Statistical Science
1996, Vol. 11, No. 4, 283–319

Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets

Roger L. Berger and Jason C. Hsu

Abstract. The bioequivalence problem is of practical importance because the approval of most generic drugs in the United States and the European Community (EC) requires the establishment of bioequivalence between the brand-name drug and the proposed generic version. The problem is theoretically interesting because it has been recognized as one for which the desired inference, instead of the usual *significant difference*, is *practical equivalence*. The concept of intersection–union tests will be shown to clarify, simplify and unify bioequivalence testing. A test more powerful than the one currently specified by the FDA and EC guidelines will be derived. The claim that the bioequivalence problem defined in terms of the ratio of parameters is more difficult than the problem defined in terms of the difference of parameters will be refuted. The misconception that size-α bioequivalence tests generally correspond to $100(1-2\alpha)\%$ confidence sets will be shown to lead to incorrect statistical practices, and should be abandoned. Techniques for constructing $100(1-\alpha)\%$ confidence sets that correspond to size-α bioequivalence tests will be described. Finally, multiparameter bioequivalence problems will be discussed.

Key words and phrases: Bioequivalence; bioavailability; hypothesis test; confidence interval; intersection–union; size; level; equivalence test; pharmacokinetic; unbiased.

1. BIOEQUIVALENCE PROBLEM

Two different drugs or formulations of the same drug are called *bioequivalent* if they are absorbed into the blood and become available at the drug action site at about the same rate and concentration. Bioequivalence is usually studied by administering dosages to subjects and measuring concentration of the drug in the blood just before and at set times after the administration. These data are then used to determine if the drugs are absorbed at the same rate.

The determination of bioequivalence is very important in the pharmaceutical industry because regulatory agencies allow a generic drug to be marketed if its manufacturer can demonstrate that the generic product is bioequivalent to the brand-name product. The assumption is that bioequivalent drugs will provide the same therapeutic effect. If the generic drug manufacturer can demonstrate bioequivalence, it does not need to perform costly clinical trials to demonstrate the safety and efficacy of the generic product. Yet, this bioequivalence must be demonstrated in a statistically sound way to protect the consumer from ineffective or unsafe drugs.

These concentration by time measurements are connected with a polygonal curve and several variables are measured. The common measurements are AUC (area under curve), $C_{\max}$ (maximum concentration) and $T_{\max}$ (time until maximum concentration). The two drugs are bioequivalent if the population means of AUC and $C_{\max}$ are sufficiently close. Descriptive statistics for $T_{\max}$ are usually provided, but formal tests are not required.

For example, let μ_T denote the population mean AUC for the generic (test) drug and let μ_R denote the population mean AUC for the brand-name (reference) drug. To demonstrate bioequivalence, the

Roger L. Berger is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 (e-mail: berger@stat.ncsu.edu). Jason C. Hsu is Professor, Department of Statistics, Ohio State University, Columbus, Ohio 43210-1247 (e-mail: jch@stat.ohio-state.edu).

following hypotheses are tested:

$$H_0: \quad \frac{\mu_T}{\mu_R} \le \delta_L \quad \text{or} \quad \frac{\mu_T}{\mu_R} \ge \delta_U$$

(1) versus

$$H_a: \quad \delta_L < \frac{\mu_T}{\mu_R} < \delta_U.$$

The values δ_L and δ_U are standards set by regulatory agencies that define how "close" the drugs must be to be declared bioequivalent. Currently, both the United States Food and Drug Administration (FDA, 1992a) and the European Community use $\delta_U = 1.25$ and $\delta_L = 0.80 = 1/1.25$ for AUC. For $C_{\max}$, the United States again uses $\delta_U = 1.25$ and $\delta_L = 0.80$, but Europe uses the less restrictive limits $\delta_U = 1.43$ and $\delta_L = 0.70 = 1/1.43$ (Hauck et al., 1995). Note that these limits for AUC and $C_{\max}$ are symmetric about 1 in the ratio scale.

Often, logarithms are taken and hypotheses (1) are stated as

$$H_0: \quad \eta_T - \eta_R \le \theta_L \quad \text{or} \quad \eta_T - \eta_R \ge \theta_U$$

(2) versus

$$H_a: \quad \theta_L < \eta_T - \eta_R < \theta_U.$$

Here, $\eta_T = \log(\mu_T)$, $\eta_R = \log(\mu_R)$, $\theta_U = \log(\delta_U)$ and $\theta_L = \log(\delta_L)$. With $\delta_U = 1.25$ and $\delta_L = 0.80$ or $\delta_U = 1.43$ and $\delta_L = 0.70$, $\theta_U = -\theta_L$, and the standards are symmetric about zero.

In a hypothesis test of (1) or (2), the Type I error rate is the probability of declaring the drugs to be bioequivalent, when in fact they are not. By setting up the hypotheses as in (1) or (2) and controlling the Type I error rate at a specified small value, say, $\alpha = 0.05$, the consumer's risk is being controlled. That (1) or (2) is the proper formulation in problems like these was recognized early on by some authors. For example, Lehmann (1959, page 88), not specifically discussing bioequivalence, says, "One then sets up the (null) hypothesis that [the parameter] does not lie within the required limits so that an error of the first kind consists in declaring [the parameter] to be satisfactory when in fact it is not." But not until Schuirmann (1981, 1987a), Westlake (1981) and Anderson and Hauck (1983) were hypotheses correctly formulated as in (1) or (2) in bioequivalence problems.

Despite the fact that bioequivalence testing problems are now correctly formulated as (1) or (2), many inappropriate statistical procedures are still used in this area. Tests that claim to have a specified size α, but are either liberal or conservative, are used. Liberal tests compromise the consumer's safety, and conservative tests put an undue burden on the generic drug manufacturer. Tests are often defined in terms of confidence intervals in statistically unsound ways. These tests, again, do not properly control the consumer's risk.

In this paper, we will describe current bioequivalence tests that have incorrect error rates. We will offer new tests that correctly control the consumer's risk. In several cases, the tests we propose are uniformly more powerful than the existing tests while still controlling the Type I error rate at the specified rate α. We will examine and criticize the current practice of defining tests in terms of $100(1 - 2\alpha)\%$ confidence sets. We will show that this only works in special cases and gives poor results in other cases. We will discuss how properly to construct $100(1 - \alpha)\%$ confidence sets that correspond to size-α tests. And we will discuss how our methods can be applied to complicated, multiparameter bioequivalence problems that have received only slight attention in the literature. The intersection–union method of testing will be found to be very useful in understanding and constructing bioequivalence tests. Section 2 provides a more detailed outline of our discussions.

Hypotheses such as (1) and (2) that specify only that population means should be close are called *average bioequivalence* hypotheses. Hypotheses that state that the whole distribution of bioavailabilities is the same for the test and reference populations are called *population bioequivalence* hypotheses. If a parametric form of these populations is assumed, then hypotheses such as (25) that specify that all population parameters (e.g., variances as well as means) should be close are population bioequivalence hypotheses. Sometimes bioequivalence is defined in terms of parameters that more directly measure equivalence of response within an individual. Good introductions to *individual bioequivalence* are given by Anderson and Hauck (1990), Hauck and Anderson (1992), Sheiner (1992), Schall and Luus (1993) and Anderson (1993). Although we do not explicitly consider individual bioequivalence in this paper, many of the concepts and techniques we describe should be applicable in that area also.

In this paper, our discussion will be entirely in terms of bioequivalence testing. But our comments and techniques apply to other problems, such as in quality assurance, in which the aim is to show that two parameters are close or that a parameter is between two specification limits. Because of this wider applicability, the methods we will discuss might more properly be referred to as *equivalence tests* and *equivalence confidence intervals*.

2. TESTS, CONFIDENCE SETS AND CURIOSITIES

Various experimental designs are used to gather data for bioequivalence trials. Chow and Liu (1992) describe parallel designs (two independent samples) and two-period and multiperiod crossover designs. The issues we discuss apply to all these different designs. For brevity, we will discuss only the simple parallel design and two period crossover design.

2.1 Difference Hypotheses

It is customary to employ lognormal models in bioequivalence studies of AUC and $C_{\max}$. See Section 2.2 for rationales for this model.

Let X^* denote a lognormal measurement from the test drug in the original scale, and let $X = \log(X^*)$. Similarly, let Y^* denote an original measurement, and let $Y = \log(Y^*)$ for the reference drug. Let (η_T, σ^2) denote the lognormal parameters for X^* and (η_R, σ^2) denote the lognormal parameters for Y^*. Then the test and reference drug means are $\mu_T = \exp(\eta_T + \sigma^2/2)$ and $\mu_R = \exp(\eta_R + \sigma^2/2)$, respectively. Therefore, the condition

$$\delta_L < \frac{\mu_T}{\mu_R} = \exp(\eta_T - \eta_R) < \delta_U$$

is equivalent to

$$\theta_L < \eta_T - \eta_R < \theta_U, \tag{3}$$

where $\theta_L = \log(\delta_L)$ and $\theta_U = \log(\delta_U)$ are known constants. Thus, the hypothesis to be tested in this lognormal model can be stated as either (1) or (2). Usually the hypotheses are stated as (2) and the test is based on log transformed data that is normally distributed with means η_T and η_R and common variance σ^2. The equivalence of (1) and (2) is dependent on the assumption of equal variances. On the other hand, if μ_T and μ_R represent the medians of X^* and Y^* and $\eta_T = \log(\mu_T)$ and $\eta_R = \log(\mu_R)$, then η_T and η_R are the medians of X and Y, respectively. So, in terms of medians, (1) and (2) are always equivalent, and the analysis can be carried out in either the original or log transformed scale. But, bioequivalence is almost always defined in terms of means rather than medians.

Westlake (1981) and Schuirmann (1981) proposed what has become the standard test of (2). It is called the "two one-sided tests" (TOST). The TOST has this general form. Let D be an estimate of $\eta_T - \eta_R$ that has a normal distribution with mean $\eta_T - \eta_R$ and variance σ_D^2. Let $\mathrm{SE}(D)$ be an estimate of σ_D that is independent of D and such that $r[\mathrm{SE}(D)]^2/\sigma_D^2$ has a χ^2 distribution with r degrees of freedom. Then

$$t = \frac{D - (\eta_T - \eta_R)}{\mathrm{SE}(D)}$$

has a Student's t-distribution with r degrees of freedom. The TOST is based on the two statistics

$$T_U = \frac{D - \theta_U}{\mathrm{SE}(D)} \quad \text{and} \quad T_L = \frac{D - \theta_L}{\mathrm{SE}(D)}. \tag{4}$$

The TOST tests (2) using the ordinary, one-sided, size-α t-test based on T_L for

$$H_{01}: \quad \eta_T - \eta_R \le \theta_L$$

(5) versus

$$H_{a1}: \quad \eta_T - \eta_R > \theta_L$$

and the ordinary, one-sided, size-α t-test based on T_U for

$$H_{02}: \quad \eta_T - \eta_R \ge \theta_U$$

(6) versus

$$H_{a2}: \quad \eta_T - \eta_R < \theta_U.$$

It rejects H_0 at level α and declares the two drugs to be bioequivalent if both tests reject, that is, if

$$T_U < -t_{\alpha, r} \quad \text{and} \quad T_L > t_{\alpha, r}, \tag{7}$$

where $t_{\alpha, r}$ is the upper 100α percentile of a Student's t-distribution with r degrees of freedom. For testing (2), all the tests we will discuss are functions of $(D, \mathrm{SE}(D))$. The distribution of $(D, \mathrm{SE}(D))$ is determined by the parameter $(\eta_T, \eta_R, \sigma_D^2)$.

In the simple parallel design, let $X_1^*, \ldots, X_m^*$ denote the independent lognormal (η_T, σ^2) measurements on m subjects from the test drug in the original scale, and let $X_1, \ldots, X_m$ denote the logarithms of these measurements. Similarly, let $Y_1^*, \ldots, Y_n^*$ and $Y_1, \ldots, Y_n$ denote the original measurements [lognormal(η_R, σ^2)] and logarithms for an independent sample of n subjects on the reference drug. If $\overline{X}$ denotes the sample mean of $X_1, \ldots, X_m$, $\overline{Y}$ denotes the sample mean of $Y_1, \ldots, Y_n$ and S^2 denotes the pooled estimate of σ^2, computed from both samples, then

$$D = \overline{X} - \overline{Y}$$

and

$$\mathrm{SE}(D) = S\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

The degrees of freedom are $r = m + n - 2$.

In bioequivalence studies, much more common than simple parallel designs are two-period, crossover designs. In a two-period, crossover design, a group of m subjects (Sequence 1) receives the reference drug and observations on the pharmacokinetic response are made. After a washout period to remove any carryover effect, this group receives the test drug and observations are again

made. A second group of n subjects (Sequence 2) receives the drugs in the opposite order. After log transformation, the response of the kth subject in the jth period of the ith sequence is modeled as

$$Y_{ijk} = \gamma + S_{ik} + P_j + F_{(i,j)} + \varepsilon_{ijk},$$

where γ is the overall mean; P_j is the fixed effect of period j; $F_{(i,j)}$ is the fixed effect of the formulation administered in period j of sequence i, that is, $F_{(1,1)} = F_{(2,2)} = F_R$ and $F_{(1,2)} = F_{(2,1)} = F_T$; S_{ik} is the random effect of subject k in sequence i; and ε_{ijk} is the random error. It is assumed that $P_1 + P_2 = F_T + F_R = 0$. The S_{ik}'s and the ε_{ijk}'s are all independent normal random variables with mean 0. The variance of S_{ik} is σ_S^2 and the variance of ε_{ijk} is σ_T^2 and σ_R^2 for the test and reference formulations, respectively. For this design,

$$D = \frac{\overline{Y}_{12\bullet} - \overline{Y}_{11\bullet} + \overline{Y}_{21\bullet} - \overline{Y}_{22\bullet}}{2}$$

is a normally distributed unbiased estimate of $F_T - F_R = \eta_T - \eta_R$ with variance

$$\sigma_D^2 = (\sigma_R^2 + \sigma_T^2)\frac{1}{4}\left(\frac{1}{m} + \frac{1}{n}\right).$$

The standard error of D is

$$\mathrm{SE}(D) = S\frac{1}{2}\sqrt{\frac{1}{m} + \frac{1}{n}},$$

where

$$S^2 = \frac{1}{m+n-2} \cdot \left[\sum_{k=1}^{m} \left(Y_{12k} - Y_{11k} - (\overline{Y}_{12\bullet} - \overline{Y}_{11\bullet})\right)^2 + \sum_{k=1}^{n} \left(Y_{21k} - Y_{22k} - (\overline{Y}_{21\bullet} - \overline{Y}_{22\bullet})\right)^2 \right].$$

The estimate D is the average of the averages of the intrasubject differences for the two sequences, and S^2 is a pooled estimate of the variance of an intrasubject difference. For this crossover design, also, the degrees of freedom are $r = m + n - 2$.

Following Lehmann (1959), we define the size of a test as

$$\text{size} = \sup_{H_0} P(\text{reject } H_0).$$

The size of the TOST is exactly equal to α, even though $P(\text{reject } H_0) < \alpha$ for every $(\eta_T, \eta_R, \sigma_D^2)$ in the null hypothesis. The supremum value of α is attained in the limit as $\eta_T - \eta_R = \theta_L$ (or θ_U) and $\sigma_D^2 \to 0$. Both the FDA bioequivalence guideline (FDA, 1992a) and the European Community guideline (EC-GCP, 1993) specify that bioequivalence be established using a 5% TOST.

The TOST is unusual in that two size-α tests are combined to form a size-α test. Often, when multiple tests are combined, some adjustment must be made to the sizes of the individual tests to achieve an overall size-α test. Why this is not necessary for the TOST is best understood through the theory of intersection–union tests (IUT's), which we describe in Section 3. In Sections 4.1 and 4.2 we will show that the IUT theory is useful for understanding the TOST. Also, the IUT theory can guide the construction of tests for (2) that have the same size α as the TOST but are uniformly more powerful than the TOST.

2.2 Ratio Hypotheses

Sometimes, a normal model should be used. In this model, the original measurements are normally distributed with means μ_T and μ_R. This model is different from the lognormal model in that now the hypothesis to be tested concerns the ratio of the means of these normal observations. That is, we wish to test (1). This problem has received less attention than (2). Dealing with the ratio μ_T/μ_R has been perceived as more difficult than dealing with the difference $\eta_T - \eta_R$.

For AUC and $C_{\max}$, the FDA (1992a) strongly recommends logarithmically transforming the data and testing the hypotheses (2). They offer three rationales for their recommendation. Based on these, the FDA (1992a, page 7) states:

> Based on the arguments in the preceding section, the Division of Bioequivalence recommends that the pharmacokinetic parameters AUC and $C_{\max}$ be log transformed. Firms are *not* encouraged to test for normality of data distribution after log transformation, nor should they employ normality of data distribution as a justification for carrying out the statistical analysis on the original scale.

The emphasis is ours.

The FDA's three rationales for log transformation are labeled "Clinical," "Pharmacokinetic" and "Statistical." The Clinical Rationale is that the real interest is in the ratio μ_T/μ_R rather than the difference $\mu_T - \mu_R$. But, the link between this fact (which we certainly do not dispute) and the log transformation of the data is based on statistical considerations. It is that a linear statistical model can be used for the transformed data to make inferences about the difference $\eta_T - \eta_R$. These inferences then can be restated in terms of μ_T/μ_R. Thus, the justification of the log transformations seems to be based mainly on the perceived difficulty in dealing with the ratio

μ_T/μ_R, rather than the difference $\eta_T - \eta_R$. If appropriate statistical procedures can be used to make inferences about the ratio μ_T/μ_R directly, then there seems to be no need for a log transformation.

The Pharmacokinetic Rationale is based on multiplicative compartmental models of Westlake (1973, 1988). The multiplicative model is changed to a linear model by the log transformation. Part of the Statistical Rationale is that, in the original scale, much bioequivalence data is skewed and appears more lognormal than normal. We agree that these two considerations suggest that the first method of analysis to be considered in bioequivalence studies is on the log transformed data, and, in most cases, this analysis will be appropriate.

The Statistical Rationale consists of the previous lognormal justification and two more points. The first is that:

> Standard parametric methods are ill-suited to making inferences about the ratio of two averages, though some valid methods do exist. Log transformation changes the problem to one of making inferences about the difference (on the log scale) of the two averages, for which the standard methods are well suited.

The second is that the small sample sizes used in typical bioequivalence studies (20–30) will produce tests for normality that have fairly low power in either the original or log scale. The FDA recommends that no check of normality be made on the log transformed data. But, if a low-power normality test rejects the hypothesis of normality for the log transformed data, then surely some caution is warranted in the use of procedures that assume normality. In this case, tests such as the TOST, based on the Student's t-distribution, are inappropriate. If normality of the log transformed data is rejected and the original data appear more normal than the log transformed data, then procedures that assume normality of the original data would seem more appropriate. In Section 4.3, we show that Sasabuchi (1980,1988a,b) described the size-α likelihood ratio test for (1). It is a simple test based on the Student's t-distribution. So the FDA's statement about ill-suited standard parametric procedures seems unfounded. We also show that the tests commonly used are liberal and have size greater than the nominal value of α. Furthermore, we show that the IUT method can be used in this problem, also, to construct size-α tests that are uniformly more powerful than the likelihood ratio test. Thus, the FDA's avoidance of (1) because of statistical difficulties is unwarranted.

An alternative test, when normality is in doubt, might be to use a Wilcoxon–Mann–Whitney analogue of the TOST [based on the original logarithmically transformed data for a parallel design, or the intrasubject between-period differences of the logarithmically transformed data, as proposed by Hauschke, Steinijans and Diletti (1990), for a crossover design].

2.3 100(1 − 2α)% Confidence Intervals

One would expect the TOST to be identical to some *confidence interval* procedure: for some appropriate $100(1-\alpha)\%$ confidence interval $[D^-, D^+]$ for $\eta_T - \eta_R$, declare the test drug to be bioequivalent to the reference drug if and only if $[D^-, D^+] \subset (\theta_L, \theta_U)$.

It has been noted (e.g., Westlake, 1981; Schuirmann, 1981) that the TOST is operationally identical to the procedure of declaring equivalence only if the ordinary $100(1-2\alpha)\%$, not $100(1-\alpha)\%$, two-sided confidence interval for $\eta_T - \eta_R$,

$$[D - t_{\alpha,r}\mathrm{SE}(D),\ D + t_{\alpha,r}\mathrm{SE}(D)], \tag{8}$$

is contained in the interval (θ_L, θ_U). In fact, both FDA (1992a) and EC-GCP (1993) specify that the TOST should be executed in this fashion.

The fact that the TOST seemingly corresponds to a $100(1-2\alpha)\%$, not $100(1-\alpha)\%$, confidence interval procedure initially caused some concern (Westlake, 1976, 1981). Recently, Brown, Casella and Hwang (1995) called this relationship an "algebraic coincidence." But many authors (e.g., Chow and Shao, 1990, and Schuirmann, 1989) have defined bioequivalence tests in terms of $100(1-2\alpha)\%$ confidence sets.

Standard statistical results, such as Theorems 3 and 4 in Section 5, give relationships between size-α tests and $100(1-\alpha)\%$ confidence intervals. In Section 5, we discuss a $100(1-\alpha)\%$ confidence interval that corresponds exactly to the size-α TOST. We also explore the relationship between $100(1-2\alpha)\%$ confidence intervals and size-α tests. We describe situations more general than the TOST in which size-α tests can be defined in terms of $100(1-2\alpha)\%$ confidence intervals. But we also give examples from the bioequivalence literature of tests that have been defined in terms of $100(1-2\alpha)\%$ confidence intervals and sets that are not size-α tests. Tests defined by $100(1-2\alpha)\%$ confidence intervals can be either liberal or conservative. Because of these potential difficulties, our conclusion is that the practice of defining bioequivalence tests in terms of $100(1-2\alpha)\%$ confidence intervals should be abandoned. If both a confidence interval and a test are required, a $100(1-\alpha)\%$ confidence interval that corresponds to the given size-α test should be used.

2.4 Multiparameter Problems

In Section 6, we discuss multiparameter bioequivalence problems. We discuss two examples in which the IUT theory can be used to define size-α tests that are uniformly more powerful than tests that have been previously proposed. These examples concern controlling the experimentwise error rate when several parameters are tested for equivalence, simultaneously.

3. INTERSECTION–UNION TESTS

Berger (1982) proposed the use of intersection–union tests in a quality control context closely related to bioequivalence testing. Tests for many different bioequivalence hypotheses are easily constructed using the IUT method. The TOST is a simple example of an IUT. Tests with a specified size are easily constructed using this method, even in complicated problems involving several parameters. And tests that are uniformly more powerful than standard tests can often be constructed using this method.

The IUT method is useful for the following type of hypothesis testing problem. Let θ denote the unknown parameter (θ can be vector valued) in the distribution of the data $\mathbf{X}$. Let Θ denote the parameter space. Let $\Theta_1, \ldots, \Theta_k$ denote subsets of Θ. Suppose we wish to test

$$(9) \qquad H_0\colon \theta \in \bigcup_{i=1}^{k} \Theta_i \quad \text{versus} \quad H_a\colon \theta \in \bigcap_{i=1}^{k} \Theta_i^c,$$

where A^c denotes the complement of the set A. The important feature in this formulation is the null hypothesis is expressed as a union and the alternative hypothesis is expressed as an intersection. For $i = 1, \ldots, k$, let R_i denote a rejection region for a test of $H_{0i}\colon \theta \in \Theta_i$ versus $H_{ai}\colon \theta \in \Theta_i^c$. Then an IUT of (9) is the test that rejects H_0 if and only if $\mathbf{X} \in \bigcap_{i=1}^{k} R_i$. The rationale behind an IUT is simple. The overall null hypothesis, $H_0\colon \theta \in \bigcup_{i=1}^{k} \Theta_i$, can be rejected only if each of the individual null hypotheses, $H_{0i}\colon \theta \in \Theta_i$, can be rejected.

Berger (1982) proved the following two theorems.

THEOREM 1. *If R_i is a level-α test of H_{0i}, for $i = 1, \ldots, k$, then the intersection–union test with rejection region $R = \bigcap_{i=1}^{k} R_i$ is a level-α test of H_0 versus H_a in* (9).

An important feature in Theorem 1 is that each of the individual tests is performed at level-α, but the overall test also has the same level α. There is no need for multiplicity adjustment for performing multiple tests. The reason there is no need for such a correction is the special way the individual tests are combined. Hypothesis H_0 is rejected only if every one of the individual hypotheses, H_{0i}, is rejected.

Theorem 1 asserts that the IUT is level-α. That is, its size is at most α. In fact, a test constructed by the IUT method can be quite conservative. Its size can be much less than the specified value α. However, Theorem 2 (a generalization of Theorem 2 in Berger, 1982) provides conditions under which the IUT is not conservative; its size is exactly equal to the specified α.

THEOREM 2. *For some $i = 1, \ldots, k$, suppose R_i is a size-α rejection region for testing H_{0i} versus H_{ai}. For every $j = 1, \ldots, k$, $j \neq i$, suppose R_j is a level-α rejection region for testing H_{0j} versus H_{aj}. Suppose there exists a sequence of parameter points $\theta_l, l = 1, 2, \ldots,$ in Θ_i such that*

$$\lim_{l\to\infty} P_{\theta_l}(\mathbf{X} \in R_i) = \alpha$$

and, for every $j = 1, \ldots, k$, $j \neq i$,

$$\lim_{l\to\infty} P_{\theta_l}(\mathbf{X} \in R_j) = 1.$$

Then the intersection–union test with rejection region $R = \bigcap_{i=1}^{k} R_i$ is a size-α test of H_0 versus H_a.

Note that in Theorem 2 the one test defined by R_i has size exactly α. The other tests defined by R_j, $j = 1, \ldots, k$, $j \neq i$, are level-α tests. That is, their sizes may be less than α. The conclusion is the IUT has size α. Thus, if rejection regions $R_1, \ldots, R_k$ with sizes $\alpha_1, \ldots, \alpha_k$ are combined in an IUT and Theorem 2 is applicable, then the IUT will have size equal to $\max_i\{\alpha_i\}$. We will discuss bioequivalence examples in which tests of different sizes are combined. The resulting test has size equal to the maximum of the individual sizes.

4. OLD AND NEW TESTS FOR DIFFERENCE AND RATIO HYPOTHESES

4.1 Two One-Sided Tests

The TOST is naturally thought of as an IUT. The bioequivalence alternative hypothesis $H_a\colon \theta_L < \eta_T - \eta_R < \theta_U$ is conveniently expressed as the intersection of the two sets,

$$\Theta_1^c = \{(\eta_T, \eta_R, \sigma_D^2)\colon \eta_T - \eta_R > \theta_L\}$$

and

$$\Theta_2^c = \{(\eta_T, \eta_R, \sigma_D^2)\colon \eta_T - \eta_R < \theta_U\}.$$

The test that rejects $H_{01}\colon \eta_T - \eta_R \leq \theta_L$ in (5) if $T_L \geq t_{\alpha, r}$ is a size-α test of H_{01}. The test that rejects $H_{02}\colon \eta_T - \eta_R \geq \theta_U$ in (6) if $T_U \leq -t_{\alpha, r}$ is a size-α

test of H_{02}. So, by Theorem 1, the test that rejects H_0 only if both of these tests reject is a level-α test of (2).

To use Theorem 2 to see that the size of the TOST is exactly α, consider parameter points with $\eta_T - \eta_R = \theta_U$ and take the limit as $\sigma_D^2 \to 0$. Such parameters are on the boundary of H_{02}. Therefore,

$$P(\mathbf{X} \in R_2) = P(T_U \le -t_{\alpha,r}) = \alpha,$$

for any $\sigma_D^2 > 0$. But,

$$P(\mathbf{X} \in R_1) = P(T_L \ge t_{\alpha,r}) \to 1 \quad \text{as } \sigma_D^2 \to 0,$$

because the power of a one-sided t-test converges to 1 as $\sigma_D^2 \to 0$ for any point in the alternative. The value $\eta_T - \eta_R = \theta_U$ is in the alternative, H_{a1}.

The advantage of considering bioequivalence problems in an IUT format is not limited to verifying properties of the TOST. Rather, other bioequivalence hypotheses, such as (1), state an interval as the alternative hypothesis. This interval can be expressed as the intersection of two one-sided intervals. So two one-sided, size-α tests can be combined to obtain a level-α (typically, size-α) test. Furthermore, as we will see in Section 6, even more complicated forms of bioequivalence can be expressed in the IUT format. This allows the easy construction of tests with guaranteed size-α for these problems.

4.2 More Powerful Tests

Despite its simplicity and intuitive appeal, the TOST suffers from a lack of power. The line labeled TOST in the top part of Table 1 shows the power function, $P(\text{reject } H_0)$, for parameter points with $\eta_T - \eta_R = \theta_U$ (or θ_L), points on the boundary between H_0 and H_a. The power function is near α for σ_D^2 near 0, but decreases as σ_D^2 grows. An unbiased test would have power equal to α for all such parameter points. The TOST is clearly biased. The bottom part of Table 1 shows the power function when the two drugs are exactly equal, $\eta_T = \eta_R$. The power is near 1 for σ_D^2 near 0, but decreases to 0 as σ_D^2 increases. Despite these shortcomings, Diletti, Hauschke and Steinijans (1991) declared that the TOST maximizes the power among all size-α tests. This is incorrect.

TABLE 1
Powers of three bioequivalence tests; $r = 30$, $\alpha = 0.05$ *and* $\theta_U = \log(1.25) = -\theta_L$

	σ_D							
	0.00	0.04	0.08	0.12	0.16	0.20	0.30	∞
	$\eta_T - \eta_R = \theta_U$ *or* θ_L							
TOST	0.050	0.050	0.050	0.031	0.003	0.000	0.000	0.000
BHM	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
New	0.050	0.050	0.050	0.047	0.049	0.050	0.050	0.050
	$\eta_T - \eta_R = 0$							
TOST	1.000	1.000	0.720	0.158	0.007	0.000	0.000	0.000
BHM	1.000	1.000	0.721	0.260	0.131	0.093	0.066	0.050
New	1.000	1.000	0.720	0.247	0.128	0.092	0.066	0.050

Anderson and Hauck (1983) proposed a test with higher power than the TOST. Whereas the TOST does not reject H_0 if $\text{SE}(D)$ is sufficiently large, the Anderson and Hauck test always rejects H_0 if D is near enough to 0, even if $\text{SE}(D)$ is large. This provides an improvement in power. However, the Anderson and Hauck test does not control the Type I error probability at the specified level α. It is liberal and the size is somewhat greater than α. Shortly after Anderson and Hauck proposed their test, Patel and Gupta (1984) and Rocke (1984) proposed the same test. This scientific coincidence was commented upon by Anderson and Hauck (1985) and Martin Andrés (1990).

Due to the seriousness of a Type I error, declaring two drugs to be equivalent when they are not, the search for a size-α test that was uniformly more powerful than the TOST continued. Munk (1993) proposed a slightly different test. Munk claims that this test is a size-α test that is uniformly more powerful than the TOST, but this claim is supported by numerical calculations, not analytic results.

Brown, Hwang and Munk (1995) constructed an unbiased, size-α test of (2) that is uniformly more powerful than the TOST. Their construction is recursive. To determine if a point $(d, \text{se}(D))$ is in the rejection region of the Brown, Hwang and Munk test, a good deal of computing can be necessary. This may limit the practical usefulness of the Brown, Hwang and Munk test. Also, sometimes the Brown, Hwang and Munk rejection region has a quite irregular shape. An example of this is shown in Figure 1.

We will now describe a new test of the hypotheses (2). This test is uniformly more powerful than the TOST. Unlike the Anderson and Hauck and Munk tests, our test is a size-α test. Our test is nearly unbiased. It is simpler to compute than the Brown, Hwang and Munk test. It will not have the irregular boundaries that the Brown, Hwang and Munk test sometimes possesses. The construction of this new test again illustrates the usefulness of the IUT method.

To simplify the notation in describing our test, we assume, without loss of generality, that $\theta_L = -\theta_U$ and we call $\theta_U = \Delta$. Following Brown, Hwang and Munk, define $S_*^2 = r[\text{SE}(D)]^2$. It is simpler to define our test in terms of the polar coordinates, centered

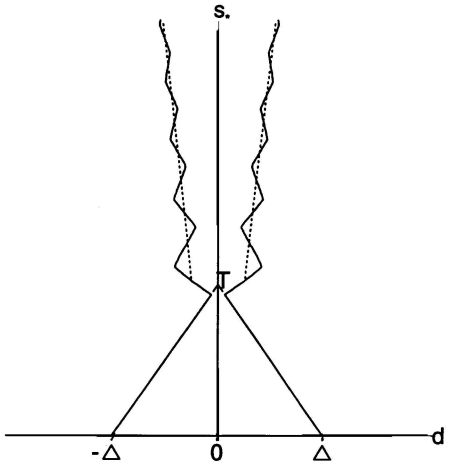

FIG. 1. *Irregular boundary of Brown, Hwang and Munk test (solid line) and smoother boundary of test from Section 4.2 (dashed line); the TOST rejection region is bounded by the triangle with vertices at* $-\Delta$, Δ *and* T. *Here* $r = 3$, $\alpha = 0.16$ *and* $-\theta_L = \theta_U = 1$.

at $(\Delta, 0)$,

$$v^2 = (d - \Delta)^2 + s_*^2$$

and

$$b = \cos^{-1}((d - \Delta)/v).$$

In the (d, s_*) space, v is the distance from $(\Delta, 0)$ to (d, s_*) and b is the angle between the d axis and the line segment joining $(\Delta, 0)$ and (d, s_*). To define a size-α test, we need the distribution of (V, B) when $\theta = \Delta$. In this case, it is easy to verify that V and B are independent. The probability density function of B is

$$f(b) = \frac{\Gamma((r+1)/2)}{\Gamma(r/2)\sqrt{\pi}}[\sin(b)]^{r-1}, \quad 0 < b < \pi,$$

which does not depend on σ_D^2. To implement our test, it is useful to note that the cumulative distribution function of B has a closed form given by

$$F(b) = \frac{b}{\pi} - \frac{1}{2\sqrt{\pi}} \cdot \sum_{k=1}^{(r-1)/2} [\sin(b)]^{2k-1} \cos(b) \frac{\Gamma(k)}{\Gamma(k + (1/2))},$$

if r is odd, and

$$F(b) = \frac{1}{2} - \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{r/2} [\sin(b)]^{2k-2} \cos(b) \frac{\Gamma(k - (1/2))}{\Gamma(k)},$$

if r is even. The probability density function of V will be denoted by $g_{\sigma_D}(v)$.

We will describe the rejection region of the new test geometrically here. Exact formulas are in the Appendix. The new test will be an IUT. We will define a size-α, unbiased rejection region, R_2, for testing (6). This R_2 will contain the rejection region of the size-α TOST and will be approximately symmetric about the line $d = 0$. Then we will define $R_1 = \{(d, s_*)\colon (-d, s_*) \in R_2\}$; R_1 is R_2 reflected across the line $d = 0$; R_1 is a size-α, unbiased rejection region for testing (5). Then $R = R_1 \cap R_2$ is the rejection region of the new test. Because R_2 is approximately symmetric about the line $d = 0$, R_1 is almost the same as R_2, and not much is deleted when we take the intersection. This foresight in choosing the individual rejection regions so that the intersection is not much smaller is always useful when using the IUT method.

The set $\{V = v\}$ is a semicircle in (d, s_*) space. For each value of v, $R_2(v) \equiv \{V = v\} \cap R_2$ is either one or two intervals of b values, that is, one or two arcs on $\{V = v\}$. These arcs will be chosen so that, for every $v > 0$,

$$\int_{R_2(v)} f(b)\, db = \alpha. \tag{10}$$

Then the rejection probability

$$P(R_2) = \int_0^\infty \int_{R_2(v)} f(b)\, db\, g_{\sigma_D}(v)\, dv = \int_0^\infty \alpha g_{\sigma_D}(v)\, dv = \alpha,$$

for every $\sigma_D > 0$ if $\eta_T - \eta_R = \Delta$. This will ensure that R_2 is a size-α, unbiased rejection region for testing (6).

We now define the arc(s) that make up $R_2(v)$. Refer to Figure 2 in this description. The rejection region of the size-α TOST, call it R_T, is the triangle bounded by the lines $s_* = 0$, $d = \Delta - t_{\alpha, r} s_*/\sqrt{r}$ (call this line l_U) and $d = -\Delta + t_{\alpha, r} s_*/\sqrt{r}$ (call this

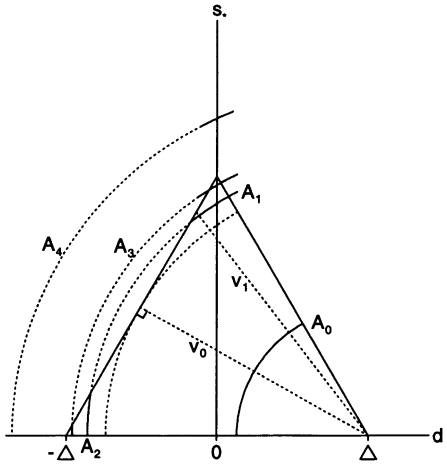

FIG. 2. *Arcs that define the rejection region* R_2.

line l_L). Let v_0 denote the distance from $(\Delta, 0)$ to l_L. In this description, we assume $1/2 > \alpha > \alpha_* \equiv 1 - F(3\pi/4)$. Brown, Hwang and Munk (1995) in their Table 1 show that if $r \geq 4$, then $\alpha = 0.05 > \alpha_*$. The new test for $\alpha \leq \alpha_*$ is given in the Appendix. Brown, Hwang and Munk did not propose any test for $\alpha \leq \alpha_*$. The condition $\alpha > \alpha_*$ ensures that the point on l_L closest to $(\Delta, 0)$ is on the boundary of R_T, as shown.

Let b_0 denote the angle between the d-axis and l_U. For $0 < v \leq v_0$, $R_2(v) = \{b\colon b_0 < b < \pi\}$. The arc A_0 in Figure 2 is an example of such an arc. So, for $v < v_0$, $R_2(v)$ is exactly the points in the TOST.

For $v_0 < v$, the semicircle $V = v$ intersects l_L at two points. Let $b_1 < b_2$ denote the angles corresponding to these two points. If $v_0 < v < 2\Delta$, let $A_2(v) = \{b\colon b_2 < b < \pi\}$. These are the points in R_T adjacent to the d-axis, and A_2 in Figure 2 is an example of such an arc. If $2\Delta \leq v$, let $A_2(v)$ be the empty set. Let $\alpha(v)$ denote the probability content of $A_2(v)$ under F. That is,

$$\alpha(v) = \begin{cases} 1 - F(b_2), & v_0 < v < 2\Delta, \\ 0, & 2\Delta \leq v. \end{cases}$$

For $v_0 < v$, $R_2(v) = A_1(v) \cup A_2(v)$, where, to ensure that (10) is true, $A_1(v)$ must satisfy

$$\int_{A_1(v)} f(b)\, db = \alpha - \alpha(v). \tag{11}$$

Let (d_1, s_{*1}) denote the point where the $\{V = v_0\}$ semicircle intersects l_U, and let v_1 denote the radius corresponding to $(-d_1, s_{*1})$. For $v_0 < v < v_1$, let b_{L1} be the angle defined by

$$F(b_1) - F(b_{L1}) = \alpha - \alpha(v), \tag{12}$$

where b_1 is as defined in the previous paragraph. Then $A_1(v) = \{b\colon b_{L1} < b < b_1\}$ is the arc that satisfies (11) whose endpoint is on l_L. For $v_0 < v < v_1$, $R_2(v) = A_1(v) \cup A_2(v)$, using this $A_1(v)$. The arcs labeled A_1 and A_2 in Figure 2 comprise such an $R_2(v)$. For $v < v_1$, the cross sections $R_2(v)$ we have defined are the same as the cross sections for the Brown, Hwang and Munk (1995) test. They now define the remainder of their rejection region recursively in terms of these arcs. We define our rejection region in a nonrecursive manner.

For $v_1 \leq v$, define two values $b_L(v) < b_U(v)$ such that $F(b_U(v)) - F(b_L(v)) = \alpha - \alpha(v)$, and the angle between the line joining $(0, 0)$ and $(v, b_L(v))$ and the s_*-axis is the same as the angle between the line joining $(0, 0)$ and $(v, b_U(v))$ and the s_*-axis. This equal angle condition is what we meant earlier by the phrase "approximately symmetric about the line $d = 0$." If $b_U(v) \geq b_1$, then $A_1(v) = \{b\colon b_L(v) < b < b_U(v)\}$. But, if $b_U(v) < b_1$, then this arc does not contain all the points in the TOST. So, if $b_U(v) < b_1$, $A_1(v) = \{b\colon b_{L1} < b < b_1\}$, where b_{L1} is defined by (12). For $v_1 \leq v$, $R_2(v) = A_1(v) \cup A_2(v)$. Recall, if $2\Delta \leq v$, $A_2(v)$ is empty, and $R_2(v)$ is the single arc $A_1(v)$. Also, for $v^2 \geq \max\{4\Delta^2, \Delta^2 + \Delta^2 r/t^2_{\alpha, r}\}$, the semicircle $\{V = v\}$ does not intersect R_T, and $R_2(v)$ is the arc defined by $b_L(v)$ and $b_U(v)$. The b_1-condition never applies in this case. In Figure 2, the solid parts of the arcs A_3 and A_4 are examples of $R_2(v)$ for $v_1 \leq v$.

The cross sections $R_2(v)$ have been defined for every $v > 0$, and this defines R_2; R_1 is the reflection of R_2 across the s_*-axis, and the rejection region of the new test is $R = R_1 \cap R_2$. This construction is illustrated in Figure 3.

In Figure 1, the rejection region R with the same size as the Brown, Hwang and Munk test is the region between the dotted lines. The boundary of R is smooth compared to the irregular boundary of the Brown, Hwang and Munk test. This smoothness results from the attempt in the construction of R to center arcs around the s_*-axis. To determine if a sample point (d, s_*^2) is in R, two arcs, $R_2(v)$ and $R_1(v) = R_2(v')$ ($v' = (-d - \Delta)^2 + s_*^2$, computed from $(-d, s_*^2)$), must be constructed. If (d, s_*^2) is on both arcs, $(d, s_*^2) \in R$. But, to determine if (d, s_*^2) is in the rejection region of the Brown, Hwang and Munk test, a starting point is selected. Then a sequence of arcs is constructed until (d, s_*^2) is passed. Then another sequence of arcs is constructed from a new starting point. This process is continued until enough arcs in the vicinity of (d, s_*^2) are obtained to approximate the boundary of the rejection region. From this it is determined if (d, s_*^2) is in the rejection region. Thus, a good deal more computation is

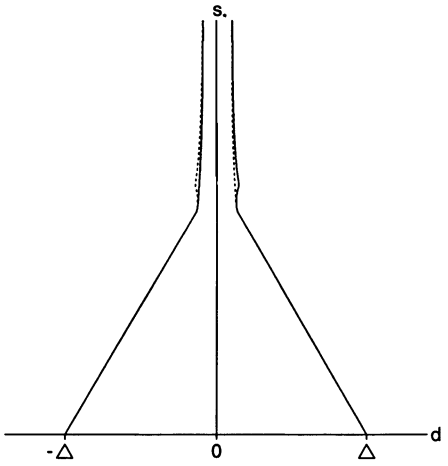

FIG. 3. *Rejection region of new test*; *region* R_2 (*between solid lines*) *and region* R_1 (*between dashed lines*); *rejection region* $R = R_1 \cap R_2$; $r = 10$ *and* $\alpha = 0.05$.

needed to implement the Brown, Hwang and Munk test. Also, the Brown, Hwang and Munk test is not defined for $\alpha \leq \alpha_*$. This smoothness, general applicability and simplicity of computation recommends R as a reasonable alternative to the Brown, Hwang and Munk test. But R is slightly biased whereas the Brown, Hwang and Munk test is unbiased.

A small power comparison of the TOST, Brown, Hwang and Munk test and our new test is given in Table 1 for $\alpha = 0.05$ and $r = 30$. In the top block of numbers, $\eta_T - \eta_R = \Delta$. For these boundary values, the power is exactly $\alpha = 0.05$ for the unbiased Brown, Hwang and Munk test. The power is also very close to 0.05 for our test, indicating it has only slight bias. But the TOST is highly biased with power much less than 0.05 for moderate and large σ_D. In the bottom block of numbers, $\eta_T - \eta_R = 0$. The drugs are equivalent. Our test and the Brown, Hwang and Munk test have very similar powers. Their powers are much greater than the TOST's power for all but small σ_D. For example, it can be seen that the power improvement is about 60% when $\sigma_D = 0.12$ and about 85% when $\sigma_D = 0.16$. Sample sizes for bioequivalence tests are often chosen so that the test has power of about 0.8 when $\eta_T = \eta_R$. In this case, Table 1 indicates there is no advantage to using the new tests over the TOST. But if the variability turns out to be larger than expected in the planning stage, the new tests offer significant power improvements.

The tests of Anderson and Hauck (1983), Brown, Hwang and Munk (1995) and our new test all have the property that, as $s_* \to \infty$, the width of the rejection region increases, eventually containing values of (d, s_*) with d outside the interval (θ_L, θ_U). There will be values (d, s_{*1}) and (d, s_{*2}) with $s_{*1} < s_{*2}$, but (d, s_{*1}) is not in the rejection region while (d, s_{*2}) is in the rejection region. This "flaring out" of the rejection region is evident in Figures 1 and 5 (see Section A.2). This counterintuitive shape was pointed out by Rocke (1984). The rejection region of any bioequivalence test that is unbiased, or approximately unbiased, must eventually contain sample points with d outside the interval (θ_L, θ_U). Some have suggested that such procedures should be truncated in the sense that the narrowest point of the rejection region be determined and then the rejection region is extended along the s_*-axis only of this width. Brown, Hwang and Munk suggest this as a possible modification of their test, although the resulting test will no longer be unbiased. We believe that notions of size, power and unbiasedness are more fundamental than "intuition" and do not recommend truncation. But for those who disagree, our new test could be truncated in this same way. The narrowest point will need to be determined numerically for all these tests, and the smoother shape of our rejection region will make this determination easier. Referring to Figure 1, a numerical routine might be fooled by the irregular shape of the Brown, Hwang and Munk test.

4.3 Tests for Ratios of Parameters

Usually, data from a bioequivalence trial is logarithmically transformed before analysis. This leads to a test of the hypotheses (2), as described in the previous section. In the model we will consider now, the original data are normally distributed. Let $X_1, \ldots, X_m$ form a random sample from a normal population with mean μ_T and variance σ^2, and let $Y_1, \ldots, Y_n$ form an independent random sample from a normal population with mean μ_R and variance σ^2. In this section, we will present our comments in terms of this simple parallel design. Yang (1991) and Liu and Weng (1995) describe models for this normally distributed data in crossover experiments.

The bioequivalence hypothesis to be tested in this case is (1), namely,

$$H_0: \quad \frac{\mu_T}{\mu_R} \leq \delta_L \quad \text{or} \quad \frac{\mu_T}{\mu_R} \geq \delta_U$$

(13) versus

$$H_a: \quad \delta_L < \frac{\mu_T}{\mu_R} < \delta_U.$$

In the past, the values of $\delta_L = 0.80$ and $\delta_U = 1.20$ were commonly used (called the ± 20 rule). However, the FDA Division of Bioequivalence (FDA, 1992a) now uses $\delta_L = 0.80$ and $\delta_L = 1.25$. These limits are symmetric about 1 in the ratio scale since $0.80 = 1/1.25$.

The parameter μ_R is positive because the measured variable, AUC or $C_{\max}$, is positive. Therefore the hypotheses (13) can be restated as

$$H_0: \quad \mu_T - \delta_L \mu_R \leq 0 \quad \text{or} \quad \mu_T - \delta_U \mu_R \geq 0$$

(14) versus

$$H_a: \quad \mu_T - \delta_L \mu_R > 0 \quad \text{and} \quad \mu_T - \delta_U \mu_R < 0.$$

The testing problem (14) was first considered by Sasabuchi (1980, 1988a, b). Let $\overline{X}$, $\overline{Y}$ and S^2 denote the two sample means and the pooled estimate of σ^2. Sasabuchi showed that the size-α likelihood ratio test of (14) rejects H_0 if and only if

$$T_1 \geq t_{\alpha, r} \quad \text{and} \quad T_2 \leq -t_{\alpha, r},$$

where

$$T_1 = \frac{\overline{X} - \delta_L \overline{Y}}{S\sqrt{1/m + \delta_L^2/n}}$$

and

$$T_2 = \frac{\overline{X} - \delta_U \overline{Y}}{S\sqrt{1/m + \delta_U^2/n}}.$$

This will be called the T_1/T_2 test.

The T_1/T_2 test is easily understood as an IUT. The usual, normal theory, size-α t-test of H_{01}: $\mu_T - \delta_L\mu_R \le 0$ versus H_{a1}: $\mu_T - \delta_L\mu_R > 0$ is the test that rejects H_{01} if $T_1 \ge t_{\alpha,r}$. Similarly, the usual, normal theory, size-α t-test of H_{02}: $\mu_T - \delta_U\mu_R \ge 0$ versus H_{a2}: $\mu_T - \delta_U\mu_R < 0$ is the test that rejects H_{02} if $T_2 \le -t_{\alpha,r}$. Because H_a is the intersection of H_{a1} and H_{a2}, these two t-tests can be combined, using the IUT method, to get a level-α test of H_0 versus H_a. Using an argument like that in Section 4.1, Theorem 2 can be used to show that the size of this test is α.

Yang (1991) and Liu and Weng (1995) proposed tests closely related to the T_1/T_2 test for the bioequivalence problem of testing (13) in a crossover experiment. Hauck and Anderson (1992) also discuss the hypotheses in the form (14), but no reference to Sasabuchi's earlier work is given. The derivation of the confidence set for μ_T/μ_R in Hsu, Hwang, Liu and Ruberg (1994) contains a mistake in the standardization. Properly corrected, their rather complicated confidence set would lead to the rejection of (14) when the simple test described above does. So, somehow, the value of this simple, size-α test seems to have been completely overlooked in the bioequivalence literature. Rather, Chow and Liu (1992) and Liu and Weng (1995) both report that the following is the standard analysis. Rewrite the hypotheses (13) or (14) as

$$\begin{aligned} &H_0\colon \mu_T - \mu_R \le (\delta_L - 1)\mu_R \\ &\qquad \text{or } \mu_T - \mu_R \ge (\delta_U - 1)\mu_R \\ &\text{versus} \\ &H_a\colon (\delta_L - 1)\mu_R < \mu_T - \mu_R < (\delta_U - 1)\mu_R. \end{aligned} \tag{15}$$

These hypotheses look like (2), but there is an important difference. In (2), θ_L and θ_U are known constants. In (15), $(\delta_L - 1)\mu_R$ and $(\delta_U - 1)\mu_R$ are unknown parameters. Nevertheless, the standard analysis proceeds to use the TOST with $(\delta_L - 1)\overline{Y}$ replacing θ_L in T_L and $(\delta_U - 1)\overline{Y}$ replacing θ_U in T_U. The standard analysis ignores the fact that a constant has been replaced by a random variable and compares these two test statistics to standard t-percentiles as in the TOST. This test will be called the T_1^*/T_2^* test.

The statistics that are actually used in this analysis are

$$\begin{aligned} T_1^* &= \frac{\overline{X} - \overline{Y} - (\delta_L - 1)\overline{Y}}{S\sqrt{1/m + 1/n}} \\ &= \frac{\overline{X} - \delta_L\overline{Y}}{S\sqrt{1/m + 1/n}} = T_1\sqrt{\frac{n + m\delta_L^2}{n + m}}, \end{aligned}$$

and

$$\begin{aligned} T_2^* &= \frac{\overline{X} - \overline{Y} - (\delta_U - 1)\overline{Y}}{S\sqrt{1/m + 1/n}} \\ &= \frac{\overline{X} - \delta_U\overline{Y}}{S\sqrt{1/m + 1/n}} = T_2\sqrt{\frac{n + m\delta_U^2}{n + m}}. \end{aligned}$$

The statistics T_1 and T_2 are properly scaled to have Student's t-distributions, but T_1^* and T_2^* are not. The T_1^*/T_2^* test is an IUT in which the two tests have different sizes. The test that rejects H_{01} if $T_1^* > t_{\alpha,r}$ has size

$$\begin{aligned} &P_{\mu_T = \delta_L\mu_R}(T_1^* > t_{\alpha,r}) \\ &\quad = P_{\mu_T = \delta_L\mu_R}\left(T_1 > \sqrt{\frac{n+m}{n+m\delta_L^2}}\,t_{\alpha,r}\right) \\ &\quad = \alpha_1 < \alpha, \end{aligned}$$

because

$$\sqrt{\frac{n+m}{n+m\delta_L^2}} > 1.$$

On the other hand, the test that rejects H_{02} if $T_2^* < -t_{\alpha,r}$ has size

$$\begin{aligned} &P_{\mu_T = \delta_U\mu_R}(T_2^* < -t_{\alpha,r}) \\ &\quad = P_{\mu_T = \delta_U\mu_R}\left(T_2 < -\sqrt{\frac{n+m}{n+m\delta_U^2}}\,t_{\alpha,r}\right) \\ &\quad = \alpha_2 > \alpha, \end{aligned}$$

because

$$\sqrt{\frac{n+m}{n+m\delta_U^2}} < 1.$$

Theorem 2 can be used to show that, as a test of the hypothesis (13), the T_1^*/T_2^* test has size $\alpha_2 > \alpha$. It is a liberal test.

The true size of the T_1^*/T_2^* test, for a nominal size of $\alpha = 0.05$, is shown in Table 2. In Table 2 it is assumed that the sample sizes from the test and reference drugs are equal, $m = n$. In this case, the size of the T_1^*/T_2^* test is simply

$$\alpha_2 = P\left(T < -\sqrt{\frac{2}{1+\delta_U^2}}\,t_{\alpha,r}\right),$$

TABLE 2
Actual size of T_1^*/T_2^* test for nominal $\alpha = 0.05$

$m = n$	5	10	15	20	30	∞
Size	0.070	0.071	0.072	0.072	0.073	0.073

where T has a Student's t-distribution with $r = 2n - 2$ degrees of freedom. It can be seen that the size of the T_1^*/T_2^* test is about 0.07 for all sample sizes. The liberality worsens slightly as the sample size increases. On the other hand, the T_1/T_2 test has size exactly equal to the nominal α. It is just as simple to implement as the T_1^*/T_2^* test. Therefore the T_1/T_2 test should replace the T_1^*/T_2^* test for testing (13).

In Section 4.2, the IUT method was used to construct a size-α test that is uniformly more powerful than the TOST. For the known σ^2 case, Berger (1989) and Liu and Berger (1995) used the IUT method to construct size-α tests that are uniformly more powerful than the T_1/T_2 test. In Figure 4, the cone-shaped region labeled R_0 is the rejection region of the T_1/T_2 test for $\alpha = 0.05$. The region between the dashed lines is the rejection region of Liu and Berger's size-α test that is uniformly more powerful. We refer the reader to Berger (1989) and Liu and Berger (1995) for the details about these tests. We believe that, for the case of σ^2 unknown, size-α tests that are uniformly more powerful than the T_1/T_2 test will be found.

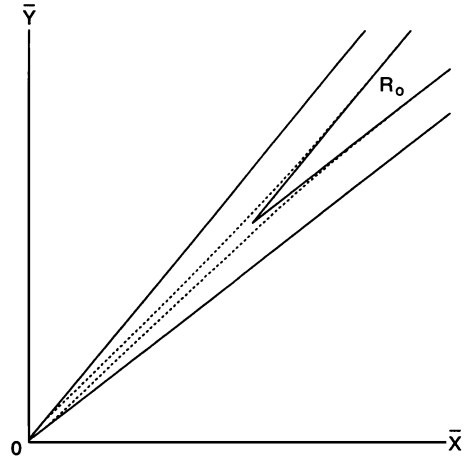

FIG. 4. *Rejection region for T_1/T_2 test is cone shaped R_0; region between dashed lines is rejection region of uniformly more powerful Liu and Berger* (1995) *test. The estimates $\overline{X}$ and $\overline{Y}$ satisfy $\delta_L < \overline{X}/\overline{Y} < \delta_U$ in the larger cone-shaped region.*

5. CONFIDENCE SETS AND BIOEQUIVALENCE TESTS

5.1 A 100(1 − α)% Confidence Interval

We will show that the $100(1-\alpha)\%$ confidence interval $[D_1^-, D_1^+]$ given by

$$(16) \quad \left[\left(D - t_{\alpha,r}\mathrm{SE}(D)\right)^-, \left(D + t_{\alpha,r}\mathrm{SE}(D)\right)^+\right]$$

corresponds to the size-α TOST for (2). Here $x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$. The $100(1-\alpha)\%$ interval (16) is equal to the $100(1-2\alpha)\%$ interval (8) when the interval (8) contains zero. But, when the interval (8) lies to the right (left) of zero, the interval (16) extends from zero to the upper (lower) endpoint of interval (8).

The confidence interval (16) has been derived by Hsu (1984), Bofinger (1985) and Stefansson, Kim and Hsu (1988) in the multiple comparisons setting, and by Müller-Cohrs (1991), Bofinger (1992) and Hsu et al. (1994) in the bioequivalence setting. Our derivation follows Stefansson, Kim and Hsu (1988) and Hsu et al. (1994), which makes the correspondence to TOST more explicit.

To see this correspondence, we use the standard connection between tests and confidence sets. Most often in statistics, this connection is used to construct confidence sets from tests via a result such as the following.

THEOREM 3 (Lehmann, 1986, page 90). *Let the data $\mathbf{X}$ have a probability distribution that depends on a parameter $\boldsymbol{\theta}$. Let Θ denote the parameter space. For each $\boldsymbol{\theta}_0 \in \Theta$, let $A(\boldsymbol{\theta}_0)$ denote the acceptance region of a level-α test of H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. That is, for each $\boldsymbol{\theta}_0 \in \Theta$, $P_{\theta=\theta_0}(\mathbf{X} \in A(\boldsymbol{\theta}_0)) \geq 1 - \alpha$. Then $C(\mathbf{x}) = \{\boldsymbol{\theta} \in \Theta: \mathbf{x} \in A(\boldsymbol{\theta})\}$ is a level $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$.*

However, in bioequivalence testing in the past, tests have often been constructed from confidence sets. A result related to this practice follows.

THEOREM 4. *Let the data $\mathbf{X}$ have a probability distribution that depends on a parameter $\boldsymbol{\theta}$. Suppose $C(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$. That is, for each $\boldsymbol{\theta} \in \Theta$, $P_\theta(\boldsymbol{\theta} \in C(\mathbf{X})) \geq 1 - \alpha$. Consider testing H_0: $\boldsymbol{\theta} \in \Theta_0$ versus H_a: $\boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \varnothing$. Then the test that rejects H_0 if and only if $C(\mathbf{X}) \cap \Theta_0 = \varnothing$ is a level-α test of H_0.*

PROOF. Let $\boldsymbol{\theta}_0 \in \Theta_0$. Then

$$P_{\theta_0}(\text{reject } H_0) \leq 1 - P_{\theta_0}(\boldsymbol{\theta}_0 \in C(\mathbf{X})) \leq \alpha. \quad \square$$

Unfortunately, Theorem 4 has not always been carefully applied in the bioequivalence area. Commonly, $100(1-2\alpha)\%$ confidence sets are used in an attempt to define level-α tests. Theorem 4 guarantees only that a level-2α test will result from a $100(1-2\alpha)\%$ confidence set. Sometimes, the size of the resulting test is, in fact, α, but this is not generally true. In this subsection we use Theorem 4 to show the correspondence between the $100(1-\alpha)\%$ confidence interval (16) and the size-α TOST. In the next subsection, we criticize the practice of using $100(1-2\alpha)\%$ confidence sets to define bioequivalence tests.

Let $\theta = \eta_T - \eta_R$. The family of size-α tests with acceptance regions

$$A(\theta_0) = \{(d, \mathrm{se}(D)): |d-\theta_0| \le t_{\alpha/2,\,r}\mathrm{se}(D)\} \tag{17}$$

leads to the usual equivariant confidence interval, which is of the form (8) but with $t_{\alpha,\,r}$ replaced by $t_{\alpha/2,\,r}$.

However, no current law or regulation states one must employ confidence sets that are equivariant over the entire real line. Using Theorem 4 and inverting the family of size-α tests defined by, for $\theta_0 \ge 0$,

$$A(\theta_0) = \{(d, \mathrm{se}(D)): d-\theta_0 \ge -t_{\alpha,\,r}\mathrm{se}(D)\} \tag{18}$$

and, for $\theta_0 < 0$,

$$A(\theta_0) = \{(d, \mathrm{se}(D)): d-\theta_0 \le t_{\alpha,\,r}\mathrm{se}(D)\} \tag{19}$$

yields the $100(1-\alpha)\%$ confidence interval (16). Technically, when inverting (18) and (19), the upper confidence limit will be open when $D + t_{\alpha,\,r}\mathrm{SE}(D) < 0$. This point is inconsequential in bioequivalence testing. The only value of the upper bound with positive probability is 0, and, in bioequivalence testing, the inference $\eta_T \ne \eta_R$ is not of interest. In terms of operating characteristics, the confidence interval with the possibly open endpoint has coverage probability $100(1-\alpha)\%$ everywhere. The confidence interval (16) also has coverage probability $100(1-\alpha)\%$ except at $\eta_T - \eta_R = 0$, where it has 100% coverage probability.

Note that the family of tests (18) contains the one-sided size-α t-test for (6), and the family of tests (19) contains the one-sided size-α t-test for (5), in contrast to the family of tests (17). The 5% TOST is equivalent to asserting bioequivalence, $\theta_L < \eta_T - \eta_R < \theta_U$, if and only if the 95% confidence interval $[D_1^-, D_1^+] \subset (\theta_L, \theta_U)$. Therefore, as pointed out by Hsu et al. (1994), it is more consistent with standard statistical theory to say that the $100(1-\alpha)\%$ confidence interval $[D_1^-, D_1^+]$, instead of the ordinary $100(1-2\alpha)\%$ confidence interval (8), corresponds to the TOST.

Pratt (1961) showed that for the $r = \infty$ case [i.e., $\mathrm{SE}(D) = \sigma_D$], when $\eta_T = \eta_R$, that is, when the test drug is indeed equivalent to the reference drug, $[D_1^-, D_1^+]$ has the smallest expected length among all $100(1-\alpha)\%$ confidence intervals for $\eta_T - \eta_R$. On the other hand, when $\eta_T - \eta_R$ is far from zero, $[D_1^-, D_1^+]$ has larger expected length than the equivariant confidence interval (8). So the bioequivalence confidence interval $[D_1^-, D_1^+]$ can be thought of as specifically constructed from Theorem 4 for more precise inference when it is expected that η_T is close to η_R. One multiparameter extension of this construction, utilized by Stefansson, Kim and Hsu (1988), gives rise to the multiple comparison with the best (MCB) confidence intervals of Hsu (1984), which eliminate treatments that are not the best and identify treatments close to the true best. In fact, the bioequivalence confidence interval (16) is an MCB confidence interval because, when only two treatments are being compared, a treatment close to the other treatment is either the true best treatment or close to the true best treatment.

This ability of a MCB confidence interval to give *practical equivalence* inference is useful in another problem. Ruberg and Hsu (1992) pointed out that whether to include certain parameters in a regression model should sometimes be formulated as a practical equivalence problem rather than a significant difference problem. In modeling the stability of a drug, for example, given the clear intent of the FDA (1987) Guideline that data from batches of a drug can be pooled only if they have practically equivalent degradation rates, the decision of which $time \times batch$ interaction terms to include in the model can logically be based on MCB confidence intervals comparing the degradation rate of each batch with the true worst degradation rate. Another problem which has not been but should be formulated as one of practical equivalence is the establishment of safety of substances such as bovine growth hormone in toxicity studies (e.g., Juskevich and Guyer, 1990), since the desired inference is practical equivalence between the treated groups and the (negative) control group (cf. Hsu, 1996, Chapter 2).

A different multiparameter extension of the same construction was utilized by Brown, Casella and Hwang (1995) to obtain the confidence region for a vector parameter $\boldsymbol{\theta}$ which has the smallest expected volume when $\boldsymbol{\theta} = \mathbf{0}$, generalizing Pratt's result. The confidence set is constructed through Theorem 4 using the family of size-α Neyman–Pearson likelihood ratio tests for H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_a: $\boldsymbol{\theta} = \mathbf{0}$. When $\widehat{\boldsymbol{\theta}}$ is multivariate normal with unknown mean vector $\boldsymbol{\theta}$ and known variance–covariance matrix Σ, the

acceptance regions are

$$A(\boldsymbol{\theta}_0) = \left\{\widehat{\boldsymbol{\theta}}\colon \boldsymbol{\theta}_0'\Sigma^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/\sqrt{\boldsymbol{\theta}_0'\Sigma^{-1}\boldsymbol{\theta}_0} > -t_{\alpha,\infty}\right\},$$

which leads to the confidence region

$$C(\widehat{\boldsymbol{\theta}}) = \left\{\boldsymbol{\theta}\colon \boldsymbol{\theta}'\Sigma^{-1}\widehat{\boldsymbol{\theta}}/\sqrt{\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta}} + t_{\alpha,\infty} > \sqrt{\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta}}\right\}. \tag{20}$$

Their paper describes and illustrates interesting geometric properties of $C(\widehat{\boldsymbol{\theta}})$.

It should be pointed out that the utility of Theorem 4 is not restricted to the construction of confidence sets which give better practical equivalence inference. Stefansson, Kim and Hsu (1988) and Hayter and Hsu (1994) used Theorem 4 to construct confidence sets associated with step-down and step-up multiple comparison methods, which are usually thought of as specifically constructed to give better significant difference inference than single-step methods.

5.2 100(1 − 2α)% Confidence Intervals

Bioequivalence tests are often defined in terms of $100(1-2\alpha)\%$ confidence sets. That is, if $\boldsymbol{\theta}$ denotes the parameter of interest, Θ_0^c denotes the set of parameter values for which the drugs are bioequivalent and $C(\mathbf{X})$ is a $100(1-2\alpha)\%$ confidence set for $\boldsymbol{\theta}$, then the drugs are declared bioequivalent if and only if $C(\mathbf{X}) \subset \Theta_0^c$. This practice seems to be based entirely on the perceived equivalence between the $100(1-2\alpha)\%$ confidence interval (8) and the size-α TOST of (2). This practice is encouraged by the fact that both FDA (1992a) and EC-GCP (1993) specify that the $\alpha = 0.05$ TOST should be executed by constructing a 90% confidence interval. In the bioequivalence literature, when used in this way, the 90% is called the *assurance* of the confidence set.

The intent of the regulating agencies is clearly to use a test with size $\alpha = 0.05$. Unfortunately, bioequivalence tests have been proposed using $100(1-2\alpha)\%$ confidence sets without any verification that the resulting tests have size α. Theorem 4 guarantees that the resulting test is a level-2α test, not size-α. In this section, we will explore the usage of $100(1-2\alpha)\%$ confidence sets. We shall show that the usual $100(1-2\alpha)\%$ confidence interval (8) results in a size-α TOST of (2) because (8) is "equal-tailed." So the relationship is deeper than the "algebraic coincidence" mentioned by Brown, Casella and Hwang (1995). Hauck and Anderson (1992) discuss this fact without proof. We shall see in examples that the use of $100(1-2\alpha)\%$ confidence sets can result in both liberal and conservative bioequivalence tests. Because there is no general guarantee that a $100(1-2\alpha)\%$ confidence set will result in a size-α test, we believe it is unwise to attempt to define a size-α test in terms of a $100(1-2\alpha)\%$ confidence set. Rather, a test with the specified Type I error probability of α should be used. Theorem 4 might be used to construct the corresponding $100(1-\alpha)\%$ confidence set.

Let $[C^-, C^+]$ denote (8), the usual $100(1-2\alpha)\%$ confidence interval for $\eta_T - \eta_R$. Why does rejecting H_0 in (2) if and only if $[C^-, C^+] \subset (\theta_L, \theta_U)$ result in a size-α test? The superficial answer is that, obviously, $C^+ < \theta_U$ is equivalent to $T_U < -t_{\alpha,r}$ and $C^- > \theta_L$ is equivalent to $T_L > t_{\alpha,r}$. Thus, the test based on $[C^-, C^+]$ is equivalent to the size-α TOST. But a more thorough understanding of this is suggested by the following result (Casella and Berger, 1990, Exercise 9.1).

THEOREM 5. *Let the data* $\mathbf{X}$ *have a probability distribution that depends on a real-valued parameter* θ. *Suppose* $(-\infty, U(\mathbf{X})]$ *is a* $100(1-\alpha_1)\%$ *upper confidence bound for* θ. *Suppose* $[L(\mathbf{X}), \infty)$ *is a* $100(1-\alpha_2)\%$ *lower confidence bound for* θ. *Then* $[L(\mathbf{X}), U(\mathbf{X})]$ *is a* $100(1-\alpha_1-\alpha_2)\%$ *confidence interval for* θ.

Now consider the $100(1-2\alpha)\%$ confidence interval $[C^-, C^+]$ for $\theta = \eta_T - \eta_R$. The interval $(-\infty, C^+]$ is a $100(1-\alpha)\%$ upper confidence bound for θ. From Theorem 4, the test that rejects H_{02} in (6) if and only if $C^+ < \theta_U$ is a level-α test of H_{02}. Likewise, $[C^-, \infty)$ is a $100(1-\alpha)\%$ lower confidence bound for θ, and the test that rejects H_{01} in (5) if and only if $C^- > \theta_L$ is a level-α test of H_{01}. Forming an IUT from these two level-α tests yields a level-α test of H_0 in (2), by Theorem 1. Thus, we see that it is not so important that $[C^-, C^+]$ is a $100(1-2\alpha)\%$ confidence interval for θ. Rather, it is the fact that $(-\infty, C^+]$ and $[C^-, \infty)$ are both $100(1-\alpha)\%$ confidence intervals that yields a level-α test. That is, it is important that $[C^-, C^+]$ is an "equal-tailed" confidence interval.

It is easy to see that $100(1-2\alpha)\%$ confidence intervals will not always yield size-α tests. Consider an "unequal-tailed" $100(1-2\alpha)\%$ confidence interval for $\theta = \eta_T - \eta_R$, $[C_1^-, C_1^+]$, defined by

$$\left[D - t_{\alpha_2,r}\mathrm{SE}(D),\ D + t_{\alpha_1,r}\mathrm{SE}(D)\right], \tag{21}$$

where $\alpha_1 + \alpha_2 = 2\alpha$. Using $(-\infty, C_1^+]$ to define a test of H_{02} yields a size-α_1 test, and using $[C_1^-, \infty)$ to define a test of H_{01} yields a size-α_2 test. Therefore, by Theorem 1, the IUT that rejects H_0 if and only if $[C_1^-, C_1^+] \subset (\theta_L, \theta_U)$ has level $\max\{\alpha_1, \alpha_2\}$. That this test has size equal to $\max\{\alpha_1, \alpha_2\}$ can be verified using Theorem 2. This relationship between the size of the test and the maximum of the one-sided error probabilities is alluded to by equation (1) in Yee (1986). The size of this test can be made arbitrarily

close to 2α by choosing α_1 close to zero and α_2 close to 2α. In this problem, the only $100(1-2\alpha)\%$ confidence interval of the form (21) that defines a size-α test happens to be the usual, equal-tailed confidence interval, $[C^-, C^+]$.

The preceding example using an unequal-tailed test simply illustrates that defining a bioequivalence test in terms of a $100(1-2\alpha)\%$ confidence interval can lead to a liberal test with size greater than α. But, no one has proposed using the interval (21) to define a bioequivalence test. So we now discuss two other examples that have been proposed in the bioequivalence literature. Both examples concern testing (1) about the ratio μ_T/μ_R.

Tests based on $100(1-2\alpha)\%$ Fieller-type confidence intervals provide examples of tests that are sometimes liberal. Mandallaz and Mau (1981), Locke (1984) and Kinsella (1989) all propose using a Fieller-type (Fieller, 1940, 1954) confidence interval to estimate μ_T/μ_R. Neither Locke nor Kinsella proposes constructing a bioequivalence test using this interval. But Mandallaz and Mau (1981), Yee (1986, 1990), Metzler (1991) and Schuirmann (1989) all propose defining a test of (1) using these Fieller confidence intervals, and all suggest that a $100(1-2\alpha)\%$ confidence interval should be used. A test defined in this way using the Locke $100(1-2\alpha)\%$ confidence interval is, in fact, a size-α test because the Locke interval is equal-tailed. However, Metzler (1991) and Schuirmann (1989) give graphs of the power function of the Mandallaz and Mau (1981) test that show that the test has size greater than the specified α. For example, Figures 3 through 9 in Metzler (1991) are graphs of $1-$ (power function) based on the Mandallaz and Mau (1981) confidence interval. At $\delta_U = 1.2$, the rejection probability is about 0.07 for the $\alpha = 0.05$ test, and the power is about 0.15 for the $\alpha = .10$ test. These figures cover a variety of sample sizes and variances, but in all cases the rejection probability exceeds the nominal α at $\delta_U = 1.2$. The same liberality of the Mandallaz and Mau test is illustrated by Figures 3–13 of Schuirmann (1989).

On the other hand, a test defined in terms of a $100(1-2\alpha)\%$ confidence set might be very conservative. An example is the test proposed by Chow and Shao (1990) for testing (1) about the ratio μ_T/μ_R. Specifically, Chow and Shao considered a two-period crossover design with no carry-over, period or sequence effects. Let $\overline{\mathbf{X}}$ denote the sample mean vector with mean $\boldsymbol{\mu} = (\mu_T, \mu_R)'$ and let S denote the sum of cross-products matrix. Let m patients receive the first sequence, let n patients receive the second sequence and let $n^* = n + m$. Then, $C = \{\boldsymbol{\mu}\colon T_1 \le F_{\alpha, 2, n^*-2}\}$ defines a $100(1-\alpha)\%$ confidence ellipse for $\boldsymbol{\mu}$, where $T_1 = n^*(n^*-2)(\overline{\mathbf{X}}-\boldsymbol{\mu})'S^{-1}(\overline{\mathbf{X}}-\boldsymbol{\mu})/2$ and $F_{\alpha, 2, n^*-2}$ is the upper 100α percentile of an F-distribution with 2 and n^*-2 degrees of freedom. Chow and Shao propose rejecting H_0 in (1) and concluding $H_a\colon \delta_L < \mu_T/\mu_R < \delta_U$ is true if and only if the 90% confidence ellipse is contained in the cone defined by H_a. They do not comment on the actual size of this test, but we assume 90% was chosen to be $100(1-2\alpha)\%$, where $\alpha = 0.05$.

Chow and Shao's test can be described much more simply by recalling the relationship between the confidence ellipse, C, and simultaneous confidence intervals for all linear functions $l'\boldsymbol{\mu}$ (Scheffé, 1959). $\boldsymbol{\mu} \in C$ if and only if

$$l'\overline{\mathbf{X}} - \sqrt{2F_{\alpha, 2, n^*-2}\, l'Sl/[n^*(n^*-2)]} \le l'\boldsymbol{\mu} \le l'\overline{\mathbf{X}} + \sqrt{2F_{\alpha, 2, n^*-2}\, l'Sl/[n^*(n^*-2)]}$$

for every vector l. In fact, the only two vectors needed to define Chow and Shao's test are $l_L = (1, -\delta_L)'$ and $l_U = (1, -\delta_U)'$. The hypotheses in (1) or (14) can be written as $H_0\colon l_L'\boldsymbol{\mu} \le 0$ or $l_U'\boldsymbol{\mu} \ge 0$ and $H_a\colon l_U'\boldsymbol{\mu} < 0 < l_L'\boldsymbol{\mu}$. Furthermore, the ellipse C is below the line $l_U'\boldsymbol{\mu} = 0$ if and only if $l_U'\overline{\mathbf{X}} + \sqrt{2F_{\alpha, 2, n^*-2}\, l_U'Sl_U/[n^*(n^*-2)]} < 0$, that is, the upper endpoint of the confidence interval for $l_U'\boldsymbol{\mu}$ is negative. Similarly, the ellipse C is above the line $l_L'\boldsymbol{\mu} = 0$ if and only if $l_L'\overline{\mathbf{X}} - \sqrt{2F_{\alpha, 2, n^*-2}\, l_L'Sl_L/[n^*(n^*-2)]} > 0$. If we define

$$T_L = \frac{l_L'\overline{\mathbf{X}}}{\sqrt{l_L'Sl_L/[n^*(n^*-2)]}}$$

and

$$T_U = \frac{l_U'\overline{\mathbf{X}}}{\sqrt{l_U'Sl_U/[n^*(n^*-2)]}},$$

then Chow and Shao's test rejects H_0 if and only if

$$T_L > \sqrt{2F_{\alpha, 2, n^*-2}} \quad \text{and} \quad T_U < -\sqrt{2F_{\alpha, 2, n^*-2}}. \tag{22}$$

This simple description of Chow and Shao's test has not appeared before. In this form, it is apparent that this test can be viewed as an IUT. A reasonable test of $H_{0L}\colon l_L'\mu \le 0$ versus $H_{aL}\colon l_L'\mu > 0$ is the test that rejects H_{0L} if $T_L > \sqrt{2F_{\alpha, 2, n^*-2}}$. A reasonable test of $H_{0U}\colon l_U'\boldsymbol{\mu} \ge 0$ versus $H_{aU}\colon l_U'\boldsymbol{\mu} < 0$ is the test that rejects H_{0U} if $T_U < -\sqrt{2F_{\alpha, 2, n^*-2}}$. Thus, Chow and Shao's test is the IUT of H_0 versus H_a formed by combining these two tests. Theorems 1 and 2 then tell us that the actual size of this test is $\alpha' = P(T > \sqrt{2F_{\alpha, 2, n^*-2}})$, where T has a Student's t-distribution with n^*-1 degrees of freedom. This is because T_L has this t-distribution if

$l'_L\boldsymbol{\mu} = 0$, and T_U has this t-distribution if $l'_U\boldsymbol{\mu} = 0$. That is, α' is the size of each of the two individual tests. We computed α' using a 90% confidence ellipse as suggested by Chow and Shao. We found that $\alpha' = 0.017$ for $m = n = 5$, 10 and 15, and $\alpha' = 0.016$ for $m = n = 20$, 30 and ∞. Thus, if the intent of using a $100(1-2\alpha)\% = 90\%$ confidence ellipse was to produce a bioequivalence test with Type I error probability of $\alpha = 0.05$, the result was very conservative.

A test of H_0 versus H_a with the desired size of α can be obtained by replacing $\sqrt{2F_{\alpha,2,n^*-2}}$ with the t-percentile, t_{α,n^*-1} in (22). Then each of the individual tests is size-α and the combined IUT also has size α. This test is uniformly more powerful than Chow and Shao's test because the rejection region of Chow and Shao's test is a proper subset of this one. This test is the analogue of the TOST for this crossover model. In fact, Yang (1991) proposed this test for this problem as an alternative to Chow and Shao's test, but Yang did not state that this test was uniformly more powerful nor quantify the conservativeness of Chow and Shao's test.

Our conclusions from the results and examples in this subsection are simple. The usage of $100(1-2\alpha)\%$ confidence sets to define bioequivalence tests should be abandoned. This practice produces tests with the appropriate size only when special, "equal-tailed" confidence intervals are used and offers no intuitive insight. The mixture of $100(1-2\alpha)\%$ confidence sets and size-α tests is only confusing. Rather, a test with the specified Type I error probability of α should be used. The IUT method can usually be used to construct such a test. Then Theorem 4 might be used to construct the corresponding $100(1-\alpha)\%$ confidence set.

6. MULTIPARAMETER EQUIVALENCE PROBLEMS

Until now, we have discussed bioequivalence testing in terms of only one parameter. In this section, we discuss two problems in which the desired inference is equivalence in terms of two parameters. These results immediately generalize to situations in which bioequivalence is defined in terms of more than two parameters.

These two examples have been discussed as multiparameter bioequivalence problems by several authors, but, in some cases, the tests that have been proposed do not have the correct size α. The proposed tests do not properly account for the multiple-testing aspect of this problem. These two multiparameter examples vividly illustrate that the IUT method can provide a simple mechanism for constructing tests with the correct size α, even in seemingly complicated bioequivalence problems. Size-α tests can be combined to obtain an overall size-α test. No adjustment for multiple testing is needed if the IUT method is used.

6.1 Simultaneous AUC and C_{max} Bioequivalence

Sections 4 and 5 discussed bioequivalence testing in terms of only one parameter. That is, the test and reference drugs are to be compared with respect to either average AUC or average $C_{\max}$. FDA (1992a) and EC-GCP (1993) consider two drugs to be bioequivalent only if they are similar in both parameters. Westlake (1988) and Hauck et al. (1995) have considered the problem of comparing AUC and $C_{\max}$ simultaneously. (Westlake actually compares three parameters, including $T_{\max}$ also, but this does not conform to current FDA guidelines.)

Assume the measurements are lognormal so that, after log transformation, we wish to consider hypotheses like (2). Let the superscripts A and C refer to the variables AUC and $C_{\max}$, respectively. For example, η_R^C denotes the mean of $\log(C_{\max})$ for the reference drug. The test and reference drugs are to be considered bioequivalent only if

$$H_a^m: \quad \theta_L < \eta_T^A - \eta_R^A < \theta_U \quad \text{and} \quad \theta_L < \eta_T^C - \eta_R^C < \theta_U. \tag{23}$$

Using current FDA guidelines, $\theta_U = \log(1.25) = -\log(0.80) = -\theta_L$. If one variable is deemed more important than another, the limits could be different for the different variables. For example, if AUC was considered more important than $C_{\max}$, then the limits θ_L^A and θ_U^A for AUC could be chosen to be narrower than the limits θ_L^C and θ_U^C for $C_{\max}$, as they are in Europe.

The statement H_a^m in (23) should be the alternative hypothesis in this multivariate bioequivalence test. The null hypothesis, H_0^m should be the negation of H_a^m. That is, H_0^m states that one or more of the four inequalities in H_a^m is false. Westlake proposed testing H_0^m versus H_a^m by doing two separate tests, one for each variable. Specifically, he proposed using the TOST to test (2) for each variable. The drugs will be declared bioequivalent only if each of the tests rejects its hypothesis. Furthermore, Westlake said a Bonferroni correction should be used, and each TOST should be performed at the $\alpha/2$ level to account for the multiple testing. (Westlake actually said $\alpha/3$ because he was considering three tests.)

Westlake's procedure is conservative. The size of Westlake's test is $\alpha/2$, not α. This is true because, although he did not use this terminology, he has

proposed an IUT. The alternative H_a^m is the intersection of two statements, one about each variable. Computing two separate TOST's and concluding that H_a^m is true only if both TOST's reject is an IUT. By Theorem 1, this test has level $\alpha/2$ if each TOST is performed at level $\alpha/2$. In fact, Theorem 2 can be used to show that this test has size equal to $\alpha/2$.

Therefore, to test H_0^m versus H_a^m, Westlake's procedure can be used except that each of the two TOSTs should be performed at size α. The resulting test has probability at most α of declaring the drugs to be bioequivalent if they are bioinequivalent.

Hauck et al. (1995) propose testing (23) using two size-α TOST's. They recognize that the Bonferroni adjustment recommended by Westlake is unnecessary, but they come to the opposite conclusion. Based on a simulation study, they conclude that this test is too conservative and suggest that the two TOST's might be performed using a higher error rate than α, and the resulting test of (23) would be size-α. (They admit that more simulations are needed to confirm this conjecture.) However, if the two TOST's are each size-α, then the test of (23) is exactly size-α. To see this, use Theorem 2 by setting $\theta_L = \eta_T^A - \eta_R^A$, $\eta_T^C = \eta_R^C$ and considering the limit as $\sigma_{D^A} \to 0$ and $\sigma_{D^C} \to 0$. Here, D^A and D^C are the estimates of $\eta_T^A - \eta_R^A$ and $\eta_T^C - \eta_R^C$, respectively. In this limit, three of the four one-sided tests will have rejection probability converging to 1, because these parameter points are in the alternative hypothesis and the corresponding standard deviations are converging to 0. The fourth one-sided test will have rejection probability exactly equal to α, for all such parameter points, because $\theta_L = \eta_T^A - \eta_R^A$ is on the boundary.

A test that is uniformly more powerful but still has size α will be obtained if the test we propose in Section 4.2 is used to perform the two tests, rather than using the two TOST's. Again, both of these tests would be performed at size α.

An alternative way of assessing the simultaneous bioequivalence of AUC and $C_{\max}$ is to inspect the Brown, Casella and Hwang (1995) confidence set (20), generalized to the Σ unknown case. Suppose $(X_i^A, X_i^C)', (Y_i^A, Y_i^C)', i = 1, \ldots, n$, are log-transformed i.i.d. observations on AUC and $C_{\max}$ under the test and reference drugs, respectively. Let $\mathbf{Z}_i = (X_i^A, X_i^C)' - (Y_i^A, Y_i^C)'$, $i = 1, \ldots, n$, which are assumed to be multivariate normal with mean $\boldsymbol{\theta} = (\eta_T^A - \eta_R^A, \eta_T^C - \eta_R^C)'$ and unknown variance–covariance matrix Σ. Let $\widehat{\boldsymbol{\theta}} = (\overline{Z}^A, \overline{Z}^C)'$ and $\widehat{\Sigma}$ be the sample mean vector and variance–covariance matrix of the $\mathbf{Z}_i$'s. Then $\boldsymbol{\theta}'\widehat{\boldsymbol{\theta}}$ is univariate normal with mean $\boldsymbol{\theta}'\boldsymbol{\theta}$ and variance $\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}/n$, while $(n-1)\boldsymbol{\theta}'\widehat{\Sigma}\boldsymbol{\theta}/\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}$ is independent of $\boldsymbol{\theta}'\widehat{\boldsymbol{\theta}}$ and has a χ^2 distribution with $n-1$ degrees of freedom. Thus, a size-α test for H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is obtained using the acceptance region

$$A(\boldsymbol{\theta}_0) = \left\{(\widehat{\boldsymbol{\theta}}, \widehat{\Sigma}): \frac{\boldsymbol{\theta}_0'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0'\widehat{\Sigma}\boldsymbol{\theta}_0/n}} > -t_{\alpha, n-1}\right\},$$

which leads to the confidence region

$$C(\widehat{\boldsymbol{\theta}}, \widehat{\Sigma}) = \left\{\boldsymbol{\theta}: \frac{\boldsymbol{\theta}'\widehat{\boldsymbol{\theta}}}{\sqrt{\boldsymbol{\theta}'\widehat{\Sigma}\boldsymbol{\theta}/n}} + t_{\alpha, n-1} > \frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{\sqrt{\boldsymbol{\theta}'\widehat{\Sigma}\boldsymbol{\theta}/n}}\right\}. \quad (24)$$

Brown, Casella and Hwang (1995) applied (20) to the simultaneous AUC and $C_{\max}$ problem for illustration, assuming Σ is known. In practice, this assumption is perhaps unrealistic considering the moderate sample size typical in bioequivalence trials.

6.2 Mean and Variance Bioequivalence

Anderson and Hauck (1990) and Liu and Chow (1992a) discuss another type of multiparameter bioequivalence. They point out that bioequivalence should not be defined only in terms of the mean responses for the two drugs. Rather, the variances of the responses of the two drugs should also be considered. If two drugs have bioequivalent means but different variances, the drug with the smaller variance might be preferred. This kind of multiparameter bioequivalence is often called population bioequivalence.

Consider a single variable, for example, AUC. Let η_T and η_R denote the means of log(AUC). Let σ_T^2 and σ_R^2 denote the intrasubject variances of the test and reference drugs, respectively. The test and reference drugs will be considered bioequivalent only if η_T and η_R are similar and σ_T^2 and σ_R^2 are similar. To demonstrate bioequivalence, we wish to test

$$H_0^m: \begin{array}{l} \eta_T - \eta_R \le \theta_L \quad \text{or} \quad \eta_T - \eta_R \ge \theta_U \\ \text{or} \\ \sigma_T^2/\sigma_R^2 \le \kappa_L \quad \text{or} \quad \sigma_T^2/\sigma_R^2 \ge \kappa_U \end{array}$$

(25) versus

$$H_a^m: \begin{array}{l} \theta_L < \eta_T - \eta_R < \theta_U \\ \text{and} \quad \kappa_L < \sigma_T^2/\sigma_R^2 < \kappa_U. \end{array}$$

The constants θ_L, θ_U, κ_L and κ_U would be chosen to define clinically important differences.

Liu and Chow (1992a) propose a size-α test of

$$H_0^\sigma: \quad \sigma_T^2/\sigma_R^2 \le \kappa_L \quad \text{or} \quad \sigma_T^2/\sigma_R^2 \ge \kappa_U$$

versus

$$H_a^\sigma: \quad \kappa_L < \sigma_T^2/\sigma_R^2 < \kappa_U.$$

Their test is an IUT composed of two size-α tests, one for testing each inequality. Wang (1994) describe an unbiased, size-α test that is uniformly more powerful than the Liu and Chow test.

The hypotheses

$$H_0^\eta: \quad \eta_T - \eta_R \le \theta_L \quad \text{or} \quad \eta_T - \eta_R \ge \theta_U$$

versus

$$H_a^\eta: \quad \theta_L < \eta_T - \eta_R < \theta_U$$

can be tested with a TOST. Because H_a^m is the intersection of H_a^η and H_a^σ, the IUT method can be used to construct a test of H_0^m versus H_a^m. The test that rejects H_0^m only if the size-α Liu and Chow test rejects H_0^σ and the size-α TOST rejects H_0^η is a size-α test of H_0^m versus H_a^m.

Liu and Chow, however, propose a more conservative combination of these two tests. Let α denote the desired size of the test of H_0^m. Let α_1 denote the size of the TOST and let α_2 denote the size of the Liu and Chow test. They say to choose α_1 and α_2 so that

$$\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2). \tag{26}$$

Liu and Chow note that the test statistics use for the TOST are independent of the test statistics used in their test, but they give no further explanation of (26). The probability that H_0^η is accepted, given that H_0^η is true, is bounded below by $1 - \alpha_1$. The probability that H_0^σ is accepted, given that H_0^σ is true, is bounded below by $1 - \alpha_2$. So the quantity α in (26) is an upper bound for the probability that at least one of the two tests rejects its null hypothesis, given that both H_0^η and H_0^σ are true. This is not the error probability of the proposed test. The error probability is the probability the both tests reject, given that either H_0^η or H_0^σ is true. Hypothesis H_0^m is the union of H_0^η and H_0^σ, not the intersection.

Again, it should be noted that a more powerful size-α test of H_0^m will be obtained if the test from Section 4.2, rather than the TOST, is used to test H_0^η and Wang's (1994) test is used to test H_0^σ.

7. CONCLUDING REMARKS

We have shown that the theory of intersection–union tests is central to bioequivalence studies. We have demonstrated the danger of incorrect association of confidence sets with such tests. Due to the traditional emphasis on *significant difference* inference in statistics, many *practical equivalence* problems have not been recognized as such, we believe. It is our hope (and anticipation) that the concepts and techniques discussed in this article will, in time, prove to be useful not only in bioequivalence studies, but in other practical equivalence problems as well.

APPENDIX
DETAILS OF NEW TEST IN SECTION 4.2

A size-α, nearly unbiased test for (2) was described geometrically in Section 4.2. In Section A.1, formulas and computational suggestions are given for the quantities that define that test. The construction in Section 4.2 is valid for $\alpha > \alpha_*$. In Section A.2 a similar construction yields a size-α, nearly unbiased test for $\alpha \le \alpha_*$. Brown, Hwang and Munk did not propose any test for $\alpha \le \alpha_*$.

A.1 Formulas for Section 4.2

Define functional notation for the transformation from rectangular to polar coordinates by

$$v(d, s_*) = \sqrt{(d - \Delta)^2 + s_*^2},$$

$$b(d, s_*) = \cos^{-1}((d - \Delta)/v(d, s_*)),$$

for $-\infty < d < \infty$ and $s_* \ge 0$. The inverse transformation is

$$d(v, b) = \Delta + v\cos(b),$$

$$s_*(v, b) = v\sin(b),$$

for $v \ge 0$ and $0 \le b \le \pi$. The point $(d, s_*) = (0, \Delta\sqrt{r}/t_{\alpha, r})$ is the vertex of the triangular region R_T. Therefore,

$$b_0 = b(0, \Delta\sqrt{r}/t_{\alpha, r}),$$
$$v_0 = 2\Delta\sin(\pi - b_0),$$
$$(d_1, s_{*1}) = (d(v_0, b_0), s_*(v_0, b_0)),$$
$$v_1 = v(-d_1, s_{*1}).$$

The line of length v_0 in Figure 2 has $b = 3\pi/2 - b_0$. Therefore, The angle b_{L1}, defined by (12), is easily found by a numeric root-finding method such as bisection.

Finally, for any point (d, s_*) on $\{V = v\}$, $s_* = \sqrt{v^2 - (d - \Delta)^2}$. For any point (d_u, s_{*u}) on $\{V = v\}$ with $d_u \le 0$, there is a unique point (d_l, s_{*l}) on $\{V = v\}$ with $d_l \ge 0$ such that the line joining (d_l, s_{*l}) and $(0, 0)$ and the s_*-axis form the same angle as the line joining (d_u, s_{*u}) and $(0, 0)$ and the s_*-axis. This point satisfies

$$\frac{d_u}{\sqrt{v^2 - (d_u - \Delta)^2}} = -\frac{d_l}{\sqrt{v^2 - (d_l - \Delta)^2}},$$

which has the solution

$$d_l = \frac{d_u(v^2 - \Delta^2)}{v^2 + 2d_u\Delta - \Delta^2}. \tag{27}$$

Using this expression for d_l in terms of d_u, the equation

$$F(b(d_u, s_u)) - F(b(d_l, s_l)) = \alpha - \alpha(v)$$

is a function of the single variable d_u. The unique solution to this equation, in the interval $\Delta - v \le d_u \le 0$ is easily found by a numeric root-finding method such as bisection. Call the solution d_U. Define d_L by (27) using $d_u = d_U$. The angles $b_U(v)$ and $b_L(v)$ are

$$b_U(v) = b\Big(d_U, \sqrt{v^2 - (d_U - \Delta)^2}\Big),$$

$$b_L(v) = b\Big(d_L, \sqrt{v^2 - (d_L - \Delta)^2}\Big).$$

A.2 New Test for $\alpha \le \alpha_*$

For small values of $\alpha \le \alpha_*$, a size-α, nearly unbiased test of (2) that is uniformly more powerful than the TOST can be constructed. The construction is very similar to and somewhat simpler than the construction in Section 4.2. The notation of Section A.1 will be used, and Figure 5 illustrates the construction.

For $\alpha \le \alpha_*$, the point on l_L closest to $(\Delta, 0)$ is the vertex of R_T, $(d_0, s_{*0}) = (0, \Delta\sqrt{r}/t_{\alpha, r})$. Let $v_0 = v(d_0, s_{*0})$. For $v \le v_0$, $R_2(v) = \{b\colon b_0 < b < \pi\}$, exactly the points in the TOST. The arc A_0 is such an arc. For $v_0 < v < 2\Delta$, $R_2(v)$ consists of two arcs;

$$R_2(v) = \{b\colon b_L(v) < b < b_U(v)\} \cup \{b\colon b_2 < b < \pi\};$$

$b_L(v)$, $b_U(v)$ and b_2 are defined as before. The two solid pieces of arc A_1 are examples of these arcs. The semicircle $\{V = v\}$ does not intersect R_T near the s_*-axis so there is no need to check that $\{b\colon b_L(v) < b < b_U(v)\}$ covers all the TOST. For $v \ge 2\Delta$, $R_2(v) = \{b\colon b_L(v) < b < b_U(v)\}$. The solid piece of arc A_3 is such an arc. In Figure 5, R_2 is outlined with a solid line, R_1 is outlined with a dashed line, and the intersection is the rejection region of the IUT.

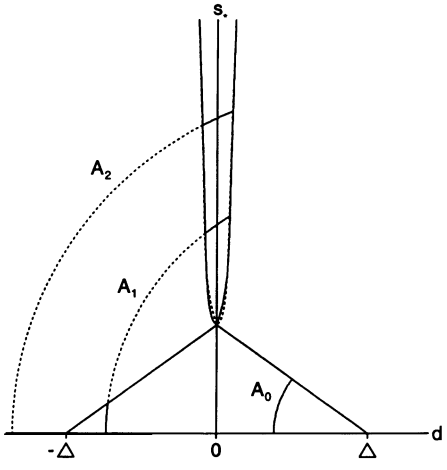

FIG. 5. *Rejection region of new test for $\alpha \le \alpha_*$: region R_2 (between solid lines) and region R_1 (between dashed lines); rejection region $R = R_1 \cap R_2$; $r = 3$ and $\alpha = 0.05$.*

ACKNOWLEDGMENTS

We thank Dr. Hans Frick and Dr. Volker Rahlfs for references on European bioequivalence guidelines.

REFERENCES

ANDERSON, S. (1993). Individual bioequivalence: a problem of switchability (with discussion). *Biopharmaceutical Report* **2** 1–11.

ANDERSON, S. and HAUCK, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Comm. Statist. Theory Methods* **12** 2663–2692.

ANDERSON, S. and HAUCK, W. W. (1985). Letter to the Editor. *Biometrics* **41** 561–563.

ANDERSON, S. and HAUCK, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **18** 259–273.

BERGER, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24** 295–300.

BERGER, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *J. Amer. Statist. Assoc.* **84** 192–199.

BOFINGER, E. (1985). Expanded confidence intervals. *Comm. Statist. Theory Methods* **14** 1849–1864.

BOFINGER, E. (1992). Expanded confidence intervals, one-sided tests, and equivalence testing. *Journal of Biopharmaceutical Statistics* **2** 181–188.

BROWN, L. D., CASELLA, G. and HWANG, J. T. G. (1995a). Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *J. Amer. Statist. Assoc.* **90** 880–889.

BROWN, L. D., HWANG, J. T. G. and MUNK, A. (1995b). An unbiased test for the bioequivalence problem. Technical report, Cornell Univ.

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference*. Duxbury, Belmont, CA.

CHOW, S.-C. and LIU, J.-P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Dekker, New York.

CHOW, S.-C. and SHAO, J. (1990). An alternative approach for the assessment of bioequivalence between two formulations of a drug. *Biometrical J.* **32** 969–976.

DILETTI, E., HAUSCHKE, D. and STEINIJANS, V. W. (1991). Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology* **29** 1–8.

EC-GCP (1993). *Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorization for Medical Products.* CPMP Working Party on Efficacy of Medical Products, Commission of the European Communities, Brussels. (Draft guideline edition.)

FDA (1987). *Guideline for Submitting Documentation for Stability Studies of Human Drugs and Biologics.* Center for Drugs and Biologics, Food and Drug Administration, Rockville, MD.

FDA (1992a). Bioavailability and bioequivalence requirements. In *U.S. Code of Federal Regulations* **21**, Chapter 320. U.S. Government Printing Office, Washington, D.C.

FIELLER, E. (1954). Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B* **16** 175–185.

FIELLER, E. C. (1940). The biological standardisation of insulin. *J. Roy. Statist. Soc. Supplement* **7** 1–64.

HAUCK, W. W. and ANDERSON, S. (1992). Types of bioequivalence and related statistical considerations. *International Journal of Clinical Pharmacology, Therapy and Toxicology* **30** 181–187.

HAUCK, W. W., HYSLOP, T., ANDERSON, S., BOIS, F. Y. and TOZER, T. N. (1995). Statistical and regulatory considerations for multiple measures in bioequivalence testing. *Clinical Research and Regulatory Affairs* **12** 249–265.

HAUSCHKE, D., STEINIJANS, V. W. and DILETTI E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies. *International Journal of Clinical Pharmacology, Therapy and Toxicology* **28** 72–78.

HAYTER, A. J. and HSU, J. C. (1994). On the relationship between stepwise decision procedures and confidence sets. *J. Amer. Statist. Assoc.* **89** 128–136.

HSU, J. C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *Ann. Statist.* **12** 1136–1144.

HSU, J. C. (1996). *Multiple Comparisons*. Chapman and Hall, London.

HSU, J. C., HWANG, J. T. G., LIU, H.-K. and RUBERG, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika* **81** 103–114.

JUSKEVICH, J. C. and GUYER, C. G. (1990). Bovine growth hormone: human food safety evaluation. *Science* **249** 875–884.

KINSELLA, A. (1989). Bootstrapping a bioequivalence measure. *The Statistician* **38** 175–179.

LEHMANN, E. L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

LEHMANN, E. L. (1986). *Testing Statistical Hypothesis*, 2nd ed. Wiley, New York.

LIU, H. and BERGER, R. L. (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Ann. Statist.* **23** 55–72.

LIU, J.-P. and CHOW, S.-C. (1992). On the assessment of variability in bioavailability/bioequivalence studies. *Comm. Statist. Theory Methods* **21** 2591–2607.

LIU, J.-P. and WENG, C.-S. (1995). Bias of two one-sided tests procedures in assessment of bioequivalence. *Statistics in Medicine* **14** 853–861.

LOCKE, C. S. (1984). An exact confidence interval from untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics* **12** 649–655.

MANDALLAZ, D. and MAU, J. (1981). Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **20** 213–222.

MARTIN ANDRÉS, A. (1990). On testing for bioequivalence. *Biometrical J.* **32** 125–126.

METZLER, C. M. (1991). Sample sizes for bioequivalence studies. *Statistics in Medicine* **10** 961–970.

MÜLLER-COHRS, J. (1991). An improvement of the Westlake symmetric confidence interval. *Biometrical J.* **33** 357–360.

MUNK, A. (1993). An improvement on commonly used tests in bioequivalence assessment. *Biometrics* **49** 1225–1230.

PATEL, H. I. and GUPTA, G. D. (1984). A problem of equivalence in clinical trials. *Biometrical J.* **26** 471–474.

PRATT, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 541–567.

ROCKE, D. M. (1984). On testing for bioequivalence. *Biometrics* **40** 225–230.

RUBERG, S. J. and HSU, J. C. (1992). Multiple comparison procedures for pooling batches in stability studies. *Technometrics* **34** 465–472.

SASABUCHI, S. (1980). A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika* **67** 429–439.

SASABUCHI, S. (1988a). A multivariate one-sided test with composite hypotheses when the covariance matrix is completely unknown. *Mem. Fac. Sci. Kyushu Univ. Ser. A* **42** 37–46.

SASABUCHI, S. (1988b). A multivariate test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Mem. Fac. Sci. Kyushu Univ. Ser. A* **42** 9–19.

SCHALL, R. and LUSS, G. H. (1993). On population and individual bioequivalence. *Statistics in Medicine* **12** 1109–1124.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

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

SCHUIRMANN, D. J. (1989). Confidence intervals for the ratio of two means from a crossover study. In *Proceedings of the Biopharmaceutical Section* 121–126. Amer. Statist. Assoc., Alexandria, VA.

SCHUIRMANN, D. L. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* **37** 617.

SHEINER, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine* **11** 1777–1788.

STEFANSSON, G., KIM, W. C. and HSU, J. C. (1988). On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2** 89–104. Springer, New York.

WANG, W. (1994). Optimal unbiased tests for bioequivalence in variability. Technical report, Cornell Univ.

WESTLAKE, W. J. (1973). The design and analysis of comparative blood-level trials. In *Current Concepts in the Pharmaceutical Sciences, Dosage Form Design and Bioavailability* (J. Swarbrick, ed.) 149–179. Lea and Febiger, Philadelphia.

WESTLAKE, W. J. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics* **32** 741–744.

WESTLAKE, W. J. (1981). Response to T.B.L. Kirkwood: bioequivalence testing—a need to rethink. *Biometrics* **37** 589–594.

WESTLAKE, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In *Biopharmaceutical Statistics for Drug Development* (Karl E. Peace, ed.) 329–352. Dekker, New York.

YANG, H.-M. (1991). An extended two one-sided tests procedure. In *Proceedings of the Biopharmaceutical Section* 157–162. Amer. Statist. Assoc., Alexandria, VA.

YEE, K. F. (1986). The calculation of probabilities in rejecting bioequivalence. *Biometrics* **42** 961–965.

YEE, K. F. (1990). Correspondence to the Editor. *The Statistician* **39** 465–466.

Comment

Walter W. Hauck and Sharon Anderson

We commend Berger and Hsu for this fine review. This paper does a very nice job of presenting the intersection–union principle and demonstrating how it can be used to develop tests and confidence intervals for various equivalence hypotheses. It will be a valuable reference for our work and that of others.

Note that we said "equivalence hypotheses," not bioequivalence hypotheses, in the above paragraph. It would be very unfortunate if the authors's decision to bill this as a bioequivalence paper kept it from being a major reference for general equivalence problems. It seems to us that the real value of this paper is not in the bioequivalence area. We have a variety of reasons for this.

First, established practice for average bioequivalence (two one-sided tests, TOST) is easily understood by the nonstatisticians who do much of the analysis and almost all the interpretation of that analysis. Since the power advantage of the almost unbiased test proposed here, as well as that of the test proposed by Brown and colleagues (Brown, Casella and Hwang, 1995; Brown, Hwang and Munk, 1995) is minimal in practical cases, there is little rationale for changing practice. That is, when the TOST approach for the hypotheses (2) yields reasonable power, there is essentially no advantage to the other procedures (for studies designed to have at least 80% power, for example). This is evident in Table 1. While the bias and power loss of the TOST are certainly real, they occur for larger variabilities than practically encountered. As the authors note, however, the power advantage could be helpful if the variability assumed for study design was lower than that obtained in the study. We do recognize that there is an obvious argument that the better (more powerful) test should be used. However, the little power gain and the need to truncate the rejection region make it a very difficult "sell."

Second, in the United States, at least, there is a movement away from average bioequivalence to individual bioequivalence. Most individual bioequivalence criteria, and specifically the one recommended by the FDA Working Group on Individual Bioequivalence (August 1996 meeting of the FDA Advisory Committee on Pharmaceutical Sciences), are aggregate and use one-sided criteria. By "aggregate" is meant that all components (formulation means, subject-by-formulation interaction and within-subject variances) are included in a single measure of inequivalence. Since there will be a single one-tailed criterion (and test), instead of the interval equivalence hypothesis for average bioequivalence, tests for these individual bioequivalence criteria do not require the intersection–union principle. References to these approaches are largely in the pharmacology and biostatistics literatures.

Third, and most important, we think equivalence testing approaches are underutilized. We often see examples where statisticians and non-statisticians are testing the wrong hypotheses, apparently stuck in a mode of thinking based on null hypotheses of no-difference. For example, one sees tests of the null hypotheses of no interaction or of equal variances when what is needed are tests of alternative hypotheses of negligible interaction and of similar variances. The authors cite Lehmann on the principles of hypothesis testing, but could also have cited Fisher (1935). It is our hope that this paper will help to stimulate better practice in this area by providing some principles for approaching the problems.

A related concern is confidence intervals and reporting so-called negative studies (i.e., studies that do not attain statistical significance). The increased emphasis on confidence intervals in recent years has helped (e.g., Simon, 1986, and Braitman, 1991), but a very common error in the clinical literature remains the equating of the lack of statistical significance with "no difference." We raise this here, since interpreting negative studies depends, at least implicitly, on equivalence notions. There is clearly some challenge to developing confidence intervals that correspond to proper tests of interval equivalence hypotheses.

Walter W. Hauck is Professor of Medicine and Head, Biostatistics Section, Division of Clinical Pharmacology, Thomas Jefferson University, 125 South Ninth Street, #403, Philadelphia, Pennsylvania 19107 (e-mail: w_hauck@hendrix.jci.tju.edu). Sharon Anderson is Director of Biostatistics and Data Management, Bristol-Myers Squibb, Princeton, New Jersey 08543-4000 (e-mail: anderson_s@bms.com), and Adjunct Associate Professor of Medicine, Thomas Jefferson University.

Comment

Michael P. Meredith and Mark A. Heise

Professors Berger and Hsu are to be congratulated for making significant contributions to the areas of hypothesis testing and confidence set estimation in bioequivalence. Their pedagogically lucid paper illustrates the statistical shortcomings of quite a few methods that have been proposed over the past 15 years in the bioequivalence literature, and then they proceed to give proper or improved solutions to these problems. In addition, the authors give a review of simultaneous AUC and $C_{\max}$ bioequivalence testing, and mean and variance bioequivalence. These are both important areas that have received inadequate statistical attention. This should prove to be an important paper for those who work in bioequivalence trials and related hypothesis testing since most of the major issues appear in this paper.

Considerable effort is devoted to developing a nearly unbiased size-α test for bioequivalence that is uniformly more powerful than the two one-sided tests (TOST) procedure. This component is along the lines of several recent papers on bioequivalence testing (Brown, Hwang and Munk, 1995; Brown, Casella and Hwang, 1995; Hsu et al., 1994). In consolidating many of the erroneous methods that have been promulgated in journals or proceedings that do not generally receive statistical peer review (as well as in some statistical journals) the authors appear to have sifted carefully through a large and varied morass of literature as reflected in their citations. The authors also point to the overlooked work of Sasabuchi (1980, 1988a, b) that supports derivation of proper bioequivalence tests for ratios of parameters.

It is worthwhile to recognize that the methods developed for the rather narrow bioequivalence focus are applicable to numerous other areas in clinical trial research. Many later-phase clinical trials are being conducted to demonstrate clinical or therapeutic equivalence directly, with rather weak statistical guidance. Anti-infectives are a common area for this kind of equivalence trial where the objective is to demonstrate that a new compound is at-least-as-good-as an existing drug that may be less useful due to evolving resistant strains of target pathogens. More generally, the TOST procedure or the herein proposed test can be applicable in positive control studies where it is unethical to include a placebo control arm. In pharmacoeconomics, managed-care providers want to mandate the lowest cost therapy that is no worse than other available therapies. This provides another important motivation to demonstrate therapeutic equivalence. The methodology described in this paper can be adapted to handle clinical endpoints that are often dichotomous or ordinal categorical.

The authors' results are also important in demonstrating clinical equivalence of a variety of products that are not suitable for traditional demonstrations of bioequivalence. Examples include formulations that are applied topically (such as corticosteroidal or analgesic ointments), are not ingested (such as therapeutic mouth rinses) or are nonsystemically available (e.g., those acting only within the gastrointestinal tract and not absorbed) as in some laxatives and antidiarrheals. In vitro assays are sometimes substituted for actual clinical use to confirm a new formulation's comparability to the original—this practice also requires similar statistical guidance for testing equivalence. Observations from many of these clinical endpoints or in vitro assays are not lognormally distributed, and results from the section on bioequivalence tests for ratios of parameters should prove useful.

Testing for bioequivalence is regarded generally as testing AUC and $C_{\max}$ for acceptable equivalence; however, this is not always the case, as described above and as follows. Sometimes the primary concern is showing that a test drug's $C_{\max}$ is 125% or *less than* that of the reference drug. For many drugs this is sensible guidance from the safety perspective. Further, there is a motivation to allow wider equivalence limits on $C_{\max}$ due to its larger variability than AUC, and the authors cover this situation in Section 6.1. Finally, there are drug formulations whose $C_{\max}$ may be of no practical importance—these are primarily extended release formulations or transdermal patch delivery systems that have a

Michael P. Meredith is Research Fellow, Biometrics and Statistical Sciences Department, Procter and Gamble Company, Cincinnati, Ohio 45242, and Adjunct Associate Professor of Biological Statistics, Biometrics Unit, Cornell University, Ithaca, New York 14853. Mark A. Heise is Statistical Scientist, Biometrics and Statistical Sciences Department, Procter and Gamble Pharmaceuticals, Cincinnati, Ohio 45242 (e-mail: heisem@pg.com).

very flat drug concentration versus time response profile. The drug's AUC becomes the primary focus for bioequivalence assessment.

Note that the measures AUC and $C_{\max}$ are derived simply by "connecting the dots" and "picking the maximum," respectively, for each individual drug concentration profile. These are clearly design dependent responses that must reflect accurately the extent and rate of bioavailability of the drug in order to test bioequivalence meaningfully. One must be assured that the time points chosen for blood samples are sufficient to define the concentration curve and yield accurate estimates of AUC and $C_{\max}$. Modeling prior pharmacokinetic data can help one develop an efficient design for selection of time points for blood collection, although the intersubject variability is sufficiently large to make this an approximation at best.

The authors claim that the TOST has greatly inferior power to the new test and to the BHM test for all but very small σ_0. The comparison of power (Table 1) is sparse and fails to illuminate sufficiently any *meaningful* distinctions between the tests. Very clearly, there is no difference between any of the tests for bioequivalence studies that are sized adequately (generally accepted to be greater than or equal to 80% power) as acknowledged in the penultimate paragraph of Section 4.3. If we compare the standard error of d for a fixed level of power, then we may consider the relative efficiency of the tests at a specified power. For example, from Table 1 directly, the tests are indistinguishable for power of 72% or greater. If we choose a low power of 50%, then the relative efficiency of the TOST procedure to the new test procedure is about 96% (expanding Table 1). Thus, about 4% more volunteers would be needed using TOST versus the new test. It is hard to argue that this places "... an undue burden on the generic drug manufacturers" as stated in Section 1. The power "advantage" occurs for trials with inadequate power (less than 50%), where the relative efficiency of the TOST procedure to the new test procedure finally begins to drop noticeably. The development and discussion of the new test is quite instructional, providing vivid interpretation of its characteristics versus those of the TOST procedure, but the practical advantages of the new test to the TOST are seen to be limited.

In Section 4.2, Berger and Hsu point out that the rejection region of their test, or any such approximately unbiased test that is uniformly more powerful than the TOST, will contain sample points for which d is outside the interval (θ_L, θ_U). They subsequently comment that "notions of size, power and unbiasedness are more fundamental than "intuition"." Certainly, these statistical notions are fundamental. But, "intuition" aside, one must first seek to meet the regulatory objectives in testing for bioequivalence. These objectives are not well met in using the proposed test since, even in the truncated version, outside the rejection region of the TOST this test concludes bioequivalence secondary to overwhelming variability. The TOST better meets the objectives precisely because it excludes such flawed conclusions. Further, if one could invert the Berger–Hsu test to obtain a $100(1-\alpha)\%$ confidence interval, the above appears to imply that the resulting confidence set for $\eta_T - \eta_R$ may, for some values of s_*, actually exclude d, the point estimate for $\eta_T - \eta_R$. We agree that such a result would not be "intuitive." The degree to which test (or confidence region) performance is nonintuitive could make the Berger–Hsu approach very difficult to sell to pharmacokineticists, physicians and the general public, who are all consumers of statistical bioequivalence assessments to one degree or another.

In Section 4.2 the following appears: "Due to the seriousness of a Type I error, declaring two drugs to be equivalent when they are not, the search for a size-α test that was uniformly more powerful than the TOST continued." The search for a size-α test that was uniformly more powerful than the TOST continues, but not due to the seriousness of any Type I errors! Note that a Type II error is failing to conclude bioequivalence when, in fact, the formulations are bioequivalent. The seriousness of a Type II error is costly only to the manufacturer and does not place any consumer at risk. We agree that a Type I error can be very serious for consumers, and the TOST is conservative (as noted in Table 1) with regard to the Type I error rate for highly variable, under-powered studies. As indicated above, the primary advantage of nearly unbiased tests is in the case of extremely variable, underpowered studies, in which case it could be considered detrimental to consumers to conclude bioequivalence.

The tutorial of Section 5 reviews the proper use of standard theorems relating confidence sets to hypothesis tests. Their demonstration of the equivalence of the $[D_1^-, D_1^+]$ interval to the TOST "... is more consistent with standard statistical theory ..." but is of questionable practical value. As in most hypothesis testing situations, there is often a practical interest in estimation as well. In general, nonstatistical consumers of bioequivalence testing will find the $100(1-2\alpha)\%$ confidence interval more informative than the proposed $100(1-\alpha)\%$ $[D_1^-, D_1^+]$ confidence interval with respect to estimation. Granted, there is the logical discontinuity between the size-α TOST and a $100(1-2\alpha)\%$ confidence

set, and the best solution may be simply to report the estimate and its associated standard error. The $[D_1^-, D_1^+]$ $100(1-\alpha)\%$ confidence interval can fail to provide a useful interval estimate, despite the logical statistical consistency regarding test size and stated confidence level. For example, if the 90% confidence interval is (1.24, 1.28), the $[D_1^-, D_1^+]$ 95% confidence interval is [1.00, 1.28]. Using the equivalence interval of [0.80, 1.25], one fails to reject the null hypothesis using the TOST. However, the $100(1-2\alpha)\%$ interval (albeit 90%) makes it clear that the primary reason for failure to show equivalence is a rather large difference between formulations. Looking at the latter interval one is unsure whether there is a large formulation difference or if perhaps the sample size was inadequate to demonstrate equivalence. Finally, a practical question to address regarding implementation of $[D_1^-, D_1^+]$ versus the $100(1-2\alpha)\%$ confidence interval is whether there can be any difference with respect to conclusions about bioequivalence, and the answer is no, as shown nicely in Section 5. This fact should not, as set forth by the authors, be taken as an endorsement for generating size-α tests from $100(1-2\alpha)\%$ confidence sets!

As an aside, in the biopharmaceutical sciences the "Min" test (Laska and Meisner, 1986, 1989), an IUT, is often referenced for testing some intersection–union hypotheses. In this case the simultaneous testing of $C_{\max}$ and AUC for bioequivalence using the TOST procedure for each could be considered an application of the Min test. Tests uniformly more powerful than the Min test have been investigated by Liu and Berger (1995).

In conclusion, Berger and Hsu have made valuable contributions by dispelling several incorrect statistical methods that have been described in the bioequivalence testing literature, in addition to illuminating the value of deriving tests and confidence sets based upon the intersection–union test methods. We believe it would be helpful to see a tenable example juxtaposing the TOST and the new test where conclusions reached by the two tests differ: that is, TOST is unable to reject the hypothesis of bioinequivalence whereas the new test can reject and conclude bioequivalence. The 95% $[D_1^-, D_1^+]$ confidence interval and the 95% confidence interval corresponding to the new test, if known, could be reported. Our greatest concern with this otherwise excellent paper is that it focuses too much attention on a new test that provides no practical advantage, and possibly some practical disadvantages, over the TOST procedure.

Comment

Jen-pei Liu and Shein-Chung Chow

1. INTRODUCTION

In the pharmaceutical industry, bioequivalence testing is usually performed as a surrogate for therapeutic equivalence in effectiveness and safety between drug products, for example, different formulations of the same drug product or an innovator drug and its generic copies. Bioequivalence is assessed based on the so-called *fundamental bioequivalence assumption* (Metzler, 1974; Chow and Liu, 1992). The fundamental bioequivalence assumption states that bioequivalent formulations or drug products are therapeutically equivalent (i.e., they have the similar therapeutic effect in terms of efficacy and safety). Hence, they can be used interchangeably. This important assumption originated from the Drug Price Competition and Patent Term Restoration Act passed by the United States (U.S.) congress in 1984. Based on this act, the U.S. Food Drug and Food Administration (FDA) was authorized to approve generic copies of an innovator drug product after the patent has expired. The sponsors are required to conduct bioequivalence trials to demonstrate that these generic copies are bioequivalent to the innovator drug product through an abbreviated new drug application (ANDA).

As indicated in Chow and Liu (1995), drug interchangeability can be classified as prescribability or switchability. Drug prescribability is referred to as the physician's choice for prescribing an appropriate drug product for his or her new patients between an innovator drug product and a num-

Jen-pei Liu is a member of the Department of Statistics, National Cheng-Kung University, Tainan, Taiwan, 70101 (e-mail: jpliu@ibm.stat.ncku.edu.tw). Shein-Chung Chow is with the Biostatistics and Data Management, Bristol-Myers Squibb Company, Plainsboro, New Jersey 08536.

ber of generic copies of the innovator drug product which have been shown to be bioequivalent to the innovator drug product. Drug prescribability is usually assessed by *population bioequivalence* (Chow and Liu, 1992). Drug switchability is related to the switch from a drug product (e.g., an innovator drug product) to an alternative drug (e.g., a generic drug product) within the same subject whose concentration of the drug product has been titrated to a steady, efficacious and safe level. To assure drug switchability, it is recommended that bioequivalence be assessed within the individual subject.

As a result, there are three types of bioequivalence, namely, average bioequivalence (ABE), population bioequivalence (PBE) and individual bioequivalence (IBE). The concept of PBE investigates the closeness between the distributions of the pharmacokinetic responses (e.g., AUC or $C_{\max}$). Chow and Liu (1992) indicate that if the pharmacokinetic responses, or their transformations, follow approximately a normal distribution, to ensure drug prescribability, requires establishing bioequivalence in both average and variability of bioavailability. On the other hand, drug switchability is for *individual bioequivalence.* The concept of individual bioequivalence is to examine the similarity between the two distributions of the pharmacokinetic responses from the same subjects. However, current regulations of the U.S. FDA, European Community (EC) and Japan only require that the evidence of *average bioequivalence* be provided in order to obtain approval of generic drugs. For detailed regulations on statistical procedures for assessment of ABE, the readers may refer to the guidance entitled *Statistical Procedures for Bioequivalence Studies Using a Two-treatment Crossover Design* issued by the U.S. FDA in July 1992 (FDA, 1992b).

Berger and Hsu provided an interesting and informative review of the application of intersection–union tests (IUT) to the problem of bioequivalence testing. Their criticisms of current statistical practices for evaluating average bioequivalence can certainly spur further research and discussion in the area of bioequivalence testing. In this Comment, we further address some scientific issues from the perspective of pharmaceutical industry. Berger and Hsu focused on the assessment of ABE. Very limited information was given regarding IBE and PBE. Note that ABE has been criticized due to its limitation for addressing drug prescribability and switchability, which will be discussed extensively in a special issue of the *Journal of Biopharmaceutical Statistics* (Chow, 1997).

2. UNIFORMLY MORE POWERFUL TESTS VERSUS UNIFORMLY MOST POWERFUL TEST

Let $(X_1, \ldots, X_n)$ be i.i.d. random variables in a sample from $N(\eta, \sigma^2)$, where σ^2 is known. Consider the one-sample version of the interval hypotheses (Chow and Liu, 1992) for equivalence for equation (2) in the article:

$$H_0: \quad \eta \le \theta_L \quad \text{or} \quad \eta \ge \theta_U \tag{1a}$$

versus

$$H_a: \theta_L < \eta < \theta_U. \tag{1b}$$

The uniformly most powerful (UMP) test exists for hypotheses (1a) and (1b) (see Roussas, 1973, page 285). In practice, however, the variance is usually unknown in a one-sample problem. In addition, for bioequivalence testing, we may encounter a two-sample problem for comparing drug products in terms of average and variability of bioavailability. As a result, under a two-sequence, two-period (2×2) crossover design as given in the article, it is a concern whether the uniformly most powerful unbiased or invariant tests (UMPU or UMPI) for equation (2) in the article exist. Note that T_L and T_U of Schuirmann's two one-sided tests (TOST) as defined in (4) (or t_L and t_U in Liu and Chow's TOST) are UMPU tests for hypotheses (5) and (6) (or one-sided hypotheses on variability) in the article, respectively. The intersection–union principle for combining these two individual UMPU tests proposed by Berger and Hsu (1996) leads to a biased test rather than an unbiased test. Therefore, it is of interest to know whether the UMPU or UMPI test can be constructed from the intersection–union principle. Suppose that there is no UMPU or UMPI test for (2); one can always derive a test for (2) which, under certain circumstances, will be more powerful than either the unbiased test (BHM) proposed in the unpublished technical report by Brown, Hwang and Munk (1995) or the nearly unbiased test (BH new) suggested in the article. The same comment is applicable to the Wang test for variability (Wang, 1994).

3. THE STANDARD ANALYSIS AND EQUIVALENCE LIMITS

Before the U.S. FDA statistical guidance on bioequivalence was issued in July 1992, the average bioequivalence was evaluated based on the ratio of average bioavailabilities through the hypotheses of (15) reformulated from (13) in the article. In this case, the bioequivalence limits involve unknown parameters. The standard analysis prior to the 1992

TABLE 1
Impact of correlation on the level of significance; sample size = 18, CV = 15%

Correlation	$\eta_T - \eta_R$	Nonparametric		Parametric	
		Standard	Liu and Weng	Standard	Liu and Weng
0.50	80	0.0380	0.0547	0.0497	0.0650
	120	0.0767	0.0577	0.0837	0.0693
0.75	80	0.0407	0.0527	0.0533	0.0633
	120	0.0787	0.0557	0.0847	0.0673
0.90	80	0.0497	0.0510	0.0847	0.0570
	120	0.0840	0.0547	0.0920	0.0630
0.95	80	0.0727	0.0575	0.0790	0.0593
	120	0.0943	0.0520	0.1027	0.0650
0.99	80	0.1647	0.0497	0.1673	0.0633
	120	0.1647	0.0500	0.1693	0.0550
0.999	80	0.3634	0.0467	0.3620	0.0580
	120	0.3423	0.0440	0.3460	0.0507
0.9999999	80	0.5120	0.0473	0.5120	0.0527
	120	0.4913	0.0537	0.4913	0.0600

Simulated data were generated from a normal distribution under a 2×2 crossover design.
Source: Liu and Weng (1995)

guidance was to substitute the unknown reference average in the limits with its least squares estimate (LSE) assuming that the resulting quantities are the true parameters. Chow and Liu (1992) and Liu and Weng (1995) reported that this was the standard analysis at the time, while academia seemed not to pay much attention to the bioequivalence problem. They, however, did not indicate that it is a correct analysis. On the contrary, Chow and Liu (1992) emphasized that the standard analysis fails to take into account the variability of the LSE of the reference average as the equivalence limits. Furthermore, Liu and Weng (1995) not only recognized this deficiency in the standard analysis but also, under a 2×2 crossover design, proposed a parametric procedure and its Wilcoxon nonparametric counterpart to overcome the drawback. Unlike the two-group parallel design used in the article, there are two correlated pharmacokinetic (PK) responses from the same subject for a bioequivalence study conducted in a 2×2 crossover design. When the correlation between the two PK responses from the same subject goes to 1, Liu and Weng (1995) showed that, theoretically and empirically, the size of the standard analysis approaches 0.5. Our Table 1 reproduces the simulation results of the impact of correlation on the level of significance from Table 1 of Liu and Weng (1995). As indicated in Table 1, if the correlation is less than 0.95, the two tests for T^*1/T^*2 of the standard parametric analysis and its nonparametric version have different sizes in the manner which was described in the article for the two independent samples. However, when the correlation exceeds 0.95, the sizes of both tests are approximately the same but are greatly inflated. On the other hand, the modified TOST, either the parametric or nonparametric version, proposed by Liu and Weng (1995) adequately controls its size at the nominal level.

For evaluation of therapeutic equivalence, unlike the pharmacokinetic responses from bioequivalence studies, the clinical endpoints are usually binary data such as cure rate or eradication rate in the antiinfective areas. The equivalence limits are determined on the estimated eradication rate of the reference drug from the previous studies. Our Table 2 gives the equivalence limits suggested by the FDA (Huque and Dubey, 1990). However, quite often, the estimated reference eradication rate, say 82%, obtained from the current study is different from that, say 77%, from previous studies. According to Table 2, an eradication rate of 82% corresponds to the equivalence limits of plus or minus 20%, while the limits of plus or minus 15% are for the eradication rate of 77%. As a result, the equivalence limits are

TABLE 2
Equivalence limits for binary responses

Response rate for the response drug	Equivalence limits
50%–80%	± 20%
81%–90%	± 15%
91%–95%	± 10%
> 95%	± 5%

Source: Huque and Dubey (1990).

to be changed from those stated in the protocol, and sample size might not be adequate to provide sufficient power because of the change of the equivalence limits. See Weng and Liu (1994) for more details. A possible approach to resolve this issue is to find a TOST for hypothesis (14) in the article, with averages replaced by eradication rates for the test and reference drug products. However, difficulty arises from the fact that the variance of binary responses is a function of the average. Further research in this area is needed.

4. POWER AND SAMPLE SIZES

Table 1 of the article provides the sizes and powers of the Schuirmann TOST, the BHM unbiased test and the new nearly unbiased BH test suggested by the article. Clearly, the Schuirmann's TOST is conservative as σ_D increases. However, the bottom line is whether the variability in Table 1 of the article is frequently encountered in bioequivalence trials. Our Table 3 converts σ_D on the logarithmic scale into the intrasubject coefficient of variation (CV) of the reference product on the original scale. In practice, a reference is classified as a highly variable drug if the intrasubject CV of its pharmacokinetic responses such as AUC exceeds 30%. From Table 3, except for $\sigma_D = 0.04$, sometimes, $\sigma_D = 0.08$ the variability used for comparison of size and power in the article is very unlikely to be encountered in practice. On the other hand, when most bioequivalence trials generate a CV under 30%, Table 1 of the article demonstrates that the size and power of the Schuirmann TOST, the BHM unbiased test and the new BH test are almost indistinguishable. The relative improvement of power over the Schuirmann TOST given in Table 1 of the article is more than 60% when σ_D is large. However, one needs to realize that the largest absolute improvement in power by both the BHM and the new BH tests is only 10.2%. Furthermore, sample size determination for bioequivalence trials is to achieve an absolute power of at least 80%. Liu and Chow (1992b) provided an approximate formula of sample size determination based on TOST which later was extended to logarithmic responses by Hauschke, Steinijans, Diletti and Burke (1992). Table 3 also gives estimated sample sizes to achieve a power of 80% for various values of σ_D at $\eta_T - \eta_R = 0$. From Table 3, unless $\sigma_D \leq 0.08$, the sample size are formidably large compared to the sample size of a typical bioequivalence trial conducted by pharmaceutical industry, which ranges from 16 to 36 subjects. Because neither Brown and colleagues (Brown, Casella and Hwang, 1995; Brown, Hwang and Munk, 1995) nor Berger and Hsu provide the formulas for sample size estimation with respect to the BHM unbiased and the new BH nearly unbiased tests, a direct comparison in savings of sample size cannot be made.

TABLE 3
Sample sizes required for Schuirmann's two one-sided test procedure for 80% power at the 5% significance level for Table 1 of the article

		Power when $\eta_T - \eta_R = 0$			Sample size
σ_D	CV	TOST	BHM	New	TOST
0.04	16.1%	1.000	1.000	1.000	12
0.08	32.8%	0.720	0.721	0.720	42
0.12	50.9%	0.158	0.260	0.247	94
0.16	71.1%	0.007	0.131	0.128	178
0.20	94.7%	0.000	0.093	0.092	312
0.30	179.5%	0.000	0.066	0.066	1112

One characteristic shared by both the BHM unbiased and the new BH tests is that the rejection region is an open region whose width increases as the estimated variability increases. This disturbing anomaly is exacerbated when sample points in the rejection region eventually lie outside the equivalence limits. On the other hand, the rejection region of the Schuirmann TOST does not share this counterintuitive shape of the rejection region because it is a triangle. Any sample points with variability greater than $\Delta\sqrt{r/t_{\alpha,r}}$ will be outside the rejection region. This conservativeness may provide a desirable consequence. Currently, all regulatory agencies in the world only require the evidence of average bioequivalence assessed by the Schuirmann TOST for approval of generic drug products. Note that σ_D^2 in Section 2.1 of the article is a function of the average of the intrasubject variabilities over the test and reference formulations. Therefore, when the usual intrasubject CV observed in bioequivalence studies is less than 30%, because the Schuirmann TOST cannot declare average bioequivalence if the $r[SE(D)]^2$ exceeds $\Delta\sqrt{r/t_{\alpha,r}}$, any difference in intrasubject variability between test and reference formulations may have less serious consequences than the BHM unbiased test and the new BH test. This may be one reason no disastrous mishap has occurred since implementation of the Schuirmann TOST by the U.S. FDA, European Community and other countries more than 10 years ago.

Construction of the BHM unbiased test is recursive and requires intensive computation. Although the new BH nearly unbiased test is simpler to compute than the BHM unbiased test, it is still much more complicated than the Schuirmann TOST. Most of all, both the BHM unbiased test and the new

BH test are based on the polar coordinates. Hence, they lack a direct intuitive interpretation for pharmacologists, clinicians or scientists to understand. Furthermore, it is more difficult to present the results from the BHM unbiased test and the BH new test than the Schuirmann TOST in the report of a bioequivalence study for nonstatistician reviewers with limited statistical background. In summary, for most of bioequivalence trials with an intrasubject CV less than 30%, the BHM unbiased test and the new BH test do not offer any real advantages over the current Schuirmann TOST. As a result, BHM and BH are of little practical importance in bioequivalence testing.

5. CONFIDENCE INTERVAL

A very important fact was established in the article: that an equal-tailed $(1 - 2\alpha)100\%$ confidence interval always yields a two one-sided test of size α for the interval hypothesis obtained from the intersection–union principle. Let L and U denote the lower and upper limits for equal-tailed $(1 - 2\alpha)100\%$ confidence interval. Then the article showed that the lower and upper limits of the $(1 - \alpha)100\%$ confidence interval corresponding to a size-α TOST are given, respectively, as

$$(2) \qquad L^- = \min(0, L) \quad \text{and} \quad U^+ = \max(0, U).$$

However, because both intervals lead to the same TOST of the same size, it follows that, as demonstrated in our Table 4, the same decision of claiming bioequivalence (or not bioequivalence) will be concluded from both intervals. Therefore, in the actual decision-making process, the conclusion will not be altered by the $(1 - \alpha)100\%$ confidence interval. In addition, the consumer's risk associated with the decision is the size of the TOST and is not 1 minus the confidence level of the interval. On the other hand, unfortunately, the article does not provide the $(1-\alpha)100\%$ confidence interval corresponding to the BHM unbiased and the BH nearly unbiased tests. Otherwise, performance of these confidence intervals could then be evaluated.

6. LOGARITHMIC TRANSFORMATION

We agree with the viewpoints about the logarithmic transformation required for AUC and $C_{\max}$ by the FDA statistical guidance on bioequivalence. After logarithmic transformation, the equivalence limits in hypothesis (2) in the article are still known constants. On the other hand, if the analysis were to be performed on the original scale, the equivalence limits in hypothesis (15) in the article are unknown constants. As a result, the reason for the logarithmic transformation is to avoid the unknown parameters as the equivalence limits in (15). However, we think that scientific integrity should not and cannot be sacrificed nor compromised for regulatory convenience. In addition, the TOST for hypothesis (15) has been proposed by Liu and Weng (1995) and others. Therefore, we agree with the article that the scale of the PK responses for the analysis cannot be dictated by regulations and should be determined by the distributions of the random components of the model such as the one in Section 2.1 of the article for a 2×2 crossover design. For detailed comparisons of TOST between the original scale and logarithmic scale, see the simulation results of Liu and Weng (1994).

Logarithmic transformation has been required by the FDA and European Community and has been

TABLE 4
Decision of claiming bioequivalence by confidence intervals in (8) *and* (16) *of the article with respect to the equivalence limits of* $\ln(1.25) = -\ln(0.8)$

Situation	(L, U)	(L, U^+)	$L > \ln(0.8)$	$U < \ln(1.25)$	Decision of BE (L, U)	Decision of BE (L, U^+)
$L < 0 < U$	(L, U)	(L, U)	Yes	Yes	BE	BE
			Yes	No	NBE	NBE
			No	Yes	NBE	NBE
			No	No	NBE	NBE
$0 < L < U$	(L, U)	$(0, U)$	Yes	Yes	BE	BE
			Yes	No	NBE	NBE
$L < U < 0$	(L, U)	$(L, 0)$	Yes	Yes	BE	BE
			No	Yes	NBE	NBE

$L = D - t_{\alpha, r}\mathrm{SE}(D)$; $U = D + t_{\alpha, r}\mathrm{SE}(D)$; $L^- = \min(0, L)$; $U^+ = \max(0, U)$.
BE = bioequivalent; NBE = not bioequivalent.

implemented by industry for quite some time. Although the FDA guidance requests that the results of analysis on the logarithmic scale also be presented in the original scale after the inverse transformation, little attention is paid to the estimation of the ratio of averages, $\exp(\eta_T - \eta_R)$. Clearly, its maximum likelihood estimator (MLE), a ratio of geometric means on the original scale, produces a positive bias. This bias could be large because the sample size of bioequivalence is quite small. Sometimes, no estimated standard error of the MLE for $\exp(\eta_T - \eta_R)$ is even given in the report. If it is provided, it is incorrect. Liu and Weng (1992) discussed the minimum variance unbiased estimator (MVUE) of $\exp(\eta_T - \eta_R)$ and its variance which should be used for bioequivalence studies. The article suggested that, when the normality assumption is in doubt, the nonparametric counterpart of TOST can be used as an alternative. However, simulations performed by Liu and Weng (1993) and Hauck et al. (1997) indicate that for evaluation of average bioequivalence the TOST based on t-statistics given in (4) in the article are quite robust to the departure from the normality assumption.

7. IUT AND INDIVIDUAL BIOEQUIVALENCE

Although the current requirement of average bioequivalence performs satisfactorily for approval of generic drugs, Chen (1997) pointed out the following limitations of average bioequivalence:

1. It only focuses on population average of test and reference formulations.
2. It ignores distribution of interest between test and reference formulations.
3. It ignores subject-by-formulation interaction.

Individual bioequivalence has the following merits for assessing equivalence between drug products:

1. It compares both averages and variances.
2. It considers subject-by-formulation interaction.
3. It addresses "switchability."
4. It provides flexible bioequivalence criteria for different drugs based on their therapeutic window.
5. It provides reasonable bioequivalence criteria for drugs with high intrasubject variability.
6. It encourages and rewards sponsors to manufacture a better formulation.

Currently, the criterion proposed by Scall and Luus (1993) is under consideration by the U.S. FDA for individual bioequivalence. However, this criterion is an aggregation of three components: square of average differences; subject-by-formulation; and difference in intrasubject variabilities. As a result, the use of an aggregate criterion in fact masks the contribution made by each component. On the other hand, inference for the aggregate criterion is quite complicated and the bootstrap technique has to be used for the pharmacokinetic responses from a bioequivalence study with sample size only from 18 to 36 because its estimators and distribution of the estimators are intractable.

However, we can consider the average, subject-by-formulation interaction and intrasubject variability as three characteristics representing the quality assurance for a drug product. It follows that these important characteristics should be examined individually, and the results then can be combined through the intersection–union principle. According to Chen (1997), to demonstrate individual bioequivalence we need to test the following: (1) intrasubject variability,

$$\text{(3a)} \qquad H_{0v}: \quad \frac{\sigma_T^2}{\sigma_R^2} \le c_L \quad \text{or} \quad \frac{\sigma_T^2}{\sigma_R^2} \ge c_U$$

versus

$$\text{(3b)} \qquad H_{av}: c_L < \frac{\sigma_T^2}{\sigma_R^2} < c_U;$$

(2) subject-by-formulation interaction,

$$\text{(4a)} \qquad H_{0i}: \sigma_I^2 \ge c_I$$

versus

$$\text{(4b)} \qquad H_{ai}: \sigma_I^2 < c_I;$$

and (3) average,

$$\text{(5a)} \qquad H_{0a}: \quad \eta_T - \eta_R \le A_L \quad \text{or} \quad \eta_T - \eta_R \ge A_U$$

versus

$$\text{(5b)} \qquad H_{aa}: A_L < \eta_T - \eta_R < A_U,$$

where c_L, c_U, c_I, A_I and A_U are chosen to define clinically important differences. One concludes individual bioequivalence if each of (3a)–(3b), (4a)–(4b) and (5a)–(5b) is rejected at the α significance level. Under a replicated design (Liu, 1995; Chow, 1996), inference for the unknown parameters in each of these three hypotheses is straightforward and is based on the exact distributions such as the F distribution. Therefore, our proposed procedure based on IUT is more intuitively appealing and easier to

implement than the aggregated method considered by the U.S. FDA.

In conclusion, the intersection–union test is an interesting concept to combine the results from individual tests for different objectives. However, the BHM unbiased and the BH new nearly unbiased methods need further evaluation before they can be used as routine practice.

ACKNOWLEDGMENTS

The authors wish to thank Professors George Casella and Paul Switzer for providing us with the opportunity to prepare this follow-up article and for constructive comments. J. P Liu's research was supported in part by Taiwan NSC Grant 86-2115-M-006-029.

Comment

Donald J. Schuirmann

The authors have written a very comprehensive paper that will be a valuable reference source for statisticians who wish to learn about the statistical aspects of bioequivalence testing. I would like to comment on three points made in the paper.

POINT 1

The authors comment on a feature that their proposed new test [of their hypotheses (2)] shares with the Anderson and Hauck test (Anderson and Hauck, 1983) and the Brown, Hwang and Munk (BHM) similar test (Brown, Hwang and Munk, 1995), namely, that beyond a certain value of s_* the width (in the d direction) of the rejection region increases as s_* continues to increase. There exist sample outcomes $(d,\ s_*)$ for which one would not reject H_0, but for the same value of d but a larger value of s_* one would reject H_0. Eventually the rejection region includes values $(d,\ s_*)$ with d outside the interval (θ_L, θ_U). The authors note that any similar or approximately similar test of hypotheses (2) must have this property.

In my personal opinion, this property renders all similar or approximately similar tests of the hypotheses (2) unacceptable. The authors dismiss these concerns as "intuition." However, there is a probabilistic argument against these tests, and it is illustrated in the authors' Table 1. For three tests, Table 1 compares the power at the endpoints, θ_U and θ_L, of the equivalence interval to the power at the midpoint of the equivalence interval, as a function of σ_D. For the case of $\sigma_D = 0.20$, the power of the new test or the BHM test at the midpoint is less than twice as much as the power at the endpoints. For $\sigma_D = 0.30$, the power at the midpoint for these two tests is only 32% higher than the power at the endpoints. In the limit, as $\sigma_D \to \infty$, the power curve is perfectly flat, as illustrated in Table 1. Even for finite σ_D, there comes a point where σ_D is large enough that there is no practical difference between the power at the midpoint and the power at the endpoints. In other words, for all intents and purposes you are no more likely to conclude equivalence if the means are as equivalent as can be than you are if the means are inequivalent. This is also true of the TOST. However, in the case of the TOST, when σ_D is large enough to produce this situation, the power is truly negligible. For the new test or the BHM test, the power is α, which is usually 5%—a nonnegligible chance of concluding equivalence from an inadequate study. For further discussion, see Schuirmann (1987b).

In my personal opinion, not only should the rejection region not get wider as s_* increases, but there should be a value of s_* beyond which we do not reject H_0 no matter what the value of d.

POINT 2

In those circumstances where it is deemed more appropriate to analyze bioavailability metrics (such as AUC and $C_{\max}$) without transformation, we are interested in testing the authors' hypotheses (1) [same as the authors' hypotheses (13)]. Restating these hypotheses as the authors' hypotheses (14) suggests the TOST, as proposed by Sasabuchi (1980), which the authors call the T_1/T_2 test. This test is clearly preferable to the test that the authors call the T_1^*/T_2^* test.

Donald J. Schuirmann is a member of the Quantitative Methods and Research Staff, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, 5600 Fishers Lane, Rockville, Maryland 20857 (e-mail: schuirmann@cder.fda.gov).

I would point out that carrying out the T_1/T_2 test when the data come from a crossover study, where the intrasubject correlation is unknown, can be tricky. The test statistics are not so simple as the T_1 and T_2 presented by the authors, which are applicable to a parallel study. Locke (1984) describes a procedure for obtaining a Fieller-type confidence set for μ_T/μ_R in the case of a standard two-period crossover study. Locke's method is easily extended to general crossover designs. Although Locke (in the 1984 paper) does not explicitly suggest using such a confidence set to carry out the TOST of hypotheses (1), he does so in a more recent paper (Locke, 1990).

In the past, the U.S. FDA routinely used the T_1^*/T_2^* test to test hypotheses (1) using untransformed data, but I can report that this is no longer the case. Most bioequivalence studies submitted to the agency are analyzed after log transformation, but when analysis of untransformed data is thought to be more appropriate, the agency now suggests basing the test on the methodology described by Locke. See, for example, the recent *Guidance—Topical Dermatologic Corticosteroids: In Vivo Bioequivalence* (FDA, 1995).

POINT 3

The authors make an important point by noting that one cannot always obtain a size-α TOST by rejecting H_0 iff a $100(1-2\alpha)\%$ confidence set is contained within the equivalence interval. This procedure only works if the $100(1-2\alpha)\%$ confidence set is "equal-tailed." Yet, as the authors point out, both the U.S. FDA and the European Community suggest that the test should be carried out by constructing a 90% confidence interval, in order to obtain a size-0.05 TOST. Fortunately, the confidence procedures currently proposed for testing the authors' hypotheses (2) after log transformation, and for testing the authors' hypotheses (1) using the methodology of Locke with untransformed data, are equal-tailed. Nevertheless, I agree with the authors that it is misleading to imply that one may always base a size-α test on a $100(1-2\alpha)\%$ confidence set.

Comment

J. T. Gene Hwang

Professors Roger Berger and Jason Hsu are to be congratulated for their interesting article, which surveys thoroughly the area of bioequivalence from a statistical perspective. This is a fast-developing research area, and before it diverges in various directions it is very useful to have this article to summarize and, to some extent, unify the important results.

Two main themes of their paper are to demonstrate that the concept of intersection–union tests "clarify, simplify and unify" bioequivalence testing, and to argue against the "misconception that size-α bioequivalence tests generally correspond to $100(1-2\alpha)\%$ confidence sets". I shall comment along these two lines.

Professor Hwang is with the Department of Mathematics, White Hall, Cornell University, Ithaca, New York 14853 (e-mail: hwang@math.cornell.edu).

1. INTERSECTION–UNION METHODS AND THE NEW TEST

I agree that the intersection–union method has a prominent position in bioequivalence tests. For one thing, the two one-sided tests procedure is one such test. As has been pointed out, the test can, however, be improved by the Brown, Hwang and Munk test (Brown, Hwang and Munk, 1995). The authors then use the intersection–union method to derive a new test which is almost as powerful as Brown, Hwang and Munk's test. The idea is quite interesting. The authors argue that the new test has the following advantages over Brown, Hwang and Munk's test:

(i) The new test is computationally less intensive.

(ii) The new test provides boundaries which are smooth, unlike the boundaries of Brown, Hwang and Munk's test, which sometimes have a quite irregular shape.

The disadvantages of the new test, as is pointed out, are that it is biased and it has slightly smaller power than Brown, Hwang and Munk's test. How-

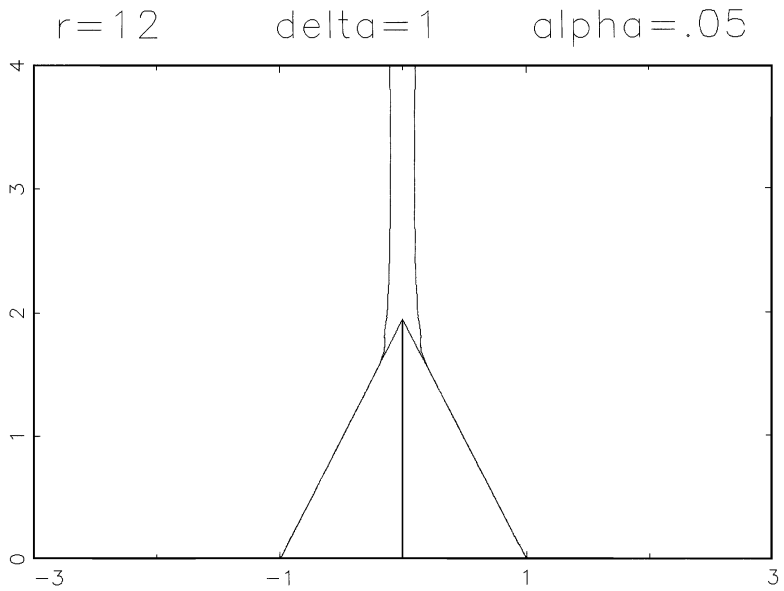

FIG. 1. *The rejection region of Brown, Hwang and Munk for* $r = 12$ *and* $\Delta = 1$. *It contains the triangle which corresponds to the two one-sided tests procedure.*

ever, the authors demonstrated in Table 1 that the loss is small.

Overall, I agree with the advantages and also the assertion that the loss in power is small. This new test is therefore theoretically useful.

From the practical side, it should be noted that the test of Brown, Hwang and Munk is not computationally intensive. It takes about 5 minutes to calculate 7200 pairs of boundary points for Figure 1 below, using a 90-mega Hertz Pentium personal computer with the Gauss program. Note that the boundary can be approximated well by a line where $D > b$ for a bound b. Furthermore, 7200 points are quite dense within $(0, b)$.

About (ii), it is true that the shape of Brown, Hwang and Munk's test is irregular when the degrees of freedom r equal 3, as shown in the authors's Figure 1. However the boundaries of Brown, Hwang and Munk's test are typically smooth, if r is not too small. See the smooth curve in Figure 1 below for $r = 12$. When $r \geq 12$, the boundary has a shape similar to Figure 1 below. In this figure, Δ is taken to be 1 without loss of generality since otherwise we may use the transformation $(D, S) \to (D/\Delta, S/\Delta)$. In applications, we typically have 24 subjects or more, and hence the degrees of freedom (taking into consideration the subject effects, etc.) is at least 20, depending on the model. Therefore Brown, Hwang and Munk's test has smooth boundaries anyway.

In conclusion, it remains to be seen whether the authors' new test would become popular in applications.

2. "THE MISCONCEPTION THAT SIZE-α TESTS CORRESPOND TO $100(1 - 2\alpha)$% CONFIDENCE SETS"

In Section 5.2, the authors argue forcefully that it is incorrect to always use a $(1 - 2\alpha)$ confidence set to construct a test without verifying that the resultant test has size α. While the approach is all right for the one-dimensional case, often it causes some problem in the higher-dimensional case. The authors give an example about the ratio problem using a two-dimensional test in the paragraph containing (22).

While I agree with the authors's assertion, it seems interesting here to point out another example relating multivariate bioequivalence hypothesis (23). We shall assume the canonical form that $X = (X_1, \ldots, X_p)'$ is a p-dimensional normal observation with mean $\sigma = (\sigma_1, \ldots, \sigma_p)'$ and covariance matrix Σ; also, S, independent of X, has a Wishart distribution with d degrees of freedom $(d > p)$. This canonical form applies to a general linear model including the crossover design with period effects and subject effects. Hence σ_i is the difference of the logarithmically transformed characteristic (such as AUC, $C_{\max}$, $T_{\max}$ etc.) of the brand name drug and the generic drug.

We focus on the p-dimensional bioequivalence hypothesis

$$H_a^m: |\sigma_i| < \Delta \quad \text{for all } i = 1, 2, \ldots, p,$$

which is a generalization of (23). The symmetry of the interval of σ_i with respect to the origin is made without loss of generality.

A $1-2\alpha$ confidence set based on Hotelling's T^2 is

$$T^2 \le \frac{dp}{d-p+1} F_{2\alpha,\, p,\, d-p+1}.$$

where

$$T^2 = d(X-\theta)' S^{-1}(X-\theta)$$

and $F_{2\alpha,\, p,\, d-p+1}$ is the 2α upper quantile of the F-distribution with p and $d-p+1$ degrees of freedom.

If we use this confidence set to construct a test, then we will declare H_a^m, that is, bioequivalence, if the confidence set is contained in H_a^m. The corresponding rejection region is recently shown in Wang, DasGupta and Hwang (1996) to be described by the inequality

$$|X_i| < \Delta - \left(\frac{p}{d-p+1} F_{2\alpha,\, p,\, d-p+1} S_{ii}\right)^{1/2} \quad \text{for all } i,$$

where X_i is the ith element of X and S_{ii} is the ith diagonal element of S. The Type I error of the test, however, is α if and only if $p = 1$.

In general, the actual size can be shown to be

$$\alpha_0 = P\left(T_d > \left(\frac{dp}{d-p+1} F_{2\alpha,\, p,\, d-p+1}\right)^{1/2}\right)$$

where T_d is a Student's-t random variable with d degrees of freedom. Note that if $p = 1$, the above probability equals

$$P\left(T_d > \left(F_{2\alpha,\, 1,\, d}\right)^{1/2}\right) = \alpha.$$

TABLE 1
Actual size α_0 for $d = 22$ when $\alpha = 0.05$

p	α_0
1	0.05
2	0.0150
3	5.18×10^{-3}
4	1.88×10^{-3}
5	6.79×10^{-4}
10	2.36×10^{-6}

However, when $p \neq 1$, Table 1 below shows that the actual size α_0 can be very small and hence the recommended test is very conservative. In this table, d is taken to be 22, corresponding to a standard 2×2 crossover design involving altogether 24 subjects with subject and period effects. This example demonstrates that using $1-2\alpha$ confidence set to derive a test may give a test of size much smaller than α as long as $p > 1$. Even for $p = 2$, the actual size already drops to 0.015 for a target size 0.05.

To achieve a correct size α, one needs to use a confidence set with coverage probability $1-a$, where a is such that

$$P\left(T_d > \left(\frac{dp}{d-p+1} F_{a,\, p,\, d-p+1}\right)^{1/2}\right) = \alpha,$$

or, equivalently,

$$\frac{dp}{d-p+1} F_{a,\, p,\, d-p+1} = t_{\alpha,\, d}.$$

Again using $a = 2\alpha$ leads to a correct size α only when $p = 1$. Values of a are given in Wang, DasGupta and Hwang (1996).

Rejoinder

Roger L. Berger and Jason C. Hsu

We thank the Editors of *Statistical Science* for soliciting these discussions of our article. All of the discussants make interesting and important points about various aspects of bioequivalence problems. We are especially pleased that the discussants represent the views of regulatory agencies, pharmaceutical companies and academics, all of whom have an interest in bioequivalence problems.

1. OTHER EQUIVALENCE PROBLEMS AND USEFULNESS OF EQUIVALENCE CONFIDENCE INTERVALS

We join Anderson and Hauck on the soap box in saying "Practical equivalence problems should be treated as such!" The drug shelf-life example mentioned in Section 5.1 is one in which the exclusion of the interaction terms in the model should be, but has not been, treated as a practical equivalence problem. The bovine growth hormone safety studies example alluded to in the same section is also one in which comparison with the negative control should be, but has not been, treated as a practical equivalence problem. We thank Meredith and Heise for pointing out additional examples. In vitro comparison of dissolution profiles of two formulations of the same drug is certainly a practical equivalence problem. However, in vivo trials with the objective of demonstrating that a new compound is at least as

good as an existing drug seem to us more appropriately formulated as one-sided inference problems.

Meredith and Heise as well as Liu and Chow seem to doubt the usefulness of insight into equivalence confidence sets. We think it would be a good reflection on the statistics profession if the official FDA documents indicated some cognizance of the equivalence confidence interval associated with the TOST, the stated decision rule. More important, insight into equivalence confidence sets in the simple two-drug problem is a reliable guide toward solving nontrivial multiple equivalence problems, as shown below using the drug shelf-life determination example from Section 5.1.

When the degradation of a drug can be represented as a simple linear regression model, both FDA (1987) and CPMP/ICH/380/95 (1993) (which applies to the United States, Europe and Japan) specify that the shelf-life be calculated as the one-sided lower 95% confidence bound on the time at which the true content reaches the lowest acceptable limit, usually 90% of the labeled amount of drug.

It is generally to the advantage of the manufacturer to establish a long shelf-life for a drug, but different batches of the same drug may degrade at different rates. When the degradation rate varies greatly from one batch to another, the guidelines *intend* that the shelf-life be calculated conservatively, from the worst degradation rate. On the other hand, the guidelines *intend* to reward a manufacturer making consistent batches with a longer shelf-life, by allowing it to be calculated from a single degradation rate based on data pooled from batches with degradation rates *practically equivalent* to the worst rate. Thus, if $\beta_1, \ldots, \beta_k$ denote the degradation rates of the k batches of the drug sampled, and rates within θ of the worst rate are practically equivalent to the worst, then data from batches i with $\beta_i - \min_{j \neq i} \beta_j \leq \theta$ can be pooled.

Currently the guidelines state that if the null hypothesis of equality of degradation rates (i.e., the hypothesis of no time×batch interaction) is accepted at the 25% level, then a reduced model with a common degradation rate (slope) is to be used with all batches pooled. This clearly violates the intent of the guidelines, as the acceptance of the no-interaction hypothesis may be due to small sample size and/or noisy data, thus rewarding with a longer shelf-life a manufacturer who does an inadequate study and/or makes inconsistent batches.

The intent of the guidelines can be met by testing the multiple hypotheses

$$(1) \qquad H_0^i\colon \beta_i - \min_{j \neq i} \beta_j > \theta, \quad i = 1, \ldots, k,$$

and pooling all batches i with H_0^i rejected. (Note the similarity between these hypotheses and the TOST hypotheses.) It is not obvious how to generalize the TOST or the more powerful tests to test (1) because an IUT would not allow for the possibility of rejecting some, but not all, hypotheses. Further, since up to $k - 1$ hypotheses in (1) may be true, there may appear to be the need for multiplicity adjustment. However, insight from Section 5.1 leads directly to 95% simultaneous equivalence confidence intervals, and pooling decisions can be based on these. Furthermore, the construction of these confidence intervals requires no multiplicity adjustment for the hypotheses in (1).

Recall that the equivalence confidence set in Section 5.1 was constructed by testing, within each half of the parameter space where η_i is smaller ($i = 1, 2$), against the alternative that the larger η_j, $j \neq i$, is larger by no more than a specified positive quantity. In the shelf-life problem, since equivalence with the worst rate is desired, within the part of the parameters space where β_i is the smallest ($i = 1, \ldots, k$), one tests against the alternative that the other rates β_j, $j \neq i$, are larger by no more than specified positive quantities. If one-sided 5% Dunnett's treatments versus control tests are used (in analogy with one-sided 5% t-tests), then the confidence intervals for $\beta_i - \min_{j \neq i} \beta_j$ that result from Theorem 3 are typically

$$(2) \qquad \left[\left(\hat{\beta}_i - \min_{j \neq i}\{\hat{\beta}_j + d_i \hat{\sigma}_{\hat{\beta}_i - \hat{\beta}_j}\}\right)^-, \left(\hat{\beta}_i - \min_{j \neq i}\{\hat{\beta}_j - d_j \hat{\sigma}_{\hat{\beta}_i - \hat{\beta}_j}\}\right)^+\right],$$

where $\hat{\beta}_i$ and $\hat{\sigma}^2_{\hat{\beta}_i - \hat{\beta}_j}$ are the usual estimates of β_i and $\mathrm{Var}(\hat{\beta}_i - \hat{\beta}_j)$, and d_i is the 5% critical value for one-sided Dunnett's test with the ith batch as the control. (The lower bounds are always as given here, but the upper bounds can be improved for some data sets. See Ruberg and Hsu, 1992.) Clearly, the intent of the guidelines is met if one pools data from batches whose upper confidence bounds in (2) are less than θ. (Logically, if no batch meets this criterion, the conclusion is the manufacturer has not done an adequate study.) This is a vivid illustration of the usefulness of the insight given in Section 5.1 toward solving more complicated equivalence problems.

Of course, pooling decisions can also be based on 90% confidence intervals which are of the form (2) but without the constraints to contain zero. For such confidence intervals to achieve 90% confidence, the critical values d_i must be increased to the 10%

critical value of the Tukey–Kramer method for all-pairwise comparisons of degradation rates (proof is as in Section 4.2.4.1 of Hsu, 1996). A calculation then shows the decision to pool batches based on these latter confidence intervals to be rather conservative, with an error rate less than 3% in the setting of the real data sets in Ruberg and Hsu (1992). This is yet another illustration of the danger of careless application of 90% confidence sets in practical equivalence problems.

2. REJECTING FOR LARGE s_*

Schuirmann, Meredith and Heise, and Liu and Chow all criticize the new test we proposed in Section 4.2 because it rejects H_0 and concludes bioequivalence for some sample points with arbitrarily large values of s_*. We want to question why one should not reject for large s_*, and, if there is good reason not to, we want to propose that this requirement be made a formal part of the problem.

Schuirmann states the claim most succinctly: "... there should be a value of s_*, beyond which we do not reject H_0 no matter what the value of d." This criticism has been made against other tests that have tried to improve the power of the TOST, such as Anderson and Hauck's (1983) test. We ask, "Why is this criticism made of bioequivalence tests when it is not made of other tests?" Consider a drug that claims to lower blood pressure. Measurements are made on subjects before and after administration of the drug, and a paired t-test is used to demonstrate that the blood pressure is lowered. This t-test uses the statistic $d/(s_d/\sqrt{n})$. This t-test will reject the null hypothesis for arbitrarily large values of s_d, but we have never seen it suggested that one should not reject H_0 if s_d is too large. Why are large values of the standard error such a concern in bioequivalence tests when they do not seem to be a concern in t-tests?

Large values of s_* suggest that σ_D is large. Presumably it is large values of σ_D that are the concern. Liu and Chow note that σ_D^2 is related to the intrasubject variances of the test and reference drugs. So, by not rejecting for large s_*, we are somehow guarding against large intrasubject variances. If control of σ_D or the intrasubject variances is really the concern, then this should be explicitly stated as part of the problem. For example, if the regulatory agency sets an upper bound of σ_{D0}, then the alternative hypothesis should be stated as

$$H_a: \quad \theta_L < \eta_T - \eta_R < \theta_U \quad \text{and} \quad \sigma_D < \sigma_{D0}.$$

Because this just adds a third condition to the alternative hypothesis, a size-α test could be constructed using the IUT method. Our new test or the BHM test could be used to test the hypothesis about $\eta_T - \eta_R$. A chi-squared test could be used to test the hypothesis about σ_D. In this way, the variability could be controlled in a well-defined way, rather than in the informal way it is now controlled by the TOST. When formulated in this way, the problem is closely related to the population bioequivalence problem of Section 6.2.

Finally, Schuirmann offers another argument why one should not reject for large values of s_*. It is that the power function of our new test or the BHM test is nearly constant at the value α for large values of σ_D. However, a t-test, as described above, has exactly this same property. So, again, why is one content to reject for large values of s_d in the t-test, but not in a bioequivalence test?

3. INDIVIDUAL BIOEQUIVALENCE

Hauck and Anderson and Liu and Chow both suggest that individual bioequivalence might be a more appropriate formulation of the problem than the average bioequivalence formulation we used. We suggested in Section 1 that the IUT method might also be useful for individual bioequivalence problems. We thank Liu and Chow for providing a concise example in which this is true. They formulate three hypotheses that place bounds on the parameters of interest. Then the IUT method is used to construct a size-α test designed to ensure that all the parameters are within their specified bounds. We think this is a very reasonable and easy to understand formulation of the individual bioequivalence problem. We are happy to see that the IUT method again provides a simple solution. A careful analysis of this problem, like the analysis that led to our new test in Section 4.2, might yield a more powerful test than the simple test proposed by Liu and Chow.

Hauck and Anderson, on the other hand, mention an aggregate criterion for individual bioequivalence that was recommended by the FDA Working Group on Individual Bioequivalence to the FDA Advisory Committee for Pharmaceutical Science at an August 1996 meeting. The aggregate individual bioequivalence criterion (IBC) proposed was

$$\text{IBC} = \frac{(\eta_T - \eta_R)^2 + c_1\sigma_I^2 + c_2(\sigma_T^2 - \sigma_R^2)}{\sigma_{R+}^2},$$

where η_T, η_R, σ_I^2, σ_T^2 and σ_R^2 are as defined by Liu and Chow, and $\sigma_{R+}^2 = \max\{\sigma_R^2, \sigma_{R0}^2\}$. To define this criterion, the regulatory agency would need to specify three constants, c_1, c_2 and σ_{R0}^2. In addition, the agency would need to specify an upper bound on

IBC to define when the two drugs were bioequivalent. We agree with some members of the FDA Advisory Committee that this criterion is very difficult to understand. We think it would be difficult to specify these four constants. We believe it would be much easier to consider each of the relevant parameters individually, as proposed by Liu and Chow. If two-sided bounds are set symmetrically, only three, rather than four, constants would need to be specified by the regulatory agency. And, we think it would be easier to specify these individual bounds than to specify constants like c_1, c_2 and σ^2_{R0} that somehow attempt to balance the relative importance of the various parameters. Note also that to achieve complete flexibility in balancing the relative importance of the various parameters, a fifth constant, c_0, to serve as a coefficient of $(\eta_T - \eta_R)^2$, is needed in the definition of the IBC. This complicates the aggregate criterion even more.

It should be noted that, at the August 1996 meeting, the FDA Advisory Committee for Pharmaceutical Science did not take any action on the Working Group's recommendation. It remains to be seen if any form of individual bioequivalence will be adopted to replace average bioequivalence. If the "disaggregate" form proposed by Liu and Chow is adopted, then IUT tests will continue to be important in the bioequivalence field.

4. ELLIPSOIDAL PEGS IN SQUARE HOLES

Hwang gives another example to illustrate that attempting to define size-α tests using $100(1-2\alpha)\%$ confidence sets is unwise. Hwang's example is similar to our Chow and Shao example in Section 5.2. Both examples use ellipsoidally shaped confidence sets. In our example, the alternative hypothesis region has a conical shape. In Hwang's example, the alternative hypothesis region is a hypercube. In both examples, the resulting test can be described in terms of a finite number of inequalities involving t statistics. And, in both examples, the resulting test is very conservative; the size of the test is much less than α. The general conclusion that one can draw from these two examples is that, when defining a test in terms of a confidence set, the confidence set should have the same shape as the alternative hypothesis region.

Hwang did not point out one interesting feature of his example. The test that he derives from the confidence ellipsoid, corrected to be size-α, is the IUT combination of size-α TOST's that we describe in Section 6.2. Using the correct t-distribution critical value, as Hwang describes in the last display of his comment, his test becomes, reject H_0 if

$$\left|\overline{X}^{(i)}\right| < \Delta - t_{\alpha,d}\sqrt{S_{ii}/d} \quad \text{for all } i.$$

This is the same as reject H_0 if $T_{Li} > t_{\alpha,d}$ and $T_{Ui} < -t_{\alpha,d}$, for all i, where

$$T_{Li} = \left(\overline{X}^{(i)} - (-\Delta)\right)/\sqrt{S_{ii}/d}$$

and

$$T_{Ui} = \left(\overline{X}^{(i)} - \Delta\right)/\sqrt{S_{ii}/d}.$$

For each i, this defines the size-α TOST for the ith parameter, and the requirement that all the TOST's reject is the IUT combination. So Hwang's example is another case in which the IUT combination of size-α tests yields a reasonable, size-α test in a bioequivalence problem. We mention in Section 6.2 that a uniformly more powerful, size-α test may be obtained by using our new test or the BHM test, rather than the TOST, for each of the p coordinates.

The confidence set described by Hwang does have one advantage over rectangular confidence sets in that its shape indicates the correlations among the variables. Thus, when the number of variables is two or three (e.g., AUC and $C_{\max}$), displaying the confidence ellipsoid may be useful, but the confidence ellipsoid does not appear useful for constructing bioequivalence tests.

5. MINOR COMMENTS

Two other points made by the discussants deserve brief comment.

Meredith and Heise thought we confused Type I and Type II errors in Section 4.2. The paragraph beginning "Due to the seriousness..." immediately follows a description of the Anderson and Hauck test. This test is more powerful than the TOST, but it is liberal; its Type I error probability is greater than α. Our next sentence meant that it was unacceptable to have a more powerful test at the expense of having size greater than α. *Due to the seriousness of a Type I error*, it is important that any proposed more powerful test strictly maintains the Type I error probability at α. That is, the *consumer's risk* is the overwhelming concern to the regulatory agency. The discussion of equivalence problems in Berger (1982) was explicitly in terms of consumer's risk.

To us, Meredith and Heise's comment that nonstatistical consumers will find the 90% equivariant confidence interval more informative than the 95% nonequivariant confidence interval for *estimation* confuses point estimation with interval estimation, the two not being mutually exclusive. We see no reason why a point estimate cannot be given

along with the equivalence confidence interval if the former is of interest, in which case, the reason for the failure to conclude bioequivalence in their example becomes apparent.

ACKNOWLEDGMENT

Again, we thank the Editors of *Statistical Science* for arranging for the illuminating discussions of our article.

ADDITIONAL REFERENCES

BRAITMAN, L. E. (1991). Confidence intervals assess both clinical significance and statistical significance. *Annals of Internal Medicine* **114** 515–517.

CHEN, M. L. (1997). Individual bioequivalence—a regulatory update. *Journal of Biopharmaceutical Statistics* **7** 5–11.

CHOW, S.-C. (1996). Statistical consideration for replicated crossover design. In *Proceedings of the FIP BIO International '96*. To appear.

CHOW, S.-C. (1997). Guest editor's note: recent issues in bioequivalence trials. *Journal of Biopharmaceutical Statistics* **7** 1–4.

CHOW, S.-C. and LIU, J. P. (1995). Current issues in bioequivalence trials. *Drug Information Journal* **29** 795–804.

CPMP/ICH/380/95 (1993). *Stability Testing Guidelines: Stability Testing of New Drugs and Products*. CPMP (Committee for Propritary Medical Products), European Agency for the Evaluation of Medical Products, London.

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

FDA (1995). *Guidance—Topical Dermatologic Corticosteroids: In Vivo Bioequivalence*. Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Government Printing Office, Washington, DC.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

HAUCK, W. W., HAUSCHKE, D., DILETTI, E., BOIS, F. Y., STEINIJANS, V. W. and ANDERSON, S. (1997). Choice of Student's t or Wilcoxon-based confidence intervals for assessment of average bioequivalence. *Journal of Biopharmaceutical Statistics*. To appear.

HAUSCHKE, D., STEINIJANS, V. W., DILETTI, E. and BURKE, M. (1992). Sample size determination for bioequivalence assessment using a multiplicative model. *Journal of Pharmacokinetics and Biopharmaceutics* **20** 557–561.

HUQUE, M. and DUBEY, S. D. (1990). A three arm design and analysis for clinical trials in establishing therapeutic equivalence with clinical endpoints. In *Proceedings of the Biopharmaceutical Section* 91–98. Amer. Statist. Assoc., Alexandria, VA.

LASKA, E. M. and MEISNER, M. J. (1986). Testing whether an identified treatment is best: the combination problem. In *Proceedings of the Biopharmaceuticals Section* 163–170. Amer. Statist. Assoc., Alexandria, VA.

LASKA, E. M. and MEISNER, M. J. (1989). Testing whether an identified treatment is best. *Biometrics* **45** 1139–1151.

LIU, J. P. (1995). Use of the repeated cross-over designs in assessing bioequivalence. *Statistics in Medicine* **14** 1067–1078.

LIU, J. P. and CHOW, S.-C. (1992b). Sample size determination for the two ones-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **20** 101–104.

LIU, J. P. and WENG, C.-S. (1992). Estimation of direct formulation effect under log-normal distribution in bioavailability/bioequivalence studies. *Statistics in Medicine* **11** 881–896.

LIU, J. P. and WENG, C.-S. (1993). Evaluation of parametric and nonparametric two one-sided tests procedures for assessing bioequivalence of average bioavailability. *Journal of Biopharmaceutical Statistics* **3** 85–102.

LIU, J. P. and WENG, C.-S. (1994). Evaluation of log-transformation in assessing bioequivalence. *Comm. Statist. Theory Methods* **23** 421–434.

LOCKE, C. S. (1990). Use of a more general model for bioavailability studies. *Comm. Statist. Theory Methods* **19** 3361–3373.

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

ROUSSAS, G. G. (1973). *A First Course in Mathematical Statistics*. Addison-Wesley, Reading, MA.

SCHUIRMANN, D. J. (1987b). A compromise test for equivalence of average bioavailability. In *Proceedings of the Biopharmaceutical Section* 137–142. Amer. Statist. Assoc., Alexandria, VA.

SIMON, R. (1986). Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine* **105** 429–435.

WANG, W., DASGUPTA, A. and HWANG, J. T. (1996). Statistical tests for multivariate bioequivalence. Technical report, Dept. Statistics, Cornell Univ.

WENG, C. S. and LIU, J. P. (1994). Some pitfalls in sample size estimation for an anti-infective study. In *Proceedings of the Biopharmaceutical Section* 56–60. Amer. Statist. Assoc., Alexandria, VA.