STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:629–648 (DOI: 10.1002/sim.1056)

Individual bioequivalence testing under 2×3 designs

Shein-Chung Chow[1], Jun Shao[2,*] and Hansheng Wang[2]

[1] *Statplus Inc., Heston Hall, Suite 206, 1790 Yardley-Langhorne Road, Yardley, PA 19067, U.S.A.*
[2] *Department of Statistics, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706, U.S.A.*

SUMMARY

In recent years, as more generic drug products become available, it is a concern not only whether generic drug products that have been approved based on the regulation of average bioequivalence will have the same quality, safety and efficacy as that of the brand-name drug product, but also whether the approved generic drug products can be used interchangeably. In its recent draft guidance, the U.S. Food and Drug Administration (FDA) recommends that individual bioequivalence (IBE) be assessed using the method proposed by Hyslop, Hsuan, and Holder to address drug switchability. The FDA suggests that a 2×4 cross-over design be considered for assessment of IBE, while a 2×3 cross-over design may be used as an alternative design to reduce the length and cost of the study. Little or no information regarding the statistical procedures under 2×3 cross-over designs is discussed in the guidance. In this paper, a detailed statistical procedure for assessment of IBE under 2×3 cross-over designs is derived. The main purpose of this paper, however, is to derive an IBE test under an alternative 2×3 design and show that the resulting IBE test is better than that under a 2×3 cross-over design and is comparable to or even better than that under a 2×4 cross-over design. Our conclusions are supported by theoretical considerations and empirical results. Furthermore, a method of determining the sample sizes required for IBE tests to reach a given level of power is proposed. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: cross-over design; Cornish–Fisher expansion; power; type I error; sample size

1. INTRODUCTION

An approved generic drug can be used as a substitute for a brand-name drug that is going off patent. In 1984, the U.S. Food and Drug Administration (FDA) was authorized to approve generic drugs through *bioavailability* and *bioequivalence* studies under the *Drug Price and Patent Term Restoration Act*. As defined in 21 CFR 320.1, bioavailability refers to the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action. *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 formulations of the same drug product or two drug products (for example, a brand-name drug and its generic copy) are equivalent in bioavailability, they will

* Correspondence to: Jun Shao, Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, WI 53706-1685, U.S.A.

Received January 2001
Accepted June 2001

reach the same therapeutic effect or they are therapeutically equivalent [1]. 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. Throughout this paper, we consider the *in vivo* bioequivalence between a reference formulation (for example, the brand-name drug) and a test formulation (for example, a generic copy).

In 1992, the FDA published its first guidance on statistical procedures for *in vivo* bioequivalence studies [2]. It requires the assessment of bioequivalence in average PK responses between the reference and test formulations, which is commonly referred to as average bioequivalence (ABE). The requirement of ABE is also indicated in the FDA's most recent guidance on bioequivalence studies for orally administered drug products [3]. The ABE approach for bioequivalence, however, has limitations for addressing drug switchability, since it focuses only on the comparison of population averages between the test and reference formulations. Drug switchability is referred to as the switch from a drug product to an alternative drug product within the same patient. To assess drug switchability, *individual bioequivalence* (IBE) testing is proposed [4–10]. Let y_{T} be the PK response from the test formulation, and y_{R} and y'_{R} be two identically distributed PK responses from an individual under the reference formulation. Then the drug switchability can be measured by

$$\theta = \begin{cases} \dfrac{E(y_{\mathrm{R}}-y_{\mathrm{T}})^2 - E(y_{\mathrm{R}}-y'_{\mathrm{R}})^2}{E(y_{\mathrm{R}}-y'_{\mathrm{R}})^2/2} & \text{if } E(y_{\mathrm{R}}-y'_{\mathrm{R}})^2/2 \geqslant \sigma_0^2 \\ \dfrac{E(y_{\mathrm{R}}-y_{\mathrm{T}})^2 - E(y_{\mathrm{R}}-y'_{\mathrm{R}})^2}{\sigma_0^2} & \text{if } E(y_{\mathrm{R}}-y'_{\mathrm{R}})^2/2 < \sigma_0^2 \end{cases} \tag{1}$$

where σ_0^2 is a given constant specified in the 2001 FDA guidance [8]. According to the 2001 FDA guidance, IBE can be claimed if the following null hypothesis H_0 is rejected at the 5 per cent level of significance:

$$H_0\colon\ \theta \geqslant \theta_U \quad \text{versus} \quad H_1\colon\ \theta < \theta_U \tag{2}$$

where θ_U is an upper limit specified in the 2001 FDA guidance.

For *in vivo* bioequivalence testing, cross-over designs (see, for example, references [1, 11]) are usually considered. For the ABE, a standard two-sequence two-period (2×2) cross-over design is recommended by reference [2]. For the IBE, however, the standard 2×2 cross-over design is not useful, because each subject only receives each formulation once, and thus it is not possible to obtain an unbiased estimator of within-subject variation $E(y_{\mathrm{R}} - y'_{\mathrm{R}})^2$ in (1). Thus, a two-sequence four-period (2×4) cross-over design, in which each subject receives each formulation twice, is recommended by reference [8] for IBE testing. A typical 2×4 cross-over design is (TRTR,RTRT) or (TRRT,RTTR), which means that subjects in two sequences receive T = test formulation and R = reference formulation in the order of TRTR and RTRT, respectively, or in the order of TRRT and RTTR, respectively. A statistical test for the IBE hypothesis (2) based on data from a 2×4 cross-over design is given in reference [8], which is based on a method proposed by Hyslop *et al* [12].

Since a 2×4 cross-over design requires four observations from each subject, it may substantially increase the length and the overall cost of the study. As an alternative, the 2001 FDA guidance suggests a 2×3 cross-over design such as (TRT,RTR) or (TRR,RTT). Another type of 2×3 design was considered in references [9, 13], which is obtained by adding an extra reference period to the 2×2 cross-over design, that is (TRR,RTR). Since an extra R

formulation is received in the third period of this design, we call it the 2×3 *extra-reference* design. The main purposes of this paper are:

1. to derive IBE tests for the previously described 2×3 designs;
2. to show that the IBE test under the 2×3 extra-reference design is not only better than that under any 2×3 cross-over design, but also comparable to or even better than the IBE test under a 2×4 cross-over design;
3. to derive a formula for determining the sample size required to achieve the desired power in IBE testing.

Statistical tests for the IBE hypothesis (2) based on data from 2×3 designs are given in Section 2, along with some discussions on why the IBE test under a 2×3 cross-over design is inefficient and why the 2×3 extra-reference design is better and comparable to 2×4 cross-over designs. Section 3 contains simulation results on the type I error probability and power of IBE tests under 2×4 and 2×3 designs. Sample size determination is considered in Section 4.

2. IBE TESTS UNDER 2×3 DESIGNS

Let y_{ijk} be the original or the log-transformation of the PK response of interest from the ith subject in the kth sequence at the jth period of the experiment, $i = 1, \ldots, n_k$, $k = 1, 2$, and $j = 1, \ldots, 3$ or 4. 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 dosing period to the next dosing period. We consider the following statistical model:

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

where μ is the overall mean, F_l is the fixed effect of the lth formulation ($l = \mathrm{T}$ or R according to the design and $F_\mathrm{T} + F_\mathrm{R} = 0$), 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), $l = \mathrm{T}$, R), S_{ikl} is the random effect of the ith subject in the kth sequence under formulation l and $(S_{ik\mathrm{T}}, S_{ik\mathrm{R}})$, $i = 1, \ldots, n_k$, $k = 1, 2$, are independent and identically distributed bivariate normal random vectors with mean 0 and an unknown covariance matrix

$$\begin{pmatrix} \sigma^2_{\mathrm{BT}} & \rho\sigma_{\mathrm{BT}}\sigma_{\mathrm{BR}} \\ \rho\sigma_{\mathrm{BT}}\sigma_{\mathrm{BR}} & \sigma^2_{\mathrm{BR}} \end{pmatrix}$$

e_{ijk}'s are independent random errors distributed as $\mathrm{N}(0, \sigma^2_{\mathrm{W}l})$, and S_{ikl}'s and e_{ijk}'s are mutually independent. Note that σ^2_{BT} and σ^2_{BR} are between-subject variances and σ^2_{WT} and σ^2_{WR} are within-subject variances. Under model (3), θ in (1) is equal to

$$\theta = \frac{\delta^2 + \sigma^2_{\mathrm{D}} + \sigma^2_{\mathrm{WT}} - \sigma^2_{\mathrm{WR}}}{\max\{\sigma^2_0, \sigma^2_{\mathrm{WR}}\}} \tag{4}$$

where $\delta = F_\mathrm{T} - F_\mathrm{R}$ and $\sigma^2_{\mathrm{D}} = \sigma^2_{\mathrm{BT}} + \sigma^2_{\mathrm{BR}} - 2\rho\sigma_{\mathrm{BT}}\sigma_{\mathrm{BR}}$ is the variance of $S_{ik\mathrm{T}} - S_{ik\mathrm{R}}$, which is referred to as the variance due to the subject-by-formulation interaction.

2.1. The method of constructing confidence bounds

Note that the hypotheses given in (2) are equivalent to

$$H_0: \gamma \geqslant 0 \quad \text{versus} \quad H_1: \gamma < 0 \tag{5}$$

where

$$\gamma = \delta^2 + \sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{WT}}^2 - \sigma_{\mathrm{WR}}^2 - \theta_U \max\{\sigma_0^2, \sigma_{\mathrm{WR}}^2\} \tag{6}$$

Therefore, it suffices to find a 95 per cent upper confidence bound $\hat{\gamma}_U$ for γ. IBE is concluded if and only if $\hat{\gamma}_U < 0$. A $\hat{\gamma}_U$, proposed by Hyslop *et al.* [12] and recommended in the 2001 FDA guidance, is based on the following result in references [14–16]. If $\gamma = \gamma_1 + \cdots + \gamma_r - \gamma_{r+1} - \cdots - \gamma_m$, where γ_j's are positive parameters, then an approximate upper confidence bound is

$$\hat{\gamma}_1 + \cdots + \hat{\gamma}_r - \hat{\gamma}_{r+1} - \cdots - \hat{\gamma}_m + \sqrt{\{(\tilde{\gamma}_1 - \hat{\gamma}_1)^2 + \cdots + (\tilde{\gamma}_m - \hat{\gamma}_m)^2\}}$$

where $\hat{\gamma}_j$ is an estimator of γ_j, $\tilde{\gamma}_j$ is a 95 per cent upper confidence bound for γ_j when $j = 1, \ldots, r$, $\tilde{\gamma}_j$ is a 95 per cent lower confidence bound for γ_j when $j = r+1, \ldots, m$, and $\hat{\gamma}_j$'s are independent.

The key in applying the method in reference [12] is that γ in (6) can be decomposed into $\gamma_1 \pm \cdots \pm \gamma_m$ so that approximately unbiased and *independent chi-square distributed* estimators of γ_j's can be obtained. Although γ in (6) is a function of δ^2, σ_{D}^2, σ_{WT}^2, and σ_{WR}^2, it is impossible to find a chi-square distributed unbiased estimator of σ_{D}^2 that is independent from estimators of other variance components. For example, under a 2×4 cross-over design, Hyslop *et al.* [12] considered the following decomposition of γ:

$$\gamma = \delta^2 + \sigma_{0.5,0.5}^2 + 0.5\sigma_{\mathrm{WT}}^2 - 1.5\sigma_{\mathrm{WR}}^2 - \theta_U \max\{\sigma_0^2, \sigma_{\mathrm{WR}}^2\} \tag{7}$$

where $\sigma_{0.5,0.5}^2$ is

$$\sigma_{a,b}^2 = \sigma_{\mathrm{D}}^2 + a\sigma_{\mathrm{WT}}^2 + b\sigma_{\mathrm{WR}}^2 \tag{8}$$

with $a = 0.5$ and $b = 0.5$, and they derived unbiased and chi-square distributed estimators of δ^2, $\sigma_{0.5,0.5}^2$, σ_{WT}^2, and σ_{WR}^2.

Under a different design, however, a different decomposition of γ may be needed.

2.2. 2×3 Cross-over designs

Without loss of generality, assume that sequence 1 has two test formulations and sequence 2 has two reference formulations. For the ith subject in sequence k, let x_{ilk} be the average of observations under formulation l and z_{ilk} be the difference between the two observations under the same formulation. Then, an unbiased estimator of δ is

$$\hat{\delta} = \frac{\bar{x}_{\mathrm{T}1} - \bar{x}_{\mathrm{R}1} + \bar{x}_{\mathrm{T}2} - \bar{x}_{\mathrm{R}2}}{2} \sim \mathrm{N}\left(\delta, \frac{\sigma_{0.5,1}^2}{4n_1} + \frac{\sigma_{1,0.5}^2}{4n_2}\right)$$

where $\bar{x}_{lk}$ is the sample mean based on x_{ilk} for a fixed (l,k) and $\sigma^2_{a,b}$ is given by (8); an unbiased estimator of $\sigma^2_{0.5,1}$ is

$$\hat{\sigma}^2_{0.5,1} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{i\mathrm{T}1} - x_{i\mathrm{R}1} - \bar{x}_{\mathrm{T}1} + \bar{x}_{\mathrm{R}1})^2 \sim \frac{\sigma^2_{0.5,1}\chi^2_{n_1-1}}{n_1 - 1}$$

an unbiased estimator of $\sigma^2_{1,0.5}$ is

$$\hat{\sigma}^2_{1,0.5} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{i\mathrm{T}2} - x_{i\mathrm{R}2} - \bar{x}_{\mathrm{T}2} + \bar{x}_{\mathrm{R}2})^2 \sim \frac{\sigma^2_{1,0.5}\chi^2_{n_2-1}}{n_2 - 1}$$

an unbiased estimator of σ^2_{WT} is

$$\hat{\sigma}^2_{\mathrm{WT}} = \frac{1}{2(n_1 - 1)}\sum_{i=1}^{n_1}(z_{i\mathrm{T}1} - \bar{z}_{\mathrm{T}1})^2 \sim \frac{\sigma^2_{\mathrm{WT}}\chi^2_{n_1-1}}{n_1 - 1}$$

and an unbiased estimator of σ^2_{WR} is

$$\hat{\sigma}^2_{\mathrm{WR}} = \frac{1}{2(n_2 - 1)}\sum_{i=1}^{n_2}(z_{i\mathrm{R}2} - \bar{z}_{\mathrm{R}2})^2 \sim \frac{\sigma^2_{\mathrm{WR}}\chi^2_{n_2-1}}{n_2 - 1}$$

where $\bar{z}_{lk}$ is the sample mean based on z_{ilk} for a fixed (l,k). Furthermore, estimators $\hat{\delta}$, $\hat{\sigma}^2_{0.5,1}$, $\hat{\sigma}^2_{1,0.5}$, $\hat{\sigma}^2_{\mathrm{WT}}$ and $\hat{\sigma}^2_{\mathrm{WR}}$ are independent. The independence of $\hat{\delta}$ and $\hat{\sigma}^2_{\mathrm{WT}}$ follows from the fact that $\mathrm{cov}(x_{i\mathrm{T}1} - x_{i\mathrm{R}1}, z_{i\mathrm{T}1}) = 0$. Independence of other estimators can be shown similarly. Decomposing γ in (6) as

$$\gamma = \delta^2 + 0.5(\sigma^2_{0.5,1} + \sigma^2_{1,0.5}) + 0.25\sigma^2_{\mathrm{WT}} - 1.75\sigma^2_{\mathrm{WR}} - \theta_U \max\{\sigma^2_0, \sigma^2_{\mathrm{WR}}\} \tag{9}$$

and applying the method in Section 2.1, we can obtain the following approximate 95 per cent upper confidence bound $\hat{\gamma}_U$ for γ. When $\sigma^2_{\mathrm{WR}} \geqslant \sigma^2_0$,

$$\hat{\gamma}_U = \hat{\delta}^2 + 0.5(\hat{\sigma}^2_{0.5,1} + \hat{\sigma}^2_{1,0.5}) + 0.25\hat{\sigma}^2_{\mathrm{WT}} - (1.75 + \theta_U)\hat{\sigma}^2_{\mathrm{WR}} + \sqrt{U} \tag{10}$$

where U is the sum of the following five quantities:

$$\left[\left(|\hat{\delta}| + t_{0.95;n_1+n_2-2}\sqrt{\left\{\frac{\hat{\sigma}^2_{0.5,1}}{4n_1} + \frac{\hat{\sigma}^2_{1,0.5}}{4n_2}\right\}}\right)^2 - \hat{\delta}^2\right]^2$$

$$0.5^2\hat{\sigma}^4_{0.5,1}\left(\frac{n_1 - 1}{\chi^2_{0.05;n_1-1}} - 1\right)^2$$

$$0.5^2\hat{\sigma}^4_{1,0.5}\left(\frac{n_2 - 1}{\chi^2_{0.05;n_2-1}} - 1\right)^2$$

$$0.25^2\hat{\sigma}_{\mathrm{WT}}^4\left(\frac{n_1-1}{\chi^2_{0.05;n_1-1}}-1\right)^2$$

$$(1.75+\theta_U)^2\hat{\sigma}_{\mathrm{WR}}^4\left(\frac{n_2-1}{\chi^2_{0.95;n_2-1}}-1\right)^2 \tag{11}$$

and $t_{a;r}$ and $\chi^2_{a;r}$ are, respectively, the ath quantiles of the t-distribution and chi-square distribution with r degrees of freedom. When $\sigma^2_{\mathrm{WR}}<\sigma_0^2$

$$\hat{\gamma}_U=\hat{\delta}^2+0.5(\hat{\sigma}^2_{0.5,1}+\hat{\sigma}^2_{1,0.5})+0.25\hat{\sigma}^2_{\mathrm{WT}}-1.75\hat{\sigma}^2_{\mathrm{WR}}-\theta_U\sigma_0^2+\sqrt{U_0} \tag{12}$$

where U_0 is the same as U except that the quantity in (11) should be replaced by

$$1.75^2\hat{\sigma}_{\mathrm{WR}}^4\left(\frac{n_2-1}{\chi^2_{0.95;n_2-1}}-1\right)^2$$

Note that the IBE test procedure does not depend on any particular choice of a 2×3 cross-over design.

For any non-negative a_1 and a_2 such that $a_1+a_2=1$, decomposition (9) is a special case of the following decomposition:

$$\gamma=\delta^2+a_1\sigma^2_{0.5,1}+a_2\sigma^2_{1,0.5}+0.5a_1\sigma^2_{\mathrm{WT}}-(1.5+0.5a_2)\sigma^2_{\mathrm{WR}}-\theta_U\max\{\sigma_0^2,\sigma^2_{\mathrm{WR}}\}$$

A confidence bound similar to that in (10) can be obtained using this decomposition. The best choice of a_i, however, depends on unknown variance components. When n_1 and n_2 are nearly the same, which is the case in most bioequivalence studies, $a_1=a_2=0.5$ (decomposition (9)) is intuitively a reasonable choice.

The confidence bound $\hat{\gamma}_U$ in (10) is referred to as the confidence bound under the reference-scaled criterion, whereas $\hat{\gamma}_U$ in (12) is referred to as the confidence bound under the constant-scaled criterion. In practice, whether $\sigma^2_{\mathrm{WR}}\geqslant\sigma_0^2$ is usually unknown. Hyslop *et al.* [12] recommended using the reference-scaled criterion or the constant-scaled criterion according to $\hat{\sigma}^2_{\mathrm{WR}}\geqslant\sigma_0^2$ or $\hat{\sigma}^2_{\mathrm{WR}}<\sigma_0^2$, which will be called the estimation method. Intuitively, the estimation method works if the true value of σ^2_{WR} is not close to σ_0^2. Alternatively, we may test the hypothesis of $\sigma^2_{\mathrm{WR}}\geqslant\sigma_0^2$ versus $\sigma^2_{\mathrm{WR}}<\sigma_0^2$ to decide which confidence bound should be used, that is, if $\hat{\sigma}^2_{\mathrm{WR}}(n_1+n_2-2)/\chi^2_{0.05;n_1+n_2-2}\geqslant\sigma_0^2$, then $\hat{\gamma}_U$ in (10) should be used; otherwise $\hat{\gamma}_U$ in (12) should be used. This will be called the test method and is more conservative than the estimation method. Some comparisons of these two methods are given in Section 3.

2.3. The 2 × 3 extra-reference design

The procedure in Section 2.2 has two disadvantages. First, the decomposition in (9) involves five unknown parameters, instead of four unknown parameters in the case where a 2×4 cross-over design is used. Second, confidence bounds for variance components of $\sigma^2_{0.5,1}$, $\sigma^2_{1,0.5}$, σ^2_{WT}, and σ^2_{WR} are constructed using chi-square distributions with degrees of freedom n_1-1 or n_2-1, instead of n_1+n_2-2 as in the case of a 2×4 cross-over design. Consequently, the power for IBE testing based on a 2×3 cross-over design is low.

Instead of trying to find a better IBE test under a 2×3 cross-over design, we now show that a better IBE test can be derived under the 2×3 extra-reference design (TRR,RTR), which requires the same number of observations as that of any 2×3 cross-over design. The number of estimable effects in the 2×3 extra-reference design is the same as that of any 2×3 cross-over design, although confounding patterns are different. Under the 2×3 extra-reference design, the number of observations under reference formulation is the same as that in a 2×4 cross-over design and, hence, σ^2_{WR} can be estimated with the same accuracy as that in a 2×4 cross-over design. Although σ^2_{WT} cannot be estimated under the 2×3 extra-reference design, the estimation of σ^2_{WT} can be avoided by considering the following decomposition of γ in (6):

$$\gamma = \delta^2 + \sigma^2_{\mathrm{I},0.5} - 1.5\sigma^2_{\mathrm{WR}} - \theta_U \max\{\sigma_0^2, \sigma^2_{\mathrm{WR}}\} \tag{13}$$

Furthermore, there are only three unknown parameters in (13).

We now show how to obtain unbiased and independent estimators of δ, $\sigma^2_{\mathrm{I},0.5}$ and σ^2_{WR}. Let x_{ilk} and z_{ilk} be the same as those in Section 2.2. Then, an unbiased estimator of δ is

$$\hat{\delta} = \frac{\bar{x}_{\mathrm{T}1} - \bar{x}_{\mathrm{R}1} + \bar{x}_{\mathrm{T}2} - \bar{x}_{\mathrm{R}2}}{2} \sim \mathrm{N}\left(\delta, \frac{\sigma^2_{\mathrm{I},0.5}}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

where $\sigma^2_{a,\,b}$ is given by (8); an unbiased estimator of $\sigma^2_{\mathrm{I},0.5}$ is

$$\hat{\sigma}^2_{\mathrm{I},0.5} = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_{i\mathrm{T}k} - x_{i\mathrm{R}k} - \bar{x}_{\mathrm{T}k} + \bar{x}_{\mathrm{R}k})^2 \sim \frac{\sigma^2_{\mathrm{I},0.5}\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 1}$$

and an unbiased estimator of σ^2_{WR} is

$$\hat{\sigma}^2_{\mathrm{WR}} = \frac{1}{2(n_1 + n_2 - 2)} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (z_{i\mathrm{R}k} - \bar{z}_{\mathrm{R}k})^2 \sim \frac{\sigma^2_{\mathrm{WR}}\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}$$

Furthermore, we can show that estimators $\hat{\delta}$, $\hat{\sigma}^2_{\mathrm{I},0.5}$ and $\hat{\sigma}^2_{\mathrm{WR}}$ are independent. Hence the method described in Section 2.1 can be applied.

When $\sigma^2_{\mathrm{WR}} \geqslant \sigma_0^2$, the reference-scaled confidence bound for γ is

$$\hat{\gamma}_U = \hat{\delta}^2 + \hat{\sigma}^2_{\mathrm{I},0.5} - (1.5 + \theta_U)\hat{\sigma}^2_{\mathrm{WR}} + \sqrt{U} \tag{14}$$

where U is the sum of the following three quantities:

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

$$\hat{\sigma}^4_{\mathrm{I},0.5}\left(\frac{n_1 + n_2 - 2}{\chi^2_{0.05;n_1+n_2-2}} - 1\right)^2$$

and

$$(1.5 + \theta_U)^2\hat{\sigma}^4_{\mathrm{WR}}\left(\frac{n_1 + n_2 - 2}{\chi^2_{0.95;n_1+n_2-2}} - 1\right)^2 \tag{15}$$

Table I.

Design	Decomposition of γ	Number of components to be estimated	Degrees of freedom for variance estimators	Variance of $2\hat{\delta}$
2×4 cross-over	Formula (7)	4	n_1+n_2-2	$\sigma^2_{0.5,0.5}(n_1^{-1}+n_2^{-1})$
2×3 cross-over	Formula (9)	5	n_1-1 or n_2-1	$\sigma^2_{0.5,1}n_1^{-1}+\sigma^2_{1,0.5}n_2^{-1}$
2×3 extra-reference	Formula (13)	3	n_1+n_2-2	$\sigma^2_{1,0.5}(n_1^{-1}+n_2^{-1})$

When $\sigma^2_{\mathrm{WR}}<\sigma^2_0$, the constant-scaled confidence interval for γ is

$$\hat{\gamma}_U=\hat{\delta}^2+\hat{\sigma}^2_{1,0.5}-1.5\hat{\sigma}^2_{\mathrm{WR}}-\theta_U\sigma^2_0+\sqrt{U_0} \tag{16}$$

where U_0 is the same as U except that the quantity in (15) should be replaced by

$$1.5^2\hat{\sigma}^4_{\mathrm{WR}}\left(\frac{n_1+n_2-2}{\chi^2_{0.95;n_1+n_2-2}}-1\right)^2$$

The methods discussed at the end of Section 2.2 can be applied to decide whether the reference-scaled bound (14) or the constant-scaled confidence bound (16) should be used.

Note that Wang [13] derived a different IBE test under the 2×3 extra-reference design. However, our proposed test procedure is constructed using the same idea in reference [12], which is recommended in the 2001 FDA guidance.

2.4. Comparison

To compare different designs, we summarize the main feature of each design as shown in Table I.

In terms of the number of components required to be estimated (the smaller the better) and the degrees of freedom for variance component estimators (the larger the better), 2×3 cross-over designs are the worst and the 2×3 extra-reference design is the best. In terms of the estimation of δ, 2×3 cross-over designs are better than the 2×3 extra-reference design if and only if $\sigma^2_{1,0.5}>\sigma^2_{0.5,1}$, which is the same as $\sigma^2_{\mathrm{WT}}>\sigma^2_{\mathrm{WR}}$, that is, the test formulation is more variable than the reference formulation, a situation which generic drug companies should try to avoid (see the discussion in Section 3). In terms of the estimation of δ, 2×4 cross-over designs are the best, since $\sigma^2_{0.5,0.5}$ is smaller than both $\sigma^2_{0.5,1}$ and $\sigma^2_{1,0.5}$. However, this comparison is somewhat unfair because 2×3 designs require only 75 per cent of the observations in a 2×4 cross-over design. If $4n_1/3$ and $4n_2/3$ are integers and are used as sample sizes in the two sequences of the 2×3 extra-reference design so that the total number of observations is the same as that of a 2×4 cross-over design having sample sizes n_1 and n_2, the 2×3 extra-reference design is more efficient than the 2×4 cross-over design when σ^2_{WR} or σ^2_{D} is large. This is because: (i) the degree of freedom for the confidence bound of σ^2_{WR} is $4(n_1+n_2)/3-2$ for the 2×3 extra-reference design and, thus, the gain in having a large degree of freedom is more when σ^2_{WR} is larger; (ii) the variance of $\hat{\delta}$ under a 2×4 cross-over design over the variance of $\hat{\delta}$ under the 2×3 extra-reference design is $4\sigma^2_{0.5,0.5}/3\sigma^2_{1,0.5}$, which is larger than one if and only if $\sigma^2_{\mathrm{D}}+0.5\sigma^2_{\mathrm{WR}}>\sigma^2_{\mathrm{WT}}$.

Therefore, the conclusion of the comparison is that 2×3 cross-over designs are not as good as the other two types of designs and the 2×3 extra-reference design is comparable to or even better than 2×4 cross-over designs. These conclusions are supported by empirical results on the type I error probability and power of the IBE tests based on these three types of designs.

3. SIMULATION RESULTS

A simulation study is carried out to investigate the following issues:

1. The type I error probability of the IBE tests. The IBE tests are based on asymptotic theory, although Cornish–Fisher's expansion is used to improve the convergence rate. Thus, it is important to empirically study the type I error probability of these tests for finite sample sizes.
2. The relative performance of the three types of designs in terms of the power of the corresponding IBE tests.
3. The relative performance of the two methods (the estimation method and the test method, see the end of Section 2.2) of determining whether the reference-scaled or the constant-scaled confidence bound should be used in IBE testing.

Four population parameters affect the performance of IBE tests, δ^2, σ_D^2, σ_{WT}^2 and σ_{WR}^2, which determine the value of γ in (6). We consider parameter values similar to those in reference [12]. That is, $\sigma_D = 0$ and 0.2, $\sigma_{WT} = 0.15$, 0.2, 0.3 and 0.5, and $\sigma_{WR} = 0.15$, 0.2, 0.3 and 0.5. We also consider the cases where $\sigma_{WT} \neq \sigma_{WR}$, which are not considered in reference [12]. For the sample sizes, we consider $n = n_1 = n_2 = 10, 15, 20, 25, 30, 35$ and 40. The values of σ_0 and θ_U are 0.2 and 2.4948, the same as those in reference [12].

Under each parameter and sample size combination, 10 000 simulations are used to compute the empirical type I error probability and power for IBE tests based on the 2×4 design (TRTR,RTRT), the 2×3 design (TRT,RTR), and the 2×3 extra-reference design. Although particular 2×3 and 2×4 cross-over designs are used, our results are applicable to all other cross-over designs, since IBE tests do not depend on choices of cross-over designs. The empirical type I error probability are given in Tables II and III for a total of 16 different combinations of parameter values. Results for other parameter combinations are similar and not reported here. In Table II, the columns under ESTI and TEST give the results obtained by using, respectively, the estimation method and the test method of determining whether the reference-scaled or the constant-scaled confidence bound should be used (see the end of Section 2.2), and the column under OPTI provides the results obtained by assuming that we know whether $\sigma_{WR}^2 \geqslant \sigma_0^2$ (which is used as a standard). However, when $\sigma_{WR} \geqslant 0.3$, simulation results under ESTI, TEST and OPTI are the same and, hence, only one column for each design is shown in Table III.

Simulated power of each test is plotted in Figures 1–4 for the parameter combinations considered in Tables II and III. When $\sigma_{WR}^2 \leqslant \sigma_0^2$, we only consider the estimation method (ESTI). Figures 1 and 2 show the power versus the sample size $n = n_1 = n_2$ with some fixed $\gamma < 0$, whereas Figures 3 and 4 show the power versus γ when the sample size n is fixed at 15 for the 2×4 design and at 20 for the 2×3 designs. Note that a 2×4 design with 15

Table II. Type I error probability of IBE tests when $\sigma_{WR} \leqslant 0.2 = \sigma_0$ (10 000 simulations).

Parameter					2×4 cross-over			2×3 cross-over			2×3 extra-reference		
δ	σ_D	σ_{WT}	σ_{WR}	n	ESTI	TEST	OPTI	ESTI	TEST	OPTI	ESTI	TEST	OPTI
0.3159	0	0.15	0.15	10	0.0471	0.0126	0.0495	0.0341	0.0065	0.0429	0.0514	0.0146	0.0535
				15	0.0533	0.0165	0.0549	0.0397	0.0076	0.0450	0.0558	0.0250	0.0574
				20	0.0506	0.0243	0.0512	0.0416	0.0070	0.0456	0.0522	0.0277	0.0525
				25	0.0535	0.0325	0.0537	0.0470	0.0106	0.0512	0.0536	0.0354	0.0539
				30	0.0538	0.0391	0.0539	0.0506	0.0155	0.0524	0.0550	0.0428	0.0551
				35	0.0516	0.0413	0.0516	0.0522	0.0188	0.0533	0.0508	0.0427	0.0508
				40	0.0536	0.0485	0.0536	0.0505	0.0222	0.0508	0.0516	0.0477	0.0516
0.2869	0	0.2	0.15	10	0.0413	0.0095	0.0432	0.0269	0.0057	0.0321	0.0460	0.0162	0.0473
				15	0.0462	0.0174	0.0477	0.0324	0.0073	0.0366	0.0511	0.0242	0.0515
				20	0.0467	0.0237	0.0474	0.0404	0.0085	0.0444	0.0551	0.0319	0.0555
				25	0.0516	0.0352	0.0518	0.0423	0.0102	0.0439	0.0536	0.0394	0.0538
				30	0.0480	0.0349	0.0481	0.0445	0.0138	0.0461	0.0483	0.0378	0.0483
				35	0.0496	0.0441	0.0496	0.0465	0.0181	0.0469	0.0492	0.0437	0.0492
				40	0.0503	0.0452	0.0503	0.0476	0.0221	0.0480	0.0488	0.0446	0.0488
0.3425	0	0.15	0.2	10	0.0553	0.0473	0.0471	0.0499	0.0446	0.0446	0.0590	0.0481	0.0480
				15	0.0619	0.0530	0.0528	0.0500	0.0459	0.0459	0.0640	0.0537	0.0534
				20	0.0541	0.0478	0.0477	0.0538	0.0498	0.0498	0.0567	0.0498	0.0498
				25	0.0594	0.0514	0.0514	0.0567	0.0518	0.0518	0.0600	0.0518	0.0516
				30	0.0600	0.0517	0.0516	0.0555	0.0504	0.0504	0.0648	0.0554	0.0552
				35	0.0554	0.0489	0.0489	0.0518	0.0482	0.0481	0.0602	0.0503	0.0503
				40	0.0579	0.0511	0.0510	0.0531	0.0494	0.0494	0.0594	0.0509	0.0507
0.3159	0	0.2	0.2	10	0.0550	0.0459	0.0456	0.0428	0.0392	0.0390	0.0557	0.0462	0.0455
				15	0.0618	0.0532	0.0530	0.0495	0.0446	0.0446	0.0610	0.0512	0.0509
				20	0.0573	0.0485	0.0483	0.0504	0.0462	0.0460	0.0608	0.0502	0.0500
				25	0.0561	0.0488	0.0486	0.0527	0.0477	0.0477	0.0593	0.0480	0.0478
				30	0.0535	0.0458	0.0457	0.0510	0.0462	0.0462	0.0543	0.0438	0.0437
				35	0.0575	0.0515	0.0515	0.0551	0.0511	0.0511	0.0597	0.0505	0.0503
				40	0.0519	0.0453	0.0453	0.0555	0.0504	0.0504	0.0589	0.0486	0.0485
0.2231	0	0.3	0.2	10	0.0465	0.0391	0.0390	0.0336	0.0307	0.0307	0.0533	0.0418	0.0416
				15	0.0463	0.0405	0.0405	0.0401	0.0356	0.0356	0.0536	0.0437	0.0433
				20	0.0524	0.0454	0.0454	0.0454	0.0408	0.0408	0.0573	0.0458	0.0454
				25	0.0574	0.0500	0.0499	0.0464	0.0414	0.0414	0.0560	0.0460	0.0456
				30	0.0517	0.0459	0.0458	0.0467	0.0426	0.0426	0.0533	0.0438	0.0433
				35	0.0506	0.0441	0.0438	0.0490	0.0444	0.0443	0.0511	0.0423	0.0412
				40	0.0546	0.0477	0.0475	0.0452	0.0413	0.0412	0.0525	0.0436	0.0431
0.2445	0.2	0.15	0.15	10	0.0447	0.0145	0.0458	0.0250	0.0073	0.0277	0.0461	0.0176	0.0472
				15	0.0425	0.0189	0.0430	0.0297	0.0071	0.0324	0.0450	0.0219	0.0453
				20	0.0448	0.0270	0.0450	0.0326	0.0076	0.0345	0.0436	0.0267	0.0437
				25	0.0458	0.0336	0.0459	0.0359	0.0117	0.0380	0.0482	0.0343	0.0482
				30	0.0516	0.0417	0.0517	0.0396	0.0147	0.0406	0.0510	0.0418	0.0510
				35	0.0455	0.0390	0.0455	0.0374	0.0152	0.0380	0.0462	0.0408	0.0462
				40	0.0497	0.0463	0.0497	0.0411	0.0208	0.0417	0.0465	0.0439	0.0465
0.2780	0.2	0.15	0.2	10	0.0526	0.0431	0.0429	0.0412	0.0364	0.0364	0.0531	0.0428	0.0425
				15	0.0554	0.0459	0.0457	0.0447	0.0404	0.0403	0.0558	0.0451	0.0449
				20	0.0581	0.0484	0.0481	0.0473	0.0437	0.0437	0.0590	0.0484	0.0478
				25	0.0586	0.0489	0.0485	0.0487	0.0437	0.0436	0.0581	0.0477	0.0474
				30	0.0554	0.0466	0.0463	0.0498	0.0431	0.0431	0.0565	0.0470	0.0466
				35	0.0539	0.0457	0.0457	0.0516	0.0455	0.0455	0.0538	0.0446	0.0445
				40	0.0603	0.0510	0.0508	0.0503	0.0456	0.0456	0.0565	0.0475	0.0473
0.2445	0.2	0.2	0.2	10	0.0487	0.0395	0.0395	0.0376	0.0330	0.0329	0.0503	0.0396	0.0393
				15	0.0555	0.0469	0.0467	0.0401	0.0363	0.0362	0.0538	0.0450	0.0446
				20	0.0530	0.0452	0.0450	0.0464	0.0427	0.0426	0.0542	0.0442	0.0439
				25	0.0568	0.0473	0.0473	0.0518	0.0463	0.0463	0.0573	0.0465	0.0460
				30	0.0578	0.0471	0.0469	0.0511	0.0455	0.0455	0.0595	0.0496	0.0490
				35	0.0584	0.0495	0.0494	0.0501	0.0456	0.0456	0.0577	0.0477	0.0476
				40	0.0547	0.0460	0.0456	0.0536	0.0488	0.0488	0.0581	0.0465	0.0463

Table III. Type I error probability of IBE tests when $\sigma_{WR} > 0.2 = \sigma_0$ (10 000 simulations).

Parameter							
δ	σ_D	σ_{WT}	σ_{WR}	n	2×4 cross-over	2×3 cross-over	2×3 extra-reference
0.5404	0	0.15	0.3	10	0.0499	0.0444	0.0507
				15	0.0533	0.0484	0.0511
				20	0.0555	0.0503	0.0562
				25	0.0514	0.0513	0.0491
				30	0.0480	0.0486	0.0495
				35	0.0526	0.0507	0.0530
				40	0.0485	0.0503	0.0490
0.4738	0	0.3	0.3	10	0.0468	0.0416	0.0447
				15	0.0503	0.0462	0.0495
				20	0.0468	0.0466	0.0480
				25	0.0478	0.0489	0.0474
				30	0.0486	0.0500	0.0491
				35	0.0474	0.0456	0.0487
				40	0.0499	0.0515	0.0511
0.2540	0	0.5	0.3	10	0.0356	0.0301	0.0431
				15	0.0390	0.0347	0.0434
				20	0.0443	0.0390	0.0438
				25	0.0448	0.0407	0.0460
				30	0.0418	0.0422	0.0439
				35	0.0453	0.0411	0.0460
				40	0.0447	0.0430	0.0417
0.9131	0	0.2	0.5	10	0.0516	0.0448	0.0516
				15	0.0534	0.0488	0.0563
				20	0.0541	0.0520	0.0537
				25	0.0504	0.0515	0.0497
				30	0.0542	0.0514	0.0574
				35	0.0513	0.0519	0.0479
				40	0.0521	0.0520	0.0513
0.7897	0	0.5	0.5	10	0.0503	0.0436	0.0484
				15	0.0495	0.0487	0.0488
				20	0.0526	0.0476	0.0502
				25	0.0497	0.0478	0.0503
				30	0.0470	0.0481	0.0453
				35	0.0512	0.0523	0.0515
				40	0.0482	0.0453	0.0482
0.4843	0.2	0.2	0.3	10	0.0486	0.0406	0.0487
				15	0.0504	0.0456	0.0519
				20	0.0554	0.0473	0.0540
				25	0.0510	0.0494	0.0494
				30	0.0495	0.0484	0.0494
				35	0.0506	0.0497	0.0497
				40	0.0484	0.0467	0.0490
0.4296	0.2	0.3	0.3	10	0.0443	0.0372	0.0449
				15	0.0456	0.0415	0.0426
				20	0.0472	0.0429	0.0508
				25	0.0490	0.0464	0.0499
				30	0.0464	0.0491	0.0476
				35	0.0484	0.0477	0.0480
				40	0.0496	0.0477	0.0462
0.8624	0.2	0.3	0.5	10	0.0507	0.0443	0.0507
				15	0.0504	0.0471	0.0522
				20	0.0511	0.0510	0.0512
				25	0.0505	0.0513	0.0506
				30	0.0547	0.0465	0.0554
				35	0.0507	0.0509	0.0493
				40	0.0507	0.0507	0.0509

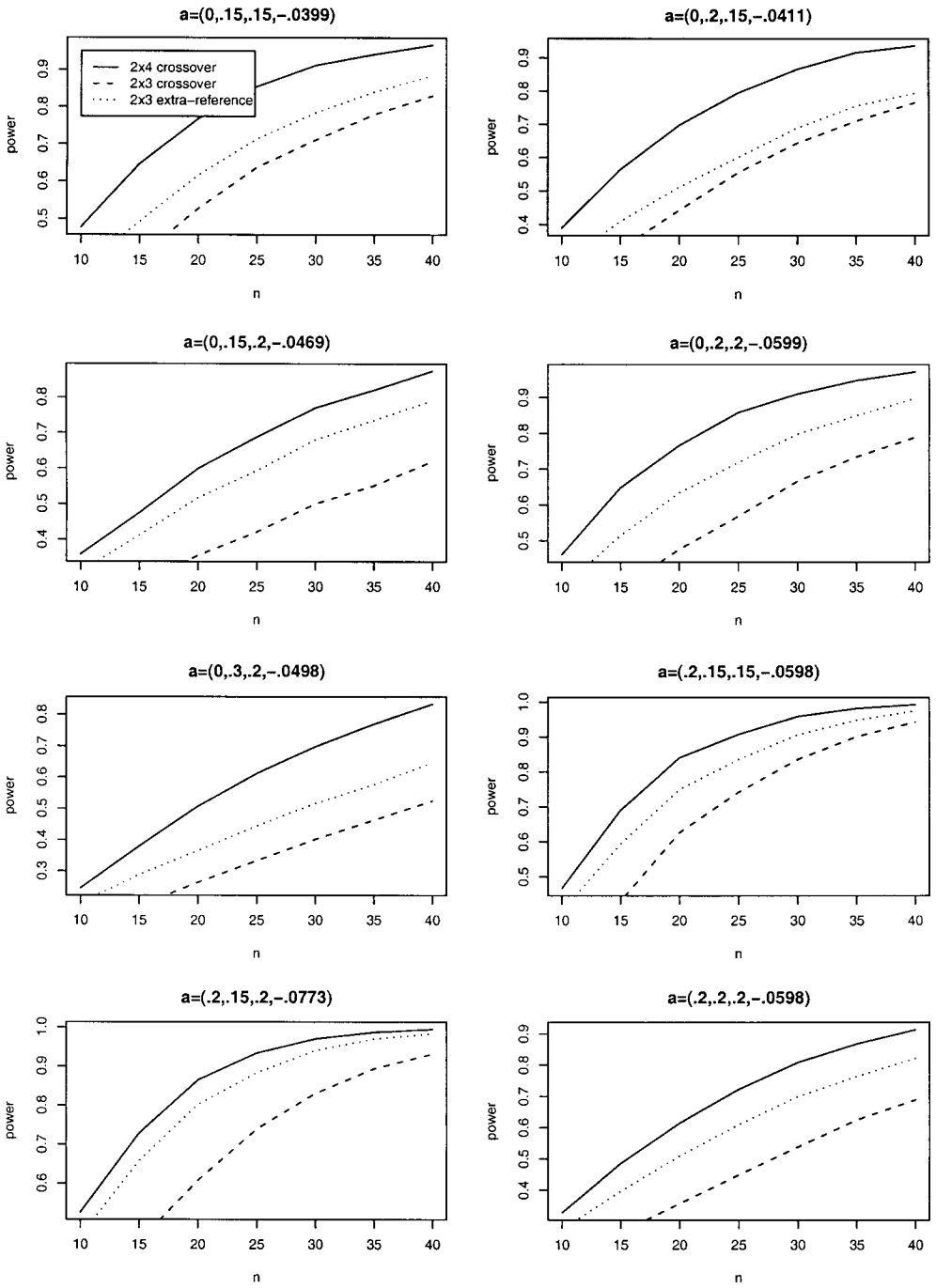

Figure 1. Power of IBE tests versus n; $a = (\sigma_D, \sigma_{WT}, \sigma_{WR}, \gamma)$.

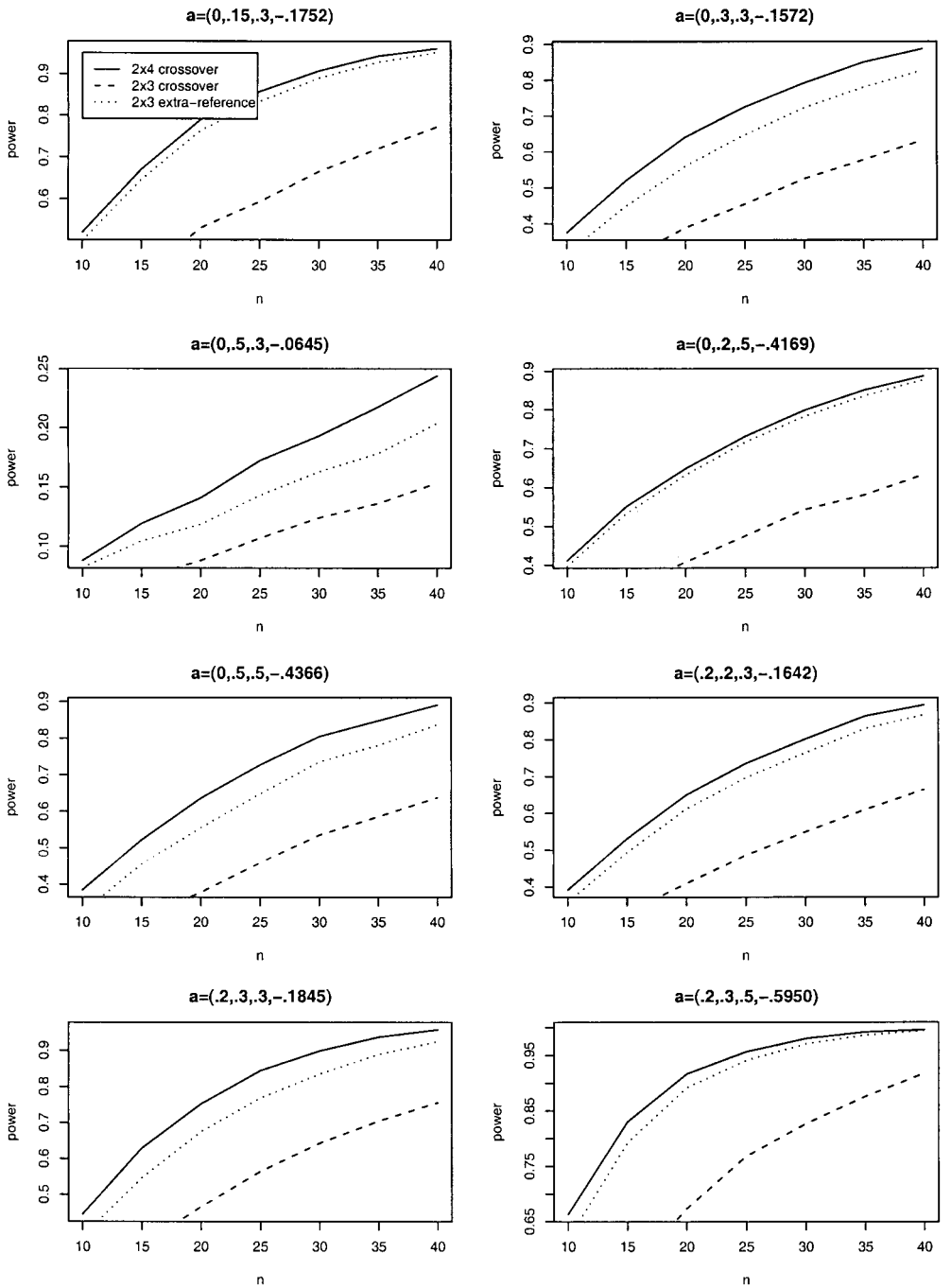

Figure 2. Power of IBE tests versus n; $a = (\sigma_D, \sigma_{WT}, \sigma_{WR}, \gamma)$.

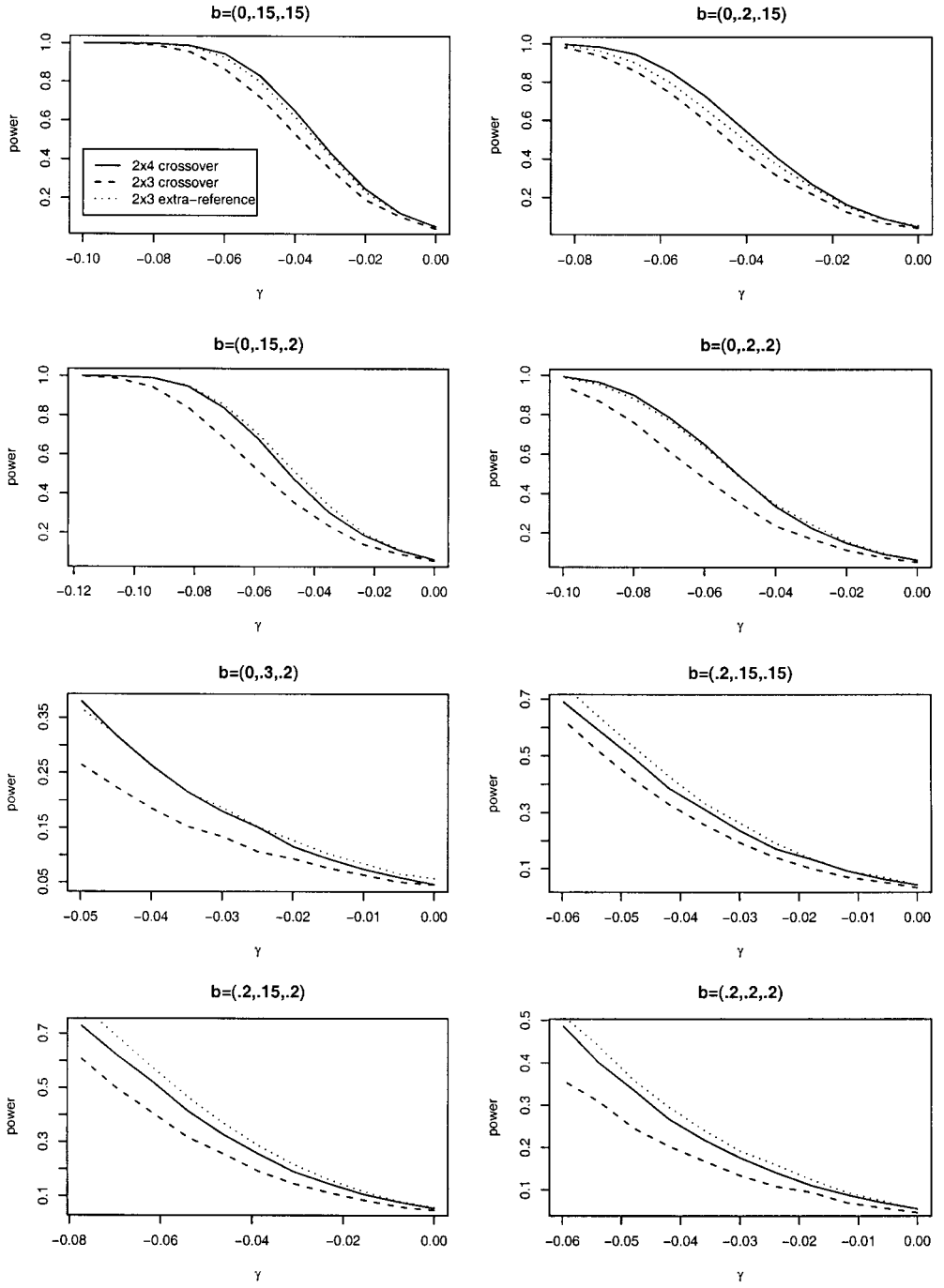

Figure 3. Power of IBE tests versus γ ($n = 15$ for 2×4 design and $n = 20$ for 2×3 designs); $b = (\sigma_D, \sigma_{WT}, \sigma_{WR})$.

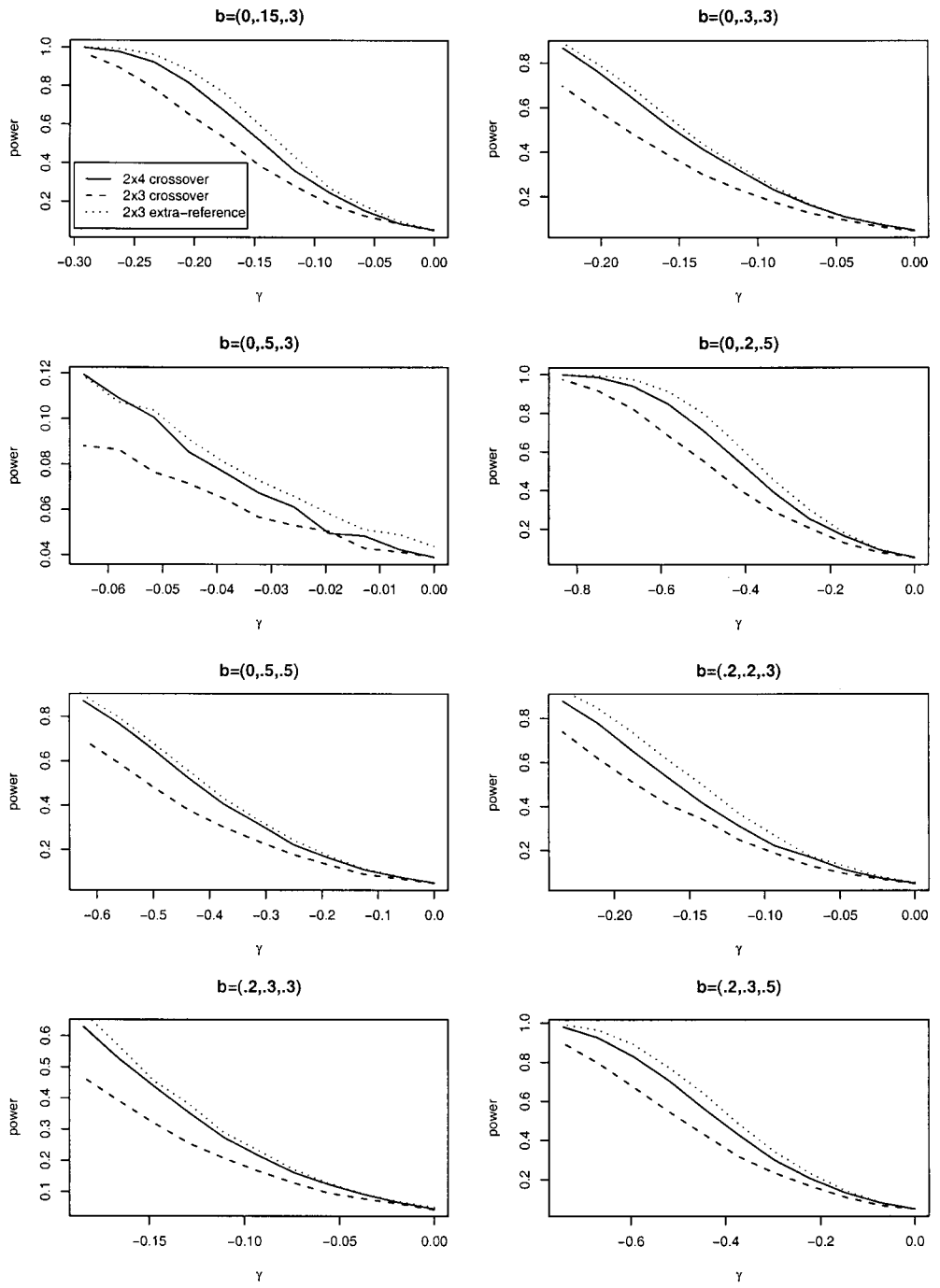

Figure 4. Power of IBE tests versus γ ($n = 15$ for 2×4 design and $n = 20$ for 2×3 designs); $b = (\sigma_D, \sigma_{WT}, \sigma_{WR})$.

subjects per sequence has the same total number of observations as that of a 2×3 design with 20 subjects per sequence.

The following is a summary of the simulation results.

1. Consider first the tests under the 2×4 cross-over design and the 2×3 extra-reference design. In terms of the type I error probability, both tests perform well. When $\sigma_{WR} > \sigma_0$ (Table III), the type I error probabilities of all tests are close to the nominal level 0.05. When $\sigma_{WR} \leqslant \sigma_0$ (Table II), the tests using the estimation approach (ESTI) of choosing the reference-scaled or the constant-scaled confidence bound can be too liberal, that is, their type I error probabilities are too large (for example, the case where $\sigma_D = 0.2$, $\sigma_{WT} = 0.15$ and $\sigma_{WR} = 0.2$), but not very substantial (type I error probabilities are $\leqslant 0.06$ in most cases). On the other hand, the test using the test approach (TEST) can be too conservative when $\sigma_{WR} = 0.15 < \sigma_0$, but is reasonably good once n is large (for example, $n > 30$). Overall, the tests using the estimation approach are better in the case of $\sigma_{WR} = 0.15$ (closer to the OPTI test), whereas the tests using the test approach are better in the case of $\sigma_{WR} = 0.2$ (closer to the OPTI test). This is because when σ_{WR} is equal or close to $0.2 = \sigma_0$, using the estimation approach is not suitable.
2. In terms of the type I error probability, the tests under the 2×3 cross-over design are clearly not as good as the tests under the other two designs. Frequently, these tests are too conservative (for example, the case where $\sigma_D = 0.2$ and $\sigma_{WT} = \sigma_{WR} = 0.15$). Although the type I error probabilities of these tests are closer to the nominal value when n is larger, the convergence speed is not as quick as those under the other two designs. This is due to the fact that $n_1 - 1$ and $n_2 - 1$ are used as the degrees of freedom in the construction of confidence bounds under the 2×3 cross-over design (see Section 2.2), instead of $n_1 + n_2 - 2$ as in the case of the 2×4 cross-over design or the 2×3 extra-reference design.
3. The power of various tests depends on the sample size as well as parameter values. For most parameter values considered in the simulation study, the power can reach 0.8 with a reasonable sample size and in some cases the power can be close to 0.9 even for the sample size of $n = 20$. However, there are situations where the power is very low even for the sample size of $n = 40$; for example, the case where $\sigma_D = 0$, $\sigma_{WT} = 0.5$ and $\sigma_{WR} = 0.3$, which indicates that when the test formulation has a higher variability than the reference formulation, it is difficult to claim IBE using the IBE tests considered in Section 2 even when the test formulation and the reference formulation are actually IBE ($\delta = 0$, $\sigma_D = 0$ and $\gamma = -0.0645$).
4. It can be seen from Figures 1 and 2 that the tests under the 2×4 design are more powerful than those under the two 2×3 designs. This comparison, however, is somewhat unfair since the 2×4 design requires more observations than the 2×3 designs. For the two 2×3 designs, it is clear that the tests under the 2×3 extra-reference design are more powerful than the tests under the 2×3 cross-over design, and the difference can be very substantial when $\sigma_{WR} > \sigma_{WT}$ or $\sigma_D > 0$. In fact, in some cases the tests under the 2×3 extra-reference design are even comparable to the tests under the 2×4 design.
5. The results in Figures 3 and 4 provide not only information about how power changes when the value of γ decreases, but also a comparison between the 2×4 design and the two 2×3 designs in the case where the total number of observations for all of these designs are the same. It can be seen that the 2×3 cross-over design is always worse

than the other two designs. Between the 2×3 extra-reference design and the 2×4 cross-over design, the former is better or comparable to the latter except for the case where $\sigma_{WT} > \sigma_{WR}$ and $\sigma_D = 0$.

4. SAMPLE SIZE DETERMINATION

We consider sample size determination for the 2×3 extra-reference design and 2×4 cross-over designs. The discussion for 2×3 cross-over designs is omitted, since 2×3 cross-over designs are not recommended based on the results in Sections 2 and 3.

Typically, we would like to choose $n = n_1 = n_2$ so that the power of the IBE test reaches a given level β when the unknown parameters are set at some initial guessing values. For the IBE test based on the confidence bound $\hat{\gamma}_U$, its power is $P(\hat{\gamma}_U < 0)$ when $\gamma < 0$.

Let $\tilde{\delta}$, $\tilde{\sigma}_D^2$, $\tilde{\sigma}_{WT}^2$ and $\tilde{\sigma}_{WR}^2$ be a set of initial values. Consider first the case where $\tilde{\sigma}_{WR}^2 > \sigma_0^2$. For the 2×3 extra-reference design, let U be defined by (14) and U_β be the same as U but with 5 per cent and 95 per cent replaced by $1 - \beta$ and β, respectively. Since

$$P(\hat{\gamma}_U < \gamma + \sqrt{U} + \sqrt{U_\beta}) \approx \beta$$

the power $P(\hat{\gamma}_U < 0)$ is approximately larger than β if $\gamma + \sqrt{U} + \sqrt{U_\beta} \leqslant 0$. Let $\tilde{\gamma}$, $\tilde{U}$ and $\tilde{U}_\beta$ be γ, U and U_β, respectively, with parameter values and their estimators replaced by $\tilde{\delta}$, $\tilde{\sigma}_D^2$, $\tilde{\sigma}_{WT}^2$ and $\tilde{\sigma}_{WR}^2$. Then, the required sample size n to have approximately power β is the smallest integer satisfying

$$\tilde{\gamma} + \sqrt{\tilde{U}} + \sqrt{\tilde{U}_\beta} \leqslant 0 \tag{17}$$

assuming that $n_1 = n_2 = n$ and $\tilde{\delta}$, $\tilde{\sigma}_D^2$, $\tilde{\sigma}_{WT}^2$, and $\tilde{\sigma}_{WR}^2$ are true parameter values. When $\tilde{\sigma}_{WR}^2 < \sigma_0^2$, the previous procedure can be modified by replacing U by U_0 in (16). If $\tilde{\sigma}_{WR}^2$ is equal or close to σ_0^2, then we recommend the use of U instead of U_0 to produce a more conservative sample size and the use of the test approach in the IBE test (see the discussion at the end of Section 2.2 and Section 3).

The procedure for 2×4 cross-over designs is the same, with U or U_0 changed to that given in reference [12] for 2×4 designs.

Note that IBE tests are based on the asymptotic theory. Thus, n should be reasonably large to ensure the asymptotic convergence. We recommend the use of the larger of 10 and the solution from (17). Hence, n is at least 10.

To study the performance of the proposed method for the sample size, we carry out a simulation study. For some given parameter values (similar to those in Section 3) and $\beta = 80$ per cent, we first compute the sample size n determined by (17) and then compute (with 10 000 simulations) the actual power P_n of the IBE test using n as the sample size for both sequences. The results for the 2×3 extra-reference design and 2×4 cross-over designs are given in Table IV. For each selected n that is smaller than 10, the power of the IBE test using $\max(n, 10)$ as the sample size, which is denoted by $P_{\max(n,10)}$, is also included. Note that $P_{\max(n,10)} = P_n$ if $n \geqslant 10$.

Table IV. Sample size n selected using (17) with $\beta = 80$ per cent and the corresponding power P_n of the IBE test based on 10 000 simulations.

Parameter					2×3 extra-reference				2×4 cross-over			
σ_D	σ_{WT}	σ_{WR}	δ	γ	n	P_n	$\max(n, 10)$	$P_{\max(n,10)}$	n	P_n	$\max(n, 10)$	$P_{\max(n,10)}$
0	0.15	0.15	0	0.0998	5	0.7226	10	0.9898	4	0.7007	10	0.9998
			0.1	−0.0898	6	0.7365	10	0.9572	5	0.7837	10	0.9948
			0.2	−0.0598	13	0.7718	13		9	0.7607	10	0.8104
0	0.2	0.15	0	−0.0823	9	0.7480	10	0.8085	7	0.7995	10	0.9570
			0.1	−0.0723	12	0.7697	12		8	0.7468	10	0.8677
			0.2	−0.0423	35	0.7750	35		23	0.7835	23	
0	0.15	0.2	0	−0.1173	9	0.8225	10	0.8723	8	0.8446	10	0.9314
			0.1	−0.1073	12	0.8523	12		10	0.8424	10	
			0.2	−0.0773	26	0.8389	26		23	0.8506	23	
0	0.2	0.2	0	−0.0998	15	0.8206	15		13	0.8591	13	
			0.1	−0.0898	20	0.8373	20		17	0.8532	17	
			0.2	−0.0598	52	0.8366	52		44	0.8458	44	
0	0.3	0.2	0	−0.0498	91	0.8232	91		71	0.8454	71	
0.2	0.15	0.15	0	−0.0598	20	0.7469	20		17	0.7683	17	
			0.1	−0.0498	31	0.7577	31		25	0.7609	25	
0.2	0.15	0.2	0	−0.0773	31	0.8238	31		28	0.8358	28	
			0.1	−0.0673	43	0.8246	43		39	0.8296	39	
0.2	0.2	0.2	0	−0.0598	59	0.8225	59		51	0.8322	51	
			0.2	−0.0498	91	0.8253	91		79	0.8322	79	
0	0.15	0.3	0	−0.2920	7	0.8546	10	0.9607	6	0.8288	10	0.9781
			0.1	−0.2820	7	0.8155	10	0.9401	7	0.8596	10	0.9566
			0.2	−0.2520	10	0.8397	10		9	0.8352	10	0.8697
			0.3	−0.2020	16	0.7973	16		15	0.8076	15	
			0.4	−0.1320	45	0.8043	45		43	0.8076	43	
0	0.3	0.3	0	−0.2245	15	0.7931	15		13	0.8162	13	
			0.1	−0.2145	17	0.7942	17		14	0.8057	14	
			0.2	−0.1845	25	0.8016	25		21	0.8079	21	
			0.3	−0.1345	52	0.7992	52		44	0.8009	44	
0	0.2	0.5	0	−0.8337	6	0.8285	10	0.9744	6	0.8497	10	0.9810
			0.1	−0.8237	6	0.8128	10	0.9708	6	0.8413	10	0.9759
			0.2	−0.7937	7	0.8410	10	0.9505	7	0.8600	10	0.9628
			0.3	−0.7437	8	0.8282	10	0.9017	8	0.8548	10	0.9239
			0.4	−0.6737	10	0.8147	10		10	0.8338	10	
			0.5	−0.5837	14	0.8095	14		14	0.8248	14	
			0.6	−0.4737	24	0.8162	24		23	0.8149	23	
			0.7	−0.3437	51	0.8171	51		49	0.8170	49	
0	0.5	0.5	0	−0.6237	15	0.7890	15		13	0.8132	13	
			0.1	−0.6137	16	0.8000	16		13	0.7956	13	
			0.2	−0.5837	18	0.7980	18		15	0.8033	15	
			0.3	−0.5337	23	0.8002	23		19	0.8063	19	
			0.4	−0.4637	32	0.8026	32		27	0.8163	27	
			0.5	−0.3737	52	0.7944	52		44	0.8045	44	
0.2	0.2	0.3	0	−0.2345	13	0.7870	13		12	0.7970	12	
			0.1	−0.2245	15	0.8007	15		14	0.8144	14	
			0.2	−0.1945	21	0.7862	21		20	0.8115	20	
			0.3	−0.1445	43	0.8037	43		40	0.8034	40	
0.2	0.3	0.3	0	−0.1845	26	0.7806	26		22	0.7877	22	
			0.1	−0.1745	30	0.7895	30		26	0.8039	26	
0.2	0.3	0.5	0	−0.7437	9	0.8038	10	0.8502	8	0.8050	10	0.8947
			0.1	−0.7337	9	0.7958	10	0.8392	9	0.8460	10	0.8799
			0.2	−0.7037	10	0.7966	10		9	0.7954	10	0.8393
			0.3	−0.6537	12	0.7929	12		11	0.8045	11	
			0.4	−0.5837	16	0.7987	16		15	0.8094	15	

The following is a summary of the results in Table IV.

1. With the sample size n determined by (17), the actual power P_n is larger than the target value 80 per cent in most cases. Only in a few cases where n determined from (17) is very small, the power P_n is lower than 75 per cent.
2. Using $\max(n, 10)$ as the sample size produces better results when n determined by (17) is very small, but in most cases it results in a power much larger than 80 per cent.
3. For each selected n, the required total number of observations is $6n$ for the 2×3 extra-reference design and $8n$ for 2×4 cross-over designs. It can be seen from Table IV that in most cases the 2×3 extra-reference design requires fewer total number of observations than a 2×4 cross-over design.

5. CONCLUSIONS

Statistical tests for IBE are derived in Section 2 for 2×3 cross-over designs and the 2×3 extra-reference design. Although a 2×3 cross-over design requires fewer observations (thus shorter duration and lower cost of the study) than a 2×4 cross-over design, we show that it does not provide an efficient IBE test as compared to those under other designs (in terms of the type I error probability and power). The 2×3 extra-reference design, which requires the same number of observations as any 2×3 cross-over design, provides a much more efficient IBE test. In addition, the IBE test under the 2×3 extra-reference design may also be better than that under a 2×4 cross-over design when both designs have the same number of observations.

We also study two methods of choosing the reference-scaled or the constant-scaled confidence bound, the estimation method and the test method. The test method is too conservative if $\sigma_{WR} < \sigma_0$ and the sample size is not very large; otherwise it performs well. The estimation method performs well except in the case where σ_{WR} is close to σ_0.

Our empirical results show that when the test formulation has a larger variability than the reference formulation, that is, $\sigma_{WT} > \sigma_{WR}$, it may be difficult to claim IBE even when the two formulations are IBE and $\sigma_D = 0$.

Finally, we propose a method of determining the required sample size for the IBE test to have a given level of power, using some initial guessing parameter values. The proposed method works well in a simulation study.

REFERENCES

1. Chow SC, Liu JP. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 2nd edn. Marcel Dekker: New York, 1999.
2. FDA. Guidance on statistical procedures for bioequivalence studies using a standard two-treatment cross-over design. Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, 1992.
3. FDA. 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, 2000.
4. Anderson S, Hauck WW. Considerations of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 1990; **8**:259–273.
5. Chen ML. Individual bioequivalence – a regulatory update. *Journal of the Biopharmaceutical Statistics* 1997; **7**:5–11.

6. Chow SC, Liu JP. *Statistical Design and Analysis in Pharmaceutical Science*. Marcel Dekker: New York, 1995.
7. Esinhart JD, Chinchilli VM. Extension to the use of tolerance intervals for assessment of individual bioequivalence. *Journal of Biopharmaceutical Statistics* 1994; **4**:39–52.
8. FDA. Guidance for industry on statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, 2001.
9. Schall R, Luus HG. On population and individual bioequivalence. *Statistics in Medicine* 1993; **12**:1109–1124.
10. Sheiner LB. Bioequivalence revisited. *Statistics in Medicine* 1992; **11**:1777–1788.
11. Jones B, Kenward MG. *Design and Analysis of Cross-Over Trials*. Chapman & Hall: London, 1989.
12. Hyslop T, Hsuan F, Holder DJ. A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* 2000; **19**:2885–2897.
13. Wang W. On testing of individual bioequivalence. *Journal of the American Statistical Association* 1999; **94**: 880–887.
14. Graybill FA, Wang CM. Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association* 1980; **75**:869–873.
15. Howe WG. Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 1974; **69**:789–794.
16. Ting N, Burdick RK, Graybill FA, Jeyaratnam S, Lu T-FC. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation* 1990; **35**:135–143.

