Shao, Kübler, Pigeot:

Consistency of the Bootstrap Procedure in Individual Bioequivalence

Projektpartner

CONSISTENCY OF THE BOOTSTRAP PROCEDURE IN INDIVIDUAL BIOEQUIVALENCE

BY JUN SHAO,

University of Wisconsin, 1210 West Dayton St., Madison, WI 53706, USA

shao@stat.wisc.edu

JÜRGEN KÜBLER,

Bayer AG, PH-PD M CDPI/PE, D–42096 Wuppertal, Germany

juergen.kuebler.jk@bayer-ag.de

AND IRIS PIGEOT

Institute of Statistics, University of Munich, Ludwigstrasse 33, D–80539 Munich, Germany

pigeot@stat.uni-muenchen.de

SUMMARY

Recently, new concepts have been proposed for assessing bioequivalence of two drug formulations, namely the so–called population and individual bioequivalence. Using moment–based and probability–based measures for evaluating the proposed bioequivalence concepts, criteria have been formulated to decide whether two formulations should be regarded as bioequivalent or not. This decision has of course to be based on an adequate statistical method where the Food and Drug Administration (FDA) guidance (1997) recommends the use of a bootstrap percentile interval.

In this paper, we discuss theoretical properties such as consistency and accuracy of the recommended bootstrap intervals. We focus our investigations on the concept of individual bioequivalence and here especially on the scaled versions of the moment–based as well as the probability–based measures as recommended by the FDA. As estimates for the former, we consider those obtained from an according analysis of variance and restricted maximum likelihood estimators under mixed effect models, where an unbiased estimator of the latter can be derived from the corresponding relative frequencies.

Some key words: Bioequivalence; Bootstrap percentile; Consistency.

1. Introduction

Currently, there is a controversial discussion on the assessment of bioequivalence of two different drug formulations (cf. Hauschke & Steinijans, 1999; for an overview of regulatory requirements and scientific issues in bioequivalence trials see e.g. Chow & Liu, 1995). Due to a draft of the FDA guidance (1997) other concepts than that of the average bioequivalence are of concern. Whereas comparing the mean bioavailability, as for instance the area under concentration versus time curve (AUC) or the observed maximum concentration $(C_{\max})$, in two populations following from administration of a reference and a test formulation for assessing average bioequivalence is the current standard, the concepts recently proposed (see e.g. Anderson & Hauck, 1990; Chen, 1997) also account for the between–subject variances of the bioavailabilities and their within–subject variances, respectively. The first aspect is considered by the concept of population bioequivalence, which is of importance in the case that a patient receives a new drug for the first time which is referred to as prescribability (Hauck & Anderson, 1992, 1994). In contrast, by assessing the so–called individual bioequivalence it should be ensured, that two formulations are said to be bioequivalent if the responses to these formulations do not differ too much in the majority of patients, which is relevant under the aspect of switchability.

Both concepts require adequate measures for their assessment. Here, various proposals can be found in the literature (e.g. Eckbohm & Melander, 1989; Anderson & Hauck, 1990; Sheiner, 1992; Holder & Hsuan, 1993; Schall & Luus, 1993; Esinhart & Chinchilli, 1994; Schall, 1995; Chinchilli, 1996; Hwang & Wang, 1997; see also Schall & Williams, 1996; for an overview see e.g. Patnaik et al., 1997), where the discussion focusses on two main ideas. The first approach measures the discrepancy between the bioavailablities of two formulations, say a reference and a test formulation, comparing the expected squared differences of the bioavailabilities following the administration of the test and twice the reference formulation, denoted with y_T, y_R, y_R'. More formally, this moment–based approach considers two formulations as bioequivalent if

$$E(y_T - y_R)^2 - E(y_R - y_R')^2 < \theta_U. \tag{1}$$

The second idea is based on comparing the probabilities of the differences of the bioavailabilities not exceeding a predefined value r. Using this probability–based approach two formulations are regarded as bioequivalent if

$$\mathrm{pr}\,(|y_T - y_R| \le r) - \mathrm{pr}\,(|y_R - y_R'| \le r) > \theta_{L_1} \tag{2}$$

or analogously if

$$\frac{\text{pr}\,(|y_T - y_R| \leq r)}{\text{pr}\,(|y_R - y_R'| \leq r)} > \theta_{L_2}. \tag{3}$$

In (1) and (2) or (3), respectively, θ_U as well as r and θ_{L_1} or θ_{L_2} are fixed constants determined by drug regulatory authorities. Depending on which bioequivalence concept is of interest and also depending of course on the assumed statistical model for the bioavailabilites, the expectations and probabilities involved in the above formulae have to be derived. For the population bioequivalence, y_T, y_R, and y_R' in (1) – (3) are considered to be independent from different subjects, although y_R and y_R' are identically distributed. For the individual bioequivalence, y_T, y_R, and y_R' are considered to be from the same subject and, thus, are dependent.

The above moment– or probability–based criteria can also be used in scaled versions, i.e. scaled with respect to the variance of the reference formulation. A decision among the resulting four criteria should depend on the within–subject variability and the therapeutic range of a drug (Schall & Williams, 1996). For the moment–based approach, FDA (1997) recommended to use the following scaled version of (1):

$$\theta < \theta_U,$$

where

$$\theta = \begin{cases} \frac{E(y_R - y_T)^2 - E(y_R - y_R')^2}{E(y_R - y_R')^2/2} & \text{if } E(y_R - y_R')^2/2 \geq \sigma_0^2 \\[2em] \frac{E(y_R - y_T)^2 - E(y_R - y_R')^2}{\sigma_0^2} & \text{if } E(y_R - y_R')^2/2 < \sigma_0^2 \end{cases} \tag{4}$$

and σ_0^2 is a given constant, from which point on the scaled criterion has a wider bioequivalence range compared to its unscaled counterpart. This value has to be determined by drug regulatory authorities, where the FDA recommended a value of 0.2 for σ_0. For the probability–based approach, Schall (1995) considered to replace r in (2) or (3) by a constant times the squared root of $E(y_R - y_R')^2$.

For evaluating the bioequivalence criteria presented above the FDA recommended using bootstrap confidence intervals, especially a bootstrap percentile interval as proposed by Schall & Luus (1993). As alternative Schall (1995) derived a bias–corrected bootstrap interval. Several studies of the properties of the bootstrap intervals for assessing individual or population bioequivalence recently appeared (e.g. Shao, Chow & Wang, 1999). For an introduction to the bootstrap see Efron & Tibshirani (1993). A detailed discussion of the

theoretical behaviour of bootstrap intervals in general can be also found for instance in Shao & Tu (1995).

In the following, we focus on investigating bootstrap percentile intervals for the assessment of individual bioequivalence. With some minor modifications, our results and discussions can be applied to bootstrap percentile intervals for the assessment of population bioequivalence. In the next section, we show that under the moment–based approach with the scaled individual bioequivalence parameter defined in (4), the standard bootstrap percentile interval and the bootstrap percentile interval proposed in FDA (1997) are consistent under fairly weak conditions. Also, both bootstrap percentile intervals are found to be inconsistent if the model parameters vary in a subset of the parameter space with zero Lebesgue measure. Furthermore, the accuracy of these bootstrap percentile intervals and the bias-corrected and accelerated bias-corrected bootstrap percentile intervals is briefly discussed. The probability–based approach with scaled or non-scaled individual bioequivalence parameter is considered in § 3. We show that the bootstrap percentile interval described in Schall & Luus (1993) is generally inconsistent. The standard bootstrap percentile interval is then derived and its consistency is established.

2. Moment–based approach
2.1 Statistical model and hypothesis

In the draft of the FDA guidance of 1997, in general, a replicated crossover design of the bioequivalence study is recommended to obtain estimates of the above bioequivalence parameters. Here, for instance, a two by three crossover design is suggested for assessing individual bioequivalence. In the first sequence, n_1 patients receive three drug treatments in the order of TRR (where T=test formulation, R=reference formulation), whereas in the second sequence, n_2 patients receive three drug treatments in the order of RTR. The total number of patients is $n = n_1 + n_2$. Note that the bootstrap procedures and our discussion of their consistency can be easily modified to the cases where other types of two by three crossover designs (e.g., TRT and RTR) or two by four crossover designs (e.g., $TRTR$ and $RTRT$) are used.

For the ith patient in the jth period and kth sequence, let y_{ijk} be the observed pharmacokinetic response (bioavailability) or its logarithm. The following statistical model is

4

assumed:

$$y_{ijk} = \mu + F_l + P_j + Q_k + W_{ljk} + S_{ikl} + e_{ijk}, \tag{5}$$

where μ is the overall mean; P_j is the fixed effect of the jth period ($\sum_j P_j = 0$); Q_k is the fixed effect of the kth sequence ($\sum_k Q_k = 0$); F_l is the fixed effect of the lth drug formulation ($l = T$ when $(j, k) = (1, 1), (2, 2)$ and $l = R$ otherwise, $F_T + F_R = 0$); W_{ljk} is the fixed effect of interaction (sum of W_{ljk}'s over any index is 0); S_{ikl} is the random effect of the ith subject in the kth sequence under drug formulation l and (S_{ikT}, S_{ikR}), $i = 1, ..., n_k$, $k = 1, 2$, are independent identically distributed random vectors with mean 0 and an unknown covariance matrix

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix};$$

e_{ijk}'s are independent random errors with mean 0 and variance σ_{Wl}^2, and (S_{ikT}, S_{ikR})'s and e_{ijk}'s are independent. Note that σ_{BT}^2 and σ_{BR}^2 are between–subject variances and σ_{WT}^2 and σ_{WR}^2 are within–subject variances.

Under model (5), one can show that θ in (4) is equal to

$$\theta = \frac{(F_R - F_T)^2 + \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} + \sigma_{WT}^2 - \sigma_{WR}^2}{\max\{\sigma_0^2, \sigma_{WR}^2\}}, \tag{6}$$

that is, θ is a nonlinear function of $F_R - F_T$ and the variance components.

Let θ_U be given by (1). Assessing individual bioequivalence can be carried out by testing the following hypotheses

$$H_0 : \theta \geq \theta_U \qquad \text{versus} \qquad H_1 : \theta < \theta_U. \tag{7}$$

It can be claimed if and only if H_0 in (7) is rejected at the 5% significance level.

2.2 THE BOOTSTRAP TEST PROCEDURES

The following bootstrap test procedure is suggested in FDA's draft guidance.

Step 1: The individual bioequivalence parameter θ in (6) is estimated by

$$\hat{\theta} = \frac{(\hat{F}_R - \hat{F}_T)^2 + \hat{\sigma}_{BT}^2 + \hat{\sigma}_{BR}^2 - 2\hat{\rho}\hat{\sigma}_{BT}\hat{\sigma}_{BR} + \hat{\sigma}_{WT}^2 - \hat{\sigma}_{WR}^2}{\max\{\sigma_0^2, \hat{\sigma}_{WR}^2\}}, \tag{8}$$

where $\hat{F}_R$, $\hat{F}_T$, $\hat{\sigma}_{BT}$, $\hat{\sigma}_{BR}$, $\hat{\rho}$, $\hat{\sigma}_{WT}$, and $\hat{\sigma}_{WR}$ are estimated using analysis of variance or the method of restricted maximum likelihood under mixed effects models. This can be done by using for instance the SAS procedure PROC MIXED.

5

Step 2: Let $Y_{ik} = (y_{i1k}, ..., y_{i3k})$ and $Y_k = (Y_{1k}, ..., Y_{n_k k})$. For each fixed k, generate a sample random sample $Y_k^{*b} = (Y_{1k}^{*b}, ..., Y_{n_k k}^{*b})$ with replacement from Y_k. In other words, Y_k^{*b} is obtained using subjects as sampling units, stratified by sequence. Repeat this process $b = 1, ..., B$ times independently to obtain $Y_k^{*1}, ..., Y_k^{*B}$, $k = 1, 2$. These are called *bootstrap* samples. In FDA's draft guidance, at least $B = 2,000$ replications are suggested.

Step 3: For each $b = 1, ..., B$, compute $\hat{F}_R^{*b}$, $\hat{F}_T^{*b}$, $\hat{\sigma}_{BT}^{*b}$, $\hat{\sigma}_{BR}^{*b}$, $\hat{\rho}^{*b}$, $\hat{\sigma}_{WT}^{*b}$, and $\hat{\sigma}_{WR}^{*b}$ by using the same methods as those in step 1 but with the data set (Y_1, Y_2) replaced by the bootstrap data set (Y_1^{*b}, Y_2^{*b}), and let

$$\tilde{\theta}^{*b} = \begin{cases} \frac{(\hat{F}_R^{*b} - \hat{F}_T^{*b})^2 + (\hat{\sigma}_{BT}^{*b})^2 + (\hat{\sigma}_{BR}^{*b})^2 - 2\hat{\rho}^{*b}\hat{\sigma}_{BT}^{*b}\hat{\sigma}_{BR}^{*b} + (\hat{\sigma}_{WT}^{*b})^2 - (\hat{\sigma}_{WR}^{*b})^2}{(\hat{\sigma}_{WR}^{*b})^2} & \text{if } \hat{\sigma}_{WR}^2 \geq \sigma_0^2 \\ \\ \frac{(\hat{F}_R^{*b} - \hat{F}_T^{*b})^2 + (\hat{\sigma}_{BT}^{*b})^2 + (\hat{\sigma}_{BR}^{*b})^2 - 2\hat{\rho}^{*b}\hat{\sigma}_{BT}^{*b}\hat{\sigma}_{BR}^{*b} + (\hat{\sigma}_{WT}^{*b})^2 - (\hat{\sigma}_{WR}^{*b})^2}{\sigma_0^2} & \text{if } \hat{\sigma}_{WR}^2 < \sigma_0^2. \end{cases} \tag{9}$$

Step 4: Let $\tilde{\theta}_{FDA}(95)$ denote the 95th percentile of $\tilde{\theta}^{*b}$, $b = 1, ..., B$. Then individual bioequivalence can be claimed if and only if $\tilde{\theta}_{FDA}(95) < \theta_U$.

This procedure will be called FDA's bootstrap procedure.

If we replace $\tilde{\theta}^{*b}$ in (9) by

$$\hat{\theta}^{*b} = \frac{(\hat{F}_R^{*b} - \hat{F}_T^{*b})^2 + (\hat{\sigma}_{BT}^{*b})^2 + (\hat{\sigma}_{BR}^{*b})^2 - 2\hat{\rho}^{*b}\hat{\sigma}_{BT}^{*b}\hat{\sigma}_{BR}^{*b} + (\hat{\sigma}_{WT}^{*b})^2 - (\hat{\sigma}_{WR}^{*b})^2}{\max\{\sigma_0^2, (\hat{\sigma}_{WR}^{*b})^2\}}, \tag{10}$$

then the previously described procedure becomes the standard bootstrap percentile method (Efron & Tibshirani, 1993, § 13.3).

2.3 THE CONSISTENCY OF THE BOOTSTRAP INDIVIDUAL BIOEQUIVALENCE TESTS

Let

$$\boldsymbol{\gamma} = (F_R, F_T, \rho, \sigma_{BT}^2, \sigma_{BR}^2, \sigma_{WT}^2, \sigma_{WR}^2),$$

$$\hat{\boldsymbol{\gamma}} = (\hat{F}_R, \hat{F}_T, \hat{\rho}, \hat{\sigma}_{BT}^2, \hat{\sigma}_{BR}^2, \hat{\sigma}_{WT}^2, \hat{\sigma}_{WR}^2)$$

and

$$\hat{\boldsymbol{\gamma}}^{*b} = (\hat{F}_R^{*b}, \hat{F}_T^{*b}, \hat{\rho}^{*b}, (\hat{\sigma}_{BT}^{*b})^2, (\hat{\sigma}_{BR}^{*b})^2, (\hat{\sigma}_{WT}^{*b})^2, (\hat{\sigma}_{WR}^{*b})^2).$$

Then θ, $\hat{\theta}$, and $\hat{\theta}^{*b}$ can be written as $\theta = g(\boldsymbol{\gamma})$, $\hat{\theta} = g(\hat{\boldsymbol{\gamma}})$, and $\hat{\theta}^{*b} = g(\hat{\boldsymbol{\gamma}}^{*b})$, respectively, for a continuous function g.

Note that if $\sigma_{WR} \neq \sigma_0$, then the function g is differentiable in a neighborhood of $\boldsymbol{\gamma}$ and, hence, the consistency of the bootstrap procedure follows if

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \to_d N(0, \Sigma) \tag{11}$$

and

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}) \to_d N(0, \Sigma), \tag{12}$$

where $n = n_1 + n_2$, Σ is an asymptotic covariance matrix for $\sqrt{n}\hat{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\gamma}}^* = \hat{\boldsymbol{\gamma}}^{*b}$ for a fixed b.

If $\hat{\boldsymbol{\gamma}}$ is obtained from an analysis of variance, then $\hat{\boldsymbol{\gamma}}$ is a differentiable function of several sample means and (11) and (12) hold under some moment conditions. In case that the restricted maximum likelihood method is used, $\hat{\boldsymbol{\gamma}}$ is obtained by solving an equation

$$\sum_{i,k} \psi(Y_{ik}, \boldsymbol{\gamma}) = 0 \tag{13}$$

with a known differentiable vector–valued function ψ. We have the following result, where the proof is given in an appendix.

Lemma 1. Suppose that $\hat{\boldsymbol{\gamma}}$ is obtained by solving equation (13) and $\hat{\boldsymbol{\gamma}}^*$ is obtained by solving the same equation with $\{Y_{ik}\}$ replaced by the bootstrap sample $\{Y_{ik}^*\}$. Assume that

(i) $n_1/n_2 \to c \in (0, 1)$;

(ii) the function $\psi(\cdot, \cdot)$ is continuous and $\psi(y, \cdot)$ is continuously differentiable for each fixed y;

(iii) $E(\psi(Y_{ik}, \boldsymbol{\gamma})) = 0$ when $\boldsymbol{\gamma}$ is the true parameter value;

(iv) there is a $\delta > 0$ such that $E\|\psi(Y_{ik}, \boldsymbol{\gamma})\|^{2+\delta} < \infty$ and, for any compact neighborhood C of $\boldsymbol{\gamma}$, $E(h_C(Y_{ik}))^{1+\delta} < \infty$, where $h_C(y) = \sup_{\beta \in C} \|\partial \psi(y, \beta)/\partial \beta\|$;

(v) $\liminf_n \lambda[n^{-1}\text{var}(s_n(\boldsymbol{\gamma}))] > 0$ and $\liminf_n \lambda[-n^{-1}E(\nabla s_n(\boldsymbol{\gamma}))] > 0$, where $s_n(\beta) = \sum_{i,k} \psi(Y_{ik}, \beta)$, $\nabla s_n(\beta) = \partial s_n(\beta)/\partial \beta$, and $\lambda[A]$ is the smallest eigenvalue of the matrix A.

If $\{\hat{\boldsymbol{\gamma}}\}$ is a consistent sequence of solutions of equation (13) and $\{\hat{\boldsymbol{\gamma}}^*\}$ is a consistent sequence of solutions of the bootstrap analogs of (13), then (11) and (12) hold.

Since the conditions required for the function ψ are satisfied for the restricted maximum likelihood estimator with some moment conditions and if for instance the error terms e_{ijk} and the random effects S_{ikl} in model (5) are assumed as normal, the consistency of the bootstrap percentile test procedure is an immediate consequence of Lemma 1.

Theorem 1. Assume the conditions in Lemma 1. Let $\hat{\theta}_{BP}(95)$ be the 95th percentile of the bootstrap distribution of $\hat{\theta}^*$. If $\sigma_{WR} \neq \sigma_0$, then

$$\mathrm{pr}\left(\hat{\theta}_{BP}(95) > \theta\right) \to 0.95. \tag{14}$$

Proof. Note that $\theta = g(\boldsymbol{\gamma})$, $\hat{\theta} = g(\hat{\boldsymbol{\gamma}})$, $\hat{\theta}^{*b} = g(\hat{\boldsymbol{\gamma}}^{*b})$, and g is a continuously differentiable function in a neighborhood of $\boldsymbol{\gamma}$ if $\sigma_{WR} \neq \sigma_0$. The result follows from Lemma 1 and standard results for the bootstrap (e.g., Shao & Tu, 1995, § 3.1.5 and § 4.2.1).

We now turn to FDA's bootstrap procedure. Suppose that $\sigma_{WR} > \sigma_0$. For the consistency of the bootstrap tests, we only need to focus on the event $\{\hat{\sigma}_{WR}^2 \geq \sigma_0^2\}$, since $\mathrm{pr}\,(\hat{\sigma}_{WR}^2 \geq \sigma_0^2) \to 1$. Comparing (9) and (10), we conclude that for the event $\{\hat{\sigma}_{WR}^2 \geq \sigma_0^2\}$, $\hat{\theta}^{*b} = \tilde{\theta}^{*b}$ and, thus, the consistency of FDA's bootstrap procedure follows from the consistency of the bootstrap percentile procedure. The same conclusion can be made if $\sigma_{WR} < \sigma_0$. Hence, under the conditions in Theorem 1, the FDA's bootstrap procedure is also consistent if $\sigma_{WR} \neq \sigma_0$. The question what happens in case that $\sigma_{WR} = \sigma_0$ is addressed in the next section.

2.4 THE INCONSISTENCY OF THE BOOTSTRAP INDIVIDUAL BIOEQUIVALENCE TEST PROCEDURES

If $\sigma_{WR} = \sigma_0$, however, both standard bootstrap percentile and FDA's bootstrap procedures are inconsistent, i.e., (14) does not hold. This is because the function g is not differentiable at $\boldsymbol{\gamma}$ when $\sigma_{WR} = \sigma_0$ and, consequently, $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ and $\sqrt{n}(\hat{\theta} - \theta)$ have different limiting distributions, which can be proved using similar arguments as in Shao (1994) or as in Example 3.10 in Shao & Tu (1995).

We may use the so called m-out-of-n bootstrap procedure (see, e.g., Shao & Tu, 1995, § 3.6) to obtain a bootstrap test that is consistent regardless of whether $\sigma_{WR} = \sigma_0$ or not. However, the m-out-of-n bootstrap procedure is not as efficient as the standard bootstrap procedure when the latter is consistent. Practically speaking, the inconsistency of the bootstrap procedures discussed here is not very serious, since the inconsistency occurs only when

the true parameter vector $\boldsymbol{\gamma}$ is in a subset of the parameter space with zero Lebesgue measure. For a further discussion on practical implications of this restriction see also § 4.

2.5 THE ACCURACY OF THE BOOTSTRAP INDIVIDUAL BIOEQUIVALENCE TEST PROCEDURES

To improve the accuracy of bootstrap confidence bounds, i.e. the convergence rate in (14), a number of different bootstrap procedures have been proposed and studied in the literature. In particular, we may apply the bias–corrected bootstrap percentile or the accelerated bias–corrected bootstrap percentile method in the individual bioequivalence problem (Efron & Tibshirani, 1993, § 22.4 & 22.5).

If $\hat{\boldsymbol{\gamma}}$ is obtained from an analysis of variance, then the accelerated bias–corrected bootstrap percentile method is second–order accurate, whereas the bootstrap percentile and bias–corrected bootstrap percentile are only first–order accurate assuming that $\sigma_{WR} = \sigma_0$. FDA's bootstrap is also first–order accurate. Although we may try to improve FDA's bootstrap procedure by using the bias–corrected or accelerated bias–corrected type of percentile method, it does not lead to any second–order accurate bootstrap procedures, because of the difference in the denominators of (9) and (10).

If $\hat{\boldsymbol{\gamma}}$ is a restricted maximum likelihood estimator, it is unknown whether the accelerated bias–corrected bootstrap percentile is second–order accurate, because of the lack of theory in Edgeworth expansions for the restricted maximum likelihood estimators.

3. PROBABILITY–BASED APPROACH

3.1 THE INDIVIDUAL BIOEQUIVALENCE PARAMETER AND BOOTSTRAP TEST PROCEDURES

We use the same notation given in § 1 and § 2. Under the probability–based approach, the individual bioequivalence parameter is defined on the left hand side of (2) or (3), i.e.,

$$\theta = P_{TR} - P_{RR} \qquad \text{or} \qquad P_{TR}/P_{RR}, \tag{15}$$

where $P_{TR} = \text{pr}\,(|y_R - y_T| \le r)$, $P_{RR} = \text{pr}\,(|y_R - y_R'| \le r)$, and r is a constant or $r = r_0 \sigma_{WR}$ with a constant r_0. Two drug formulations are regarded as individually bioequivalent if and only if $\theta > \theta_L$. Assessing individual bioequivalence can be carried out by testing

$$H_0 : \theta \le \theta_L \qquad \text{versus} \qquad H_1 : \theta > \theta_L. \tag{16}$$

9

From now on we focus on the case of $\theta = P_{TR}/P_{RR}$, since the discussion for the case of $\theta = P_{TR} - P_{RR}$ is exactly the same.

We again consider the two by three crossover design described in § 2.1 and assume model (5). For each subject (fixed i and k), taking the difference between two observations under formulation R gives

$$y_{i21} - y_{i31} = \delta_{231} + e_{i21} - e_{i31} \quad (k = 1)$$

and

$$y_{i12} - y_{i32} = \delta_{132} + e_{i12} - e_{i32} \quad (k = 2),$$

where $\delta_{231} = P_2 - P_3 + W_{R21} - W_{R31}$ and $\delta_{132} = P_1 - P_3 + W_{R12} - W_{R32}$. Unfortunately, the distribution of $y_{i21} - y_{i31}$ (or $y_{i12} - y_{i32}$) is not the same as that of $y_R - y_R'$ unless $\delta_{231} = 0$ (or $\delta_{132} = 0$), which is true when there is no period or interaction effect. Similarly, the differences

$$y_{i11} - y_{i21} = F_T - F_R + \delta_{121} + S_{i1T} - S_{i1R} + e_{i11} - e_{i21}$$

$$y_{i11} - y_{i31} = F_T - F_R + \delta_{131} + S_{i1T} - S_{i1R} + e_{i11} - e_{i31}$$

$$y_{i22} - y_{i12} = F_T - F_R + \delta_{212} + S_{i2T} - S_{i2R} + e_{i22} - e_{i12}$$

and

$$y_{i22} - y_{i32} = F_T - F_R + \delta_{232} + S_{i2T} - S_{i2R} + e_{i22} - e_{i32}$$

do not have the same distribution as that of $y_T - y_R$, unless $\delta_{121} = \delta_{131} = \delta_{212} = \delta_{232} = 0$, where $\delta_{121} = P_1 - P_2 + W_{T11} - W_{R21}$, $\delta_{131} = P_1 - P_3 + W_{T11} - W_{R31}$, $\delta_{212} = P_2 - P_1 + W_{T22} - W_{R12}$, and $\delta_{232} = P_2 - P_3 + W_{T22} - W_{R32}$.

We consider first the case where r is a known constant (the case of $r = r_0 \sigma_{WR}$ is considered in § 3.3). Let $\delta = (\delta_{121}, \delta_{131}, \delta_{212}, \delta_{232}, \delta_{231}, \delta_{132})$. If δ is known, then

$$\hat{P}_{RR}(\delta) = \frac{1}{n} \left[\sum_{i=1}^{n_1} I\left(|y_{i21} - y_{i31} - \delta_{231}| \leq r \right) + \sum_{i=1}^{n_2} I\left(|y_{i12} - y_{i32} - \delta_{132}| \leq r \right) \right]$$

and

$$\hat{P}_{TR}(\delta) = \frac{1}{2n} \left[\sum_{i=1}^{n_1} I\left(|y_{i11} - y_{i21} - \delta_{121}| \leq r \right) + I\left(|y_{i11} - y_{i31} - \delta_{131}| \leq r \right) \right.$$

$$\left. + \sum_{i=1}^{n_2} I\left(|y_{i22} - y_{i12} - \delta_{212}| \leq r \right) + I\left(|y_{i22} - y_{i32} - \delta_{232}| \leq r \right) \right]$$

are unbiased estimators of P_{RR} and P_{TR}, respectively, where $I(A)$ is the indicator function of an event A.

Since δ is usually unknown, Schall & Luus (1993) have proposed the following bootstrap test procedure, where δ is replaced by an appropriate estimator based on the original data set.

Step 1: Let $Y_{ik} = (y_{i1k}, ..., y_{i3k})$ and $Y_k = (Y_{1k}, ..., Y_{n_k k})$. For each fixed k, generate a simple random sample $Y_k^{*b} = (Y_{1k}^{*b}, ..., Y_{n_k k}^{*b})$ with replacement from Y_k. Repeat this process $b = 1, ..., B$ times independently.

Step 2: Let $\hat{\delta}$ be an estimator of δ obtained from an analysis of variance or using the restricted maximum likelihood method. Let $\hat{P}_{TR}^{*b}(\delta)$ and $\hat{P}_{RR}^{*b}(\delta)$ be the same as $\hat{P}_{TR}(\delta)$ and $\hat{P}_{RR}(\delta)$, respectively, but with the y_{ijk} replaced by the bootstrap data y_{ijk}^{*b}. Define

$$\tilde{\theta}^{*b} = \hat{P}_{TR}^*(\hat{\delta})/\hat{P}_{RR}^*(\hat{\delta}). \tag{17}$$

Step 3: Let $\tilde{\theta}_{SL}(5)$ be the 5th percentile of $\tilde{\theta}^{*b}$, $b = 1, ..., B$. Then, individual bioequivalence can be claimed if and only if $\tilde{\theta}_{SL}(5) > \theta_L$.

However, it is shown in the next section that this bootstrap test procedure is inconsistent in the sense that

$$\mathrm{pr}\,(\tilde{\theta}_{SL}(5) < \theta) \not\rightarrow 0.95. \tag{18}$$

If we replace $\tilde{\theta}^{*b}$ in (17) by

$$\hat{\theta}^{*b} = \hat{P}_{TR}^{*b}(\hat{\delta}^{*b})/\hat{P}_{RR}^{*b}(\hat{\delta}^{*b}),$$

where $\hat{\delta}^{*b}$ is the same as $\hat{\delta}$ but is computed based on the bootstrap data set (Y_1^{*b}, Y_2^{*b}), then the previously described procedure becomes the standard bootstrap percentile method. We show in the next section that the bootstrap percentile test procedure is consistent.

3.2 THE CONSISTENCY OF BOOTSTRAP INDIVIDUAL BIOEQUIVALENCE TESTS

We first establish the consistency of the bootstrap percentile method. Let $\hat{P}_{TR}^*(t)$ (or $\hat{P}_{RR}^*(t)$) be $\hat{P}_{TR}^{*b}(t)$ (or $\hat{P}_{RR}^{*b}(t)$) with a fixed b. We need the following result.

Lemma 2. Assume that $n_1/n_2 \to c \in (0,1)$ and that the distribution functions of $y_T - y_R$ and $y_R - y_R'$ are continuous and differentiable. Then

$$\sup_{\|t-\delta\| \leq cn^{-1/2}} |\hat{P}_{TR}(t) - \hat{P}_{TR}(\delta) - P_{TR}(t) + P_{TR}(\delta)| = o_p(n^{-1/2}) \tag{19}$$

11

and

$$\sup_{\|t-\delta\|\leq cn^{-1/2}} |\hat{P}^*_{TR}(t) - \hat{P}^*_{TR}(\delta) - \hat{P}_{TR}(t) + \hat{P}_{TR}(\delta)| = o_p(n^{-1/2}) \tag{20}$$

for any constant $c > 0$, where $P_{TR}(t) = E[\hat{P}_{TR}(t)]$ for any t and $P_{TR}(\delta) = P_{TR}$. Results (19) and (20) also hold if P_{TR} is replaced by P_{RR}.

Proof. Since $\hat{P}_{TR}(t)$ is a type of empirical distribution estimator of $P_{TR}(t)$, result (19) is a standard result in probability theory (e.g., Serfling, 1980, Lemma E, p. 97). Result (20) is a bootstrap analog of (19) and is proved in Shao & Chen (1998, Lemma 1).

Theorem 2. Assume the conditions in Lemmas 1 and 2. Let $\hat{\theta}_{BP}(5)$ be the 5th percentile of the bootstrap distribution of $\hat{\theta}^* = \hat{P}^*_{TR}(\hat{\delta}^*)/\hat{P}^*_{RR}(\hat{\delta}^*)$. Then, it holds that

$$\text{pr}\,(\hat{\theta}_{BP}(5) < \theta) \to 0.95.$$

Proof. Following the proof of Lemma 1 we obtain that

$$\sqrt{n}(\hat{\delta} - \delta) \to_d N(0, \Sigma_\delta)$$

and

$$\sqrt{n}(\hat{\delta}^* - \hat{\delta}) \to_d N(0, \Sigma_\delta),$$

where Σ_δ is a covariance matrix. From result (19) in Lemma 2, we have

$$\hat{P}_{TR}(\hat{\delta}) - P_{TR}(\delta) = \hat{P}_{TR}(\hat{\delta}) - \hat{P}_{TR}(\delta) + \hat{P}_{TR}(\delta) - P_{TR}(\delta)$$

$$= P_{TR}(\hat{\delta}) - P_{TR}(\delta) + \hat{P}_{TR}(\delta) - P_{TR}(\delta) + o_p(n^{-1/2})$$

$$= P'_{TR}(\delta)(\hat{\delta} - \delta) + \hat{P}_{TR}(\delta) - P_{TR}(\delta) + o_p(n^{-1/2}),$$

which converges in law to a normal distribution with mean 0. The same result holds if P_{TR} is replaced by P_{RR}. Hence, it suffices to show that $\hat{P}^*_{TR}(\hat{\delta}^*) - \hat{P}_{TR}(\hat{\delta})$ has the same limiting distribution as that of $\hat{P}_{TR}(\hat{\delta}) - P_{TR}(\delta)$. From results (19) and (20) in Lemma 2, we obtain that

$$\hat{P}^*_{TR}(\hat{\delta}^*) - \hat{P}_{TR}(\hat{\delta}) = \hat{P}^*_{TR}(\hat{\delta}^*) - \hat{P}^*_{TR}(\hat{\delta}) + \hat{P}^*_{TR}(\hat{\delta}) - \hat{P}_{TR}(\hat{\delta})$$

$$= \hat{P}_{TR}(\hat{\delta}^*) - \hat{P}_{TR}(\hat{\delta}) + \hat{P}^*_{TR}(\hat{\delta}) - \hat{P}_{TR}(\hat{\delta}) + o_p(n^{-1/2})$$

$$= P_{TR}(\hat{\delta}^*) - P_{TR}(\hat{\delta}) + \hat{P}^*_{TR}(\hat{\delta}) - \hat{P}_{TR}(\hat{\delta}) + o_p(n^{-1/2})$$

$$= P'_{TR}(\hat{\delta})(\hat{\delta}^* - \hat{\delta}) + \hat{P}^*_{TR}(\hat{\delta}) - \hat{P}_{TR}(\hat{\delta}) + o_p(n^{-1/2})$$

12

$$= P'_{TR}(\delta)(\hat{\delta}^* - \hat{\delta}) + \hat{P}^*_{TR}(\delta) - \hat{P}_{TR}(\delta) + o_p(n^{-1/2}),$$

which has the same limiting distribution as

$$P'_{TR}(\delta)(\hat{\delta} - \delta) + \hat{P}_{TR}(\delta) - P_{TR}(\delta). \tag{21}$$

The same result holds if P_{TR} is replaced by P_{RR}. This completes the proof.

The inconsistency of the bootstrap test procedure proposed by Schall & Luus (1993) (i.e., result (18)) follows from the consistency of the standard bootstrap percentile test procedure (Theorem 2) and the fact that $\hat{P}^*_{TR}(\hat{\delta}^*)$ and $\hat{P}^*_{TR}(\hat{\delta})$ have different limiting distributions. Under the conditions of Theorem 2, it follows from Lemma 2 that

$$\hat{P}^*_{TR}(\hat{\delta}) - \hat{P}_{TR}(\hat{\delta}) = \hat{P}^*_{TR}(\delta) - \hat{P}_{TR}(\delta) + o_p(n^{-1/2}),$$

which has the same limiting distribution as that of $\hat{P}_{TR}(\delta) - P_{TR}(\delta)$, not that of the quantity given in (21) in the proof of Theorem 2. Hence, the asymptotic distributions of $\hat{P}^*_{TR}(\hat{\delta})$ and $\hat{P}^*_{TR}(\hat{\delta}^*)$ are different.

3.3 THE CASE OF $r = r_0\sigma_{WR}$

Schall (1995) suggests the use of $r = r_0\sigma_{WR}$ instead of a constant r in the definition of the individual bioequivalence parameter θ in (15). Define

$$\hat{P}_{RR}(\delta, \sigma_{WR}) = \frac{1}{n} \left[\sum_{i=1}^{n_1} I\left(|y_{i21} - y_{i31} - \delta_{231}| \leq r_0\sigma_{WR} \right) + \sum_{i=1}^{n_2} I\left(|y_{i12} - y_{i32} - \delta_{132}| \leq r_0\sigma_{WR} \right) \right]$$

and

$$\hat{P}_{TR}(\delta, \sigma_{WR}) = \frac{1}{2n} \left[\sum_{i=1}^{n_1} I\left(|y_{i11} - y_{i21} - \delta_{121}| \leq r_0\sigma_{WR} \right) + I\left(|y_{i11} - y_{i31} - \delta_{131}| \leq r_0\sigma_{WR} \right) \right.$$

$$\left. + \sum_{i=1}^{n_2} I\left(|y_{i22} - y_{i12} - \delta_{212}| \leq r_0\sigma_{WR} \right) + I\left(|y_{i22} - y_{i32} - \delta_{232}| \leq r_0\sigma_{WR} \right) \right].$$

If δ and σ_{WR} are known, then these estimators are unbiased for P_{RR} and P_{TR}, respectively. Let $\hat{\sigma}_{WR}$ be an estimator of σ_{WR} obtained from an analysis of variance or using the restricted maximum likelihood method. Let

$$\hat{\theta}^{*b} = \hat{P}^{*b}_{TR}(\hat{\delta}^*, \hat{\sigma}^*_{WR})/\hat{P}^{*b}_{RR}(\hat{\delta}^*, \hat{\sigma}^*_{WR}),$$

13

where $\hat{P}_{TR}^{*b}$, $\hat{P}_{RR}^{*b}$, $\hat{\delta}^*$, and $\hat{\sigma}_{WR}^*$ are bootstrap analogs of $\hat{P}_{TR}$, $\hat{P}_{RR}$, $\hat{\delta}$, and $\hat{\sigma}_{WR}$, respectively, and the bootstrap samples are obtained as described in § 3.1. Let $\hat{\theta}_{BP}(5)$ be the 5th percentile of $\hat{\theta}^{*b}$, $b = 1, ..., B$. Then, the standard bootstrap percentile test rejects H_0 (i.e., claims bioequivalence) if and only if $\hat{\theta}_{BP}(5) > \theta_L$.

Following the proof of Theorem 2 with δ replaced by (δ, σ_{WR}), we conclude that this bootstrap percentile test procedure is consistent. On the other hand, the bootstrap test using the 5th percentile of $\tilde{\theta}^{*b} = \hat{P}_{TR}^{*b}(\hat{\delta}, \hat{\sigma}_{WR})/\hat{P}_{RR}^{*b}(\hat{\delta}, \hat{\sigma}_{WR})$, $b = 1, ..., B$, is inconsistent for the same reason discussed at the end of § 3.2.

Thus, it is crucial to achieve consistency of the probability–based bootstrap procedure that the unknown parameters σ_{WR} and also δ are estimated from each bootstrap sample within the algorithm and not only once using the original data set. This is in contrast to the moment–based approach investigated in § 2, the consistency of which also holds if the decision about the reference scaled or constant scaled version of the bioequivalence parameter is only based on the original data set.

4. Discussion

As seen from the proofs in the preceding sections the FDA's bootstrap procedure using the moment–based approach yields a consistent test procedure in case that the within–subject variance does not equal the point from which on the scaled criterion has a wider bioequivalence range than the unscaled. In turn this means that the procedure is inconsistent only if $\sigma_{WR} = \sigma_0$. Note that this restriction is, however, of nearly no practical relevance since on the one hand it only concerns a single point, an event which should occur only seldom, and on the other hand there is some degree of freedom in the choice of σ_0. This value is fixed by a regulatory decision based on reflections such as in the draft of the FDA guidance, Appendix A (1997) or in Schall & Williams (1996). The latter for instance have recommended that σ_0 might be chosen smaller than 0.3 because a within–subject coefficient of variation taking a value of 30% has already been commonly regarded as highly variable. The derivation of the FDA has come up with a value of 0.2.

The proposal by Schall & Luus (1993) based on a comparison of the probabilities of the difference between corresponding bioavailabilities instead of moments, however, turns out to be inconsistent, although it can be easily modified to get a consistent bootstrap test procedure, namely the standard bootstrap percentile method as discussed in § 3.2.

14

Of course, one should be aware of the fact that the above results are asymptotic in nature. Thus, additional investigations based on real–data examples and simulation studies are called for to judge their applicability in finite situations. This does not only concern their actual coverage probabilities, but also power considerations and sample size determinations. Comparable analyses can be found for instance in Schall (1995) for the bias–corrected bootstrap, Kimanani & Potvin (1997) for a parametric confidence interval, Hauschke et al. (1999) for a confidence interval from untransformed, normally distributed data for the ratio of two means, and in Shao, Chow & Wang (1999) for the bootstrap percentile method, the bias–corrected bootstrap, and the hybrid bootstrap. Examples for case studies of individual bioequivalence are among others given in Schall & Luus (1993), Endrenyi (1994), Schall (1995), and in Shumaker & Metzler (1998).

Nevertheless, the obtained results may help to take away some of the discomfort related to the use of bootstrap intervals for statistically evaluating individual or population bioequivalence.

Appendix: Proof of Lemma 1

The proofs of (11) and (12) follow standard arguments in asymptotic theory. Since the proofs are similar, we only prove (12).

Let $s_n^*(\beta) = \sum_{i,k} \psi(Y_{ik}^*, \beta)$ and $\nabla s_n^*(\beta) = \partial s_n^*(\beta)/\partial \beta$. From the mean–value theorem and the fact that $s_n^*(\hat{\gamma}^*) = 0$,

$$-s_n^*(\hat{\gamma}) = \left[\int_0^1 \nabla s_n^*(\hat{\gamma} + t(\hat{\gamma}^* - \hat{\gamma}))dt \right] (\hat{\gamma}^* - \hat{\gamma}).$$

It follows from conditions (i)–(v) and the consistency of $\hat{\gamma}$ and $\hat{\gamma}^*$ that

$$\frac{1}{n} \left\| \int_0^1 \nabla s_n^*(\hat{\gamma} + t(\hat{\gamma}^* - \hat{\gamma}))dt - \nabla s_n^*(\hat{\gamma}) \right\| = o_p(1).$$

Let E_* be the bootstrap conditional expectation, given Y_k, $k = 1, 2$, and let $H_n = E_*(\nabla s_n^*(\hat{\gamma}))$. By the law of large numbers and conditions (i)–(v),

$$\frac{1}{n} \left\| \nabla s_n^*(\hat{\gamma}) - H_n \right\| = o_p(1)$$

15

and, consequently,

$$-H_n^{-1} s_n^*(\hat{\boldsymbol{\gamma}}) = \{1 + o_p(1)\}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}).$$

Thus, (12) follows if we can show that

$$V_n^{-1/2} H_n^{-1} s_n^*(\hat{\boldsymbol{\gamma}}) \to_d N(0, I), \tag{22}$$

where $V_n = E_*[s_n^*(\hat{\boldsymbol{\gamma}}) s_n^*(\hat{\boldsymbol{\gamma}})']$ (note that $E_*[s_n^*(\hat{\boldsymbol{\gamma}})] = s_n(\hat{\boldsymbol{\gamma}}) = 0$). Using Liapunov's central limit theorem, result (22) follows from

$$\frac{1}{(l'V_n l)^{1+\delta/2}} \sum_{i,k} E_* |l' H_n^{-1} \psi(Y_{ik}^*, \hat{\boldsymbol{\gamma}})|^{2+\delta} = o_p(1) \tag{23}$$

for any nonzero vector l with the same dimension as $\boldsymbol{\gamma}$. The proof is completed since result (23) follows from conditions (i)–(v).

References

ANDERSON, S. & HAUCK, W. W. (1990). Consideration of individual bioequivalence. *J. Pharmacokin. Biopharm.* **18**, 259–273.

CHEN, M.-L. (1997). Individual bioequivalence – a regulatory update. *J. Biopharm. Statist.* **7**, 5–11.

CHINCHILLI, V. M. (1996). The assessment of individual and population bioequivalence. *J. Biopharm. Statist.* **6**, 1–14.

CHOW, S. C. & LIU, J. P. (1995). Current issues in bioequivalence trials. *Drug Inf. J.* **29**, 795–804.

EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

EKBOHM, G. & MELANDER, H. (1989). The subject–by–formulation interaction as a criterion of interchangeability of drugs. *Biometrics* **45**, 1249–1254.

ENDRENYI, L. (1994). A method for the evaluation of individual bioequivalence. *Int. J. Clin. Pharmacol. Ther.* **32**, 497–508.

ESINHART, J. D. & CHINCHILLI, V. M. (1994). Extension to the use of tolerance intervals for the assessment of individual bioequivalence. *J. Biopharm. Statist.* **4**, 39–52.

FDA (1997). *In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches.* Guidance for Industry, Center for Drug Evaluation and Research, Food and Drug Administration (Draft, date: October 1997).

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

HAUCK, W. W. & ANDERSON, S. (1994). Measuring switchability and prescribability: when is average bioequivalence sufficient? *J. Pharmacokin. Biopharm.* **22**, 551–564.

HAUSCHKE, D., KIESER, M., DILETTI, E. & BURKE, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Stat. Med.* **18**, 93–105.

HAUSCHKE, D. & STEINIJANS, V. W. (1999). The U.S. draft guidance regarding population and individual bioequivalence approaches: comments by research–based pharmaceutical company. To appear in: *Stat. Med.*.

HOLDER, D. J. & HSUAN, F. (1993). Moment–based criteria for determining bioequivalence. *Biometrika* **80**, 835–846.

HWANG, J. T. G & WANG, W. (1997). The validity of the test of individual equivalence ratios. *Biometrika* **84**, 893–900.

KIMANANI, E. K. & POTVIN, D. (1997). A parametric confidence interval for a moment–based scaled criterion for individual bioequivalence. *J. Pharmacokin. Biopharm.* **25**, 595–614.

PATNAIK, R. N., LESKO, L. J., CHEN, M. L. & WILLIAMS, R. L. (1997). Individual bioequivalence. New concepts in the statistical assessment of bioequivalence metrics. *Clin. Pharmacokinet.* **33**, 1–6.

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

SCHALL, R. & LUUS, H. G. (1993). On population and individual bioequivalence. *Stat. Med.* **12**, 1109–1124.

SCHALL, R. & WILLIAMS, R. L. (1996). Towards a practical strategy for assessing individual bioequivalence. *J. Pharmacokin. Biopharm.* **24**, 133–149.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

SHAO, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the Amer. Math. Soc.* **122**, 1251–1262.

SHAO, J. & CHEN, Y. (1998). Bootstrapping sample quantiles based on complex survey data under hot deck imputation. *Statistica Sinica* **8**, 1071–1085.

SHAO, J., CHOW, S.-C. & WANG, B. (1999). The bootstrap procedure in individual bioequivalence. To appear in: *Stat. Med.*.

SHAO, J. & TU, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

SHEINER, L. B. (1992). Bioequivalence revisited. *Stat. Med.* **11**, 1777–1788.

SHUMAKER, R. C. & METZLER, C. M. (1998). The Phenytoin trial is a case study of 'individual' bioequivalence. *Drug Inf. J.* **32**, 1063–1072.