha_1 + (1-\tau)\alpha_2$ for some $\tau \in [0, 1]$. Thus, we obtain

$$\left|T(\alpha_1) - T(\alpha_2)\right| = \left|\{1 - \dot{\omega}(\alpha^*)\}(\alpha_1 - \alpha_2)\right| = \left|1 - \dot{\omega}(\alpha^*)\right|\left|\alpha_1 - \alpha_2\right| < \left|\alpha_1 - \alpha_2\right|.$$

Then, using Kirszbraun theorem (see Federer, 2014), we can extend the function $T(\gamma)$ with respect to $\gamma \in \mathcal{A}$ to a contraction map from $\mathbb{R}$ to $\mathbb{R}$. Thus, Banach fixed point theorem ensures that $T\left(\alpha^{*(k)}\right)$ converges as $k \to \infty$. We then define the limit of the sequence $\left(\alpha^{*(k+1)}\right)_{k\in\mathbb{N}}$ as $\alpha^*$, which is the unique fixed point of the function $T(\gamma)$. Indeed, we have

$$\alpha^* = T\left(\alpha^*\right) = \alpha + \alpha^* - \omega\left(\alpha^*\right).$$

By rearranging terms, we have

$$\alpha^* = \underset{\gamma\in\mathcal{A}}{\text{argzero}}\ \omega\left(\gamma\right) - \alpha = \underset{\gamma\in[\alpha,0.5)}{\text{argzero}}\ \omega\left(\gamma\right) - \alpha,$$

concluding the convergence of the sequence $\left(\alpha^{*(k+1)}\right)_{k\in\mathbb{N}}$. As a result, there exists some $0 < \epsilon < 1$ such that for $k \in \mathbb{N}$ we have

$$\left|\alpha^{*(k+1)} - \alpha^*\right| < \epsilon^k\left|\alpha^* - \alpha\right| < \frac{1}{2}\exp(-bk),$$

for some positive constant $b$.

## C. Comparison of the $\alpha$-TOST with Cao and Mathew (2008) Method

In Figure 7, the critical values for different values of $\nu$, obtained by Cao and Mathew (2008) (Table 1) and the ones obtained using the $\alpha$-TOST, are compared. One can see that the method of Cao and Mathew (2008) appears to be an approximation of the $\alpha$-TOST, evaluated asymptotically, i.e., at $\nu \to \infty$.

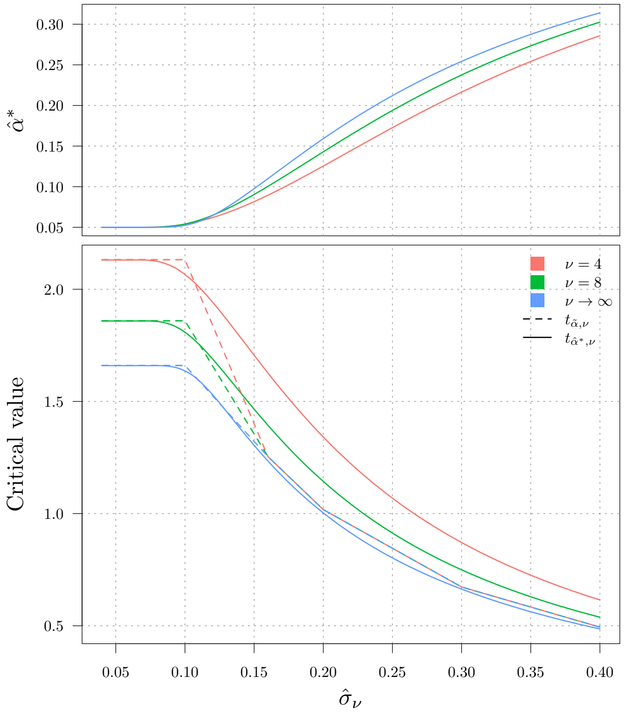

Figure 7: Upper panel: values of $\widehat{\alpha}^*$ of the $\alpha$-TOST (y-axis) as a function of $\widehat{\sigma}_\nu$ (x-axis) for different values of $\nu$ (coloured lines). Lower panel: comparison of the critical values (y-axis) obtained by the method of Cao and Mathew (2008) (dashed lines showing $t_{\tilde{\alpha},\nu}$) and of the $\alpha$-TOST (solid lines showing $t_{\widehat{\alpha}*,\nu}$) as a function of $\widehat{\sigma}_\nu$ (x-axis) for different values of $\nu$ (coloured lines). Note that for all values of $\nu$ considered here and for values of $\widehat{\sigma}_\nu$ above 0.1, the critical values of Cao and Mathew (2008) correspond to a piece-wise version of the critical values of the $\alpha$-TOST obtained when $\nu$ is large. Therefore, their correction appears to be an approximation of the $\alpha$-TOST, evaluated asymptotically, i.e., at $\nu \to \infty$.