Penultimate version. If citing, please refer instead to the published version in Statistics in Medicine, DOI: 10.1002/sim.7446.

Simultaneous confidence regions for multivariate bioequivalence

Philip Pallmann, Thomas Jaki

Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics,
Lancaster University, Lancaster, UK

1 Introduction

Statistical assessment of bioequivalence has many applications in pharmacological research and production e.g., when the mode of administration is altered or when the production site is changed Patterson and Jones (2006). The most prominent usage, however, is in the development of generic medicinal products. The formal approval of generics can be substantially abbreviated if a sponsor is able to demonstrate that the me-too product is bioequivalent to its well-investigated brand-name counterpart, in the sense that both are comparable in bioavailability. The standard framework for such an investigation is the two-treatment, two-period, two-sequence ($2 \times 2 \times 2$) crossover design Jones and Kenward (2015): study participants are randomly assigned to receive the test drug first and then the reference (sequence TR), or *vice versa* (sequence RT), and blood samples are taken at a series of time points after administration of the compound. The measured concentration values are conventionally summarised by pharmacokinetic (PK) parameters. Since there is no single PK measure that is thought to capture all aspects of bioavailability, it is common practice to investigate multiple proxy measures like the area under the concentration-time curve from zero to the last observed time point (AUC_{0-t}) or extrapolated to infinity ($AUC_{0-\infty}$), or the maximum observed concentration (C_{max}) Cawello (1999).

All relevant guidelines from regulatory bodies in the US and Europe require multiple PK parameters to be analysed. The FDA's general guidance on bioequivalence U.S. Food & Drug (2003) recommends *three* PK parameters in single-dose studies: AUC_{0-t}, $AUC_{0-\infty}$ (both measuring total exposure), and C_{max} (as a measure of peak exposure). This statement was reiterated in two recent draft guidelines U.S. Food & Drug (2013, 2014). The FDA guidance on statistical aspects of bioequivalence U.S. Food & Drug (2001) does not differentiate between AUC_{0-t} and $AUC_{0-\infty}$ but rather only requests AUC and C_{max}. The EMA's guideline on bioequivalence European Medicines (2010) demands *two* PK parameters in single-dose studies, AUC_{0-t} (or, on rare occasions, AUC_{0-72h} truncated at 72h) and C_{max}, and so do the relevant guidelines for Canada Health (2012) and Japan Japan Generic Medicines (2012).

There is no doubt these quantities are correlated as we calculate them using the same data of the same samples from the same individuals. None of the guidelines, however, mentions this fact, and neither do they make any recommendations about multivariate analyses. As a consequence, most researchers analyse all PK measures individually. The generally accepted statistical analysis for a single PK parameter is the two one-sided tests (TOST) procedure Schuirmann (1987) with the type I error rate controlled at level α, usually chosen as 0.05. The same test decision can be attained with an ordinary $100(1-2\alpha)\%$ confidence interval (CI): the inclusion approach i.e., checking whether the CI lies within prespecified equivalence boundaries, is the most common way of showing bioequivalence to this day.

In this paper we go beyond such univariate considerations and discuss how bioequivalence can and should be demonstrated for two or more PK parameters simultaneously. The TOST procedure itself is easy to extend by applying the intersection-union principle Berger (1982), and other multivariate equivalence tests are available as well Hoffelder et al. (2015); Hua et al. (2015), but there have been controversies about how to construct a simultaneous confidence region around the vector of estimated PK measures. A number of ideas have been put forward in the past two decades: various authors suggested methods tailored to (bio-)equivalence problems Brown et al. (1995); Wang et al. (1999); Munk and Pflüger (1999); Quan et al. (2001); Tseng (2002), but also approaches that were not specifically designed for application in pharmaceutical statistics Casella and Hwang (1983); Tseng and Brown (1997) may prove useful. For three and more PK parameters, methods that involve shrinkage (see Casella and Hwang (2012) for an overview) seem appealing due to Stein's paradox Stein (1956).

Despite this wealth of theoretical options, joint confidence sets for multiple PK measures have not gained wide acceptance to the present day, and we believe this is because of two major obstacles: the lack of decision guidance for practical data problems, and the lack of accessible software. Moreover, several of the proposed methods rely on conditions that are rarely met in the real world, like known variance or independence. Some have even more blatant downsides such as non-existing or infinitely large confidence regions in some cases. Another practical issue is how to translate a joint confidence region into marginal simultaneous CIs for the single PK measures, especially when the region has some irregular shape. And seemingly favourable properties like a small volume do not necessarily entail good (marginal) operating characteristics.

We are not aware of any comprehensive study that contrasts the available methods with regard to their usefulness under realistic circumstances i.e., smallish sample sizes, unknown and possibly heterogeneous variances, and high correlation between several PK measures. Many of the methods for multi-parameter confidence regions do not extend smoothly to the unknown variance case Efron (2006), and approximations have to be put up with; hence the performance of different methods will have to be compared by simulation rather than through analytical arguments. In this paper we present and discuss simulation-based comparisons of methods for multiple PK parameters. The focus of our work lies on frequentist ideas that yield simultaneous confidence sets for multivariate average bioequivalence, but we acknowledge that there are sophisticated Bayesian approaches Ghosh and Gönen (2008); Molina de Souza et al. (2009) and methods for individual and population bioequivalence Chinchilli and Elswick Jr. (1997); Chervoneva et al. (2007) as well.

The remainder of this article is organised as follows. We illustrate the multivariate bioequivalence problem using a data example of ticlopidine hydrochloride in Section 2. Then we review various methods for confidence regions of normal means and their application to bioequivalence in Section 3. Statistical properties of these methods are compared via simulation in Section 4. We evaluate the ticlopidine hydrochloride data in Section 5. A discussion and some practical recommendations in Section 6 conclude the paper.

2 Example: ticlopidine hydrochloride

Marzo *et al.* Marzo et al. (2002) investigated two formulations of ticlopidine hydrochloride, an agent known to inhibit platelet aggregation and prevent thromboembolic disorders. They set up a single-dose study of 250 mg of active ingredient administered as a tablet in a $2 \times 2 \times 2$ crossover design. 24 healthy male volunteers were randomised to either sequence AB or BA of the commercial reference product Tiklid (A) and a test formulation developed by the study sponsor (B), with a washout period of three weeks in between.

Several standard PK parameters were calculated for each of the 24 individuals using a non-compartmental approach Cawello (1999); these data are presented in Table II of Marzo et al. (2002). We focus our attention on C_{max}, AUC_{0-t} and $AUC_{0-\infty}$, the three measures required to be shown bioequivalent according to the FDA's guidance. The graphical display in Figure 1 reveals that the three of them are highly correlated. The correlation of AUC_{0-t} and $AUC_{0-\infty}$ is close to unity ($\rho = 0.973$), but also their respective correlations with C_{max} are high ($\rho = 0.698$ and $\rho = 0.808$).

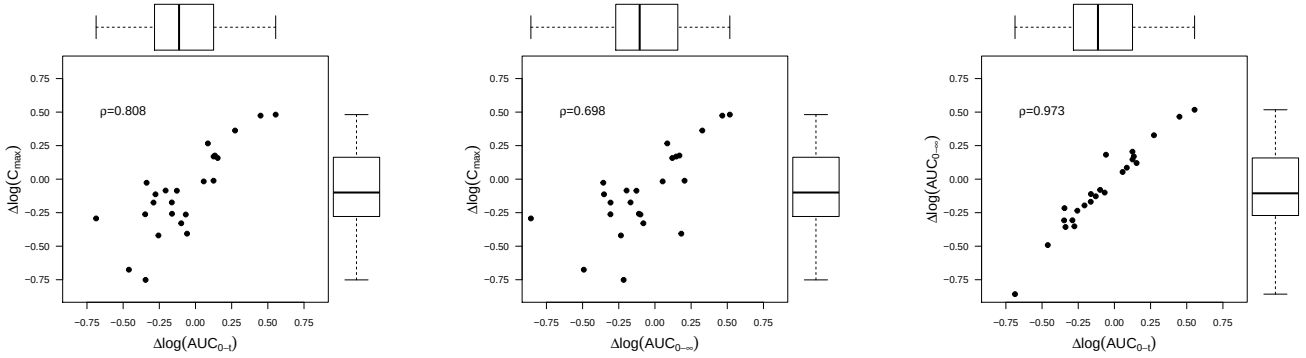

Figure 1: Ticlopidine hydrochloride data: scatterplots and marginal boxplots for the differences (test minus reference) of the logarithms of individual pharmacokinetic measurements ($n = 24$). Left: C_{max} vs. AUC_{0-t}; middle: C_{max} vs. $AUC_{0-\infty}$; right: $AUC_{0-\infty}$ vs. AUC_{0-t}; ρ: Pearson correlation coefficient.

Table 1 summarises results of the univariate analyses carried out separately for each of the three PK parameters. We see that average AUC_{0-t}, $AUC_{0-\infty}$, and C_{max} are estimated to be 7.7%, 6.6%, and 9.0% lower, respectively, for T in comparison to R. The standard errors are fairly similar for all three parameters. The TOST p-values of 0.012 for both AUC measures and 0.031 for C_{max} indicate significance at the 5% level, and the univariate 90% CIs are well within the conventional [80%, 125%] margins. All 90% ordinary CIs contain the point of exact equivalence, therefore they are identical to the 95% "expanded" CIs Berger and Hsu (1996).

We will revisit this data example in Section 5 to demonstrate the application of simultaneous confidence regions for two or three PK parameters.

Table 1: Ticlopidine hydrochloride data: results of univariate analyses separately for AUC_{0-t}, $AUC_{0-\infty}$, and C_{max}: point estimates, standard errors, TOST p-values, lower and upper bounds of 90% confidence intervals. Estimates and interval bounds are given on the logarithmic scale (in round brackets on the original scale).

	$\widehat{\theta}$	SE($\widehat{\theta}$)	p	lower	upper
AUC_{0-t}	–0.080 (0.923)	0.059	0.012	–0.181 (0.834)	0.020 (1.021)
$AUC_{0-\infty}$	–0.068 (0.934)	0.064	0.012	–0.178 (0.837)	0.042 (1.042)
C_{max}	–0.094 (0.910)	0.066	0.031	–0.207 (0.813)	0.018 (1.019)

3 Methods

3.1 Assessing bioequivalence in crossover designs

Suppose that a bioequivalence trial of a test (T) and a reference (R) compound has been conducted with n individuals in a $2 \times 2 \times 2$ crossover design. The goal is to establish the equivalence of T and R simultaneously with regard to $p \geq 1$ pharmacokinetic parameters (such as AUC_{0-t}, $AUC_{0-\infty}$, C_{max}, ...) whose values have been estimated from a set of concentration measurements separately for each individual, using either compartmental or non-compartmental techniques. Compartmental modelling Källén (2008) can be intricate because it involves choosing and fitting a nonlinear mixed-effects model Davidian and Giltinan (1995). On the other hand, non-compartmental estimation is straightforward with minimal assumptions Cawello (1999) and has recently also been extended to sparse and incomplete sampling schemes Wolfsegger and Jaki (2009); Jaki and Wolfsegger (2012); Jaki et al. (2013). We will not dwell on estimation techniques here but simply work with the resulting estimates.

We denote the ith PK parameter for the test and reference group by μ_i^T and μ_i^R, with $i = 1, \ldots, p$. The pharmacological question, "can T safely be considered equivalent to R?", translates into the statistical question whether the ratio μ_i^T / μ_i^R is within a pre-defined acceptable range, typically chosen as [0.80, 1.25], simultaneously for all p PK parameters.

Quantities like AUC and C_{max} are typically assumed to be log-normal, and therefore logarithmised before further processing them with statistical methods that demand normality. So the quantity to be analysed is

$$\theta_i = \log\left(\frac{\mu_i^T}{\mu_i^R}\right) = \log(\mu_i^T) - \log(\mu_i^R)$$

and is assumed to be normal with variance σ_i^2. The conventional choice for the equivalence range $[-\Delta, \Delta]$ on the log scale then uses $\Delta = \log(1.25)$ so that the resulting interval $[-0.223, 0.223]$ is symmetric around zero.

3.2 Single parameter analysis

For the simplest case where just one PK parameter θ is to be shown bioequivalent, the generally accepted procedure is the two one-sided tests (TOST) suggested by Schuirmann Schuirmann (1987). The null hypothesis space is partitioned into two disjoint sub-spaces, and each of these partial null hypotheses is tested separately at level α. In practice one computes a one-sample t-test for

$$H_{01}\colon \theta \leq -\Delta \qquad \text{vs.} \qquad H_{A1}\colon \theta > -\Delta$$

at level α, and another one for

$$H_{02}\colon \theta \geq \Delta \qquad \text{vs.} \qquad H_{A2}\colon \theta < \Delta$$

also at level α. If both tests reject their respective nulls, we can reject the joint null

$$H_0 \equiv H_{01} \cup H_{02}$$

in favour of the alternative

$$H_A \equiv H_{A1} \cap H_{A2} : \ -\Delta < \theta < \Delta$$

with the type I error rate controlled at level α, due to the intersection-union principle Berger (1982).

An operationally equivalent approach is to construct a two-sided $100(1-2\alpha)\%$ CI

$$\left[\widehat{\theta} - SE(\widehat{\theta})t_{1-\alpha,\nu}, \widehat{\theta} + SE(\widehat{\theta})t_{1-\alpha,\nu}\right]$$

where $SE(\widehat{\theta})$ is the standard error of $\widehat{\theta}$, and $t_{1-\alpha,\nu}$ the $100(1-\alpha)\%$ quantile of Student's t-distribution with ν degrees of freedom. One may reject H_0 and claim bioequivalence whenever the CI is wholly contained in $[-\Delta, \Delta]$. This procedure yields equivalent decisions to the TOST; however, the peculiar fact that a $100(1-2\alpha)\%$ CI matches with a level-α test procedure has been noted by several statisticians. Berger and Hsu Berger and Hsu (1996), among others, pointed out that this relationship holds only because the ordinary CI is equi-tailed. They also clarified that there exists a $100(1-\alpha)\%$ CI that is compatible with the TOST's decision:

$$\left[\min\left(0, \widehat{\theta} - SE(\widehat{\theta})t_{1-\alpha,\nu}\right), \max\left(0, \widehat{\theta} + SE(\widehat{\theta})t_{1-\alpha,\nu}\right)\right].$$

This CI was derived in various contexts Hsu (1984); Stefansson et al. (1988); Müller-Cohrs (1991); Hsu et al. (1994) and is a special case of the "expanded" CIs discussed by Bofinger Bofinger (1985, 1992). It always contains zero, and in consequence it has coverage probability 1 at $\theta = 0$ and $(1-\alpha)$ elsewhere.

Brown *et al.* Brown et al. (1997) suggested a sharper test whose critical region, however, has a very irregular shape. Simplified versions of this test with smoother boundaries of the critical region were developed as well Berger and Hsu (1996); Munk et al. (2000).

3.3 Multiple PK parameters

Extending the assessment of bioequivalence to the multi-parameter case ($p \geq 2$) corresponds to testing the pair of hypotheses

$$H_0\text{: } |\theta^{(i)}| \geq \Delta_i \text{ for at least one } i$$
$$H_A\text{: } |\theta^{(i)}| < \Delta_i \text{ for all } i$$

where Δ_i is in the simplest case the same for all PK parameters $i = 1, \ldots, p$. Finding a suitable test procedure appears trivial at first sight: since the goal is to demonstrate equivalence for *all* PK measures simultaneously, an extension of the level-α TOST using the intersection-union principle will control the overall type I error rate at α, and no correction for multiplicity is required. In practice, p separate TOSTs will be carried out for the p parameters, each at level α. For the CI inclusion approach both the ordinary $100(1-2\alpha)\%$ and the "expanded" $100(1-\alpha)\%$ intervals can be used.

Unfortunately, this procedure will be conservative and have poor coverage probability (CP) unless the p parameters are perfectly correlated Phillips (2009); Tsai et al. (2014). When they are independent, the test size will be $0.05^2 = 0.0025$ (left-hand side of Figure 2). Even with typical values for the correlation between AUC and C_{max}, such as $0.7 < \rho < 0.9$, the test will have a size of only 2–3% rather than the intended 5%. Likewise, nominal coverage will only be achieved under perfect correlation of the parameters (right-hand side of Figure 2). When both PK parameters are uncorrelated, the joint CP of the ordinary $100(1-2\alpha)\%$ CIs is $100(1-2\alpha)^2\%$, and the joint CP of the "expanded" $100(1-\alpha)\%$ CIs is $100(1-\alpha)^2\%$.

Multiple alternatives to the simultaneous TOSTs have been published, but it is far from obvious what the "best" solution is. In this paper we focus on methods that provide confidence regions for the interesting PK parameters rather than (only) p-values and perhaps CIs. The basic idea is to construct a p-dimensional simultaneous confidence region around a p-variate normal mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ that is estimated as $\widehat{\boldsymbol{\theta}}$ and has covariance matrix $\boldsymbol{\Sigma}$, and then possibly derive the marginal simultaneous CIs for the p parameters, or alternatively to construct the simultaneous CIs directly. Joint confidence regions allow for a much more insightful interpretation than marginal CIs, let alone p-values, when the task is to estimate the likely location and variability of the parameter vector $\boldsymbol{\theta}$ Douglas (1993).

In the following we review several methods for joint confidence regions in the context of multi-parameter bioequivalence. While some of them were developed for this very purpose, others originated from the more general problem of building confidence regions around a multivariate normal mean.

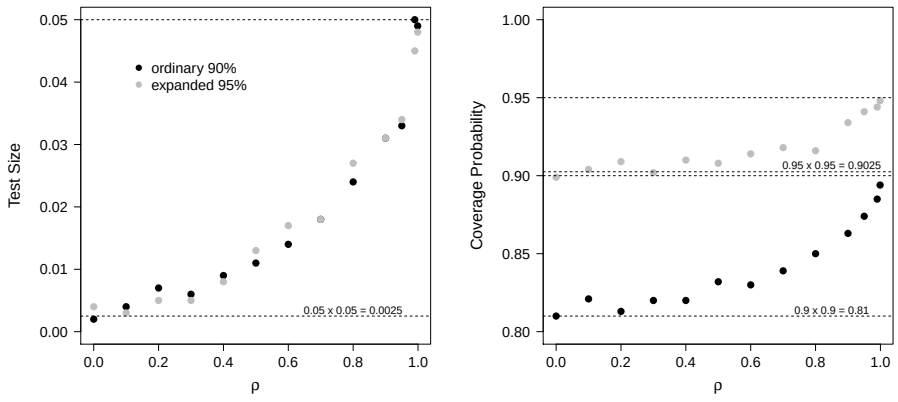

Figure 2: Test size (left) and coverage probability (right) of simultaneous TOST procedures for $p = 2$ parameters with $\theta_1 = \theta_2 = \log(1.25)$, $\sigma_1^2 = \sigma_2^2 = 0.1$, and total sample size $n = 20$ depending on the correlation ρ. Black: ordinary 90% confidence intervals; grey: "expanded" 95% confidence intervals (10,000 simulations).

3.3.1 The standard confidence region:

The standard $100(1-\alpha)\%$ simultaneous confidence region Chew (1966) for $\boldsymbol{\theta}$ is the ellipse

$$C^0(X) = \left\{\boldsymbol{\theta}\text{: } (X-\boldsymbol{\theta})'\boldsymbol{\Lambda}^{-1}(X-\boldsymbol{\theta}) \leq \sigma^2\chi^2_{1-\alpha,p}\right\}$$

where X is a random variable following a p-variate normal distribution $\mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\chi^2_{1-\alpha,p}$ is the $100(1-\alpha)\%$ quantile of the χ^2 distribution with p degrees of freedom, and covariance $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Lambda}$ with arbitrary (positive-definite and known) $\boldsymbol{\Lambda}$ and known σ^2.

When σ^2 is unknown and estimated by s^2 where $\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$ is independent of $\boldsymbol{\theta}$, the standard region becomes

$$C^0(X,s) = \left\{\boldsymbol{\theta}\text{: } (X-\boldsymbol{\theta})'\boldsymbol{\Lambda}^{-1}(X-\boldsymbol{\theta}) \leq \frac{s^2 p}{\nu}F_{1-\alpha,p,\nu}\right\}$$

where $F_{1-\alpha,p,\nu}$ is the $100(1-\alpha)\%$ quantile of the F distribution with p and ν degrees of freedom. Neither of these two regions is sufficiently general to incorporate unknown $\boldsymbol{\Lambda}$, which is the standard case in practice. If we ignore the correlation among the θ_i, the left-hand side reduces to $||X-\boldsymbol{\theta}||^2$ where $||\cdot||$ denotes the Euclidian norm.

Wang *et al.* Wang et al. (1999) put forward a $100(1-\alpha)\%$ simultaneous confidence region that is given by

$$C^0(X,\widehat{\boldsymbol{\Sigma}}) = \left\{\boldsymbol{\theta}\text{: } n(X-\boldsymbol{\theta})'\widehat{\boldsymbol{\Sigma}}^{-1}(X-\boldsymbol{\theta}) \leq \frac{\nu p}{\nu-p+1}F_{1-\alpha,p,\nu-p+1}\right\}.$$

The left-hand part of the inequality is Hotelling's T^2 statistic Hotelling (1931). In comparison to the ellipses $C^0(X)$ and $C^0(X,s)$, it allows to include an estimate $\widehat{\boldsymbol{\Sigma}}$ of the unknown covariance matrix.

Using this confidence set to assess bioequivalence with an inclusion approach as described in 3.2 has been shown to be conservative Wang et al. (1999); Munk and Pflüger (1999). Figure 3 displays the actual sizes α^* of a simultaneous test procedure induced by the p-dimensional 90% confidence set. For a single PK parameter ($p = 1$), we get the well-known result that the ordinary 90% CI corresponds to a test size of 5%. Already in the bivariate case ($p = 2$), the size is well below 0.02, even for very large n.

3.3.2 The limaçon of Pascal:

Brown *et al.* Brown et al. (1995) derived a confidence region for $\boldsymbol{\theta}$ whose expected volume is minimised at a prespecified point $\boldsymbol{\theta}_0$. The natural choice for $\boldsymbol{\theta}_0$ in a bioequivalence setting is $\mathbf{0}$, the point of exact equivalence.

If the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be known, the $100(1-\alpha)\%$ simultaneous confidence region is

$$C^{Lim}(X) = \left\{\boldsymbol{\theta}\text{: } \frac{\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}X}{\sqrt{\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}}} + z_{1-\alpha} > \sqrt{\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}}\right\}$$

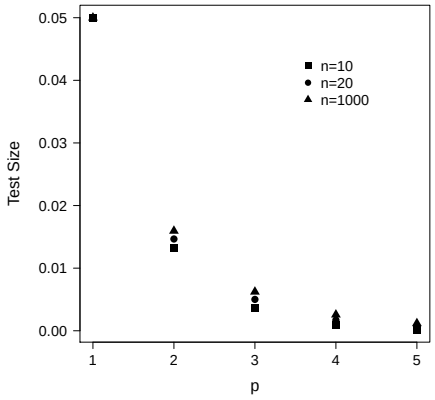

Figure 3: Actual size α^* of the simultaneous test procedure induced by the standard 90% confidence set, given the number of PK measures p and total sample size n.

where $z_{1-\alpha}$ is the $100(1-\alpha)\%$ standard normal quantile. This region always contains $\boldsymbol{\theta}_0 = \mathbf{0}$, and its boundary has the shape of the main lobe of the limaçon of Pascal; see the examples in the top rows of Figures 4–7.

Berger and Hsu Berger and Hsu (1996) outlined how to transfer this result to the case of unknown covariance, estimated as $\widehat{\boldsymbol{\Sigma}}$. Now the $100(1-\alpha)\%$ simultaneous confidence region is

$$C^{Lim}(X, \widehat{\boldsymbol{\Sigma}}) = \left\{ \boldsymbol{\theta}\colon \frac{\boldsymbol{\theta}' X}{\sqrt{\boldsymbol{\theta}' \widehat{\boldsymbol{\Sigma}} \boldsymbol{\theta}/n}} + t_{1-\alpha,\nu} > \frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{\sqrt{\boldsymbol{\theta}' \widehat{\boldsymbol{\Sigma}} \boldsymbol{\theta}/n}} \right\}.$$

It seems this confidence region for unknown $\boldsymbol{\Sigma}$ has never been investigated in detail; instead for instance Munk and Pflüger (1999) plugged in $\widehat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$ in the formula for $C^{Lim}(X)$, falsely assuming known covariance, but this will achieve nominal coverage probability only for large sample sizes, as we show in Section 4. Brown *et al.* themselves Brown et al. (1995) only described the univariate case where the confidence set reduces to the "expanded" CI of 3.2. They further wrote: "Generalizations to higher dimensions should also be of interest. Presumably, the limaçon will not appear here." We can confirm this is true, but the result is another peculiar shape where the inner lobe of the limaçon appears to be everted, like yeast cells budding; see the examples in the bottom rows of Figures 4–7.

We illustrate the basic properties of the limaçon confidence regions with a few graphics. Figure 4 visualises one obvious problem: the volume of the confidence region depends on the choice of $\boldsymbol{\theta}_0$. Setting $\boldsymbol{\theta}_0 = \mathbf{0}$ is the natural choice in a bioequivalence context and yields minimal volume at the origin; however, small regions would be particularly desirable when the θ_i are near the equivalence boundaries. With the estimated value $\widehat{\boldsymbol{\theta}}$ moving away from $\boldsymbol{\theta}_0$, the volume of the region blows up like a balloon.

If prior information about $\boldsymbol{\theta}$ is available (e.g., from a pilot study), one might think about using it for $\boldsymbol{\theta}_0$. Assume a pilot study yielded $\widehat{\theta}_1 = \widehat{\theta}_2 = 0.1$, then it could be wise to set $\boldsymbol{\theta}_0 = (0.1, 0.1)$ to achieve minimum volume around the probable value of the mean rather than around $\mathbf{0}$. Whatever the choice of θ_0, it has to be made before seeing the data.

Increasing the total sample size n will reduce the volume of the confidence regions (Figure 5); however, since the region must always contain both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, its diameter can never be smaller than the Euclidian distance between the two of them. Changes in $\widehat{\boldsymbol{\theta}}$ and n do not only have an impact on the volume of the limaçon confidence regions but also on their shape. The same is true when changing the variances σ_1^2 and σ_2^2 and the correlation ρ (Figures 6 and 7).

3.3.3 Tseng's method:

Tseng Tseng (2002) considered a confidence region that is centered about the origin $\boldsymbol{\theta} = \mathbf{0}$ and minimises the expected effective length

$$\ell_{eff}\left(C(X)\right) \equiv 2 \sup_{y \in C(X)} ||y||,$$

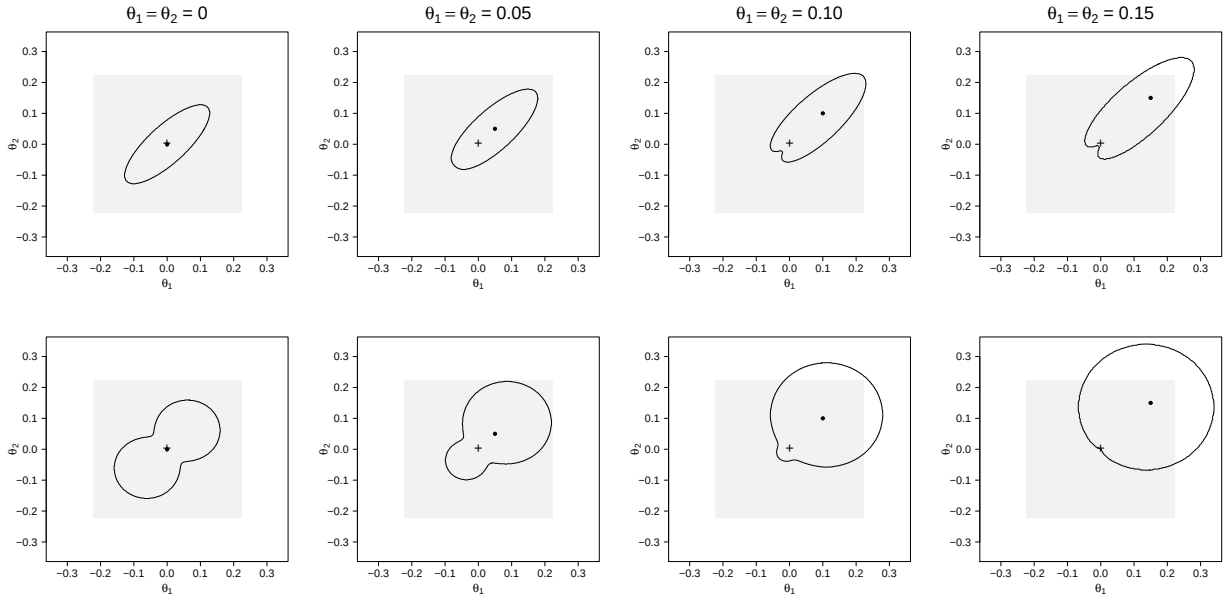

Figure 4: Limaçon-type simultaneous 90% confidence regions for bivariate normal data with varying means θ_1 and θ_2, variances $\sigma_1^2 = \sigma_2^2 = 0.1$, correlation $\rho = 0.8$, and total sample size $n = 10$. The area of bioequivalence (80–125% for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$. Top row: asymptotic regions assuming $\boldsymbol{\Sigma}$ is known (with $\widehat{\boldsymbol{\Sigma}}$ plugged in for $\boldsymbol{\Sigma}$); bottom row: finite-sample regions allowing for unknown $\boldsymbol{\Sigma}$.

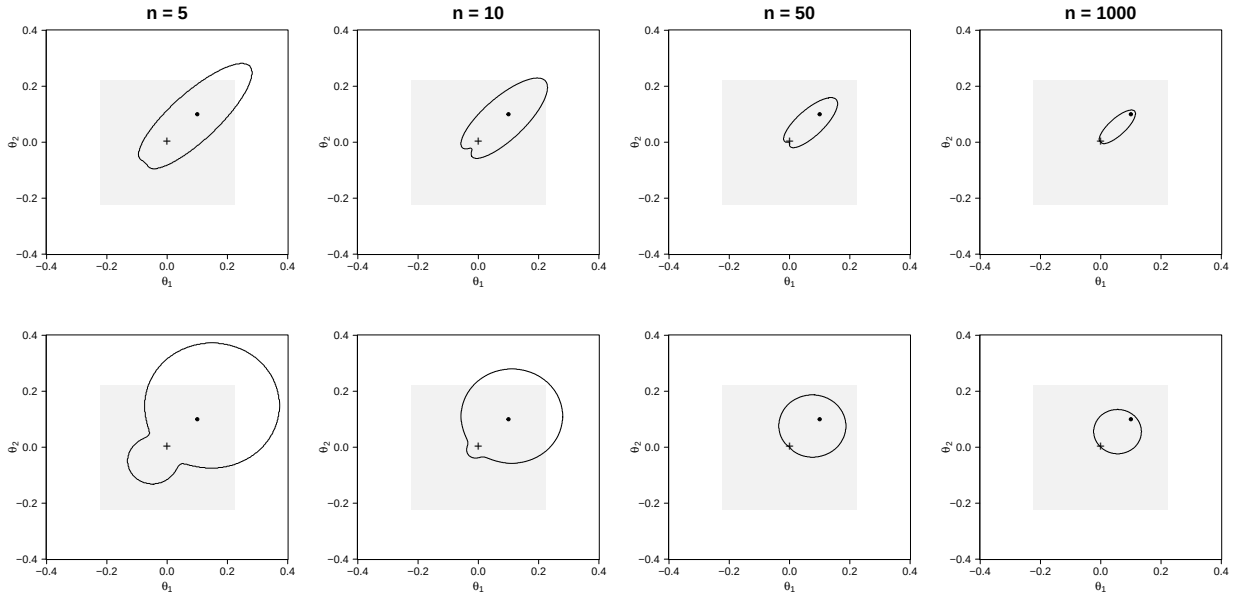

Figure 5: Limaçon-type simultaneous 90% confidence regions for bivariate normal data with means $\theta_1 = \theta_2 = 0.1$, variances $\sigma_1^2 = \sigma_2^2 = 0.1$, correlation $\rho = 0.8$, and varying total sample sizes n. The area of bioequivalence (80–125% bioequivalence for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$. Top row: asymptotic regions assuming $\boldsymbol{\Sigma}$ is known (with $\widehat{\boldsymbol{\Sigma}}$ plugged in for $\boldsymbol{\Sigma}$); bottom row: finite-sample regions allowing for unknown $\boldsymbol{\Sigma}$.

rather than the expected volume, at the origin. Tseng's confidence set is ill-conditioned in that it can be empty; however, this does not impair the validity of the associated test.

Assuming that $\boldsymbol{\Sigma}$ is known to be the identity matrix $\mathbf{I}$, the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ is

$$C^{Tse}(X) = \left\{\boldsymbol{\theta}\colon ||X||^2 \geq \chi^2_{1-\alpha,p}(||\boldsymbol{\theta}||^2)\right\}$$

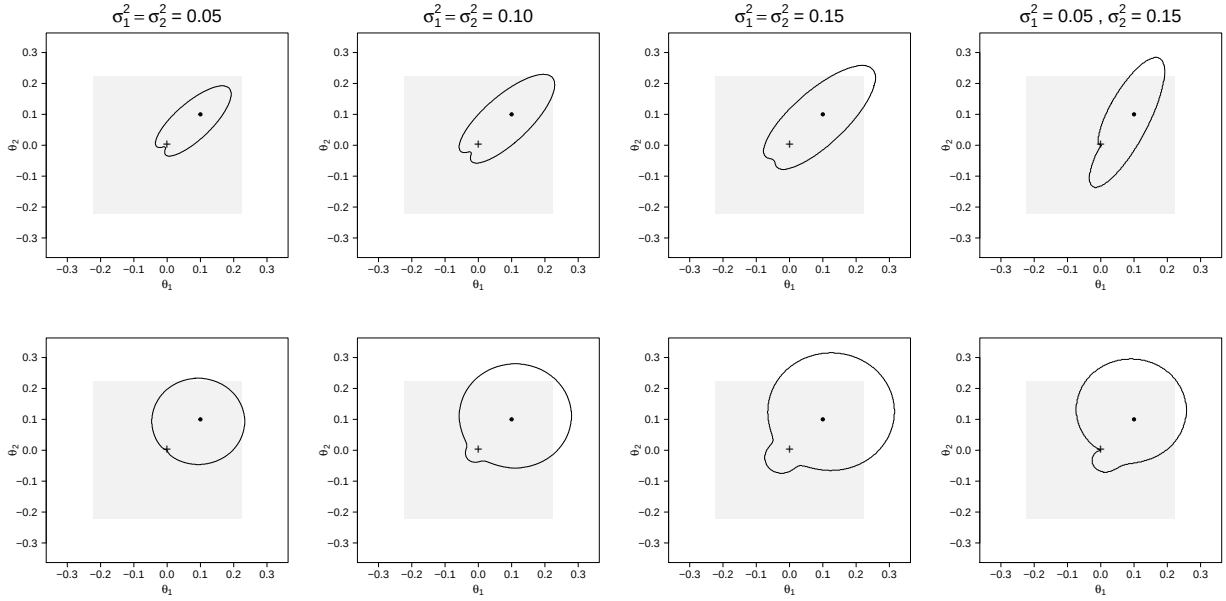

Figure 6: Limaçon-type simultaneous 90% confidence regions for bivariate normal data with means $\theta_1 = \theta_2 = 0.1$, varying variances σ_1^2 and σ_2^2, correlation $\rho = 0.8$, and total sample size $n = 10$. The area of bioequivalence (80–125% bioequivalence for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$. Top row: asymptotic regions assuming $\boldsymbol{\Sigma}$ is known (with $\widehat{\boldsymbol{\Sigma}}$ plugged in for $\boldsymbol{\Sigma}$); bottom row: finite-sample regions allowing for unknown $\boldsymbol{\Sigma}$.

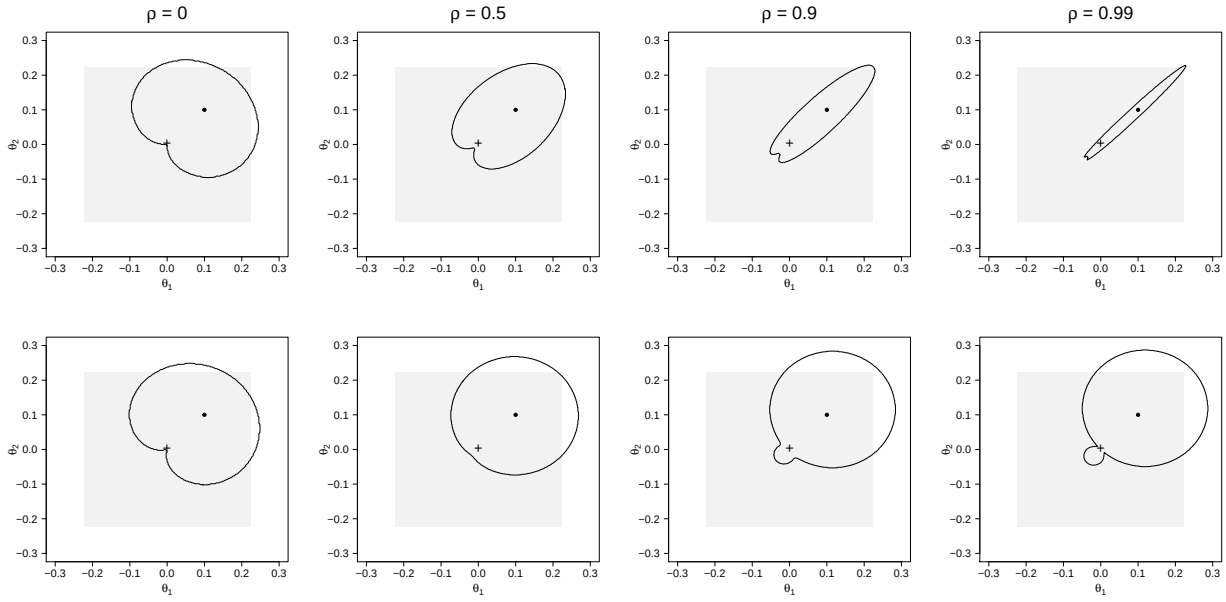

Figure 7: Limaçon-type simultaneous 90% confidence regions for bivariate normal data with means $\theta_1 = \theta_2 = 0.1$, variances $\sigma_1^2 = \sigma_2^2 = 0.1$, varying correlation ρ, and total sample size $n = 10$. The area of bioequivalence (80–125% bioequivalence for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$. Top row: asymptotic regions assuming $\boldsymbol{\Sigma}$ is known (with $\widehat{\boldsymbol{\Sigma}}$ plugged in for $\boldsymbol{\Sigma}$); bottom row: finite-sample regions allowing for unknown $\boldsymbol{\Sigma}$.

where $\chi^2_{1-\alpha,p}(\lambda)$ is the $100(1-\alpha)\%$ quantile of the χ^2-distribution with p degrees of freedom and noncentrality parameter λ.

An approximate confidence set for the case of unknown variance i.e., $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ with $\mathbf{I}$ known is given by

$$C^{Tse}(X,s) = \left\{\boldsymbol{\theta}\colon \frac{||X||^2}{ps^2} \geq F_{1-\alpha,p,\nu}\left(\frac{||\boldsymbol{\theta}||^2}{s^2}\right)\right\}$$

where $F_{1-\alpha,p,\nu}(\lambda)$ is the $100(1-\alpha)\%$ quantile of the F-distribution with p and ν degrees of freedom and noncentrality parameter λ.

3.3.4 Tseng & Brown's method:

For dimensions $p > 2$ Tseng and Brown Tseng and Brown (1997) proposed the pseudo-empirical Bayes $100(1-\alpha)\%$ confidence region

$$C^{TB}(X) = \left\{\boldsymbol{\theta}\colon ||X - \boldsymbol{\theta}\left(1+\gamma(||\boldsymbol{\theta}||^2)\right)||^2 \leq \chi^2_{1-\alpha,p}\left(||\boldsymbol{\theta}||^2\gamma^2(||\boldsymbol{\theta}||^2)\right)\right\}$$

where $\gamma(||\boldsymbol{\theta}||^2) = \frac{1}{A+B||\boldsymbol{\theta}||^2}$, and $\chi^2_{1-\alpha,p}(\lambda)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with p degrees of freedom and noncentrality parameter λ.

One practical hurdle for using this method, besides the assumption of known $\boldsymbol{\Sigma} = \mathbf{I}$, is that it is somewhat unclear how to choose values for A and B. Tseng and Brown derived necessary and sufficient conditions under which the confidence region is connected and dominates the standard region asymptotically. They used $A = 1$ together with $B = \frac{1}{p-2}$ or $B = \frac{1}{2(p-2)}$ in their numerical illustrations and also provided a table with values for A that ensure connectedness and depend on α, p, and B. We will not consider this method in our simulation study in Section 4.

3.3.5 Casella & Hwang's method:

Casella and Hwang Casella and Hwang (1983) developed an empirical Bayes approach that leads to a confidence region centred at the positive-part James-Stein estimator James and Stein (1961)

$$\delta^+(\widehat{\boldsymbol{\theta}}, s) = \left(1 - \frac{\nu(p-2)s^2}{(\nu+2)n||\widehat{\boldsymbol{\theta}}||^2}\right)^+ \widehat{\boldsymbol{\theta}}$$

where $(x)^+$ indicates $\max(0,x)$. This estimator entails shrinkage towards $\mathbf{0}$ whenever $p \geq 3$; for $p = 2$ it equals the maximum likelihood (ML) estimator $\widehat{\boldsymbol{\theta}}$.

The $100(1-\alpha)\%$ simultaneous confidence region recentred at $\delta^+(\widehat{\boldsymbol{\theta}}, s)$ is

$$C^{CH}(X,s) = \left\{\boldsymbol{\theta}\colon ||\boldsymbol{\theta} - \delta^+(\widehat{\boldsymbol{\theta}}, s)|| \leq \frac{s}{\sqrt{n}} v_E\left(\frac{n||X||}{s}\right)\right\}$$

with

$$v_E^2\left(\frac{n||X||}{s}\right) = \begin{cases} \left(1 - \frac{a}{pF_{1-\alpha,p,\nu}}\right)\left[pF_{1-\alpha,p,\nu} - p\log\left(1 - \frac{a}{pF_{1-\alpha,p,\nu}}\right)\right] & \text{if } \frac{n||X||^2}{s^2} \leq pF_{1-\alpha,p,\nu} \\ \left(1 - \frac{as^2}{n||X||^2}\right)\left[pF_{1-\alpha,p,\nu} - p\log\left(1 - \frac{as^2}{n||\widehat{\boldsymbol{\theta}}||^2}\right)\right] & \text{if } \frac{n||X||^2}{s^2} > pF_{1-\alpha,p,\nu} \end{cases}$$

and $a = \frac{\nu(p-2)}{\nu+2}$.

In the bivariate case when $a = 0$ and $v_E^2 = pF_{\alpha,p,\nu}$, the confidence region $C^{CH}(X,s)$ reduces to the standard region and is centred at the ML estimator $\widehat{\boldsymbol{\theta}}$. Related approaches have been proposed that do not use empirical Bayes arguments to obtain the radius function v_E^2 but Taylor series or parametric bootstrapping Samworth (2005) or a piecewise cubic Hermite interpolating polynomial function Abeysekera and Kabaila (2017).

3.3.6 Nonparametric bootstrap and kernel density estimation:

We propose an approach that employs nonparametric bootstrapping Efron and Tibshirani (1993); Davison and Hinkley (1997) and hence does not rely on any parametric assumption such as (multivariate) normality of the $\boldsymbol{\theta}$. Starting from the $n \times p$ matrix $\mathbf{Y}$ whose jth row contains the p PK measures of the jth individual ($j = 1,\ldots,n$), we perform bootstrapping separately for each column vector $\mathbf{y}_i = (y_{i1},\ldots,y_{in})'$ and generate B bootstrap samples $\mathbf{y}_i^*$ by drawing randomly with replacement. B is a large integer, typically 1000 or 10,000. We calculate the bootstrap mean of the

ith PK parameter and bth bootstrap replication as $\bar{y}^*_{ib} = \sum_{j=1}^{n} y^*_{jib}/n$, and these $\bar{y}^*_{ib}$ can be assembled in B bootstrap mean vectors $\bar{\mathbf{y}}^*_b = (\bar{y}^*_{1b}, \ldots, \bar{y}^*_{pb})'$. We compute a p-dimensional binned kernel density estimate Wand and Jones (1995) of the $\bar{\mathbf{y}}^*_b$ with a bivariate Gaussian kernel and select a bandwidth using a direct plug-in approach Sheather and Jones (1991). A $100(1-\alpha)\%$ confidence region is given by the area that covers $100(1-\alpha)\%$ of the estimated kernel density. For $p = 2$ dimensions this procedure is implemented in the R package `KernSmooth` Wand (2015).

3.4 Marginal (simultaneous) confidence intervals

So far we have only considered simultaneous confidence regions where combinations of values for the single parameters $\theta_1, \ldots, \theta_p$ are assessed *jointly*. For practical interpretation, however, it can be desirable to marginalise the p-dimensional joint region and obtain (simultaneous) CIs for the single PK measures. This can be done in different ways. Projecting a $100(1-\alpha)\%$ region's boundary onto the axes is simple but leads to a set of CIs whose joint coverage probability is greater than $(1-\alpha)$.

If the confidence region is symmetric in shape (like an ellipse in 2D, an ellipsoid in 3D, etc.), a (hyper-)rectangle can be constructed whose edges represent the marginalised CIs. With oddly-shaped regions like the limaçon, however, it is unclear how to derive marginal simultaneous CIs that are an improvement over the boundary's projection onto the axes.

In some cases it is possible to derive CIs without taking a detour via projections of the confidence region: for the empirical Bayes region of 3.3.5, He He (1992) showed how to construct corresponding simultaneous CIs for the p parameters directly. This method has recently attracted some attention in the context of selected parameters Qiu and Hwang (2007); Hwang and Zhao (2013), and was extended to the unknown variance case Hwang et al. (2009).

4 Simulation study

To characterise the different methods presented in 3.3 and appraise their usefulness in practical situations, we simulated a variety of scenarios that are relevant for real bioequivalence trials and looked into basic statistical properties like test size, power, joint coverage probability, and average width. Such numerical investigations are necessary as there are few theoretical results for the interesting case of unknown σ^2 Casella and Hwang (2012). All simulations were run in R version 3.1.3 R Core (2015) and with 1000 replications.

We consider the case of $p = 2$ PK parameters (e.g., $\log(AUC)$ and $\log(C_{max})$) that are jointly distributed as multivariate normal

$$\mathcal{N}\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

with mean parameters $\theta_1 = \theta_2 \in \{\log(1.00), \log(1.01), \ldots, \log(1.30)\}$, variances $(\sigma_1^2, \sigma_2^2) \in \{(0.20, 0.20), (0.05, 0.05), (0.20, 0.05)\}$, correlations $\rho \in \{0, 0.5, 0.9\}$, and total sample sizes $n \in \{20, 50, 1000\}$. In this section we present only simulation results for $\boldsymbol{\sigma}^2 = (0.05, 0.05)$ with $\rho = 0.9$ for brevity (and some additional scenarios for power); results for the other parameter settings are available as supplementary material.

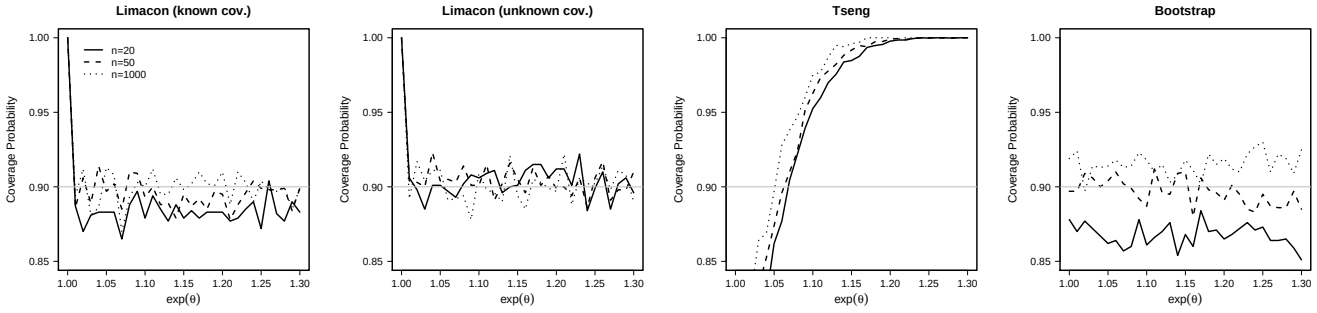

Figure 8: Simulated joint coverage probabilities of various 90% joint confidence regions for variances $\boldsymbol{\sigma}^2 = (0.05, 0.05)$, correlation $\rho = 0.9$, and total sample size n (1000 runs).

Simulated joint coverage probabilities (CPs) are displayed in Figure 8. The asymptotic limaçon region assuming known covariance is slightly liberal (87–89% CP) for samples of $n = 20$ and 50 but achieves the desired CP of 90% with $n = 1000$, whereas the finite-sample variant has CP very close to the nominal 90% for all choices of n. This also holds true for $\rho = 0$ as well as for larger and unequal variances, as can be seen in supplemental Figures S1–S3. Both limaçon-type regions have 100% CP at $\boldsymbol{\theta}_0 = \mathbf{0}$ by definition. The Tseng region is liberal for values of θ_1 and θ_2 that are close to θ_0 and conservative otherwise. The liberalism may be due to the method not incorporating the correlation properly. When the PK parameters are uncorrelated, Tseng's region maintains 90% CP for values of $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0 = \mathbf{0}$, as can be seen in supplemental Figures S1 and S2. The conservatism can be explained with the fact that the confidence region is grossly inflated as $\boldsymbol{\theta}$ moves away from θ_0. And under some circumstances Tseng's region can be an empty set Tseng (2002), which adds to the oddity of its behaviour. The bootstrap approach appears to be liberal for small and a bit conservative for large sample sizes. The standard elliptical region has exact CP under our multivariate normal model, given that the correlation is incorporated; this need not be shown by simulation. Failing to account for the correlation, however, can lead to deviations from the nominal CP in either direction, depending on the true value of ρ (Figure 9).

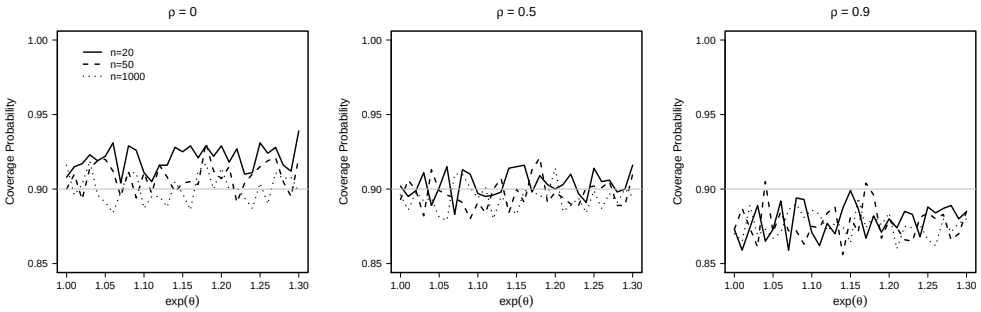

Figure 9: Simulated joint coverage probabilities of the standard 90% joint confidence region ignoring correlation for variances $\boldsymbol{\sigma}^2 = (0.05, 0.05)$, correlation ρ, and total sample size n (1000 runs).

The power simulations are summarised in Figure 10, where power is defined as the probability of declaring bioequivalence for both PK parameters. We used the TOST procedure as a benchmark to compare the powers of the different confidence regions against. It is evident from the curves in Figure 10 that confidence regions entail a power loss in comparison to the TOST at level $\alpha = 0.1$. In some situations (e.g., equal variances and high correlation), the power loss is marginal, especially for the limaçon region assuming known variance. Unfortunately, its power breaks down when the variances are unequal, which is common in reality (e.g., C_{max} tends to be more variable than AUC). There is no confidence region that has consistently good power across all simulated scenarios; see also supplemental Figures S4–S6. With $n = 1000$ the limaçon-type and Tseng regions perform very poorly, as by definition they must contain both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ and hence cannot be shrunken indefinitely. The power of all methods is reduced when $\rho = 0$, which is in line with the effect illustrated in Figure 2.

The average widths of the regions in θ_1 and θ_2 direction are presented in Figure 11. The standard and bootstrap methods yield regions with constant width for all choices of θ_1 and θ_2. On the other hand, the Tseng and limaçon-type regions' average widths go up as θ_1 and θ_2 move away from $\theta_0 = 0$. The rise is steeper for the finite-sample version of the limaçon than for its asymptotic counterpart, and even much steeper for Tseng's method; see also supplemental Figures S7–S9. With unequal variances, the widths of some of the regions are different in θ_1 and θ_2 direction (supplemental Figure S9; cf. also the rightmost panes of Figure 6).

Average widths for $\boldsymbol{\theta} = \mathbf{0}$ are listed in Table 2. We see the limaçon regions are substantially narrower at $\boldsymbol{\theta}_0 = \mathbf{0}$ than those of all other methods. Unsurprisingly, the average width decreases with increasing n for all methods.

Table 2: Simulated average widths of the 90% joint confidence regions at $\boldsymbol{\theta} = \mathbf{0}$ for variances $\boldsymbol{\sigma}^2 = (0.05, 0.05)$ and correlation $\rho = 0.9$ (1000 runs).

n	Standard/Tseng	Limaçon (asy.)	Limaçon (fin.)	Bootstrap
20	0.050	0.037	0.038	0.045
50	0.031	0.023	0.024	0.029
1000	0.007	0.005	0.005	0.007

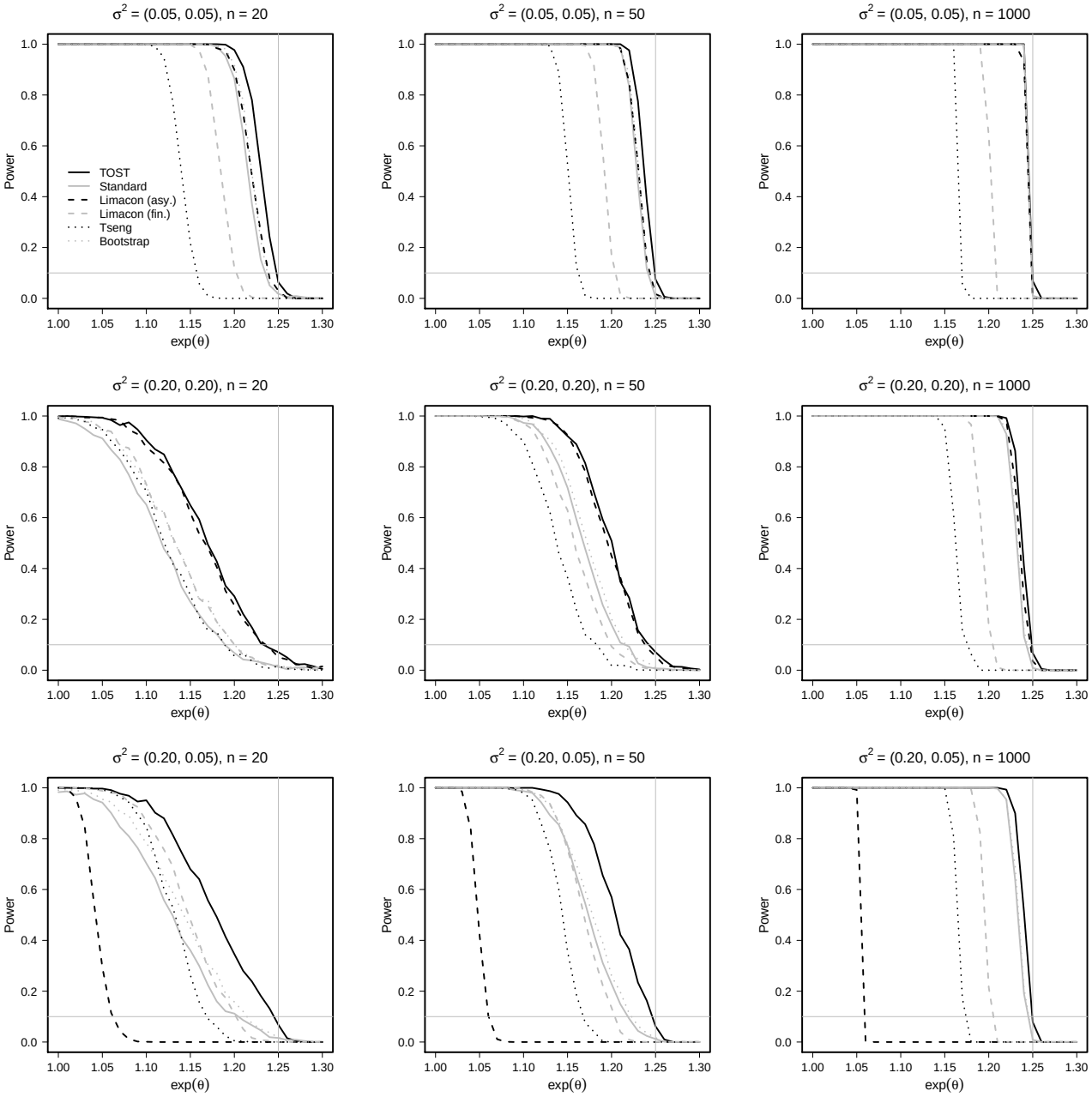

Figure 10: Simulated probabilities of declaring both PK measures bioequivalent using various 90% joint confidence regions and the TOST for variances $\boldsymbol{\sigma}^2$, correlation $\rho = 0.9$, and total sample size n (1000 runs).

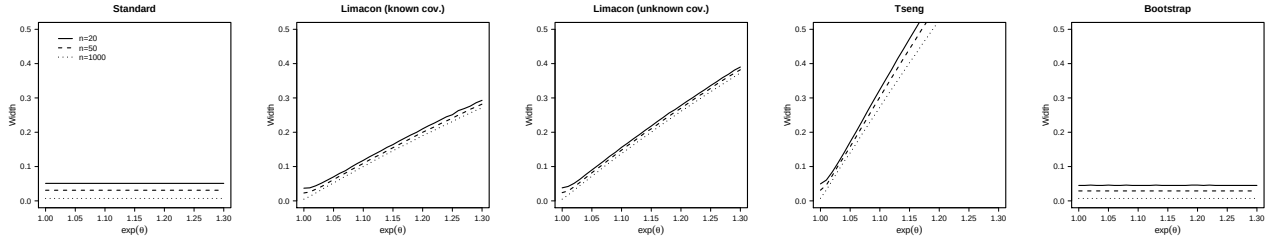

Figure 11: Simulated average widths of various 90% joint confidence regions for variances $\boldsymbol{\sigma}^2 = (0.05, 0.05)$, correlation $\rho = 0.9$, and total sample size n (1000 runs).

5 Example: ticlopidine hydrochloride (continued)

We introduced an example dataset from a bioequivalence study of ticlopidine hydrochloride in Section 2 and summarised the results of the univariate TOSTs for AUC_{0-t}, $AUC_{0-\infty}$, and C_{max} in Table 1. Now we reanalyse these data using the simultaneous confidence regions we reviewed in Section 3.

5.1 Two PK parameters: AUC_{0-t} and C_{max}

In our first multivariate analysis we focus on AUC_{0-t} and C_{max}, the two PK parameters that must be shown bioequivalent according to the EMA's, Health Canada's, and the Japanese guidelines European Medicines (2010); Health (2012); Japan Generic Medicines (2012). Rather than evaluating each of them individually (as done in Section 2), we propose a truly bivariate analysis of both parameters simultaneously.

Figure 12 shows simultaneous 90% confidence regions for AUC_{0-t} and C_{max} using different methods presented in Section 3; the corresponding marginal(ised) simultaneous CIs obtained by projecting the simultaneous regions' boundaries onto the axes are listed in Table 3. The standard regions do not allow to conclude bioequivalence, and neither do the Casella & Hwang region (because it simply reduces to the standard region ignoring correlation in the bivariate case) and the bootstrap region. On the other hand, both limaçon-type regions and the Tseng region are clearly within [–0.223, 0.223] for both AUC_{0-t} and C_{max}, although the Tseng region is substantially bigger because it is designed to be symmetric around $\mathbf{0}$. The limaçon region with unknown $\boldsymbol{\Sigma}$ looks almost circular, but we have seen in the illustrations that this can happen under various circumstances (cf. Figures 4–7); in fact, it has a little dent and is not symmetric around $\widehat{\boldsymbol{\theta}}$. The angularity of the bootstrap region is due to the low number of samples that are being bootstrapped, hence it cannot be "smoothed out" by increasing B, the number of bootstrap replications.

It should be noted that the three 90% regions that lie within the [80%, 125%] margins (limaçon and Tseng) are those that are asymmetric about $\widehat{\boldsymbol{\theta}}$ i.e., those for which a 90% confidence set does not necessarily correspond to a test level of 5% Berger and Hsu (1996). The latter can only be ensured with 95% regions (Figure 13): the asymptotic limaçon and the Tseng region still allow us to conclude bioequivalence whereas the limaçon-type region assuming unknown covariance now protrudes beyong the lower equivalence threshold for C_{max}.

Table 3: Ticlopidine hydrochloride data: boundaries of various simultaneous 90% confidence regions projected onto the axes (and 90% TOST confidence intervals for comparison) for AUC_{0-t} and C_{max}.

	AUC_{0-t}	C_{max}	Bioequivalent?
TOST	[0.834, 1.021]	[0.813, 1.019]	✓
Standard (uncorrelated)	[0.803, 1.061]	[0.792, 1.046]	-
Standard (known covariance)	[0.806, 1.057]	[0.782, 1.059]	-
Standard (unknown covariance)	[0.806, 1.057]	[0.782, 1.059]	-
Limaçon (known covariance)	[0.854, 1.033]	[0.836, 1.029]	✓
Limaçon (unknown covariance)	[0.823, 1.047]	[0.813, 1.034]	✓
Tseng	[0.825, 1.212]	[0.825, 1.212]	✓
Casella & Hwang	[0.803, 1.061]	[0.792, 1.046]	-
Bootstrap	[0.820, 1.045]	[0.791, 1.045]	-

5.2 Three PK parameters: AUC_{0-t}, $AUC_{0-\infty}$, and C_{max}

We extend our analysis and consider $AUC_{0-\infty}$ in addition to AUC_{0-t} and C_{max}, as required by the FDA U.S. Food & Drug (2003), and perform a genuinely trivariate analysis. Table 4 lists the marginal(ised) 90% simultaneous CIs obtained from using the methods described in Section 3 by projecting the boundaries onto the axes. The intervals for AUC_{0-t} and C_{max} are slightly wider than in the bivariate analysis (Table 3), and the interval bounds of $AUC_{0-\infty}$ are very similar to those of AUC_{0-t}, which is not much of a surprise given their correlation of $\rho = 0.973$. The conclusions regarding bioequivalence are the same as with the bivariate analysis.

The Casella-Hwang region now differs from the standard one in that it is shifted: the maximum likelihood estimate is $\widehat{\boldsymbol{\theta}} = (0.923, 0.934, 0.910)'$ whereas the James-Stein-type estimate $\delta^+(\widehat{\boldsymbol{\theta}}, s) = (0.936, 0.946, 0.925)'$ is "shrunken" towards $\mathbf{1}$ (in fact it is shrunken towards $\mathbf{0}$ on the log scale).

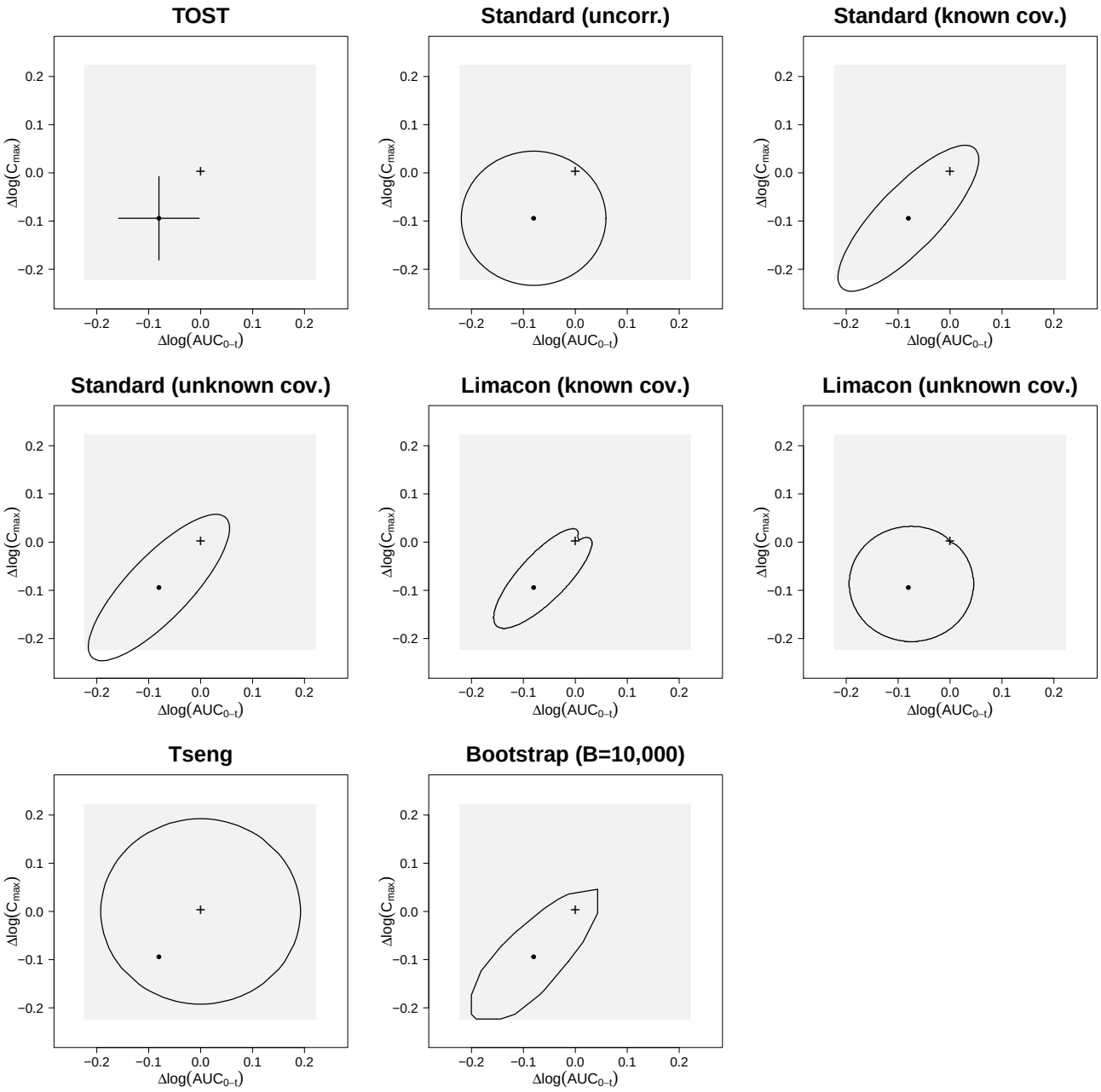

Figure 12: Ticlopidine hydrochloride data: Simultaneous 90% confidence intervals and regions for the differences (test minus reference) of the logarithms of AUC_{0-t} and C_{max}. The area of bioequivalence (80–125% for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$.

6 Discussion

In this paper we have reviewed a variety of simultaneous confidence regions for application in multi-parameter bioequivalence studies. Statistical methods for confidence sets around normal mean vectors have been published on several occasions during the past three decades, with different optimality criteria such as minimum expected volume Brown et al. (1995) or minimum expected effective length Tseng (2002) being proposed, but most of these papers focused on mathematical theory and disregarded practical issues. As a consequence, there are a number of methods that work well in academic scenarios e.g., asymptotically, or with known variance, or when the PK parameters are uncorrelated, but whose performance in more realistic settings has been largely unclear, and hence these developments have not made any impact on the way PK data are evaluated in practice.

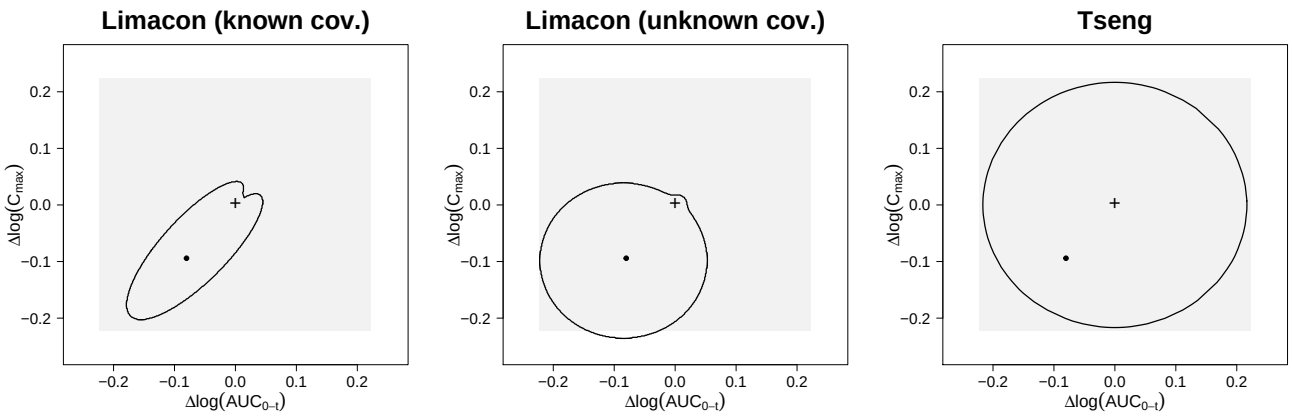

Figure 13: Ticlopidine hydrochloride data: Simultaneous 95% confidence intervals and regions for the differences (test minus reference) of the logarithms of AUC_{0-t} and C_{max}. The area of bioequivalence (80–125% for each PK parameter) is shaded grey. The dot indicates the estimate $\widehat{\boldsymbol{\theta}}$, the cross is at $\boldsymbol{\theta}_0 = \mathbf{0}$.

Table 4: Ticlopidine hydrochloride data: boundaries of various simultaneous 90% confidence regions projected onto the axes (and 90% TOST confidence intervals for comparison) for AUC_{0-t}, $AUC_{0-\infty}$, and C_{max}.

	AUC_{0-t}	$AUC_{0-\infty}$	C_{max}	Bioequivalent?
TOST	[0.834, 1.021]	[0.837, 1.042]	[0.813, 1.019]	✓
Standard (uncorrelated)	[0.783, 1.088]	[0.793, 1.101]	[0.772, 1.073]	-
Standard (known covariance)	[0.787, 1.082]	[0.786, 1.111]	[0.762, 1.087]	-
Standard (unknown covariance)	[0.784, 1.086]	[0.782, 1.116]	[0.758, 1.093]	-
Limaçon (known covariance)	[0.848, 1.049]	[0.847, 1.072]	[0.830, 1.052]	✓
Limaçon (unknown covariance)	[0.806, 1.062]	[0.809, 1.071]	[0.798, 1.059]	✓
Tseng	[0.818, 1.222]	[0.818, 1.222]	[0.818, 1.222]	✓
Casella & Hwang	[0.800, 1.096]	[0.808, 1.107]	[0.791, 1.083]	-

We have characterised properties of simultaneous confidence sets in situations that we consider relevant for real-world bioequivalence analyses: smallish sample sizes, unknown and potentially heterogeneous variances, and highly correlated PK measures. We have investigated the limaçon-type region with unknown covariance matrix for the first time; it turns out to have a non-convex boundary shape that resembles budding yeast cells, and it can have considerably larger volume than the limaçon with known covariance, as illustrated in Figures 4–7. The latter, however, achieves its nominal CP only for very large sample sizes when the covariance is unknown, although the deviations from the nominal CP are usually small. The Tseng region, unable to account for unknown variance and correlation, almost never maintains its nominal CP.

When it comes to assessing (bio-)equivalence, the use of multi-dimensional regions will hardly ever bring about a power gain as compared to the conventional TOST procedure. In our simulations we found not a single instance where two-dimensional confidence regions outperformed the TOST in terms of power. They are more informative than TOST CIs Douglas (1993), however, and may be worthwhile considering, perhaps as an additional tool, to study the probable location and variability of the estimate $\widehat{\boldsymbol{\theta}}$.

The joint alternative space defined by $[-\Delta, \Delta]$ for each PK parameter is (hyper-)rectangular, which is suboptimal from a power perspective for joint confidence regions. If the study target were to show overall equivalence, rather than equivalence at all PK parameters, a more powerful ellipsoidal alternative could be motivated as well Munk and Pflüger (1999); Hoffelder et al. (2015).

The limaçon-type regions may seem unsuitable for real-world applications due to their curious shapes. As a quick-and-dirty remedy, we suppose the convex hulls around these regions could be used as a (conservative) confidence regions that are often still smaller than the classical ellipse but perhaps easier to appreciate than the non-convex boundaries.

We have focused on the simultaneous assessment of $p = 2$ PK parameters because this is the most common case in practice. For $p \geq 3$ shrinkage estimation comes into play: then recentring the confidence set at a James-Stein-type estimate may have a greater impact on the inferential properties than a reduction in volume Casella and Hwang

(2012), not least because, as Efron Efron (2006) pointed out, "reduced volume by itself offers no guarantee of superior performance."

By the same token, a large volume does not necessarily entail a low probability of rejecting H_0, as we could see in our analysis of the ticlopidine hydrochloride data: the Tseng region has a much larger volume than its competitors but still leads to the conclusion of equivalence whereas much smaller regions do not (Figure 12). This example data analysis also illustrated the potential advantage of the limaçon-type confidence regions: they are entirely within the bioequivalence region and therefore allow to claim bioequivalence whereas the standard regions do not—even though the estimated mean $\widehat{\boldsymbol{\theta}} = (-0.080, -0.094)$ is not particularly close to $\boldsymbol{\theta}_0 = \mathbf{0}$. We are not sure, however, whether 90% regions are appropriate for the asymmetric limaçon and Tseng methods when the goal is to limit α at 5%. The FDA's and EMA's principal guidelines on bioequivalence U.S. Food & Drug (2001); European Medicines (2010) are not particularly helpful here either: they demand that 90% CIs be constructed, but they also state explicitly that these CIs correspond to hypothesis tests of size 0.05. Berger and Hsu Berger and Hsu (1996) have clarified that this slightly peculiar relationship holds for equi-tailed CIs like those from the TOST, but not in general. So if we employ methods that produce non-equi-tailed CIs (by projection onto the axes) or otherwise asymmetric regions, or use "expanded" $100(1-\alpha)\%$ rather than $100(1-2\alpha)\%$ CIs for the TOST, and want to control the type I error rate at level 0.05, then the joint CP to be aimed at should probably be 95% rather than 90%.

The limaçon-type regions yield substantially smaller confidence sets near $\boldsymbol{\theta}_0$ than the other methods. On the other hand, the limaçon and Tseng regions can become ginormous if $\boldsymbol{\theta}$ is far away from $\boldsymbol{\theta}_0$ but this is not too critical in practice because in that case (bio-)equivalence is unlikely anyway. The performance of the limaçon-type regions and also the Tseng region critically depends on the choice of $\boldsymbol{\theta}_0$. Setting $\boldsymbol{\theta}_0 = \mathbf{0}$ is the obvious thing to do in the absence of any further prior knowledge. If however there is information available e.g., from a previous or pilot study, one can hope to reduce the regions' volume by choosing $\boldsymbol{\theta}_0$ according to this prior information.

A limitation of several methods is that they involve unrealistic assumptions about $\boldsymbol{\Sigma}$, which will hardly ever have the form $\sigma^2\mathbf{I}$ and be known for real-world PK data. Our simulation study showed that incorporating the correlation is vital, and failing to do so may result in either a conservative or liberal procedure (Figures 8 and 9). To overcome this problem, Casella and Hwang Casella and Hwang (1983) suggested to transform the data $\mathbf{x}$ into

$$\mathbf{x}^* = \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{x}$$

where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Lambda}$ and σ^2 may be unknown. The new $\mathbf{x}^*$ is uncorrelated and has unit variance. This idea seems appealing at first sight, but the crux is that we must still assume $\boldsymbol{\Lambda}$ to be known. In practice, we wish to retransform the obtained CI boundaries $\mathbf{l}^*$ and $\mathbf{u}^*$ to get interpretable confidence limits

$$\mathbf{l} = \mathbf{l}^*\boldsymbol{\Lambda}^{\frac{1}{2}} \qquad \text{and} \qquad \mathbf{u} = \mathbf{u}^*\boldsymbol{\Lambda}^{\frac{1}{2}}.$$

This works out in theory and asymptotically but the performance with realistically small sample sizes can be poor.

As the assessment of multi-parameter bioequivalence is a multivariate problem, we recommend treating it as such by studying joint confidence regions for the joint parameter vector $\boldsymbol{\theta}$ rather than marginal (simultaneous) CIs for individual θ_i's, at least as an (additional) exploratory tool. In principle, the boundary of a joint region can be used directly for inference with respect to pre-defined margins $[-\Delta, \Delta]$, and there is no inherent need to derive univariate intervals. When CIs for the single PK parameters are nonetheless desired, the easiest way of translating a simultaneous confidence region into marginal parameter-specific confidence limits is by projecting the region's boundary onto the axes, but this makes the marginalised simultaneous CIs conservative Nickerson (1994). Superior solutions (such as direct interval construction) are available for spheres and ellipsoids, but rather hard to imagine for the limaçon-type regions as they are both non-convex and asymmetric.

To facilitate the multivariate analysis of bioequivalence data, we provide an R package `jocre` Pallmann (2017) with functions to draw simultaneous confidence regions and calculate marginal (simultaneous) CIs.

Due to our focus on bioequivalence in this paper, we only covered methods that comply with the all-or-nothing criterion where success is defined by *all* PK parameters being equivalent, as required by the major regulatory agencies. Outside the framework of PK analysis, however, there are also applications where equivalence of some but not all endpoints is interesting. The step-up procedure of Quan *et al.* Quan et al. (2001) yields simultaneous CIs for this case, and for one-sided problems of non-inferiority we refer to Hasler and Hothorn Hasler and Hothorn (2013).

Acknowledgements

This report is independent research arising from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. The authors were also supported by the MRC Network of Hubs for Trials Methodology Research (MR/L004933/1-R/N/P/B1) and the MRC North West Hub for Trials Methodology Research (MR/K025635/1).

We are grateful to Dominic Magirr who provided R code for the limaçon region. We would also like to thank the Associate Editor and an anonymous referee for some very helpful suggestions, especially with regard to the simulation study.

References

W. Abeysekera and P. Kabaila. Optimized recentered confidence spheres for the multivariate normal meanl. *Electron J Stat*, 11(1):1798–1826, 2017.

R. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982. doi:10.2307/1267823.

R. Berger and J. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat Sci*, 11(4): 283–319, 1996. doi:10.1214/ss/1032280304.

E. Bofinger. Expanded confidence intervals. *Commun Stat Theory Methods*, 14(8):1849–1864, 1985. doi:10.1080/03610928508829017.

E. Bofinger. Expanded confidence intervals, one-sided tests, and equivalence testing. *J Biopharm Stat*, 2(2):181–188, 1992. doi:10.1080/10543409208835038.

L. Brown, G. Casella, and J. Hwang. Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *J Am Stat Assoc*, 90(431):880–889, 1995. doi:10.2307/2291322.

L. Brown, J. Hwang, and A. Munk. An unbiased test for the bioequivalence problem. *Ann Stat*, 25(6):2345–2367, 1997. doi:10.1214/aos/1030741076.

G. Casella and J. Hwang. Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *J Am Stat Assoc*, 78(383):688–698, 1983. doi:10.2307/2288139.

G. Casella and J. Hwang. Shrinkage confidence procedures. *Stat Sci*, 27(1):51–60, 2012. doi:10.1214/10-STS319.

W. Cawello, editor. *Parameters for Compartment-Free Pharmacokinetics: Standardisation of Study Design, Data Analysis and Reporting*. Shaker Verlag, Aachen, Germany, 1999. ISBN 978-3-8265-4767-5.

I. Chervoneva, T. Hyslop, and W. Hauck. A multivariate test for population bioequivalence. *Stat Med*, 26(6):1208–1223, 2007. doi:10.1002/sim.2605.

V. Chew. Confidence, prediction, and tolerance regions for the multivariate normal distribution. *J Am Stat Assoc*, 61 (315):605–617, 1966.

V. Chinchilli and R. Elswick Jr. The multivariate assessment of bioequivalence. *J Biopharm Stat*, 7(1):113–123, 1997. doi:10.1080/10543409708835173.

M. Davidian and D. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC, Boca Raton, FL, 1995. ISBN 0-412-98341-9.

A. Davison and D. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, New York, NY, 1997. ISBN 978-0-521-57391-2.

J. Douglas. Confidence regions for parameter pairs. *Am Stat*, 47(1):43–45, 1993. doi:10.2307/2684784.

B. Efron. Minimum volume confidence regions for a multivariate normal mean vector. *J R Stat Soc Series B Stat Methodol*, 68(4):655–670, 2006. doi:10.1111/j.1467-9868.2006.00560.x.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993. ISBN 978-0-412-04231-2.

A. European Medicines. Guideline on the Investigation of Bioequivalence, Jan. 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf. Accessed July 27, 2017.

P. Ghosh and M. Gönen. Bayesian modeling of multivariate average bioequivalence. *Stat Med*, 27(13):2402–2419, 2008. doi:10.1002/sim.3160.

M. Hasler and L. Hothorn. Simultaneous confidence intervals on multivariate non-inferiority. *Stat Med*, 32(10): 1720–1729, 2013. doi:10.1002/sim.5633.

K. He. Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Stat Decis*, 10(1–2):121–132, 1992. doi:10.1524/strm.1992.10.12.121.

C. Health. Guidance Document: Conduct and Analysis of Comparative Bioavailability Studies, Feb. 2012. http://www.hc-sc.gc.ca/dhp-mps/alt_formats/pdf/prodpharma/applic-demande/guide-ld/bio/gd_cbs_ebc_ld-eng.pdf. Accessed July 27, 2017.

T. Hoffelder, R. Gössl, and S. Wellek. Multivariate equivalence tests for use in pharmaceutical development. *J Biopharm Stat*, 25(3):417–437, 2015. doi:10.1080/10543406.2014.920344.

H. Hotelling. The generalization of Student's ratio. *Ann Math Stat*, 2(3):360–378, 1931. doi:10.1214/aoms/1177732979.

J. Hsu. Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann Stat*, 12(3): 1136–1144, 1984. doi:10.1214/aos/1176346732.

J. Hsu, J. Hwang, K. Liu, and S. Ruberg. Confidence intervals assocaited with tests for bioequivalence. *Biometrika*, 81(1):103–114, 1994. doi:10.1093/biomet/81.1.103.

S. Hua, S. Xu, and R. D'Agostino Sr. Multiplicity adjustments in testing for bioequivalence. *Stat Med*, 34(2):215–231, 2015. doi:10.1002/sim.6247.

J. Hwang and Z. Zhao. Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *J Am Stat Assoc*, 108(502):607–618, 2013. ISSN 0162-1459. doi:10.1080/01621459.2013.771102.

J. Hwang, J. Qiu, and Z. Zhao. Empirical Bayes confidence intervals shrinking both means and variances. *J R Stat Soc Series B Stat Methodol*, 71(1):265–285, 2009. doi:10.1111/j.1467-9868.2008.00681.x.

T. Jaki and M. Wolfsegger. Non-compartmental estimation of pharmacokinetic parameters for flexible sampling designs. *Stat Med*, 31(11–12):1059–73, 2012. ISSN 1097-0258. doi:10.1002/sim.4386.

T. Jaki, P. Pallmann, and M. Wolfsegger. Estimation in AB/BA crossover trials with application to bioequivalence studies with incomplete and complete data designs. *Stat Med*, 32(30):5469–5483, 2013. doi:10.1002/sim.5886.

W. James and C. Stein. Estimation with quadratic loss. In , editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379. University of California Press;, Berkeley, CA, 1961.

A. Japan Generic Medicines. Guideline for Bioequivalence Studies of Generic Products, 2012. http://www.jga.gr.jp/english/wp-content/uploads/sites/4/2015/03/976d589c1980222c707651f7adac45d3.pdf. Accessed July 27, 2017.

B. Jones and M. Kenward. *Design and Analysis of Cross-Over Trials*. CRC Press, Boca Raton, FL, 3rd edition, 2015. ISBN 978-1-4398-6142-4.

A. Källén. *Computational Pharmacokinetics*. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 978-1-4200-6065-2.

A. Marzo, L. Dal Bo, A. Rusca, and P. Zini. Bioequivalence of ticlopidine hydrochloride administered in single dose to healthy volunteers. *Pharmacol Res*, 46(5):401–407, 2002. doi:10.1016/S1043-6618(02)00084-1.

R. Molina de Souza, J. Achcar, and E. Martinez. Use of Bayesian methods for multivariate bioequivalence measures. *J Biopharm Stat*, 19(1):42–66, 2009. doi:10.1080/10543400802513676.

J. Müller-Cohrs. An improvement over the Westlake symmetric confidence interval. *Biom J*, 33(3):357–360, 1991. doi:10.1002/bimj.4710330319.

A. Munk and R. Pflüger. $1–\alpha$ equivariant confidence rules for convex alternatives are $\alpha/2$-level tests—with applications to the multivariate assessment of bioequivalence. *J Am Stat Assoc*, 94(448):1311–1319, 1999. doi:10.1080/01621459.1999.10473883.

A. Munk, J. Hwang, and L. Brown. Testing average equivalence – finding a compromise between theory and practice. *Biom J*, 42(5):531–552, 2000. doi:10.1002/1521-4036(200009)42:5<531::AID-BIMJ531>3.3.CO;2-Y.

D. Nickerson. Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *Am Stat*, 48(2):120–124, 1994. doi:10.2307/2684261.

P. Pallmann. *jocre: Joint confidence regions. R package version 0.3.3;* , 2017. URL `http://cran.r-project.org/package=jocre`.

S. Patterson and B. Jones. *Bioequivalence and Statistics in Clinical Pharmacology*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-530-6.

K. Phillips. Power for testing multiple instances of the two one-sided tests procedure. *Int J Biostat*, 5(1):article 15, 2009. doi:10.2202/1557-4679.1169.

J. Qiu and J. Hwang. Sharp simultaneous confidence intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63(3):767–776, 2007. ISSN 0006341X. doi:10.1111/j.1541-0420.2007.00770.x.

H. Quan, J. Bolognese, and W. Yuan. Assessment of equivalence on multiple endpoints. *Stat Med*, 20(21):3159–3173, 2001. doi:10.1002/sim.985.

T. R Core. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; , 2015.

R. Samworth. Small confidence sets for the mean of a spherically symmetric distribution. *J R Stat Soc Series B Stat Methodol*, 67(3):343–361, 2005. ISSN 13697412. doi:10.1111/j.1467-9868.2005.00505.x.

D. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokin Biopharm*, 15(6):657–680, 1987. doi:10.1007/BF01068419.

S. Sheather and M. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Series B Methodol*, 53(3):683–690, 1991.

G. Stefansson, W. Kim, and J. Hsu. On confidence sets in multiple comparisons. In S. Gupta and J. Berger, editors, *Statistical Decision Theory and Related Topics IV, Volume 2*, pages 89–104. Springer;, New York, NY, 1988. ISBN 978-1-4612-8365-2.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In , editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206. University of California Press;, Berkeley, CA, 1956.

C. Tsai, C. Huang, and J. Liu. An approximate approach to sample size determination in bioequivalence testing with multiple pharmacokinetic responses. *Stat Med*, 33(19):3300–3317, 2014. doi:10.1002/sim.6182.

Y. Tseng. Optimal confidence sets for testing average bioequivalence. *Test*, 11(1):127–141, 2002. doi:10.1007/BF02595733.

Y. Tseng and L. Brown. Good exact confidence sets for a multivariate normal mean. *Ann Stat*, 25(5):2228–2258, 1997. doi:10.1214/aos/1069362396.

A. U.S. Food & Drug. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence, Jan. 2001. `http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf`. Accessed July 27, 2017.

A. U.S. Food & Drug. Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products – General Considerations, Mar. 2003. `https://www.fda.gov/ohrms/dockets/ac/03/briefing/3995B1_07_GFI-BioAvail-BioEquiv.pdf`. Accessed July 27, 2017.

A. U.S. Food & Drug. Guidance for Industry (Draft): Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA, Dec. 2013. `http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm377465.pdf`. Accessed July 27, 2017.

A. U.S. Food & Drug. Guidance for Industry (Draft): Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs – General Considerations, Mar. 2014. `http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm389370.pdf`. Accessed July 27, 2017.

M. Wand. *KernSmooth: Functions for kernel smoothing supporting Wand & Jones (1995). R package version 2.23.15;* , 2015. URL `http://cran.r-project.org/package=KernSmooth`.

M. Wand and M. Jones. *Kernel Smoothing*. Chapman & Hall, London, UK, 1995. ISBN 978-0-412-55270-1.

W. Wang, J. Hwang, and A. DasGupta. Statistical tests for multivariate bioequivalence. *Biometrika*, 86(2):395–402, 1999. doi:10.1093/biomet/86.2.395.

M. Wolfsegger and T. Jaki. Non-compartmental estimation of pharmacokinetic parameters in serial sampling designs. *J Pharmacokin Pharmacodyn*, 36(5):479–494, 2009. ISSN 1567567X. doi:10.1007/s10928-009-9133-9.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.