Biometrical Journal **42** (2000) 5, 531–552

Testing Average Equivalence — Finding a Compromise Between Theory and Practice

AXEL MUNK

Ruhr-Universität Bochum
Fakultät für Mathematik
Germany

J. T. GENE HWANG

Department of Statistics
Cornell University
Ithaca
USA

LAWRENCE D. BROWN

Department of Statistics
The Wharton School of Buisiness
Philadelphia
USA

Summary

Recently, BROWN, HWANG, and MUNK (1998) proposed and unbiased test for the average equivalence problem which improves noticeably in power on the standard two one-sided tests procedure. Nevertheless, from a practical point of view there are some objections against the use of this test which are mainly adressed to the 'unusual' shape of the critical region. We show that *every* unbiased test has a critical region with such an 'unusual' shape. Therefore, we discuss three (biased) modifications of the unbiased test. We come to the conclusion that a suitable modification represents a good compromise between a most powerful test and a test with an appealing shape of its critical region. In order to perform these tests figures are given containing the rejection region. Finally, we compare all tests in an example from neurophysiology. This shows that it is beneficial to use these improved tests instead of the two one-sided tests procedure.

Key words: Bioequivalence problem; Average equivalence; Unbiased testing; Two one-sided tests procedure; Exteroceptive suppression; Foundations of statistics.

1. Introduction

Since the pioneering work of WESTLAKE (1972, 1974, 1976, 1979, 1981), METZLER (1974) and many others bioequivalence assessment has become a broadly

applied tool in pharmaceutical research because it allows one to prove statistically the *similarity* (rather than the difference) of two formulations of a drug. Typically, a test formulation (T) (i.e. a new dosage form) of the active ingredient is to be compared with a reference formulation (R) (which is usually the original manufacturer's formulation already on the market). As recommended by the FDA (1992) and other drug administrations (WHO, 1987) the experiment should be conducted using a two period cross-over design in order to guarantee a small variability and lack of bias. The rate and extent of absorption of the active ingredient are assessed by pharmacokinetic parameters such as $C_{\max}$ (maximum concentration of the ingredient) or *AUC* (Area Under the blood concentration Curve).

The data are assumed to follow a lognormal distribution. Therefore, a logarithmic transformation is applied to each individual measurement (such as *AUC*) and hence the transformed data is normally distributed with means μ_T or μ_R corresponding to the test or reference formulation, respectively. This logarithmic transformation reduces the problem involving a ratio of *means* in the original scale to a problem involving the difference $\mu_T - \mu_R$. The interest is then in testing

$$H : |\mu_T - \mu_R| > \Delta \quad \text{versus} \quad K : |\mu_T - \mu_R| \leq \Delta \tag{1}$$

where e.g. the tolerance limit $\Delta = \log 1.25$ for the *AUC* is widely accepted by drug authorities in order to guarantee that at least 80% and no more than 125% of of the ingredient is absorbed in the same time. In the mean time the criterion of average bioequivalence (1) has been criticised by various authors (cf. ENDRENYI (1995), ENDRENYI and SCHULZ (1993), HAUCK and ANDERSON (1992), HOLDER and HSUAN (1993), LIU and CHOW (1994), SCHALL (1995), SHEINER (1992) or WELLEK (1993) among many others) and different bioequivalence criteria (various types of population and individual bioequivalence) have been suggested, which have been recognized as more reasonable for bioequivalence assessment by many authors. This is highlighted in a recent draft guidance for industry entitled "Average, Population and Individual Approaches to Establishing Bioequivalence" (U.S. Dep. of Health and Human Services, Food and Drug Administration, CDER, Rockville, MD, 1999).

In this paper, however, we focus only on the testing problem (1) because in the mean time it has received great interest in various other fields where the assessment of equivalence is of interest (see e.g. the comment about the potential use of equivalence tests in various medical applications made by Hauck and Anderson in BERGER and HSU (1996), ROGERS, HOWARD, and VESSEY (1993) in psychology, ROY (1997) in chemistry, MCBRIDE (1998) for an application in environmental statistics and the data example in Section 3 for an application in neurophysiology.)

In the following we discuss in more detail how this problem is mathematically equivalent to one in which the observed data are $(D, \hat{\sigma}^2)$ where D is independent of $\hat{\sigma}^2$ and

$$D \sim N(\theta, \sigma^2) \quad \text{and} \quad \nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu\,, \tag{2}$$

i.e. D is distributed according to a normal distribution with mean θ and variance σ^2 and $\nu\hat{\sigma}^2/\sigma^2$ is distributed according to a central χ^2 distribution with ν degrees of freedom. In this formulation a simple scale transformation (devide each datum by Δ) of the data allows us to always choose $\Delta = 1$. Therefore, the bioequivalence problem is to test the hypothesis

$$H : |\theta| > 1 \quad \text{versus} \quad K : |\theta| \leq 1\,, \tag{3}$$

where $\theta = \mu_T - \mu_R$. This is made explicit in Section 2. Observe, that this implies that D is an unbiased estimator for θ and $\hat{\sigma}^2$ is an unbiased estimator for the variance of D.

Throughout this paper we denote $\nu\hat{\sigma}^2 = S^2$. The degrees of freedom ν and the variance σ^2 depend on the particular choice of the experimental design and the sample size [for a careful discussion of various designs see CHOW and LIU (1992)]. Note that every two sample ANOVA model can be reduced to this setting, including the two-period cross-over design. The most important special cases are discussed explicetly in the following.

Model I (2×2-*crossover*): In the particular case of a two period crossover design we observe (CHOW and LIU, 1992, p. 34)

$$\text{sequence 1} \begin{cases} Y_{iR1} = S_{i1} + \mu_R + \pi_1 + \varepsilon_{iR1} \\ Y_{iT1} = S_{i1} + \mu_T + \pi_2 + \varepsilon_{iT1} \end{cases}$$

$$\text{sequence 2} \begin{cases} Y_{iT2} = S_{i2} + \mu_T + \pi_1 + \varepsilon_{iT1} \\ Y_{iR2} = S_{i2} + \mu_R + \pi_2 + \varepsilon_{iR2} \end{cases}$$

where Y_{ijk} is the response of the ith subject in the kth sequence for the jth formulation in which $j = R, T$, $k = 1, 2$, $i = 1, 2, \ldots, n_k$, μ_j is the fixed effect for the jth formulation, π_1 and π_2 are fixed period effects with $\pi_1 + \pi_2 = 0$, S_{ik} is the random subject effect, ε_{ijk} is the intrasubject random error in observing Y_{ijk}. Observe that in sequence 2 R and T are administered in reversed order, the subsequent definition of V_{ik}, however, does not take this into account. To reduce this setting to the general notation above in (2), set

$$V_{ik} = (Y_{iRk} - Y_{iTk})/2\,, \qquad \bar{V}_{.k} = 1/n_k \sum_{i=1}^{n_k} V_{ik}\,,$$

$$D = V_{.1} + V_{.2}\,, \qquad \hat{\sigma}^2 = \frac{\dfrac{1}{n_1} + \dfrac{1}{n_2}}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (V_{i1} - \bar{V}_{.1})^2 + \sum_{i=1}^{n_2} (V_{i2} - \bar{V}_{.2})^2 \right).$$

Under the assumption that ε_{ijk} are independently normally distributed with zero mean and common variance σ_ε^2, D and $\hat{\sigma}^2$ satisfy (2) with $\nu = n_1 + n_2 - 2$ and $\sigma^2 = \dfrac{1}{2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\sigma_\varepsilon^2$, $\theta = \mu_R - \mu_T$.

Model 2 *(parallel group design)*: Another example of interest is a parallel group design where

$$X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} N(\mu_R, \sigma_0^2), \qquad Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} N(\mu_T, \sigma_0^2).$$

Here

$$D = \bar{X} - \bar{Y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \right).$$

Again $\sigma^2 = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \sigma_0^2$ and $\nu = n_1 + n_2 - 2$. Note, that whenever in bioequivalence testing crossover effects cannot be excluded, e.g. when the drug has a rather long half-life, this design is more appropriate [cf. CHOW and LIU (1992), DETTE and MUNK (1997)].

Model 3 *(prepost design)*: Here we observe n independent outcomes

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} N((\mu_T, \mu_R)^t, \Sigma),$$

where $D_i = X_i - Y_i$, $i = 1, \ldots, n$,

$$D = \frac{1}{n} \sum_{i=1}^{n} D_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (D_i - D)^2$$

and hence $\nu = n - 1$. Here $\sigma^2 = \frac{1}{n} (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})$.

The FDA-guidance (1992) recommends to conduct the two one $-$ sided tests procedure (SCHUIRMANN, 1987). Therefore, in this paper we call this test the standard (bioequivalence) test which is sometimes also refered to as a double-t test because its rejection region is the intersection of two α-level one sided t-rejection regions for the two null hypotheses $H_1 : \theta > \Delta$ and $H_2 : \theta < -\Delta$, respectively. MÜLLER-COHRS (1990) and others [see MUNK (1993) or HSU et al. (1994)] noted that this test is biased, especially as the underlying variance σ^2 increases.

The question whether a (nontrivial) unbiased test for the bioequivalence problem exists was implicitly raised in a paper by HODGES and LEHMANN (1954). Recently, this problem was solved by BROWN et al. (1998) under some minor constraints (cf. Section 2) on the degrees of freedom and the nominal level α. For a discussion of the remaining cases cf. MUNK (1999a, b). Since their rejection region contains that of the standard test [cf. Theorem 1 of BROWN et al. (1998)] the unbiased test is always more poweful than the standard test. Moreover, a numerical comparison of the power functions shows that this test sometimes exhibits a noticeable improvement on the double t test (see also Figure 2 in Section 2). Unfortunately, the shape of the critical region of the unbiased test has some properties which do not seem reasonable from a practical point of view.

In the following, we summarize the objections of many colleagues to use of the unbiased test. We mention that these are objections against the use of this *test*, not

objections against the formulation of the *testing problem*, as briefly discussed in the Introduction. In particular, when this test was presented by Hwang in an invited session at the ASA Annual meeting on Bioequivalence, 1993 in San Francisco and by Munk at the Conference on Statistical and Regularory Issues on Bioequivalence, 1995 in Düsseldorf the subsequent discussions motivated us to recommend a modification of the unbiased test procedure which will be presented in the next Section of this paper. In summary, it was suggested that the following five requirements concerning the shape of the critical region should be satisfied:

- **R1.** Every average equivalence test $\varphi(D, S)$ should be **symmetric** in D, i.e. $\varphi(D, S) = \varphi(-D, S)$, where φ denotes the critical function of the test.
- **R2.** Every average equivalence test should be D-**homogeneous**, i.e. the critical region conditioned on $S = s$ should be a (possibly degenerate or unbounded) interval in D.
- **R3.** Every average equivalence test should be S-**homogeneous**, i.e. the critical region conditioned on $D = d$ should be a (possibly degenerate or unbounded) interval in S.
- **R4.** The critical region of an average equivalence test should be D-**bounded**, i.e. for any S, the D-coordinate should be entirely contained in an interval $[-d_0, d_0]$, where $d_0 < \infty$ is a prespecified bound. A good choice for d_0 is $d_0 = \Delta$, which means that we never decide for equivalence whenever the magnitude of the UMVU-estimator D is larger than the equivalence limit. In this case, the rejection region is said to be Δ-bounded.
- **R5.** Finally, the test should be S-**bounded**. This means that the hypothesis should not be rejected for s too large. This requirement is intended to force the experimenter to keep small the variability of the experiment.

We will comment on these properties.

R1. Without any doubt, the restriction to symmetric tests is very natural from two points of view. Firstly the testing problem (3) remains invariant with respect to the group of reflections at $\theta = 0$ which induces the same invariance property for $\varphi(\cdot, S)$. Secondly the decision of the test should remain invariant under relabelling of the treatments in any two sample design.

R2. For symmetric tests D-homogenity is equivalent to the following: If we declare average equivalence for some $D = d$ then we will do the same for all d's with smaller absolute value, when observing the same s. In particular, this property guarantees that whenever a test for the equivalence problem is similar it is also unbiased.

R3. If a test is not S-homogeneous, we allow for assessing average equivalence for $D = d$ when observing a large value of the standard deviation $s = \sqrt{\nu}\,\hat{\sigma}$, say s_3 or a small value s_1 while for an intermediate estimation of the standard deviation s_2 average equivalence cannot be concluded. To some colleagues this property seems unreasonable and undesirable.

R4. This requirement seems to be reasonable because otherwise the statistician is left in the paradoxical position of rejecting $H : |\theta| > \Delta$ while at the same time estimating a value θ by the value $|D| > \Delta$. See also SCHUIRMANN (1987) for a careful discussion of this property. Asymptotic theory will also require that when ν (and hence the sample size) approaches infinite every consistent (sequence) of test (s) will never allow for rejection for values of D, s.t. $|D| > \Delta$.

R5. It is argued, that if the variability is large (which is indicated by a large observed s) we reject the hypothesis only because we are allowed to exhaust the given type I error although there may not exist much evidence in the data to prefer the alternative. It can even be shown that there exists no equivalence test for (1) with a larger power than the nominal level α when the variance tends to infinity. In other words, in this case the information provided by the data becomes arbitrary small. Although plausible from the above reasoning, R5 is finally not shared by us. This will become clear in Section 3 where we illustrate in an example some unappealing implications of S-boundedness. Observe in particular, that the classical t-test could not be applied if we accept objections of type R5. For a similar argument see also BERGER and HSU (1996).

Tests which fulfill R1. and R2. will be called symmetric D-homogeneous. Certainly, these requirements are so convincing, that we do restrict our considerations in the following solely to these tests.

The aim of this paper is twofold. First, we prove that a symmetric D-homogeneous unbiased test *cannot* have *any* of the properties R3.–R5. (c.f. Theorem 2.1). Secondly, we suggest a (biased) modification of the unbiased test which fulfills the requirements R1.–R4. but sill noticeably improves on the standard test. This is then illustrated by an example.

The paper is organized as follows. In the next Section we present the unbiased test φ_u of BROWN et al. (1998) and three (biased) modifications. These tests are strictly ordered in the sense that their critical regions are decreasingly ordered, and hence their power functions are also descreasingly ordered. The test φ_u is uniformly more powerful than all the other tests. However, this test does not fulfill any of the requirements R3.–R5. The first modification φ_Δ of the unbiased test is Δ-bounded but neither S-homogeneous nor S-bounded. The second modification $\varphi_{\Delta S}$ is Δ-bounded and S-homogeneous but not S-bounded. Finally, we investigate a test $\varphi_{\Delta SS}$ which has all three properties. A numerical investigation shows that the loss in power of all these modified tests compared with the unbiased test is neglegible. The only exception is the S-bounded test, which improves only slightly on the standard test. Hence each of the other tests may improve significantly on the standard test. Nevertheless, it turns out that for large power values (such as $\beta = 0.8$ or 0.9) all tests almost the same power as the standard test. In particular, this indicates that we cannot hope to reduce the sample size for planning a powerful average equivalence study by using tests alternative to the standard procedure. Nevertheless, the use of one of these tests may become still favourable as is illus-

trated in an example in Section 3. In this example the standard test does not lead to the conclusion of equivalence although there is strong evidence of this because the difference of the means is estimated as nearly 0.

In summary, we recommend the use of the *S*-homogeneous but *S*-unbounded modification $\varphi_{\Delta S}$. This represents a good compromise between a test which is still powerful and one with an appealing critical region.

Those readers who are interested in performing the presented tests can obtain the SAS-IML code for the iterative construction of the critical region from the first author on request. Additionally, ascii-files containing the critical rejection region are available on the world-wide-web under HTTP://WWW.RUHR-UNI-BOCHUM.DE/MATHEMATIK3/MUNK.HTLM

2. The Unbiased Test and Modifications

The standard test. We start with a brief description of the standard test (SCHUIRMANN, 1987). This rejects the hypothesis of nonequivalence (and thus councludes that equivalence holds) whenever

$$\Delta \geq |D| + t_{1-\alpha}\hat{\sigma}\,, \tag{4}$$

where $t_{1-\alpha}$ denotes the upper $1-\alpha$-quantile of a central t-distribution F_ν with ν degrees of freedom. Sometimes this test is referred to as the confidence interval inclusion rule [MANDALLAZ and MAU (1981)] because we can represent the critical region (4) as those values of (D, S) for which the symmetric confidence interval $[D - t_{1-\alpha}\hat{\sigma},\ D + t_{1-\alpha}\hat{\sigma}]$ is entirely contained in the equivalence region $[-\Delta, \Delta]$. From this representation it is obvious that φ_t fullfills all the requirements R1.–R5. See BROWN, CASELLA, and HWANG (1994), HSU et al. (1994), BAUER and KIESER (1996) and BERGER and HSU (1996) for a general discussion of such rules. As mentioned in the introduction this test is always biased. In particular, the power function tends to zero for increasing variances, independently of the mean θ.

The unbiased test. BROWN et al. (1998, Sect. 4, 5) suggested an unbiased test φ_u. This was shown to be uniformly more powerful than the standard test. The unbiased test is *D*-homogeneous. Its critical region can be written as

$$C_u := \{(D, \hat{\sigma}) : |D/\Delta| \leq f(\hat{\sigma}/\Delta)\}$$

where $f(\cdot)$ denotes a function which has to be evaluated iteratively by numerically computing the solution to equation (4.11) in (BROWN et al., 1998). Note, that the construction of this unbiased test is only valid for large enough values of the nominal level α and degrees of freedom ν. For example, when $\alpha = 0.01,\ 0.025,\ 0.05,\ 0.1$ we require at least $\nu = 9,\ 6,\ 5,\ 3$ degrees of freedom [cf. Table 1 in BROWN et al. (1998)]. When we consider a two period crossover design (cf. model I) with equal sample sizes $n_1 = n_2 = n$ in each sequence these degrees of freedom

would correspond to a required sample size of $n = 6, 4, 4, 3$. Therefore, this constraint is certainly quite minor in any practical application. Note, that in the remaining cases for small degrees of freedom different constructions were suggested by MUNK (1992, 1999a, 1999b) and BERGER and HSU (1996) which improve on the standard test.

An algorithm of the construction of the unbiased test is carefully described in (BROWN et al., 1998) and an implementation in SAS-IML can be obtain from the authors on request. The numerical effort for the iterative computation of the boundary function f is considerable. Therefore, we display in Appendix B the boundary functions f of the critical region C_u for various degrees of freedom ν at a nominal level $\alpha = 0.05$, where $D \geq 0$. The region for negative D is obtained by reflection. We remind the reader that in these plots we are assuming $\Delta = 1$. To apply these figures in sitations involving other values of Δ the data must be appropriately rescaled. The rejection region is the region between the $\hat{\sigma}$-axis and the graph of the boundary function f (solid line). The dashed line denotes the boundary of the critical region of the standard test. These coincide with the critical region of the unbiased test (solid line) for small values of $\hat{\sigma}$. Observe, that the critical region of the unbiased test entirely contains the critical region of the standard test.

The figures involve only degrees of freedom ν with $\nu \leq 30$ because asymptotically as $\nu \to \infty$, a good approximation for the critical region of the unbiased test is

$$|F_\nu((D - \Delta)/\hat{\sigma}) - F_\nu((-D - \Delta)/\hat{\sigma})| \leq \alpha \tag{5}$$

which was suggested as a test for the bioequivalence problem by ANDERSON and HAUCK (1983). For small samples the actual level of this test exceeds the nominal level significantly. In the large sample case ($\nu \geq 30$) we draw from FRICK (1987) that the approximation of the nominal level is quite accurate. We found numerically in accordance with the table presented in FRICK (1987) that the size of this test satisfies

$$\sup_{\sigma > 0} \alpha_\nu(\sigma) \leq 0.055 \quad \text{if} \quad \nu \geq 30\,,$$

which is certainly acceptable for practical purposes. The critical region of the unbiased test can be approximated for large values of $\hat{\sigma}$ ($\hat{\sigma}^2 \geq 2\Delta^2$) by the straight line $\hat{\sigma} = |D|/t_{\nu,(1+\alpha)/2}$ which follows from the formulas (4.12) and (4.13) in BROWN et al. (1998). This allows us to finish the figures for large values of $\hat{\sigma}$.

In the following we will describe the above mentioned tests explicitely. All tests are performed as follows. For a given equivalence bound Δ for the absolute value $|\theta|$ of $\theta = E[D]$ start with a transformation of the observed $(D, \hat{\sigma})$ into $(|D|/\Delta, \hat{\sigma}/\Delta)$. Hence we may assume $\Delta = 1$ and in the following we denote these transformed values again as $(D, \hat{\sigma})$ as long as no confusion is possible. Then choose the degrees of freedom ν in accordance to the particular model (e.g. Model 1, 2 or 3) and apply the corresponding figure in Appendix B.

The Δ-bounded test. If one decides for the use of a Δ-bounded test we suggest to cut off the critical region where $|D| > 1$. (This part of region is not displayed in the figures because in this case we may use the asymptotical straight line $\hat{\sigma} = |D|/t_{\nu,(1+\alpha)/2}$). This describes the test φ_{Δ}.

The *S*-homogeneous test. We now construct $\varphi_{\Delta S}$. Consider for the moment only those cases where $\nu \geq 13$. Let D^* denote the smallest D on the boundary of the critical region of the unbiased test. Denote the corresponding value of $\hat{\sigma}$ as $\hat{\sigma}^*$ (cf. Figure 1).

To obtain the *S*-homogeneous test $\varphi_{\Delta S}$ simply remove the points in the rejection region of φ_u which are beyond the vertical line trough $(D, \hat{\sigma}^*)$. Hence, $\varphi_{\Delta S}$ equals φ_u if $\hat{\sigma} \leq \hat{\sigma}^*$ and

$$\varphi_{\Delta S} \text{ rejects if and only if } D \leq D^*, \text{ whenever } \hat{\sigma} > \hat{\sigma}^* .$$

Observe, that the *S*-homogeneous modification is also Δ-bounded.

The *S*-bounded test. Finally, we obtain an *S*-bounded test $\varphi_{\Delta SS}$ when the rejection region is additionally truncated at $\hat{\sigma} = \hat{\sigma}^*$. Hence this test equals $\varphi_{\Delta S}$ with the following exception:

$$\varphi_{\Delta SS} \text{ does not reject for } \hat{\sigma} > \hat{\sigma}^* .$$

Note that this test is also *S*-homogeneous and Δ-bounded. However, as mentioned in the Introduction we do not share the position of some colleagues that *S*-bound-

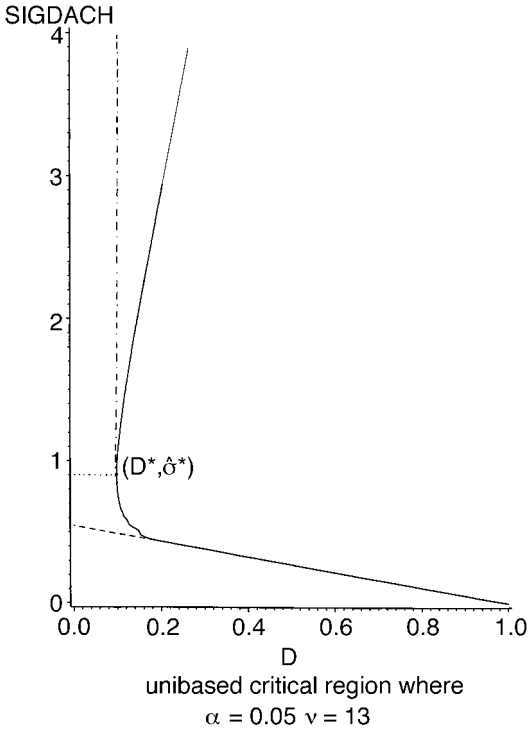


Fig. 1. The critical region of the standard test φ_t (dashed line), the unbiased test φ_u (solid line), the Δ-bounded test (not explicetely displayed) the *S*-homogeneous test $\varphi_{\Delta S}$ (– – . – – – . – –) and the *S*-bounded test $\varphi_{\Delta SS}$ (·····) in the case $\alpha = 0.05$, $\nu = 13$

edness is a necessary condition for an equivalence test to satisfy. This is in accordance with the reasoning of BERGER and HSU (1996). Note further that any *S*-bounded modification is somewhat subjective in determining the cut off point. For example, when $\nu = 13$, the corresponding *S*-homogeneous test $\varphi_{\Delta S}$ rejects according to the figure as long as $D \leq 0.1$ if $\hat{\sigma} > 0.85$. If $\hat{\sigma} < 0.85$ the rejection region remains the same as that of the unbiased test. Finally, the *S*-bounded test equals φ_u as long as $\hat{\sigma} < 0.85$ but never rejects for $\hat{\sigma} > 0.85$.

If $\nu \leq 12$ we found numerically (which is reflected by the figures in Appendix B) that the above described procedure does not necessarily lead to *S*-homogeneous modifications. Here an *S*-homogeneous modification can be obtained when one rejects in the union of the critical region of the standard test and the rectangle $|D| \leq D^*$. One could also use the lower envelope of the boundary of the unbiased test. However, in practical applications degrees of freedom ≤ 13 certainly occur very rarely.

We mentioned before that the unbiased test [as well as the test suggested by ANDERSON and HAUCK (1983)] although being symmetrical *D*-homogeneous, fails to satisfy the conditions R3.–R5. The following Theorem shows that each of these conditions even contradicts the property of unbiasedness.

Theorem 2.1: *Assume the setting (2) of the average equivalence problem (3).*

1. *No unbiased test for the average equivalence problem is S-bounded or D-bounded.*
2. *Any symmetrical unbiased D-homogeneous critical region with* $\alpha < 1/2$ *is not S-homogeneous.*

(For the proof see Appendix A).

The Theorem shows an interesting and perhaps suprising fact that unbiasedness – although reasonable from a formal statistical point of view – contradicts the requirements R3.–R5. The implications of this result are serious. The same comments would of course apply to any UMP or UMPU test if these optimal tests exist (which is still an open problem).

On the other hand there is no reason to prefer the two one sided *t* tests procedure when better tests are available which fullfill R3.–R5. Recall, that all modifications suggested above are uniformly more powerful than the standard procedure. We have performed an extensive numerical study where the power functions of all these tests are evaluated for various degrees of freedom ν and levels α by means of a Gaussian quadrature formula as is described in (BROWN et al., 1998, Sect. 3), which was controlled by a Monte Carlo simulation with 10000 replications in each setting. The power function was evaluated at $\theta = 0, 0.1, \ldots, 1.5$ and the figures were obtained using cubic spline interpolation. Note, that for the numerical calculation of the power and hence of the required sample size when planning an experiment it is convenient to write the power function of any of these tests as

$$P_{\theta,\sigma^2}(|D| \leq f(S)) = \int_0^\infty \left\{ \Phi\left(\frac{f(s)-\theta}{\sigma}\right) - \Phi\left(\frac{-f(s)-\theta}{\sigma}\right) \right\} P^S_{\sigma,\nu}(\mathrm{d}s)\,.$$

Here $P^{S}_{\sigma,\nu}$ denotes the distribution of S with density of $\mathbb{R}^+$

$$g_{\nu,\sigma}(s) = \frac{(1/2)^{\nu/2-1}}{\Gamma(\nu/2)} \exp\left(-\frac{s^2}{2\sigma^2} + (\nu-1)\log s - \nu \log \sigma\right)$$

and Φ denotes the cdf. of a standard normal random variable.

To be brief here we display in Figure 2 only the power of the tests standard test, the unbiased test the S-bounded test and the S-homogeneous test as a function of $|\theta|$ for the case $\nu = 7,\ 12,\ 22$ and $\alpha = 0.05$ at $\sigma = 1/3,\ 1/2,\ 2/3,\ 3/4$ where $\Delta = 1$. Figures for the Δ-bounded test are supressed (because its shape and power is very closed to the unbiased test). Power figures of this test can also be

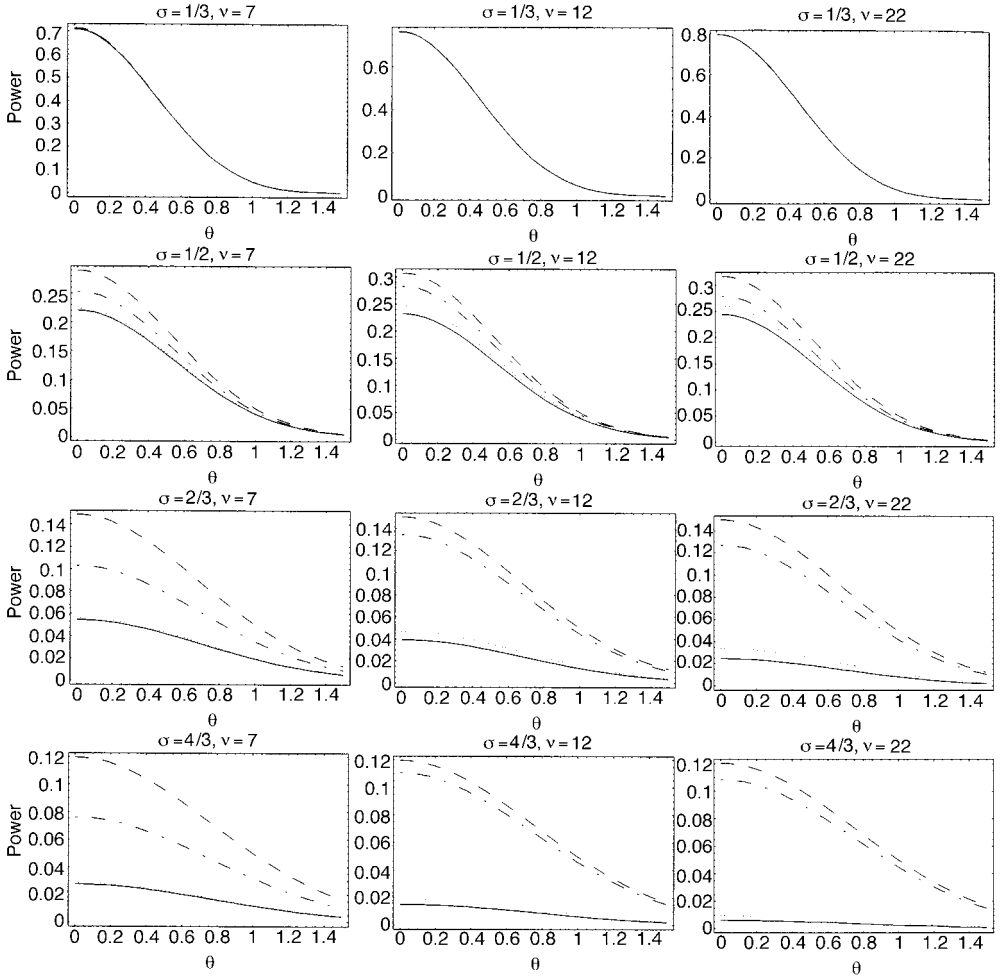


Fig. 2. The power functions of the unbiased trest (dashed line $- - -$) the S-homogeneous modification (dashed/dotted line $- \cdot -$), the standard test (solid line ——) and the S-bounded test (dotted $\cdot \cdot \cdot \cdot \cdot$) where $\nu = 7,\ 12,\ 22$, $\alpha = 0.05$ and $\sigma = 1/3,\ 1/2,\ 2/3,\ 3/4$. The x-axis is labeled by $|\theta|$

found in BROWN et al. (1998). Note, that the power surface of each test $\beta(\theta, \sigma)$ is always unimodal in the first coordinate with maximum at $\theta = 0$ and a decreasing function in the second coordinate.

Our numerical results may be summarized as follows. Provided $\sigma/\Delta \ll 0.4$ (here always $\Delta = 1$) we find that all tests have approximately the same power. When σ/Δ becomes larger the maximum deviation between the power function of the unbiased test and that of the standard test increases where we observed a maximum difference of 0.12 at $\sigma = 3/4$, $\nu = 22$. We could not find larger difference for other degrees of freedom and variances σ^2. For variances σ^2 tending to infinity the unbiased test improves on the standard test by an amount of α because the power of the last named test converges to 0. The *S*-bounded test $\varphi_{\Delta S}$ is always very close to the standard test, whereas the *S*-homogeneous test improves significantly in power on the standard test provided σ is not too small. Here, we have chosen the above described modification of the *S*-bounded test which is also *S*-homogeneous. Recall, that for ν small, this is not necessarily the case. Of course, larger choices of truncation lead to more powerful tests. Note, that the actual level and the power of all test, besides of the unbiased test, tends to 0 if $\sigma \to \infty$. Finally, it is interesting to note that the parameter σ (actually σ/Δ, but here $\Delta = 1$) serves as a rather good indication of the amount of power for any of these tests, independently of ν. Recall that σ^2 is the variance of D, the unbiased estimator of θ and hence depends on the sample size via the particular model (cf. model I–III in Section 1).

Discussion. These observations suggest to prefer the *S*-homogeneous (and Δ-bounded) test $\varphi_{\Delta S}$ because it fullfills requirements R1.–R4. and its power improves significantly on that of the standard test for those parameters of the alternative where the variance is not too large or not too small. Moreover, the additional improvement in power which wouls result from using the unbiased test or the Δ-bounded test is suprisingly small. Hence we are not in the difficult situation that we are led to use this test (recall that its critical region violates R3.–R5.) by means of its great superiority in power.

Sometimes it is argued that the obtained power improvement by the unbiased test or one of its modifications such as $\varphi_{\Delta S}$ on the standard test is irrelevant. The basis of this argument is the assertion that the power function β of a well planned experiment should be rather large, say $\beta > 0.8$, because only in these cases we are interested in the required sample size in order to guarantee a preassigned probability of type II error. But in this realm the decision of φ_t and $\varphi_{\Delta S}$ say, will agree with high probability and the tests will have almost identical power. However, in practice experiments are sometimes performed in which the power is smaller than this ideal and in which φ_t and the above discussed tests may well differ in power. This can for example result from poor overall planning or from properly delineated planning which is, however, based on incorrect a-priori estimates of the relevant parameter or in situations where time or cost compel experiments with

smaller power than this ideal. The following example describes a recent experiment in which the modifications of the unbiased test notably differ from the standard test. Moreover, this examples illustrates that it is not always reasonable to measure the improvement on a test solely by means of comparing the power functions.

3. Example

The essential drawback of the standard test procedure is illustrated by the following example. Observe, that this transfers directly to the bioequivalence setting.

STEINHOFF, FANGMEIER, and PAULUS (1995) investigated the influence of epilepsy on exteroceptive suppression (ES) of temporalis (chewing) muscle activity. They compared the ES of muscle activity of 31 epileptic patients and 20 normal controls where measurements were taken from each subject at the left and right temporalis, respectively. It was conjectured that ES is *not* a suitable method to discriminate between epileptic and non − epileptic subjects. Therefore, it was considered as appropriate to perform an equivalence test.

The results are displayed in Table 1. Here $\bar{x}$ denotes the sample mean and *sde* the sample standard deviation in each group for the left and right muscle, respectively. Also let sde_T denote the observed standard deviation of the data obtained by averaging the measurements of the right and left temporalis of each person. See Table 1 for the summary statistics. We were particularly interested in the following:

I. We want to compare the response of normal and epileptic persons based on the averaged data of the right and left chewing muscle, respectively. Comparing the standard deviations in each group justifies the assumption of homogenity of the variances. Assuming Model 2 in the Introduction we obtain the pooled total standard deviation as 10.972 (not displayed in the table) and estimate the mean difference as $d = \bar{z}_R - \bar{z}_E = 42.94 - 42.195 = 0.745$. Here $\nu = n_1 + n_2 - 2 = 49$. Therefore we may apply the approximation in (5). To compare the tests discussed in the last section we evaluate for each test the smallest equivalence bound Δ^* for which the hypothesis $H : |\theta| > \Delta^*$ is rejected given the outcomes (d, sde_T), i.e.

$$\Delta^*(d, s_T) := \inf\left\{\Delta : \varphi\left(d/\Delta, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2} sde_T/\Delta\right) = 1\right\}.$$

Table 1

Means and standard deviations of the muscle activities at the right and left temporalis of epileptic and reference patients

	right nerve		left nerve			total mean
mean/st. dev.	$\bar{x}$	*sde*	$\bar{x}$	*sde*	sde_T	
Reference ($n_1 = 20$)	43.10	9.23	42.78	10.33	9.255	42.94
Epileptic ($n_2 = 31$)	42.93	11.88	41.46	12.18	11.67	42.195

We obtain from (4) that $\Delta^* = 6.02$ for the standard test whereas the unbiased test (and the modified tests φ_Δ, $\varphi_{\Delta S}$, $\varphi_{\Delta SS}$) allow to conclude equivalence when $\Delta^* = 5.17$, which is $\leq 10\%$ of the range of the estimated means.

II. It was also of interest to show that there is no difference in ES between the right and left temporalis in the control group. Observe, that this corresponds to model 2 in Section 1. Therefore we evaluate the estimated mean difference as $\bar{x}_{R(r)} - \bar{x}_{R(l)} = d = 0.32$ and we draw from Table 1 the pooled standard deviation as $s_T = 9.255$. Further $\nu = n_1 - 1 = 19$. Now we obtain for the standard test $\Delta^* = 3.9$ and the other tests allow for assessing equivalence at the bound $\Delta^* = 2.97$. The situation becomes more drastic when d trends to zero, e.g. assume that we had observed $d = 0.1$. Then we would obtain $\Delta^* = 3.68$ for the standard test but for the unbiased test $\Delta^* = 0$. This reflects that the unbiased test is in fact not Δ-bounded. The S-bounded test leads to $\Delta^* = 2.31$ and the S-homogeneous test gives $\Delta^* = 1$. Here a reduction of the equivalence limit of more than 75% can be obtained by using the S-homogeneous test instead of the standard test!

4. Conclusions

The observations in the last example can be explained as follows. Whenever the estimated standard error is too large (which happens with high probability when σ is large, of course), the S-bounded test and the standard test φ_t never reject H even for $d = 0$ (cf. case II in the last example). In part II of Example 3 it was found that the sample standard deviation in each group is only $\approx 1/4$ of the *means* (and hence there is no indication that the true variance is unusually large, i.e. argument R4. of Section 2 does not apply) and we obtained for the scaled *differences* of the means $d/s \approx 1/30$. Therefore, in case II of the last example it becomes apparent that the data may give reason to decide for equivalence although the estimated standard deviation is rather large relative to the estimated *difference* average. In other words, in this example the standard deviation appears to be large because the mean difference was found to be unexpectedly small. Therefore, the S-bounded modification and the standard test were found to be insufficient.

In summary, we draw from this example that it is reasonable to seek to improve on the standard test even though we cannot achieve a noticeable difference in power for large values of the power function β, larger than 0.8 say. In particular, we found that one can reject for a much larger range of D values than the standard test without contradicting requirements R1.–R4. Taking into account this observation and the discussion in Section III we finally recommed the S-homogeneous test $\varphi_{\Delta S}$ as a compromise between a test which allows for concluding the equivalence for large values of the sample variance whenever D/Δ is small but with an appealing critical region from a practical point of view.

As pointed out by a referee good statistical practice should entail more than reporting on a pure test decision (whatever the underlying test is). Therefore, it is considered an advantage of the *TOST* that it automatically forces us to compute a $(1-2\alpha)$-confidence interval. Furthermore, *P*-value are easy to obtain.

We mention that the same holds for the *S*-homogeneous test, too. Interestingly, a confidence interval can be obtained by inverting the acceptance region (cf. LEHMANN (1986) for a description of this method) of the *S*-homogeneous test. To this end let $A_s(\Delta)$ denote the acceptance region of the *S*-homogeneous test conditioned on $S = s$ for a fixed boundary value Δ of the hypothesis. Then

$$C(D) = \{\theta : D \in A_s(\theta),\ |\theta| \geq 0\}$$

defines a confidence interval around D with confidence level $1-\alpha$. We mention that it can be seen easily that this confidence interval is uniformly shorter than the $1-2\alpha$ confidence interval associated with the *TOST*. To this end simply recall, that the rejection region of the *TOST* is entirely in that of the *S*-homogeneous test for any Δ. Furthermore, note that given $S = s$ the resulting interval is simply connected and symmetric around D due to the *S*-homogeneity. This property fails to hold for the confidence interval associated with the unbiased test, which reflects again the unapealing shape of this test. We will, however, not pusue this topic here and postpone a more detailed discussion to a subsequent paper.

Acknowledgements

We are indebted to B. J. Steinhoff, B. Fangmeyer, B. Paulus and J. Baudewig at the Department of Clinical Neurophysiology, Göttingen University for providing the ES-data at our disposal. Parts of this paper were written while A. Munk was visiting The Wharton School of Buisiness, Philadelphia, PA and Cornell University, Ithaca, NY. A. Munk acknowledges the Deutsche Forschungsgemeinschaft for making this visit possible. We would like to thank L. Pralle for computational assistance. Further, we acknowledge very helpful comments of three referees which led to an improved version of an earlier draft of this paper.

Appendix A: Proofs

We start with a technical Lemma which will be used in the proof of Theorem 2.1. Let $\bar{C}$ denote the closure of a set $C \subset \mathbb{R}^2$, i.e. the smallest closed subset containing C.

Lemma A.1.: *Let C be the critical region of an unbiased test at level* $\alpha > 0$ *for the bioequivalence problem. Then we have* $(\Delta, 0),\ (-\Delta, 0) \in \bar{C}$.

Proof: Assume $(\Delta, 0) \notin \bar{C}$. Then there exists an open, nonempty neighborhood U of $(\Delta, 0)$ with the property

$$U \cap C = \emptyset\,. \tag{6}$$

This leads to

$$\lim_{\sigma \to 0} P_{\theta,\sigma}(U) = \delta_{(\theta,0)}(U) = 1\,, \quad \text{for} \quad \theta = \Delta\,,$$

where $\delta_x(\cdot)$ denotes the dirac measure at x. Finally, we obtain from (6) that $\lim_{\sigma \to 0} P_{\Delta,\sigma}(C) = 0$ which contradicts the unbiasedness.

Lemma A.2.: Assume that C is S-homogeneous and symmetric D-homogeneous. If (d_1, s_1) and $(d_2, s_2) \in C$, s.t. $d_1 \le d_2$, $s_1 \le s_2$, then the rectangle

$$\{(d, s) : 0 \le d \le d_1,\, s_1 \le s \le s_2\} \subset C\,.$$

Proof: By symmetry and D-homogeneity it follows that the range of D for a given $S = s$ is empty or it forms an interval containing 0. Hence the line segment joining $(0, s_2)$ and (d_2, s_2) is included in C. This implies that (d_1, s_2) is in C. Finally by S-homogeneity, the line segment joining (d_1, s_2) and (d_1, s_1) is in C. By D-homogeneity again, the rectangle is in C.

Proof of Theorem 2.1:

1.) We shall first show that an unbiased rejection region C cannot be S_0-bounded. Otherwise as $\sigma \to \infty$

$$P_{\theta,\sigma}(C) \le P_{\theta,\sigma}(s \le s_0) = P_{\theta,\sigma}\left(\chi^2 \le \frac{s_0^2}{\sigma^2}\right) \to 0$$

where χ^2 is a χ^2 random variable with v degrees of freedom. This shows that the rejection probability at $\theta = \Delta$ cannot be α, contracting the unbiasedness of C.

Similarly, C cannot be D_0-bounded. Otherwise as $\sigma \to \infty$

$$P_{\theta,\sigma}(C) \le P(|D| \le d_0) = \Phi\left(\frac{d_0 - \theta}{\sigma}\right) - \Phi\left(-\frac{d_0 + \theta}{\sigma}\right) \to 0$$

where Φ denotes the cumulative distribution function of the standard normal distribution. This again contradicts the unbiasedness of C and hence C cannot be d_0-bounded.

2.) Under the assumptions that $0 < \alpha < \frac{1}{2}$ and C is unbiased symmetric D-homogeneous, we shall prove that C is not S-homogeneous. Suppose that C is S-homogeneous. We shall show that this leads to a contradiction. We work with $\bar{C}$ first, which can be shown by using Lemma A.2 to be D-homogeneous and S-homogeneous since C is. Also since C is not d-bounded, neither is $\bar{C}$. Since $\bar{C}$ is not D-bounded, there exists $s_* > 0$ and $d_* > \Delta$ such that $(d_*, s_*) \in C$. (Note that if s_1 is always zero for $d_1 > \Delta$, the region $\bar{C}$ is basically d-bounded almost surely and have zero probability as $\sigma \to \infty$ by an argument similar to 1.). This implies that C also has zero probability as $\sigma \to \infty$, contradicting to the assumption that C is unbiased.) Also from Lemma A.1, $(\Delta, 0) \in \bar{C}$. Now by Lemma A.2,

$$\{(d, s) : 0 \le d \le \Delta,\, 0 \le s \le s_*\} \subset \bar{C}\,,$$

which implies that $(\Delta, 0)$, $(\Delta, s*) \in \bar{C}$. Returning to C, there exist two sequences (d_n, s_n) and (d'_s, s'_n) in C such that they approach $(\Delta, 0)$ and $(\Delta, s*)$ respectively, i.e., as $n \to \infty$

$$(d_n, s_n) \to (\Delta, 0)$$

and

$$(d'_n, s'_n) \to (\Delta, s*)\,.$$

Lemma (A.2.) then implies that

$$\{(d, s) : 0 \le d \le \min(d_n, d'_n)\,, \quad \text{and} \quad \min(s_n, s'_n) \le s \le \max(s_n, s'_n)\} \subset C\,.$$

Letting $n \to \infty$ concludes that

$$\{(d, s) : 0 \le d < \Delta \quad \text{and} \quad 0 < s < s^*\} \subset C\,,$$

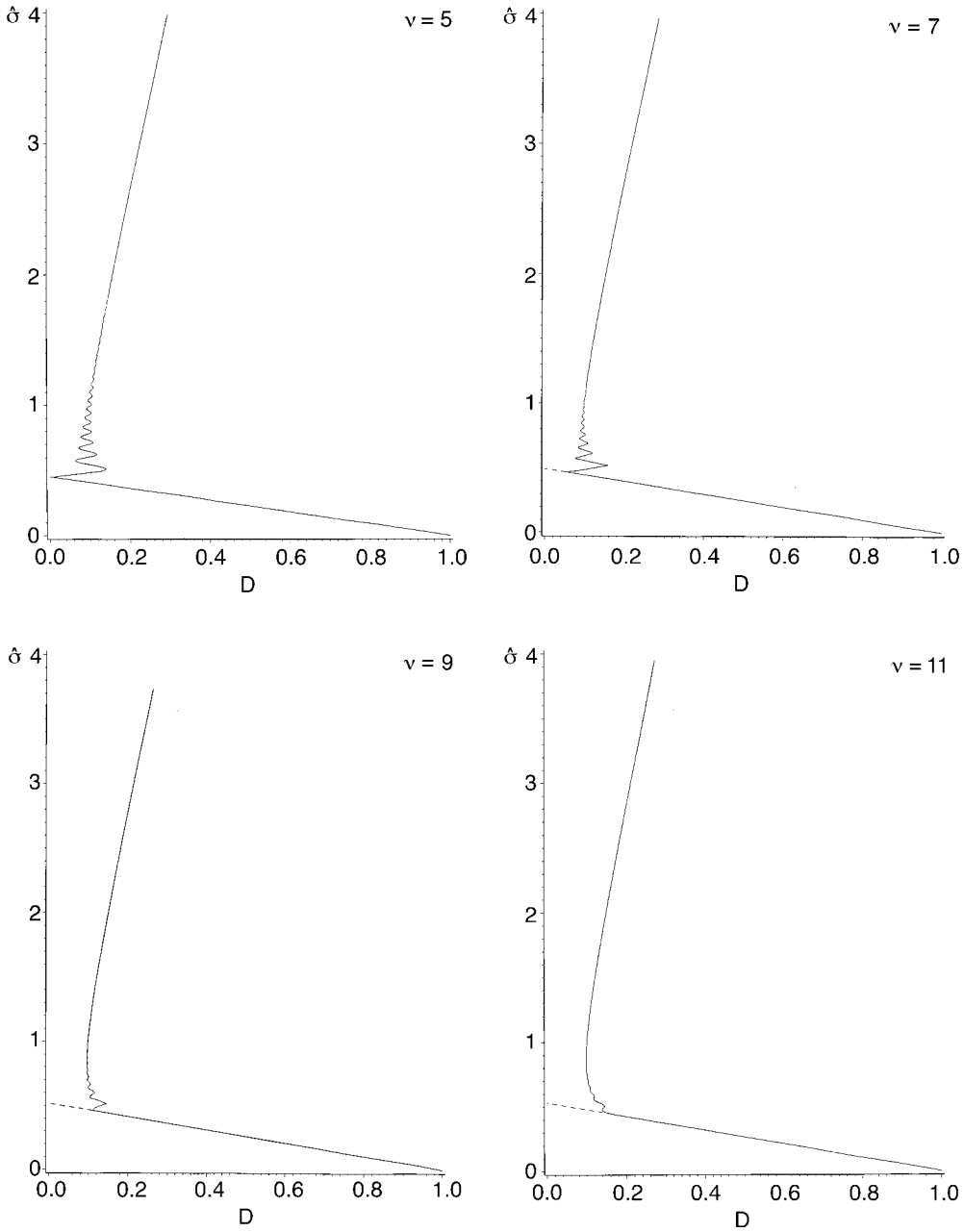
and hence

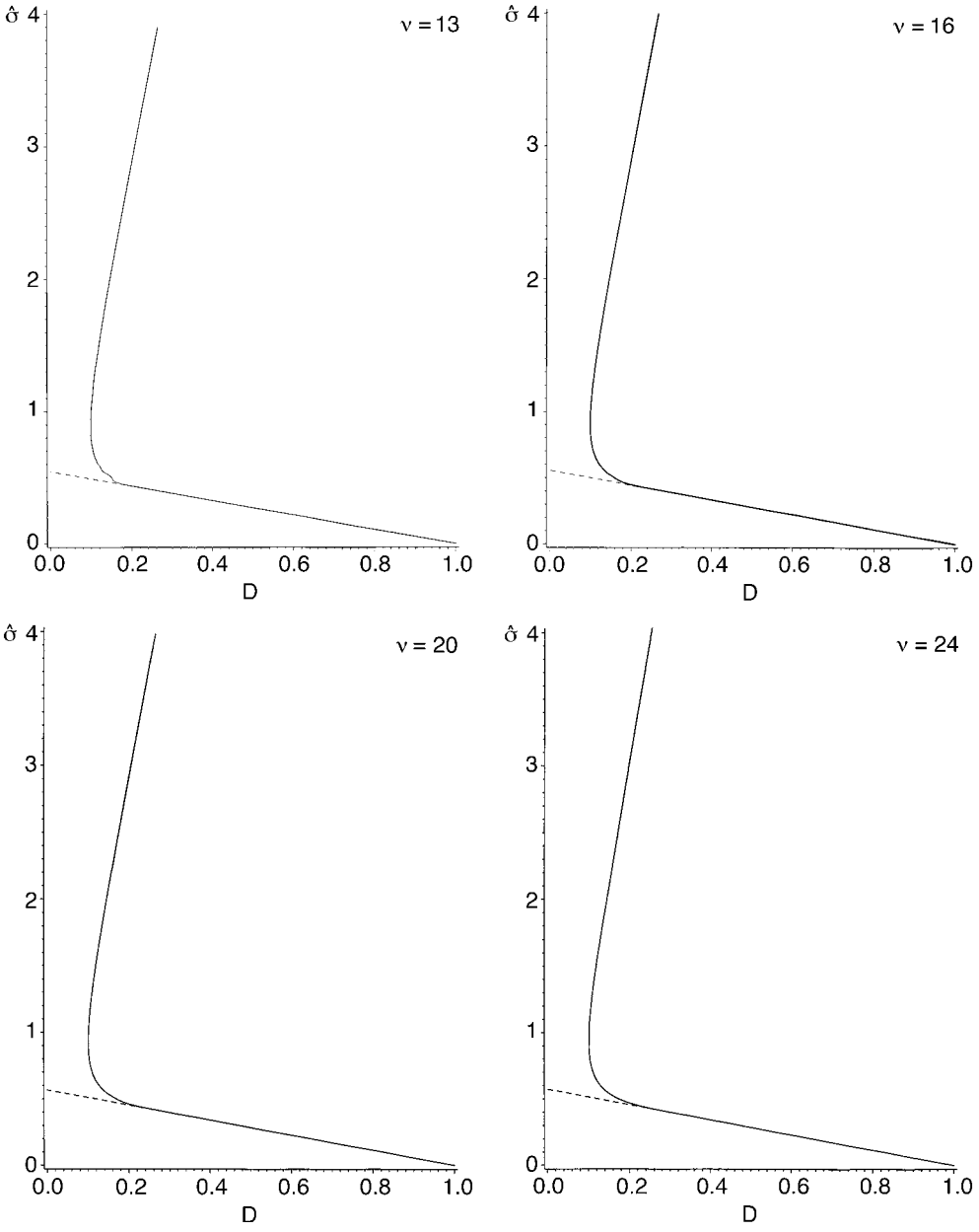
$$\begin{aligned} P(C) &\ge P(0 \le D < \Delta)\, P(0 < s < s^*) \\ &= P(0 \le D \le \Delta)\, P(0 \le s < s^*) \to \frac{1}{2} \quad \text{as} \quad \sigma \to 0\,. \end{aligned}$$

This contradicts to the assumption that $\alpha < \frac{1}{2}$.

Appendix B

The critical regions of the unbiased test for $\alpha = 0.05$ and $\Delta = 1$ in the $(D, \hat{\sigma})$-plane where $\nu = 5, 7, 9, 11, 13, 16, 20, 24$





References

ANDERSON, S. and HAUCK, W., 1983: A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun. Statist.-Theor. Meth.* **12**, 2663–2691.

BAUER, P. and KIESER, M., 1996: A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83**, 934–937.

BERGER, R. L. and HSU, J. C., 1996: Bioequivalence trials, intersection union tests and equivalence confidence sets. With discussion. *Statistical Science* **11**, 283–319.

BROWN, L. D., CASELLA, G., and HWANG, J. T. G., 1995: Optimal confidence sets, bioequivalence and the limacon of Pascal. *Journ. Americ. Statist. Assoc.* **90**, 880–889.

BROWN, L. D., HWANG, J. T. G., and MUNK, A., 1998: An unbiased test the bioequivalence problem. *Annals of Statistics* **25**, 2345–2367.

CHOW, S. C. and LIU, J. P., 1992: *Design and Analysis of Bioequivalence Studies*. STATISTICS: textbooks and monographs. Marcel Dekker, Inc.

DETTE, H. and MUNK, A., 1997: Optimal allocation of sample size in Welch's test for equivalence assessment. *Biometrics* **53**, 1143–1150.

ENDRENYI, L., 1995: A simple approach for the evaluation of individual bioequivalence. *Drug. Inf. Journ.* **29**, 847–855.

ENDRENYI, L. and SCHULZ, M., 1993: Individual variation and the acceptance of average bioequivalence. *Drug Inf. Journ.* **27**, 195–201.

FDA, 1992: *Draft recommendation on statistical procedures for bioequivalence studies using the standard treatment cross over design*. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD.

FRICK, H., 1987: On level and power of Anderson and Hauck's procedure for testing equivalence in comparative bioavailability. *Commun. Statist.-Theor. Meth.* **16**, 2771–2778.

HAUCK, W. W. and ANDERSON, S., 1992: Types of bioequivalence and related statistical considerations. *Int. J. Clin, Pharmacol. Ther. Toxicol.* **30**, 181–187.

HOLDER and HSUAN, 1993: Moment based criteria for determining bioequivalence. *Biometrika* **80**, 835–846.

HODGES, J. L. and LEHMANN, E. L., 1954: Testing the approximative validity of statistical hypotheses. *Journal of Royal Statistical Society* B **16**, 261–268.

HSU, J. C., HWANG, J. T. G., LIU, H.-K., and RUBERG, S. J., 1994: Confidence intervals associated with tests for bioequivalence. *Biometrika* **811**, 103–114.

LEHMANN, E., 1986: *Testing Statistical Hypotheses*, 2nd edition. Wiley, New York.

LIU, J. P. and CHOW, S. C., 1994: A two one-sided tests procedure for assessment of individual bioequivalence. *Statistics in Medicine to appear.*

MANDALLAZ, D. and MAU, J., 1981: Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **37**, 213–222.

MCBRIDE, G., 1998: Equivalence tests can enhance environmental science and management. *Austral. and New Zealand Journ. Statist.* **41**, 19–29.

METZLER, C. M., 1974: Bioavailability: A problem in equivalence. *Biometrics* **30**, 309–317.

MUNK, A., 1992: Äquivalenztests in Exponentiellen Familien. *Unpublished Thesis, Department of Mathematics, Göttingen University.*

MUNK, A., 1993: An improvement on commonly used tests in bioequivalence assessment. *Biometrics* **49**, 1225–30. (See also the correspondence in *Biometrics* **50**, 884–886 (1994)).

MUNK, A., 1999a: An unbiased test for the bioequivalence problem – the the small sample case. *Journ. Statist. Plann. Inference* to appear.

MUNK, A., 1999b: An unbiased test for the equivalence problem – another christmas tree. *Statist. & Prob. Lett.* **41**, 401–406.

MÜLLER-COHRS, J., 1990: The power of the Anderson-Hauck test and the double *t*-test. *Biometrical Journal* **32**, 259–266.

ROGERS, J. L., HOWARD, K. I., and VESSEY, J. T., 1993: Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin* **113**, 553–565.

ROY, T., 1997: Calibrated nonparametric confidence sets. *Journ. Mathemat. Chemistry* **21**, 103–109.

SCHALL, R., 1995: Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics* **51**, 615–626.

SCHUIRMANN, D. L., 1987: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacocinetics and Biopharmaceutics* **15**, 657–680.

SHEINER, L., 1992: Bioequivalence revisited. *Statist. Medicine* **11**, 1777–1788.

STEINHOFF, B. J., FANGMEYER, B., PAULUS, W., 1995: Exteroceptive suppression of temporalis nuscle activity in epilepsy. *Technical Report, Department of Clinical Neurophysiology, Georg August University, Göttingen.*

WELLEK, S., 1993: Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. *Biom. Journal* **35**, 47–55.

WESTLAKE, W. J., 1972: Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of the Pharmaceutical Sciences* **61**, 1340–41.

WESTLAKE, W. J., 1974: The use of balanced incomplete block designs in comparative bioavailability trials. *Biometrics* **30**, 319–327.

WESTLAKE, W. J., 1976: Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **32**, 741–744.

WESTLAKE, W. J., 1979: Statistical aspects of comparative bioavailability trials. *Biometrics* **35**, 319–327.

WESTLAKE, W. J., 1981: Bioequivalence testing – a need to rethink. (Reader reaction response). *Biometrics* **37**, 591–593.

WHO, 1986: Guidelines for the Investigation of Bioavailability. *Copenhagen*.

AXEL MUNK
Ruhr-Universität Bochum
Fakultät für Mathematik
Gebäude NA 3/74
Universitätsstraße 150
44780 Bochum
Germany
Email: Axel.Munk@ruhr-uni-bochum.de

Received, October 1998
Revised, November 1999
Revised, January 2000
Accepted, February 2000