Statistica Sinica **13**(2003), 539-554

STATISTICAL TESTS FOR POPULATION BIOEQUIVALENCE

Shein-Chung Chow, Jun Shao and Hansheng Wang

StatPlus, Inc., University of Wisconsin and Peking University

Abstract: In its 2001 guidance, the U.S. Food and Drug Administration (FDA) recommends that population bioequivalence (PBE) and individual bioequivalence (IBE) be assessed to address respectively the prescribability and switchability between a brand-name drug product and its new formulation or generic copy. For IBE, the FDA recommends a 2×4 crossover design and a statistical test procedure proposed by Hyslop, Hsuan and Holder (2000). The same method is also recommended in FDA (2001) for assessment of PBE under the 2×4 crossover design. However, we note that, asymptotically, FDA's PBE test has a size smaller than the nominal level and thus has a low power to detect PBE. In addition, the 2001 FDA guidance does not provide any statistical procedure for PBE under commonly used 2×2 or 2×3 crossover designs. In this paper, an asymptotically valid statistical test is derived for PBE under the 2×2, 2×3 or 2×4 crossover design, using the method of moments and linearization. A method of determining the sample size required to achieve a desired power of the PBE test is also proposed. Simulation results are provided to examine the performance of the proposed PBE test and FDA's test. Finally, an example is presented for illustration.

Key words and phrases: Crossover design, drug prescribability, linearization, power, nominal level, sample size.

1. Introduction

When a brand-name drug is going off patent, the innovator drug company will usually develop a new formulation to extend its exclusivity in the marketplace. At the same time generic drug companies may file new, abbreviated drug applications for generic drugs approval. *In vivo* bioequivalence testing is usually considered as a surrogate for clinical evaluation of drug products based on the *Fundamental Bioequivalence Assumption* that when two drug formulations are equivalent in bioavailability, they will reach the same therapeutic effect or they are therapeutically equivalent (Chow and Liu (1999)). Pharmacokinetic (PK) responses such as area under the blood or plasma concentration-time curve (AUC) and maximum concentration ($C_{\max}$) are usually considered the primary measures for bioavailability. In 1992, the U.S. Food and Drug Administration (FDA) published its first guidance on statistical procedures for *in vivo* bioequivalence studies (FDA (1992)), which requires that the evidence of bioequivalence in average PK responses between the reference formulation (e.g., the brand-name) and the test formulation (e.g., a new formulation or a generic copy) be provided. Bioequivalence in average PK responses is referred to as *average bioequivalence* (ABE), which is also required in the FDA most recent guidance on bioequivalence studies for orally administered drug products (FDA (2000)). The ABE approach for bioequivalence, however, has limitations for addressing drug interchangeability, since it focuses only on the comparison of population averages between the test and reference formulations. Drug interchangeability can be classified as either drug prescribability or drug switchability. Drug prescribability is referred to as the physician's choice for prescribing an appropriate drug for his/her new patients among the drug products available, while drug switchability is related to the switch from a drug product to an alternative drug product within the same patient. To assess drug prescribability and switchability, *population bioequivalence* (PBE) and *individual bioequivalence* (IBE) are proposed (see, for example, Anderson and Hauck (1990), Esinhart and Chinchilli (1994), Sheiner (1992), Schall and Luus (1993), Chow and Liu (1995), and Chen (1997)).

In its 2001 guidance (FDA (2001)), the FDA recommends a statistical test procedure for IBE, which is based on a 2 × 4 crossover design and a method proposed by Hyslop, Hsuan, and Holder (2000). The same method is also recommended in FDA (2001) for assessment of PBE under the 2 × 4 crossover design. However, the method proposed by Hyslop, Hsuan, and Holder (2000) is not directly applicable to PBE testing due to the violation of the primary assumption of independence among the estimated components of the PBE criterion. In this paper we focus on assessing PBE, which is recommended by FDA (2001) for new drug application during the investigational phase of drug development. After an introduction of the PBE criterion, in Section 2 we show that the size of the PBE test recommended in FDA (2001) is asymptotically smaller than the nominal level.

Although FDA (2001) indicates that a standard 2 × 2 crossover design may be used for assessment of PBE, little information regarding the statistical test procedure is provided. Using the method of moments and linearization, we derive in Section 3 asymptotically valid statistical tests for PBE under commonly used 2×2, 2×3 or 2×4 crossover designs. Also included in Section 3 is a method of sample size determination for the use of the proposed PBE test. Some simulation results are given in Section 4 to examine the finite sample performance of the proposed methods and FDA's PBE test. Finally, Section 5 contains an example for illustration.

2. FDA's PBE Test

2.1. Design, model, and criterion

Let y_T be the PK response (or its log-transform) from the test formulation, y_R and y'_R be two identically distributed PK responses (or their log-transforms) from the reference formulation, where y_T, y_R and y'_R are independent observations from different subjects. Then the drug prescribability can be measured by

$$\theta = \begin{cases} \dfrac{E(y_R - y_T)^2 - E(y_R - y'_R)^2}{E(y_R - y'_R)^2/2} & \text{if } E(y_R - y'_R)^2/2 \geq \sigma_0^2 \\ \dfrac{E(y_R - y_T)^2 - E(y_R - y'_R)^2}{\sigma_0^2} & \text{if } E(y_R - y'_R)^2/2 < \sigma_0^2 \end{cases} \tag{1}$$

(FDA (2001)), where σ_0^2 is a constant specified by the FDA. According to FDA (2001), PBE can be claimed if the following null hypothesis H_0 is rejected at the 5% level of significance:

$$H_0 : \theta \geq \theta_U \quad \text{versus} \quad H_1 : \theta < \theta_U, \tag{2}$$

where θ_U is an upper limit specified by the FDA. According to FDA (2001), sponsors or applicants wishing to use the PBE approach should contact the Agency (FDA) for further information on σ_0 and θ_U.

For *in vivo* bioequivalence testing, crossover designs (see, for example, Jones and Kenward (1989), Chow and Liu (1999)) are usually considered. Let y_{ijk} be the original or the log-transform of the PK response of interest from the ith subject in the kth sequence at the jth period of the experiment, where $i = 1, \ldots, n_k$, $k = 1, 2$, $j = 1, \ldots, p$, and p is the number of periods of the crossover design. A sufficient length of washout between dosing periods is usually applied to wear off the possible residual effect that may be carried over from one period to the next. The following statistical model is commonly considered:

$$y_{ijk} = \mu + F_l + W_{ljk} + S_{ikl} + e_{ijk}, \tag{3}$$

where μ is the overall mean; F_l is the fixed effect of the lth formulation ($l = T$ or R according to the design and $F_T + F_R = 0$); W_{ljk}'s are fixed period, sequence, and interaction effects ($\sum_k \bar{W}_{lk} = 0$, where $\bar{W}_{lk}$ is the average of W_{ljk}'s with fixed (l, k), $l = T$, R); S_{ikl} is the random effect of the ith subject in the kth sequence under formulation l and (S_{ikT}, S_{ikR}), $i = 1, \ldots, n_k$, $k = 1, 2$, are independent and identically distributed bivariate normal 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 distributed as $N(0, \sigma_{Wl}^2)$, and S_{ikl}'s and e_{ijk}'s are mutually 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 (3), θ in (1) is

$\theta = (\delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2)/\max\{\sigma_0^2, \sigma_{TR}^2\}$, where $\delta = F_T - F_R$, $\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2$. Let $\lambda = \delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2 - \theta_U \max\{\sigma_0^2, \sigma_{TR}^2\}$. Then the hypothesis in (2) is equivalent to

$$H_0 : \lambda \geq 0 \quad \text{versus} \quad H_1 : \lambda < 0. \tag{4}$$

If $\hat{\lambda}_U$ is a 95% upper confidence bound for λ, then a level 5% test rejects H_0 (i.e., concludes PBE) if and only if $\hat{\lambda}_U < 0$ (Lehmann (1986), pp.90-91).

Note that θ or λ is an aggregated measure, i.e., different combinations of the values for δ, σ_{TT}^2 and σ_{TR}^2 can result in the same value of λ. In some cases, it may not be a suitable measure for assessing PBE (see, e.g., Chow (1999)). As a partial remedy, FDA (2001) recommends that PBE be claimed if and only if H_0 in (4) be rejected at level 5% *and* the observed difference of means (for the test and reference formulations) is within ±0.223. In this paper, we focus on testing hypothesis (4), i.e., on the construction of an approximate 95% upper confidence bound $\hat{\lambda}_U$.

2.2. FDA's PBE test under the 2 × 4 design

The 2001 FDA guidance adopts the method by Hyslop, Hsuan, and Holder (2000) in the construction of the upper confidence bound for PBE under the 2×4 crossover design. This method, based on results in Howe (1974), Graybill and Wang (1980) and Ting, Burdick, Graybill, Jeyaratnam, and Lu (1990), can be described as follows. Suppose

$$\lambda = \lambda_1 + \cdots + \lambda_r - \lambda_{r+1} - \cdots - \lambda_m, \tag{5}$$

where the λ_j's are positive parameters. Suppose $\hat{\lambda}_j$ is a point estimator of λ_j and $\tilde{\lambda}_j$ is a 95% upper confidence bound for λ_j when $j = 1, \ldots, r$, and $\tilde{\lambda}_j$ is a 95% lower confidence bound for λ_j when $j = r+1, \ldots, m$. If $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ are independent, then an approximate 95% upper confidence bound for λ is

$$\hat{\lambda}_U = \hat{\lambda}_1 + \cdots + \hat{\lambda}_r - \hat{\lambda}_{r+1} - \cdots - \hat{\lambda}_m + \sqrt{(\tilde{\lambda}_1 - \hat{\lambda}_1)^2 + \cdots + (\tilde{\lambda}_m - \hat{\lambda}_m)^2}. \tag{6}$$

Note that (5) is a decomposition of λ into a linear function with components λ_j, and $\hat{\lambda}_U$ in (6) is an aggregated confidence bound in terms of individual exact confidence bounds for the λ_j's. The independence of $\hat{\lambda}_j$'s is a key condition in applying this method. Hyslop, Hsuan, and Holder (2000) successfully applied it to IBE testing.

For PBE testing, FDA (2001) decomposed the λ in (4) into $\lambda_1 = \delta^2$, $\lambda_2 = \sigma_{TT}^2$ and $\lambda_3 = \sigma_{TR}^2 + \theta_U \max\{\sigma_0^2, \sigma_{TR}^2\}$. Then applied (6) with $\hat{\lambda}_1 = \hat{\delta}^2$, $\hat{\lambda}_2 = \hat{\sigma}_{TT}^2$, and $\hat{\lambda}_3 = \hat{\sigma}^2_{TR} + \theta_U \max\{\sigma^2_0, \hat{\sigma}^2_{TR}\}$, where

$$\hat{\delta} = \frac{\bar{x}_{T1} + \bar{x}_{T2}}{2} - \frac{\bar{x}_{R1} + \bar{x}_{R2}}{2};$$

$$\hat{\sigma}^2_{TT} = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} [(x_{Tki} - \bar{x}_{Tk})^2 + (z_{Tki} - \bar{z}_{Tk})^2/4];$$

$$\hat{\sigma}^2_{TR} = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} [(x_{Rki} - \bar{x}_{Rk})^2 + (z_{Rki} - \bar{z}_{Rk})^2/4];$$

x_{lki} and z_{lki} are the average and difference, respectively, of the two observations from the ith subject in the kth sequence under drug treatment l in a 2×4 crossover design, $i = 1, \ldots, n_k$, $k = 1, 2$, $l = T, R$; $\bar{x}_{lk}$ is the sample mean of x_{lki}, $i = 1, \ldots, n_k$; $\bar{z}_{lk}$ is the sample mean of z_{lki}, $i = 1, \ldots, n_k$; and n_k is the sample size in the kth sequence. It follows from (6) that FDA's PBE test rejects H_0 in (4) if and only if $\hat{\lambda}_U < 0$, where

$$\hat{\lambda}_U = \hat{\lambda}_1 + \hat{\lambda}_2 - \hat{\lambda}_3 + \sqrt{U_1 + U_2 + U_3};$$

$$U_1 = \left[\left(|\hat{\delta}| + t_{0.95;n_1+n_2-2} \frac{\hat{\sigma}_{0.5,0.5}}{2} \sqrt{n_1^{-1} + n_2^{-1}}\right)^2 - \hat{\delta}^2\right]^2;$$

$$U_2 = \hat{\sigma}^4_{TT} \left(\frac{n_1 + n_2 - 2}{\chi^2_{0.05;n_1+n_2-2}} - 1\right)^2;$$

$$U_3 = \hat{c}^2 \hat{\sigma}^4_{TR} \left(\frac{n_1 + n_2 - 2}{\chi^2_{0.95;n_1+n_2-2}} - 1\right)^2;$$

$$\hat{\sigma}^2_{0.5,0.5} = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_{iTk} - x_{iRk} - \bar{x}_{Tk} + \bar{x}_{Rk})^2; \tag{7}$$

$\hat{c} = 1 + \theta_U$ if $\hat{\sigma}^2_{TR} \geq \sigma^2_0$, $\hat{c} = 1$ if $\hat{\sigma}^2_{TR} < \sigma^2_0$; and $t_{a;n_1+n_2-2}$ and $\chi^2_{a;n_1+n_2-2}$ are the 100ath percentile of the central t-distribution and chi-square distribution, respectively, with $n_1 + n_2 - 2$ degrees of freedom.

Although $\hat{\delta}$ is independent of $(\hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR})$ (see the Appendix), the two variance estimators $\hat{\sigma}^2_{TT}$ and $\hat{\sigma}^2_{TR}$ are not independent. This was first noticed by Quiroz, Ting, Wei, and Burdick (2000). In fact, it is shown in the Appendix that

$$\text{Cov}\,(\hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR}) = 2\rho^2 \sigma^2_{BT} \sigma^2_{BR}/(n_1 + n_2 - 2). \tag{8}$$

Thus, the test recommended in FDA (2001) is statistically inappropriate in the sense that its size may be very different from the nominal level 5%. In Section 4 we study the type I error and power of FDA's test empirically. Here, we present an asymptotic result assuming that $n_1 = n_2 = n$ and n is large. It can be shown that, as $n \to \infty$,

$$\text{the asymptotic size of FDA's PBE test } = \Phi\left(\frac{z_{0.05}}{\sqrt{1 - 2c\rho^2\sigma_{BT}^2\sigma_{BR}^2/\sigma_\lambda^2}}\right),$$

where Φ is the standard normal distribution function, $z_{0.05}$ is the 5th percentile of the standard normal distribution,

$$\sigma_\lambda^2 = 2\delta^2\sigma_{0.5,0.5}^2 + 0.25\sigma_{WT}^4 + 0.25c^2\sigma_{WR}^4 + (\sigma_{BT}^2 + 0.5\sigma_{WT}^2)^2 + c^2(\sigma_{BR}^2 + 0.5\sigma_{WR}^2)^2,$$

$c = 1 + \theta_U$ if $\sigma_{TR}^2 \geq \sigma_0^2$, $c = 1$ if $\sigma_{TR}^2 < \sigma_0^2$, and $\sigma_{0.5,0.5}^2$ is given by

$$\sigma_{a,b}^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} + a\sigma_{WT}^2 + b\sigma_{WR}^2 \tag{9}$$

with $a = 0.5$ and $b = 0.5$.

This indicates that the asymptotic size of FDA's test for PBE is always less than the nominal level 5% unless $\rho = 0$ (which is impractical). Note that if the size of a PBE test is less than the nominal level 5%, it means that this test is too conservative and has unnecessarily low power.

Although the 2001 FDA guidance indicates that a 2 × 2 crossover design may be used for assessment of PBE, no detailed test procedure is provided.

3. PBE Tests Based on Moment Estimators and Linearization

In this section, we propose PBE tests of asymptotic size 5%, using the method of moments and linearization under model (3) in Section 2.1.

3.1. The 2 × 2 crossover design

For any of the 2 × 2, 2 × 3, or 2 × 4 crossover designs, let x_{lki} and z_{lki} be as defined in Section 2, except that when there is a single observation under a given sequence-formulation combination, x_{ilk} is the same as the original observation and z_{ilk} is defined to be 0. Let $\hat{\delta}$, $\hat{\sigma}_{TT}^2$, and $\hat{\sigma}_{TR}^2$ be as defined in Section 2. We now derive an asymptotic (as $n_k \to \infty$) 95% upper confidence bound for λ by applying linearization to the moment estimator

$$\hat{\lambda} = \hat{\delta}^2 + \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2 - \theta_U \max\{\sigma_0^2, \hat{\sigma}_{TR}^2\}. \tag{10}$$

The resulting PBE test is asymptotically of size 5%. Its performance is studied by simulation in Section 4.

When it is known that $\sigma_{TR}^2 \geq \sigma_0^2$, the proposed upper confidence bound is

$$\hat{\lambda}_U = \hat{\delta}^2 + \hat{\sigma}_{TT}^2 - (1 + \theta_U)\hat{\sigma}_{TR}^2 + t_{0.95;n_1+n_2-2}\sqrt{V}, \tag{11}$$

where V is an estimated variance of $\hat{\delta}^2 + \hat{\sigma}^2_{TT} - (1+\theta_U)\hat{\sigma}^2_{TR}$ of the form $V = (2\hat{\delta}, 1, -(1+\theta_U))C(2\hat{\delta}, 1, -(1+\theta_U))'$ and C is an estimated variance-covariance matrix of $(\hat{\delta}, \hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR})$. Since $\hat{\delta}$ and $(\hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR})$ are independent,

$$C = \begin{pmatrix} \dfrac{\hat{\sigma}^2_{1,1}}{4}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) & (0,0) \\ (0,0)' & \dfrac{(n_1-1)C_1}{(n_1+n_2-2)^2} + \dfrac{(n_2-1)C_2}{(n_1+n_2-2)^2} \end{pmatrix}, \qquad (12)$$

where $\hat{\sigma}^2_{1,1} = \frac{1}{n_1+n_2-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(x_{iTk} - x_{iRk} - \bar{x}_{Tk} + \bar{x}_{Rk})^2$ (an estimator of $\sigma^2_{1,1}$ given by (9) with $a = 1$ and $b = 1$), C_1 is the sample covariance matrix of $((x_{iT1} - \bar{x}_{T1})^2, (x_{iR1} - \bar{x}_{R1})^2)$, $i = 1, \ldots, n_1$, and C_2 is the sample covariance matrix of $((x_{iR2} - \bar{x}_{R2})^2, (x_{iT2} - \bar{x}_{T2})^2)$, $i = 1, \ldots, n_2$.

When $\sigma^2_{TR} < \sigma^2_0$, the upper confidence bound for λ should be modified to

$$\hat{\lambda}_U = \hat{\delta}^2 + \hat{\sigma}^2_{TT} - \hat{\sigma}^2_{TR} - \theta_U\sigma^2_0 + t_{0.95;n_1+n_2-2}\sqrt{V_0}, \qquad (13)$$

where $V_0 = (2\hat{\delta}, 1, -1)C(2\hat{\delta}, 1, -1)'$.

The confidence bound in (11) is referred to as the confidence bound under the reference-scaled criterion, whereas the confidence bound in (13) is referred to as the confidence bound under the constant-scaled criterion. In practice, whether $\sigma^2_{TR} \geq \sigma^2_0$ or not is unknown. There are two methods of determining whether the reference-scaled criterion or the constant-scaled criterion should be used. The first method, which is used by Hyslop, Hsuan and Holder (2000) and FDA (2001), applies the reference-scaled criterion or the constant-scaled criterion according as $\hat{\sigma}^2_{TR} \geq \sigma^2_0$ or $\hat{\sigma}^2_{TR} < \sigma^2_0$. This method is referred to as the estimation method and, intuitively, it works well if the true value of σ^2_{TR} is not close to σ^2_0. The second method is based on a test of $\sigma^2_{TR} \geq \sigma^2_0$ versus $\sigma^2_{TR} < \sigma^2_0$: if $\hat{\sigma}^2_{TR}(n_1+n_2-2)/\chi^2_{0.05;n_1+n_2-2} \geq \sigma^2_0$, then the reference-scaled criterion should be used; otherwise the constant-scaled criterion should be used. This method is referred to as the test method. The test method is more conservative than the estimation method. A comparison of the two is given in Section 4.

An important practical issue is the determination of sample sizes n_1 and n_2 for achieving a desired power (e.g., 80%) of the PBE test, for a given set of parameter values. It follows from the calculation in the Appendix that when $n_1 = n_2 = n$, an approximate formula to determine n is given by

$$n \geq \frac{2\delta^2\sigma^2_{1,1} + \sigma^4_{TT} + c^2\sigma^4_{TR} - 2c\rho^2\sigma^2_{BT}\sigma^2_{BR}}{\lambda^2}(z_{0.95} + z_\beta)^2 \qquad (14)$$

for a set of given values of δ, $\sigma^2_{1,1}$, σ^2_{TT}, σ^2_{TR}, σ^2_{BT}, σ^2_{BR} and ρ, where z_t is the tth quantile of the standard normal distribution, and β is the desired power.

3.2. The 2 × 4 crossover design

When the study also considers IBE, the design has to be a higher order crossover design. If a higher order crossover design is used, a test procedure for PBE can be obtained by using the same idea as described in Section 3.1, but with more data to increase accuracy.

Consider the 2 × 4 crossover design as recommended by FDA (2001) for IBE testing: each subject receives two formulations exactly twice, and subjects in different sequences receive different formulations at any given period. Let $\hat{\delta}$, $\hat{\sigma}^2_{TT}$, and $\hat{\sigma}^2_{TR}$ be as defined in Section 2. A test for PBE can be obtained by using $\hat{\lambda}_U$ in (11) or (13) with

$$C = \begin{pmatrix} \dfrac{\hat{\sigma}^2_{0.5,0.5}}{4}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right) & (0,0) \\ (0,0)' & \dfrac{(n_1-1)C_1}{(n_1+n_2-2)^2}+\dfrac{(n_2-1)C_2}{(n_1+n_2-2)^2}+\dfrac{C_0}{2(n_1+n_2-2)} \end{pmatrix},$$

where $\hat{\sigma}^2_{0.5,0.5}$ is defined by (7) (an estimator of $\sigma^2_{0.5,0.5}$ given by (9) with $a = b = 0.5$), C_1 and C_2 are the same as those in Section 3.1,

$$C_0 = \begin{pmatrix} \hat{\sigma}^4_{WT} & 0 \\ 0 & \hat{\sigma}^4_{WR} \end{pmatrix},$$

$$\hat{\sigma}^2_{Wl} = \frac{1}{2(n_1+n_2-2)}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(z_{ilk}-\bar{z}_{lk})^2.$$

3.3. The 2 × 3 crossover design

As indicated in FDA (2001), a 2 × 3 crossover design may be used as an alternative to the 2 × 4 crossover design for assessment of IBE. The standard 2 × 3 crossover design is the same as the 2 × 4 crossover design with the last period removed. Assume that sequence 1 has two test formulations and sequence 2 has two reference formulations. Let $\hat{\delta}$, $\hat{\sigma}^2_{Tl}$, and $\hat{\sigma}^2_{Wl}$ be the same as those in Section 3.2. A test for PBE can be obtained by using $\hat{\lambda}_U$ in (11) or (13) with

$$C = \begin{pmatrix} \dfrac{\hat{\sigma}^2_{0.5,1}}{4n_1}+\dfrac{\hat{\sigma}^2_{1,0.5}}{4n_2} & (0,0) \\ (0,0)' & \dfrac{(n_1-1)C_1}{(n_1+n_2-2)^2}+\dfrac{(n_2-1)C_2}{(n_1+n_2-2)^2}+\dfrac{C_0}{2(n_1+n_2-2)} \end{pmatrix},$$

where $\hat{\sigma}^2_{0.5,1} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{iT1}-x_{iR1}-\bar{x}_{T1}+\bar{x}_{R1})^2$, $\hat{\sigma}^2_{1,0.5} = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{iT2}-x_{iR2}-\bar{x}_{T2}+\bar{x}_{R2})^2$, C_1 and C_2 are the same as those in Section 3.1, and

$$C_0 = \frac{1}{n_1+n_2-2}\begin{pmatrix} (n_1-1)\hat{\sigma}^4_{WT} & 0 \\ 0 & (n_2-1)\hat{\sigma}^4_{WR} \end{pmatrix}.$$

3.4. The 2 × 3 extra-reference design

For IBE testing, Schall and Luus (1993) and Chow, Shao and Wang (2002) considered a 2 × 3 extra-reference design obtained by adding an extra reference period to the 2 × 2 crossover design, i.e., the third periods of both sequences are under the reference formulation. Let $\hat{\delta}$, $\hat{\sigma}^2_{WR}$, and $\hat{\sigma}^2_{TR}$, be the same as previously defined. Then, a test for PBE can be obtained by using $\hat{\lambda}_U$ in (11) or (13) with

$$C = \begin{pmatrix} \frac{\hat{\sigma}^2_{1,0.5}}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}\right) & (0,0) \\ (0,0)' & \frac{(n_1-1)C_1}{(n_1+n_2-2)^2}+\frac{(n_2-1)C_2}{(n_1+n_2-2)^2}+\frac{C_0}{2(n_1+n_2-2)} \end{pmatrix},$$

where $\hat{\sigma}^2_{1,0.5} = \frac{1}{n_1+n_2-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(x_{iTk}-x_{iRk}-\bar{x}_{Tk}+\bar{x}_{Rk})^2$, C_1 and C_2 are the same as those in Section 3.1, and $C_0 = \begin{pmatrix} 0 & 0 \\ 0 & \hat{\sigma}^4_{WR} \end{pmatrix}$.

3.5. Discussion

The method derived by Hyslop, Hsuan, and Holder (2000) and FDA's PBE test depend on the normality assumption on y_{ijk}'s in (3). For non-normal y_{ijk}'s, their tests are not asymptotically valid. Since our proposed PBE tests are based on moment estimators and linearization, they are still asymptotically valid when y_{ijk}'s are non-normal. It can be seen from the proofs in the Appendix that when y_{ijk}'s are non-normal, the estimated variance-covariance matrix C in (12) is still a consistent estimator of the variance-covariance matrix of $(\hat{\delta}, \hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR})$.

4. Simulation Results

4.1. The 2 × 2 Design

A simulation study was carried out to examine the type I error probability and power of the PBE tests under the 2 × 2 crossover design when the sample size $n_1 = n_2 = n$ is 10, 20, 30, 40, 50, or 60. Values of σ_{BT}, σ_{BR}, σ_{WT}, and σ_{WR} are chosen from 0.1, 0.4, and 0.6, and values of ρ are 0.75 and 1. For the computation of the type I error probability, we considered the situation where the value of δ is chosen such that $\lambda = 0$. According to the 2001 FDA guidance, the values of σ_0 and θ_U are chosen to be 0.2 and 1.74, respectively. For each parameter and sample size combination, 10,000 simulation runs were used to compute the empirical type I error probability. Normal random variates were generated according to (3), using the Fortran subroutine Random.f90 in the Department of Statistics, University of Wisconsin-Madison.

Table 1 reports the empirical type I error probabilities when the test method is used to decide whether the reference-scaled or the constant-scaled criterion should be used (see Section 3.1). In terms of these type I error probabilities, the PBE test generally performs well. In most cases, the type I error probability is under the nominal value 5%.

Table 1. Type I Error Probability of the PBE Test Under 2 × 2 Crossover Design. ($\lambda = 0$; Nominal Level 5%; 10,000 Simulations)

		$\rho = .75$						$\rho = 1$					
		σ_{WT}, σ_{WR}						σ_{WT}, σ_{WR}					
σ_{BT}, σ_{BR}	n	.1,.1	.1,.4	.4,.4	.4,.6	.6,.4	.6,.6	.1,.1	.1,.4	.4,.4	.4,.6	.6,.4	.6,.6
.1,.1	10	.0406	.0325	.0406	.0360	.0527	.0465	.0351	.0304	.0478	.0370	.0538	.0453
	20	.0508	.0322	.0434	.0365	.0477	.0426	.0527	.0335	.0408	.0380	.0544	.0453
	30	.0583	.0346	.0427	.0426	.0454	.0464	.0534	.0375	.0427	.0388	.0493	.0456
	40	.0576	.0368	.0410	.0389	.0463	.0443	.0554	.0374	.0449	.0382	.0492	.0431
	50	.0601	.0390	.0464	.0404	.0489	.0401	.0550	.0387	.0426	.0376	.0453	.0430
	60	.0556	.0412	.0458	.0430	.0508	.0453	.0549	.0366	.0430	.0398	.0463	.0471
.1,.4	10	.0301	.0306	.0393	.0319	.0419	.0403	.0276	.0296	.0340	.0366	.0528	.0406
	20	.0304	.0331	.0367	.0376	.0416	.0391	.0307	.0309	.0329	.0376	.0442	.0399
	30	.0335	.0357	.0374	.0355	.0445	.0395	.0306	.0333	.0408	.0358	.0402	.0404
	40	.0355	.0324	.0405	.0360	.0400	.0442	.0317	.0312	.0371	.0328	.0477	.0406
	50	.0352	.0367	.0362	.0434	.0415	.0398	.0348	.0358	.0411	.0373	.0413	.0436
	60	.0346	.0364	.0391	.0369	.0431	.0439	.0352	.0385	.0408	.0397	.0425	.0421
.4,.4	10	.0343	.0361	.0408	.0372	.0507	.0401	.0288	.0327	.0367	.0346	.0459	.0395
	20	.0353	.0361	.0355	.0368	.0465	.0413	.0297	.0305	.0372	.0351	.0475	.0383
	30	.0350	.0345	.0339	.0382	.0442	.0449	.0282	.0317	.0374	.0366	.0474	.0400
	40	.0360	.0367	.0389	.0388	.0441	.0429	.0331	.0331	.0393	.0385	.0458	.0400
	50	.0387	.0371	.0408	.0359	.0447	.0409	.0328	.0334	.0394	.0376	.0432	.0436
	60	.0393	.0390	.0425	.0376	.0444	.0431	.0338	.0378	.0393	.0365	.0494	.0437
.4,.6	10	.0312	.0337	.0367	.0313	.0420	.0420	.0273	.0295	.0344	.0346	.0415	.0373
	20	.0320	.0320	.0353	.0347	.0401	.0383	.0252	.0290	.0331	.0362	.0374	.0359
	30	.0306	.0335	.0358	.0386	.0402	.0372	.0272	.0300	.0371	.0367	.0381	.0390
	40	.0335	.0341	.0358	.0344	.0440	.0404	.0274	.0333	.0344	.0354	.0351	.0408
	50	.0317	.0366	.0385	.0399	.0384	.0379	.0316	.0352	.0388	.0348	.0455	.0387
	60	.0366	.0389	.0355	.0364	.0445	.0425	.0298	.0358	.0383	.0367	.0366	.0430
.6,.4	10	.0524	.0406	.0499	.0413	.0584	.0434	.0422	.0402	.0476	.0400	.0556	.0454
	20	.0507	.0427	.0418	.0386	.0502	.0447	.0414	.0347	.0413	.0385	.0486	.0413
	30	.0507	.0353	.0473	.0399	.0493	.0454	.0446	.0360	.0433	.0389	.0503	.0423
	40	.0462	.0419	.0433	.0407	.0475	.0452	.0378	.0377	.0431	.0365	.0511	.0394
	50	.0470	.0405	.0441	.0413	.0436	.0416	.0388	.0361	.0433	.0385	.0482	.0445
	60	.0443	.0445	.0468	.0419	.0452	.0427	.0417	.0391	.0414	.0398	.0457	.0438
.6,.6	10	.0313	.0366	.0408	.0382	.0472	.0412	.0315	.0309	.0367	.0333	.0421	.0401
	20	.0345	.0369	.0388	.0334	.0410	.0419	.0301	.0291	.0354	.0362	.0421	.0357
	30	.0382	.0350	.0396	.0362	.0447	.0381	.0298	.0325	.0363	.0384	.0414	.0396
	40	.0357	.0334	.0335	.0380	.0424	.0419	.0314	.0354	.0387	.0399	.0431	.0402
	50	.0404	.0347	.0414	.0388	.0414	.0415	.0330	.0353	.0351	.0379	.0440	.0409
	60	.0402	.0372	.0418	.0372	.0440	.0430	.0324	.0359	.0382	.0375	.0451	.0397

Results of using the estimation method to decide which criterion should be used were also obtained, but the results are the same as those in Table 1 except for two cases. When $\sigma_{BT} = \sigma_{BR} = \sigma_{WT} = \sigma_{WR} = 0.1$ and $\rho = 0.75$, the type I error probabilities are 0.0723, 0.0620, 0.0607, 0.0582, 0.0601, 0.0556 for $n = 10$, 20, 30, 40, 50, 60, respectively. When $\sigma_{BT} = \sigma_{BR} = \sigma_{WT} = \sigma_{WR} = 0.1$ and $\rho = 1$, the type I error probabilities are 0.0680, 0.0630, 0.0564, 0.0555, 0.0551, 0.0549 for $n = 10$, 20, 30, 40, 50, 60, respectively. Thus, the test method for deciding whether the reference-scaled or the constant-scaled criterion should be used performs better than the estimation method, although the two methods produce identical results when $\sigma_{TR} > \sigma_0 = 0.2$.

For all $n = 10, \ldots, 60$ and some selected combinations of parameter values (for which $\lambda < 0$), the empirical power of the PBE test is also obtained but not reported here. The general finding is that when $\sigma_{TT} \leq \sigma_{TR}$, a reasonably large power (e.g., 80%) can usually be reached with reasonable values of n and λ; when $\sigma_{TT} > \sigma_{TR}$, it is difficult to claim PBE (i.e., the power is low) even when the two formulations are actually PBE.

Simulation results are also obtained for the performance of the proposed formula (14) for sample size determination. For some combinations of the parameter values used in Table 1 and $\beta = 80\%$, we first compute the sample size n determined by (14) and then compute (with 10,000 simulations) the actual power P_n of the PBE test using n as the sample size for both sequences. The results are not reported here. The general findings are

1. The proposed formula (14) works well, i.e., the power P_n corresponding to each selected n is larger than the target value 80%, although the sample size produced by formula (14) is conservative since P_n is much larger than 80% in some cases.

2. When the variation of the test formulation is larger than that of the reference formulation, it is difficult to claim PBE even when the two formulations are both PBE and ABE. When $\delta = 0$ and $(\sigma_{BT}, \sigma_{BR}, \sigma_{WT}, \sigma_{WR}) = (0.4, 0.4, 0.6, 0.4)$ or $(0.6, 0.4, 0.4, 0.4)$, for example, the required sample size n to claim PBE ranges from 35 to 46.

4.2. The 2 × 4 design

In this section we report results from a simulation study that investigates the type I error probability and the power of FDA's PBE test, and the proposed test in Section 3.2, under the 2×4 crossover design with $n_1 = n_2 = 20$. Values of the parameters δ, σ_{BT}, σ_{BR}, σ_{WT}, σ_{WR}, and ρ are given in Table 2. The value of θ_U is 1.74.

Table 2. Type I error probability and power of PBE tests under 2 × 4 crossover design with $n_1 = n_2 = 20$.

(Nominal Level 5%; 10,000 Simulations)

					$\rho = 0.75$		$\rho = 1.00$	
σ_{BT}	σ_{BR}	σ_{WT}	σ_{WR}	δ	PBET	FDA	PBET	FDA
0.4	0.4	0.1	0.1	0.4373	0.0335*	0.0143*	0.0255*	0.0000*
				0.1956	0.7539	0.5747	0.9977	0.7853
				0.1383	0.8747	0.7191	1.0000	0.9561
				0	0.9461	0.8330	1.0000	0.9998
0.4	0.4	0.2	0.1	0.4016	0.0401*	0.0169*	0.0309*	0.0001*
				0.1796	0.5738	0.3819	0.9180	0.3542
				0.1270	0.6977	0.4982	0.9751	0.5469
				0	0.8110	0.6075	0.9970	0.7411
0.4	0.4	0.3	0.3	0.5303	0.0388*	0.0268*	0.0336*	0.0143*
				0.2372	0.7440	0.6568	0.8848	0.7300
				0.1677	0.8578	0.7790	0.9623	0.8672
				0	0.9443	0.8888	0.9918	0.9468
0.4	0.6	0.2	0.1	0.7657	0.0311*	0.0303*	0.0257*	0.0135*
				0.2421	0.9933	0.9909	1.0000	1.0000
				0	0.9998	0.9997	1.0000	1.0000
0.4	0.6	0.3	0.3	0.8404	0.0296*	0.0322*	0.0274*	0.0205*
				0.2658	0.9941	0.9947	0.9997	0.9997
				0	1.0000	0.9998	1.0000	1.0000
0.6	0.4	0.1	0.2	0.2345	0.0509*	0.0116*	0.0530*	0.0000*
				0.1049	0.1319	0.0331	0.2701	0.0006
				0.0742	0.1482	0.0377	0.3211	0.0007
				0	0.1674	0.0402	0.3692	0.0011
0.6	0.4	0.3	0.3	0.2850	0.0524*	0.0170*	0.0518*	0.0031*
				0.1275	0.1491	0.0554	0.2072	0.0204
				0.0901	0.1661	0.0673	0.2347	0.0249
				0	0.1892	0.0755	0.2760	0.0309
0.6	0.6	0.1	0.2	0.6928	0.0321*	0.0168*	0.0263*	0.0000*
				0.2191	0.9249	0.8119	1.0000	0.9916
				0	0.9753	0.9054	1.0000	0.9997
0.6	0.6	0.2	0.1	0.6215	0.0344*	0.0122*	0.0256*	0.0000*
				0.1965	0.7977	0.5979	0.9994	0.8000
				0	0.8937	0.7253	1.0000	0.9553
0.6	0.6	0.3	0.3	0.7115	0.0373*	0.0223*	0.0332*	0.0022*
				0.2250	0.8747	0.7558	0.9931	0.8892
				0	0.9471	0.8646	0.9997	0.9670

FDA: FDA's PBE test.

PBET: The PBE test proposed in Section 3.2.

*: Type I error probability

The type I error probability and power for FDA's test and the proposed test are listed in Table 2. The results are based on 10,000 simulation runs. Since 5% is the nominal significant level, it is clear that FDA's test is too conservative and the proposed test is more powerful. The performance of FDA's test becomes worse when ρ becomes larger or the between subject variances are larger than within subject variances.

5. An Example

A single center, randomized, single-blind, 2×2 crossover study was conducted to compare the liquid HSA-free formulation (test formulation) and the standard reconstituted powder formulation (reference formulation) of a drug product intended for treating multiple sclerosis patients. Forty healthy volunteers ($n_1 = n_2 = 20$) were randomly assigned to receive one of the two formulations at Day 1 and Day 14, respectively, after a washout period of 13 days. Blood samples were taken over a 7-day period following each treatment. That is, blood samples were drawn between 5 to 15 minutes pre-dose, at 2, 4, 6, 9, 12, 18, and 21 hours post-dose, and at 24, 30, 36, 48, 72, 96, and 168 hours post-dose. Serum human interferon-beta concentrations were determined by means of a validated assay.

Two pharmacokinetic responses, area under the curve from 0 to 168 hours (AUC) and peak concentration (Cmax), are considered. Statistics for the PBE test are provided in Table 3 for $\theta_U = 1.125$ (the most conservative PBE bound) and $\theta_U = 1.74$, as suggested by FDA (1999). In any case $\hat{\lambda}_U < 0$ and, hence, PBE can be claimed in terms of either AUC or Cmax.

Table 3. Statistics for the PBE and ABE Tests (Section 5).

					The PBE Test		The ABE Test
Variable	$\hat{\delta}$	$\hat{\sigma}^2_{1,1}$	$\hat{\sigma}^2_{TT}$	$\hat{\sigma}^2_{TR}$	$\hat{\lambda}^*_U$	$\hat{\lambda}^{**}_U$	$(\hat{\delta}_-, \hat{\delta}_+)$
AUC	.1868	.4615	.4528	.8539	-.5788	-.7374	(.0057, .3678)
Cmax	.1843	.4101	.2510	.3998	-.2040	-.2820	(.0430, .3255)

$\hat{\lambda}^*_U = \hat{\lambda}_U$ with $\theta_U = 1.125$
$\hat{\lambda}^{**}_U = \hat{\lambda}_U$ with $\theta_U = 1.74$

It is interesting to compare the PBE analysis with the ABE analysis. According to the 1992 or the 2000 FDA guidance, ABE can be claimed if and only if the 95% confidence interval $(\hat{\delta}_-, \hat{\delta}_+)$ is within $(-0.223, 0.223)$, where $\hat{\delta}_\pm = \hat{\delta} \pm t_{0.95;n_1+n_2-2}\hat{\sigma}_{1,1}(\frac{1}{4n_1} + \frac{1}{4n_2})^{1/2}$. Statistics for ABE testing are included in Table 3. It turns out that for both AUC and Cmax, ABE cannot be claimed.

Note that the ABE approach is not suitable for assessment of bioequivalence for highly variable drug products (i.e., the intrasubject CV is greater than 30%).

The ABE approach also penalizes the test product that has smaller variability as compared to the reference product, which is the case for this example (Table 3). As indicated in this example, the PBE analysis provides a more reliable assessment of bioequivalence.

Acknowledgement

The authors are grateful to two referees for their helpful comments and suggestions.

Appendix

1. Proof of the independence of $\hat{\delta}$ and $(\hat{\sigma}^2_{TT}, \hat{\sigma}^2_{TR})$

It suffices to show that $\bar{y}_{11}$ and $\sum_{i=1}^{n_1}(y_{i21} - \bar{y}_{21})^2$ are independent, where $\bar{y}_{jk}$ is the average of y_{ijk}, $i = 1, \ldots, n_k$. Since y's are normally distributed, the result follows from the fact that for each i, $\mathrm{Cov}\,(\bar{y}_{11}, y_{i21} - \bar{y}_{21}) = \mathrm{Cov}\,(\bar{y}_{11}, y_{i21}) - \mathrm{Cov}\,(\bar{y}_{11}, \bar{y}_{21}) = \frac{1}{n_1}\mathrm{Cov}\,(y_{i11}, y_{i21}) - \frac{1}{n_1^2}\sum_{t=1}^{n_1}\mathrm{Cov}\,(y_{t11}, y_{t21}) = 0$.

2. Proof of (8) From the definition of $\hat{\sigma}^2_{TT}$ and $\hat{\sigma}^2_{TR}$, it suffices to show that if (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) bivariate normal random vectors with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}, \tag{15}$$

then $\mathrm{Cov}\,(S_x^2, S_y^2) = 2\rho^2\sigma_x^2\sigma_y^2/(n-1)$, where S_x^2 and S_y^2 are the sample variances of X's and Y's, respectively. Using orthogonal transformations, we can show that

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n-1} a_i^2 \quad \text{and} \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n-1} b_i^2,$$

where (a_i, b_i), $i = 1, \ldots, n-1$, are i.i.d. bivariate normal random vectors with mean 0 and covariance matrix Σ given by (15). Hence, the result follows from the following lemma.

Lemma. *Let (X, Y) be a bivariate normal random vector with mean 0 and covariance matrix Σ given by (15). Then $\mathrm{Cov}\,(X^2, Y^2) = 2\rho^2\sigma_x^2\sigma_y^2$.*

Proof. Let $a = (\sigma_x^{-1}X + \sigma_y^{-1}Y)/\sqrt{2}$ and $b = (\sigma_x^{-1}X - \sigma_y^{-1}Y)/\sqrt{2}$. Then $\mathrm{Var}\,(a) = (1+\rho)$, $\mathrm{Var}\,(b) = (1-\rho)$, and $\mathrm{Cov}\,(a, b) = 0$. Also, $X^2Y^2 = 0.25\sigma_x^2\sigma_y^2(a^2 - b^2)^2$. Hence,

$$\begin{aligned} E(X^2Y^2) &= 0.25\sigma_x^2\sigma_y^2 E(a^4 + b^4 - 2a^2b^2) \\ &= 0.25\sigma_x^2\sigma_y^2[3(1+\rho)^2 + 3(1-\rho)^2 - 2(1+\rho)(1-\rho)] \\ &= 0.25\sigma_x^2\sigma_y^2(6 + 6\rho^2 - 2 + 2\rho^2) = \sigma_x^2\sigma_y^2(1 + 2\rho^2) \end{aligned}$$

Therefore, $\operatorname{Cov}(X^2, Y^2) = E(X^2Y^2) - E(X^2)E(Y^2) = 2\rho^2\sigma_x^2\sigma_y^2$.

3. Derivation of formula 14 for sample size determination

Assume that $n_1 = n_2 = n$. It follows from the previous proof that

$$C_k \to 2\begin{pmatrix} \sigma_{TT}^4 & \rho^2\sigma_{BT}^2\sigma_{BR}^2 \\ \rho^2\sigma_{BT}^2\sigma_{BR}^2 & \sigma_{TR}^4 \end{pmatrix} \quad \text{in probability.}$$

Hence, $nV_c \to V_\infty = 2\delta^2\sigma_{1,1}^2 + \sigma_{TT}^4 + c^2\sigma_{TR}^4 - 2c\rho^2\sigma_{BT}^2\sigma_{BR}^2$ in probability, where $V_c = V$ when $c = 1 + \theta_U$ and $V_c = V_0$ when $c = 1$. Let $\hat{\lambda}$ be given by (10). Then $\hat{\lambda}_U = \hat{\lambda} + t_{0.95;2n-2}\sqrt{V_c}$ and

$$\begin{aligned}\beta &\approx P\left(\hat{\lambda} - t_{\beta;2n-2}\sqrt{V_c} \le \lambda\right)\\ &= P\left(\sqrt{n}(\hat{\lambda}_U - \lambda) \le (t_{0.95;2n-2} + t_{\beta;2n-2})\sqrt{nV_c}\right)\\ &\approx P\left(\sqrt{n}(\hat{\lambda}_U - \lambda) \le (z_{0.95} + z_\beta)\sqrt{V_\infty}\right),\end{aligned}$$

since $t_{\beta;2n-2} \to z_\beta$. Since the power of the PBE test is $P(\hat{\lambda}_U < 0) = P(\sqrt{n}(\hat{\lambda} - \lambda) < -\sqrt{n}\lambda)$, it is asymptotically no smaller than β if $-\sqrt{n}\lambda \ge (z_{0.95} + z_\beta)\sqrt{V_\infty}$, which is the same as (14) since $\lambda < 0$.

References

Anderson, S. and Hauck, W. W. (1990). Considerations of individual bioequivalence. *J. Pharmacokinetics and Biopharmaceutics* **8**, 259-273.

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

Chow, S. C. (1999). Individual bioequivalence — a review of the FDA draft guidance. *Drug Inform. J.* **33**, 435-444.

Chow, S. C. and Liu, J. P. (1995). *Statistical Design and Analysis in Pharmaceutical Science.* Marcel Dekker, New York, New York.

Chow, S. C. and Liu, J. P. (1999). *Design and Analysis of Bioavailability and Bioequivalence Studies.* 2nd edition. Marcel Dekker, New York.

Chow, S. C., Shao, J. and Wang, H. (2002). Individual bioequivalence testing under 2 × 3 designs. *Statist. Medicine* **21**, 629-648.

Esinhart, J. D. and Chinchilli, V. M. (1994). Extension to the use of tolerance intervals for assessment of individual bioequivalence. *J. Biopharm. Statist.* **4**, 39-52.

FDA (1992). *Guidance on Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design.* Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

FDA (2000). *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations.* Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

FDA (2001). *Guidance for Industry on Statistical Approaches to Establishing Bioequivalence.* Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

Graybill, F. A. and Wang, C. M. (1980). Confidence intervals on nonnegative linear combinations of variances. *J. Amer. Statist. Assoc.* **75**, 869-873.

Howe, W. G. (1974). Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *J. Amer. Statist. Assoc.* **69**, 789-794.

Hyslop, T., Hsuan, F. and Holder, D. J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Statist. Medicine* **19**, 2885-2897.

Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Cross-Over Trials*. Chapman and Hall, London.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Springer, New York.

Quiroz, J., Ting, N., Wei, G. C. G. and Burdick, R. K. (2000). A modified large sample approach in assessment of population bioequivalence. *J. Biopharm. Statist.* **10**, 527-544.

Schall, R. and Luus, R. E. (1993). On population and individual bioequivalence. *Statist. Medicine* **12**, 1109-1124.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statist. Medicine* **11**, 1777-1788.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S. and Lu, T.-F. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *J. Statist. Comput. Simulation* **35**, 135-143.

StatPlus, Inc., Heston Hall, Suite 206, 1790 Yardley-Langhorne Road, Yardley, PA 19067.

Department of Statistics, University of Wisconsin, 1210 Wst Dayton Street, Madison, WI 53706-1685.

E-mail: shao@stat.wisc.edu

E-mail: hansheng@stat.wisc.edu

(Received August 2001; accepted August 2002)