Journal of Pharmacokinetics and Biopharmaceutics, Vol. 24, No. 1, 1996

Towards a Practical Strategy for Assessing Individual Bioequivalence

Robert Schall[1,4] and Roger L. Williams[2] for the Food and Drug Administration Individual Bioequivalence Working Group[3]

Received August 1, 1995—Final February 21, 1996

Bioequivalence of two drug formulations is currently defined by drug regulatory authorities in terms of the mean responses following administration of the test and reference formulations (average bioequivalence). However, the various potential shortcomings of average bioequivalence are now understood, and switchability, and thus individual bioequivalence, has become a reasonable expectation when changing from one pharmaceutically equivalent drug product to another. Progress has been made in developing criteria for individual bioequivalence, and an overview and classification of most of the different approaches to the assessment of individual bioequivalence have been achieved. As a consequence of this classification, the different character of scaled and unscaled bioequivalence measures has been recognized and, in turn, this leads to the proposal, made in this paper, of using both scaled and unscaled criteria for bioequivalence assessment of different classes of drugs, depending on their within-subject variability and therapeutic range. This strategy addresses the shortcomings of average bioequivalence, and, when applied to data sets from bioequivalence studies with four-period replicate crossover designs, turns out to have some satisfactory properties. Open questions and areas for further research are discussed.

KEY WORDS: bootstrap; individual bioequivalence; moment-based criteria; population bioequivalence; probability-based criteria; scaled criteria; unscaled criteria.

INTRODUCTION

Bioequivalence of two drug formulations is currently defined by drug regulatory authorities in terms of the mean responses following

[1] Division of Biometry, Department of Pharmacology, University of the Orange Free State, P.O. Box 339 (G6), Bloemfontein 9300, South Africa.
[2] CDER, Food and Drug Administration, 5600 Fishers Lane, Rockville, Maryland 20587.
[3] Members of the working group: Mei-Ling Chen and Rabi Patnaik (Co-chairmen), Fred Balch, Keith Chan, Larry Lesko, Tom Ludden, Stella Machado, Donald Schuirmann, and Roger Williams.
[4] To whom correspondence should be addressed.

0090-466X/96/0200-0133$09.50/0

administration of the test and reference formulations (1,2). Specifically, for two formulations to be bioequivalent the "test/reference" ratio of the geometric means (in the population) of the bioavailability characteristic under investigation must lie in the bioequivalence range of 80–125%. The definition of bioequivalence in terms of the means of bioavailability characteristics has been termed "average bioequivalence."

It has been recognized that the concept of average bioequivalence may be deficient. Hwang *et al.* (3) pointed out that (average) bioequivalence may not imply switchability of formulations. More recently, Anderson and Hauck (4), and independently Ekbohm and Melander (5), initiated discussion and research in the areas of population and individual bioequivalence. New approaches for the assessment of bioequivalence have also been considered (6–12) and additional work in the field continues to appear.

With the development of criteria and statistical methods to assess individual bioequivalence, a new goal is now apparent: to choose, among the several different new criteria and methods a methodology that both satisfactorily addresses the potential shortcomings of average bioequivalence and can be practically implemented. In this paper, we discuss this goal, while focusing on the case of individual bioequivalence (a parallel development is possible for the case of population bioequivalence), along the following lines:

First, it is useful to bring some order into the collection of various methods for the assessment of individual bioequivalence that have been proposed in recent years. Here we use a cross-classification of bioequivalence criteria into moment-based vs. probability-based criteria, and unscaled vs. scaled criteria (13). This yields four different approaches to assess individual bioequivalence.

Second, we discuss how far the four different classes of bioequivalence criteria satisfactorily address the shortcomings of average bioequivalence.

Third, we propose a methodology for the assessment of individual bioequivalence that has both some satisfactory properties and the potential to be practically implemented.

In this analysis, it is apparent that no single class of bioequivalence criteria can satisfy both the shortcomings of average bioequivalence and the requirements of practicality. The solution is to use a combination of an unscaled and a scaled criterion. Selection of one of these two classes of criteria will depend on the variability and the therapeutic range of the drug under investigation. Thus the classification of bioequivalence criteria into unscaled and scaled criteria corresponds to a classification of drugs into those with low variability vs. drugs with high variability, and into drugs with a narrow therapeutic range vs. drugs with a wide therapeutic range.

CLASSIFICATION OF BIOEQUIVALENCE CRITERIA

Notation and Statistical Model for Bioavailability Data

We use the following model for the bioavailability characteristics Y_T from the test and Y_R from the reference formulation (4,6,7,9):

$$\begin{aligned} Y_T &= \mu_T + b_T + e_T \\ Y_R &= \mu_R + b_R + e_R \end{aligned} \tag{1}$$

Here μ_i is the mean response ($i = R, T$) or population average for the ith formulation, b_i is the mean deviation from the population average of a given individual so that

$$m_i = \mu_i + b_i$$

is the individual's mean response to the test and reference formulation respectively, and e_i represents the within-subject variation, that is, a deviation from the individual's mean response on different occasions. We assume that the b_i are independent from the e_i and denote the within-subject variance of Y_T and Y_R by

$$\text{var}(e_i) = \sigma_{Wi}^2$$

and the between-subject variance by

$$\text{var}(b_i) = \sigma_{Bi}^2$$

For a given subject, b_R and b_T may be correlated, and the variance of the within-subject difference of b_T and b_R is denoted by

$$\text{var}(b_T - b_R) = \sigma_D^2$$

Model 1 is additive. For some bioavailability characteristics, such as the area under the drug concentration vs. time curve (AUC), a multiplicative model is considered appropriate (14). Such data, however, are conventionally analyzed after logarithmic transformation, so that on the log scale an additive model again applies. Thus the random variables Y_T and Y_R are, in the following, meant to represent log-transformed data if bioavailability characteristics such as AUC are analyzed. If Y_T and Y_R represent log-transformed data, then the random variables X_T and X_R represent the untransformed data, that is $Y_T = \log(X_T)$ and $Y_R = \log(X_R)$.

Criterion for Average Bioequivalence

According to current FDA and EC guidelines (1,2) two formulations are bioequivalent if the ratio of the means of the bioavailability characteristics X_T and X_R of the test and reference lies in the interval 0.8 to 1.25,

that is, if

$$0.8 \leq \frac{E(X_T)}{E(X_R)} = \frac{\tau_T}{\tau_R} \leq 1.25 \tag{2}$$

The definition of bioequivalence in terms of mean bioavailability characteristics has recently been referred to as "average bioequivalence" to distinguish it from individual and population bioequivalence.

In terms of the population means μ_T and μ_R of Y_T and Y_R the statistical criterion of average bioequivalence after log transformation is

$$-\Delta_{av} \leq \mu_T - \mu_R \leq \Delta_{av} \tag{3}$$

The constant Δ_{av} in criterion [Eq. (3)] determines the bioequivalence range for average bioequivalence. Definition [Eq. (3)] is consistent with definition [Eq. (2)] when $\Delta_{av} = \log(1.25)$.

Cross-classification of Criteria for Individual Bioequivalence

Most of the new criteria for individual bioequivalence may be cross-classified into moment-based vs. probability-based criteria, and unscaled vs. scaled criteria (13). With this approach, four different classes of criteria for individual bioequivalence exist: moment-based unscaled; moment-based scaled; probability-based unscaled; and probability-based scaled. Four bioequivalence measures, respectively representing each of these four classes, are listed in Table I, together with the relevant publications. Most of the

Table I. Measures for Individual Bioequivalence

Moment-based, unscaled[a]

$$M_{mu} = (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2$$

Moment-based, scaled[b]

$$M_{ms} = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2}{\sigma_{WR}^2}$$

Probability-based, unscaled[c]

$$M_{pu} = Pr[|m_T - m_R| \leq r], \qquad r = \log(1.25) \text{ (for example)}$$

Probability-based, scaled[d]

$$M_{ps} = Pr\left[\frac{|Y_T - Y_R|}{\sigma_{WR}} \leq \gamma\right], \quad \gamma = \sqrt{2} \text{ (for example)}$$

[a] Schall and Luus (7), Holder and Hsuan (9) (a special case).
[b] Sheiner (6), Schall and Luus (7), Ekbohm and Melander (5) (a special case), Endrenyi (8) (a special case).
[c] Anderson and Hauck (4).
[d] Schall (11), Hauck and Anderson (12).

recently published approaches to bioequivalence assessment are covered by this cross-classification. The corresponding criteria (definitions of bioequivalence) for individual bioequivalence are:

1. Moment-based, unscaled

$$M_{\text{mu}} = (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 \leq \Delta_{\text{mu}}^2 \tag{4}$$

2. Moment-based, scaled

$$M_{\text{ms}} = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2}{\sigma_{WR}^2} \leq \Delta_{\text{ms}}^2 \tag{5}$$

3. Probability-based, unscaled

$$M_{\text{pu}} = Pr[|m_T - m_R| \leq r] \geq MINP_u \tag{6}$$

4. Probability-based, scaled

$$M_{\text{ps}} = Pr\left[\frac{|Y_T - Y_R|}{\sigma_{WR}} \leq \gamma\right] \geq MINP_s \tag{7}$$

The constants Δ_{mu}^2, Δ_{ms}^2, $MINP_u$ and $MINP_s$ determine the bioequivalence range for the bioequivalence measure in question. (These constants are understood to be fixed numbers to be determined by regulatory authorities.) A bioequivalence measure together with a bioequivalence range is called a bioequivalence criterion.

The above cross-classification is quite general, but the four bioequivalence measures and corresponding criteria [Eqs. (4)–(7)] should be viewed as examples of criteria from those categories. Alternative bioequivalence measures from each of those classes might yet emerge.

DO NEW BIOEQUIVALENCE CRITERIA MEET SHORTCOMINGS OF AVERAGE BIOEQUIVALENCE?

Shortcomings of Average Bioequivalence

Average bioequivalence focuses only on the population averages of bioavailability characteristics of the test and reference, and from this fact two potential shortcomings follow (4): (i) by considering only averages, differences between the test and reference in the variability of bioavailability characteristics, and, in general, differences in their distribution, are not taken into account; and (ii) by considering only population averages one does not necessarily ensure that the bioavailability of the test and reference is close within individuals.

A third shortcoming of average bioequivalence as it is conventionally applied has been mentioned in the context of highly variable drugs; it seems reasonable to have a wider bioequivalence range (in terms of the mean difference $\mu_T - \mu_R$) for highly variable drugs than for drugs with relatively low variation (Generic Drugs Advisory Committee, Silver Spring, February 1993): (iii) the bioequivalence range (Δ_{av}) for average bioequivalence is fixed; this does not take into account the "natural" within-formulation variability.

To remedy these potential shortcomings of average bioequivalence, a criterion for individual bioequivalence should have the following properties:

1. The criterion must take into account more than population means; ideally it will compare distributions of bioavailabilities, but as a first step it should take into account variances in addition to means.
2. The new criterion must ensure switchability; it must not only ensure that means are close, but that individual bioavailabilities are close when switching from one formulation to another.
3. In particular for highly variable drugs, the bioequivalence range for the new criterion should reflect the high variability; in effect, one should have a wider bioequivalence range (for $\mu_T - \mu_R$, and more generally, for the moment-based unscaled measure M_{mu}) for highly variable drugs than for drugs with moderate or low variability.

Properties of Moment-Based Criteria

Details about moment-based bioequivalence measures can be found in the relevant publications (see Table I). A review of these measures and of relationships and differences between them is given by Schall (13). Here we consider whether the new criteria satisfy the requirements listed above.

Because $(\mu_T - \mu_R)^2 + \sigma_D^2 = E(m_T - m_R)^2$ the moment-based unscaled bioequivalence measure M_{mu} can be written as

$$M_{mu} = E(m_T - m_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2$$

Thus the measure will be small, and indicate individual bioequivalence, if the individual means m_T and m_R of test and reference are close, and if the within-subject variance of the test is close to the reference. The individual means are close if the population means of test and reference are close, and if σ_D^2 is small; σ_D^2 is the variance of the difference $m_T - m_R$ between the individual means and is related to the subject by formulation interaction (15). We note that for a given mean difference $\mu_T - \mu_R$, and given σ_D^2, it becomes easier to satisfy the bioequivalence criterion as the variance of the test formulation becomes smaller relative to the variance of the reference

formulation. In most cases, this is clinically desirable and encourages formulations with small variance. Because of these properties, the moment-based unscaled criterion satisfies requirements 1 and 2 set out above: The criterion takes into account more than population means, namely, also the within-subject variances; and it assesses switchability by measuring the closeness of individual means rather than merely the closeness of population means.

The moment-based unscaled criterion does not, however, satisfy the third requirement. This can be seen most easily by considering the special case when $\sigma_D^2 = 0$ and $\sigma_{WT}^2 = \sigma_{WR}^2$. In this case the criterion reduces to $M_{mu} = (\mu_T - \mu_R)^2$, which is essentially the criterion for average bioequivalence. The bioequivalence range Δ_{mu}^2 for $M_{mu} = (\mu_T - \mu_R)^2$ is fixed and does not depend on the variability of the drug under investigation.

In contrast, the moment-based scaled criterion satisfies requirement 3. The numerator of the moment-based scaled criterion is the same as that of the moment-based unscaled criterion, but the denominator is the within-subject variance of the reference. One declares bioequivalence using the moment-based scaled criterion if $M_{ms} = M_{mu}/\sigma_{WR}^2 \leq \Delta_{ms}^2$, which is equivalent to $M_{mu} \leq \sigma_{WR}^2 \cdot \Delta_{ms}^2$. Thus, using the moment-based scaled criterion is equivalent to using the moment-based unscaled criterion, but with a bioequivalence range that is proportional to the within-subject variance of the reference formulation.

Properties of Probability-Based Criteria

The probability-based unscaled measure is the probability that the individual means of test and reference are close. As such it ensures switchability, as far as the closeness of the individual means is concerned, but it does not take into account possible differences in the within-subject variability of the test and reference. Thus the measure satisfies requirement 2, but neither requirement 1 or 3.

The probability-based scaled measure is the probability that the scaled difference between the test and reference bioavailability is small. Schall (11) has shown that this measure is closely related to the moment-based scaled measure. Thus the probability-based scaled measure satisfies all three requirements set out above.

A PRACTICAL STRATEGY FOR ASSESSING INDIVIDUAL BIOEQUIVALENCE

In view of the cross-classification of bioequivalence criteria outlined above, the choice is between moment-based and probability-based criteria, and between scaled and unscaled criteria.

Moment-based Versus Probability-based

The choice between moment-based and probability-based criteria is largely a matter of interpretability of the criterion, statistical convenience, and preference (13). The moment-based measures express the test/reference and reference/reference discrepancy in terms of expected squared differences, while the probability-based measures express the test/reference and reference/reference closeness in terms of probabilties. Holder and Hsuan (9) have pointed out the close relationship between the moment-based unscaled and the probability-based unscaled measures, and similarly, Schall (11) has shown that the moment-based scaled and the probability-based scaled measures are equivalent under certain assumptions. It is unlikely, therefore, that a bioequivalence assessment using moment-based measures would come to radically different conclusions from an assessment using probability-based measures.

Below we suggest a mixed strategy (using both scaled and unscaled bioequivalence measures) to assess bioequivalence. Because we have a completely satisfactory probability-based criterion only for the scaled, but not for the unscaled, case, the moment-based criteria allow full implementation of the mixed scaled/unscaled strategy, as proposed below. Thus, for the moment, we prefer and concentrate on the moment-based measures; once a satisfactory probability-based unscaled measure becomes available to complement the probability-based scaled measure, the mixed scaled/scaled could also be implemented using probability-based measures.

Scaled Versus Unscaled: Mixed Strategy for Individual Bioequivalence

The choice between a scaled and an unscaled criterion is fundamental and seemingly difficult to make. Scaled criteria are attractive because, as pointed out above, they alone satisfy all three requirements for an appropropriate bioequivalence criterion, in particular, only they lead to wider bioequivalence ranges for highly variable drugs, which is desirable. However, they lead to narrower bioequivalence ranges (narrower than the unscaled criterion) for drugs with low variability, which is not always desirable. For example, for a drug with low variability but wide therapeutic window it makes no sense to use the scaled criterion which would effectively make the bioequivalence range narrower than that used at present. Thus a dilemma exists; only scaled criteria satisfy all three requirements on new bioequivalence criteria, but use of a scaled criterion in all cases implies use of bioequivalence ranges that, for some drugs, are neither practical nor justifiable on scientific grounds.

The surprisingly simple solution to this dilemma is that either criterion can be used, depending on the variability and the therapeutic window of the

drug under investigation. According to this approach, four general categories of drugs are possible: low variability/narrow therapeutic range; low variability/wide therapeutic range; high variability/wide therapeutic range; and high variability/narrow therapeutic range. The scaled and unscaled approaches would then be used in the different categories in the following way:

Class 1. Low intrasubject variability/high toxicity (narrow therapeutic range): For these drugs, scaling would always be used. This approach causes the new formulation to meet stringent standards relative to the reference formulation, more stringent in general than the current standard of average bioequivalence. Drugs that fall into this category might be ones that carry a recommendation to adjust dose based on plasma/blood concentration levels. Examples include: theophylline, digoxin, certain antiepileptic drugs (phenytoin), and certain antiarrhythmic drugs.

Class 2. Low intrasubject variability/low toxicity (wide therapeutic range): The unscaled approach would be used. Most drugs either fall in this, or in the following category.

Class 3. High intrasubject variability/low toxicity (wide therapeutic range): For these drugs, the scaled approach is used. Effectively, a wider bioequivalence range than at present is allowed for this class of drugs.

Class 4. High variability/high toxicity (narrow therapeutic range): This category is likely to have few if any members because drugs with high variability and narrow therapeutic range will pass through safety and efficacy trials only with difficulty (16). For drugs in this category the unscaled approach would be used.

In practice, one could proceed as follows: for drugs in Class 1 only the scaled approach is used; for drugs in classes 2 and 3 a scientifically and practically satisfactory definition of bioequivalence is that two formulations are bioequivalent if they satisfy either the unscaled, or the scaled criterion. With this mixed strategy, and an appropriate choice of bioequivalence range for the scaled and unscaled criteria (see the next section) the definition of bioequivalence is never stricter for Classes 2 and 3 than at present (as far as the bioequivalence range is concerned), even for drugs with low variability, but is less strict than at present for highly variable drugs with wide therapeutic range. Finally for drugs in Class 4 only the unscaled approach is used.

Bioequivalence Ranges

As pointed out above, in the special case when $\sigma_D^2 = 0$ and $\sigma_{WT}^2 = \sigma_{WR}^2$ the criterion reduces to $M_{mu} = (\mu_T - \mu_R)^2$, which is essentially the criterion for average bioequivalence for which the bioequivalence range

is $\Delta^2_{av} = [\log(1.25)]^2$. By analogy with the case of average bioequivalence, the bioequivalence range for M_{mu} can thus be chosen as $\Delta^2_{mu} = \Delta^2_{av} = [\log(1.25)]^2$.

The bioequivalence range for the moment-based scaled criterion can be considered through an analysis of the shape of the bioequivalence region (Fig. 1) of the mixed strategy outlined above, where the acceptable values (indicating bioequivalence) for M_{mu} are indicated as a function of σ^2_{WR}, the within-subject variance of the reference. The line $M_{mu} = \Delta^2_{mu}$ indicates the bioequivalence range for the unscaled criterion; this line runs parallel to the σ^2_{WR} axis at a distance of Δ^2_{mu}. That means that all values of $M_{mu} \leq \Delta^2_{mu}$ (actually independent of σ^2_{WR}) indicate bioequivalence. The line $M_{mu} = \Delta^2_{ms} \cdot \sigma^2_{WR}$ indicates the bioequivalence region for the scaled criterion; this line runs through the origin with a slope of Δ^2_{ms}. This means that all values of $M_{mu} \leq \Delta^2_{ms} \cdot \sigma^2_{WR}$ indicate bioequivalence, which is equivalent to $M_{ms} = M_{mu}/\sigma^2_{WR} \leq \Delta^2_{ms}$, which in turn is the moment-based scaled criterion.

The union of the two areas circumscribed by the lines $M_{mu} = \Delta^2_{mu}$ and $M_{mu} = \Delta^2_{ms} \cdot \sigma^2_{WR}$ is the total bioequivalence region of the mixed strategy, because bioequivalence is given when either the unscaled or the scaled criterion are satisfied.

The line $M_{mu} = \Delta^2_{mu}$ crosses the line $M_{mu} = \Delta^2_{ms} \cdot \sigma^2_{WR}$ at σ^2_{W0}. This is the within-subject variance of the reference formulation from which point on the scaled criterion has effectively a wider bioequivalence range than the unscaled criterion. By fixing σ^2_{W0}, the slope of line $M_{mu} = \Delta^2_{ms} \cdot \sigma^2_{WR}$, and thus Δ^2_{ms} is fixed through the relationship $\Delta^2_{ms} = \Delta^2_{mu}/\sigma^2_{W0}$. Thus the choice of the bioequivalence range Δ^2_{ms} for the moment-based scaled criterion effectively is made through the choice of σ^2_{W0}.

This choice is a regulatory decision, but one can get an idea what σ^2_{W0} should be by recalling the motivation for introducing the scaled approach: the scaled approach is used because this effectively leads to a wider bioequivalence range for highly variable drugs. The within-subject variance of the reference from which point on the scaled criterion has indeed a wider bioequivalence range than the unscaled criterion is σ^2_{W0}. Conventionally, a within-subject coefficient of variation of 30% has been viewed as 'highly variable,' which would imply that σ_{W0} would not be larger than about 0.3 (note that a coefficient of variation of 30% for the bioavailability characteristic on the original scale corresponds to a standard deviation of about 0.3 on the logarithmic scale). Because at a CV of 30% one would like to have a bioequivalence range that is already rather wider than the range for the unscaled criterion, σ_{W0} might be chosen smaller than 0.3.

Another approach to assign a value for σ_{W0} would be to recall that the scaled approach will also be used for drugs in Class 1, namely, low variability/narrow therapeutic range, for which drugs the bioequivalence

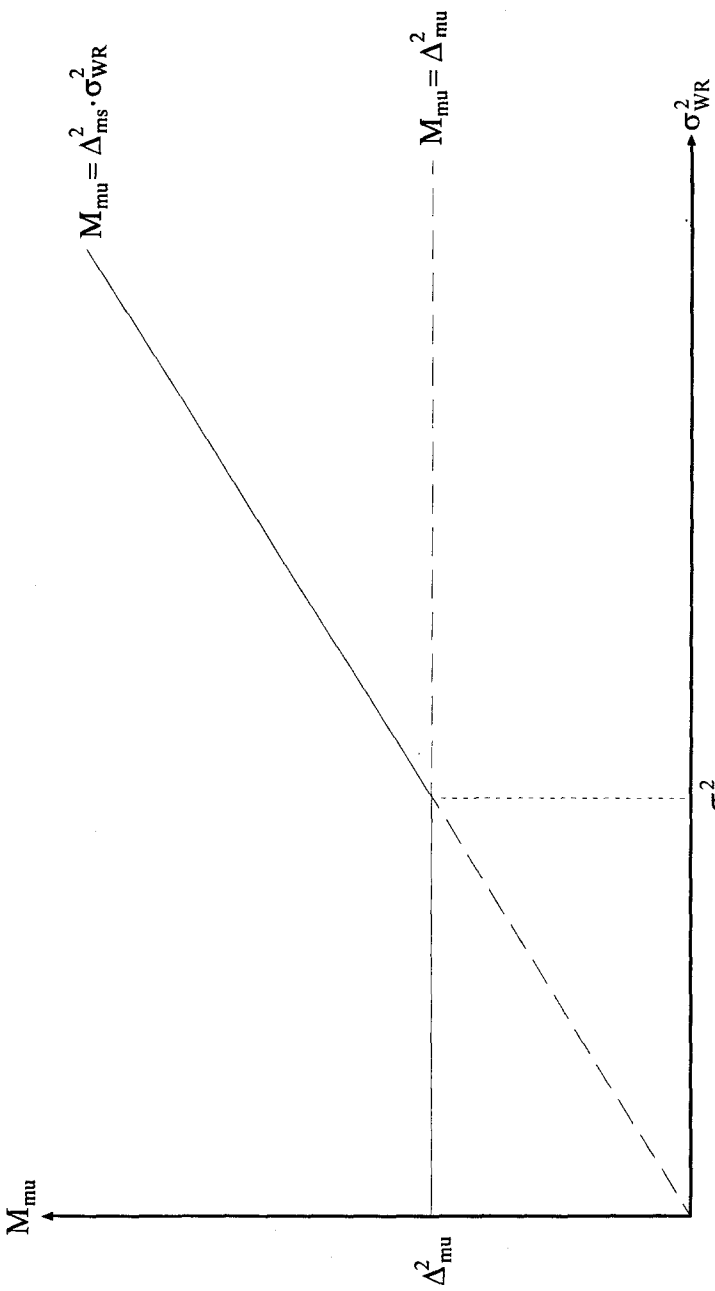

Fig. 1. Bioequivalence region of the mixed strategy.

range should be no wider than at present. Thus σ_{W0} would have to be greater than or equal to the within-subject variability of narrow therapeutic range drugs such as theophylline, digoxin and phenytoin.

STATISTICAL METHODS

As with the analysis of average bioequivalence, one can, in the analysis of individual bioequivalence, use confidence intervals to derive statistical decision rules. This involves constructing a one-sided $100 \cdot (1-\alpha)\% = 95\%$ confidence interval for M_{mu} (M_{ms}). If the upper bound of the interval is less than or equal to Δ^2_{mu} (Δ^2_{ms}), bioequivalence is declared.

The problem of statistically assessing bioequivalence is therefore essentially one of constructing one-sided confidence intervals for the bioequivalence measures M_{mu} and M_{ms}. However, this is not easy, not even in the seemingly simple case of average bioequivalence where, strictly, one is required to construct a one-sided confidence interval for $|\mu_T - \mu_R|$. Exact, let alone closed-form, solutions for constructing confidence intervals for the bioequivalence measures discussed in this paper are not available.

Various approximate methods, however, have been proposed for the construction of confidence intervals for new bioequivalence measures. These include (i) approximate F statistics (5); (ii) maximum likelihood based methods (6); (iii) bootstrap methods (7,11). At present, a thorough investigation with respect to power (sample size requirements) and significance levels (actual coverage probability of confidence intervals) seems to have been carried out only for the bootstrap method (11). Additional research in the area of statistical methods for the construction of confidence intervals (or hypothesis tests) for the different bioequivalence measures is necessary.

EXAMPLES OF APPLICATION

The moment-based approaches for the assessment of individual bioequivalence were applied to data from three bioequivalence studies that involved moderately to highly variable drugs (for reasons of confidentiality the drug names cannot be given). In each study, a two-treatment, four-period cross-over design was used.

Table II reports summary statistics relating to the assessment of average and individual bioequivalence of this data, namely estimates of the test/reference mean ratio τ_T/τ_R, σ_D, σ_{WR}, and σ_{WT}. Furthermore, a two-sided 90% confidence interval (CI) for τ_T/τ_R is reported, as well as one-sided 95% CI for M_{mu} and M_{ms}. The estimates of τ_T/τ_R, σ_D, σ_{WR}, and σ_{WT}, and the 90% CI for τ_T/τ_R, were calculated from an ANOVA of the log-transformed data (AUC and C_{max} as indicated), as is conventional. For the purposes of

Table II. Summary Statistics for the Assessment of Average and Individual Bioequivalence Using Four-Period Replicate Design Studies

Study variable	N	Estimate of τ_R/τ_R	σ_D	σ_{WR}	σ_{WT}	Average BE 90% *CI* for τ_T/τ_R(%)	Individual BE: Unscaled 95% *CI* for M_{mu}(%)	Individual BE: Scaled 95% *CI* for M_{ms}(%)
Study 1								
AUC	38	0.96	0.00	0.43	0.26	88–106	115	105
AUC[a]	38	1.04	0.00	0.26	0.43	94–114	160	137
Study 2								
C_{max}	24	0.99	0.00	0.52	0.43	84.116	141	113
Study 3								
AUC	34	1.08	0.27	0.21	0.22	101–117	144	140

[a]Test and reference labels reversed.

these examples, the *CI* for M_{mu} and M_{ms} were calculated using the bias-corrected bootstrap method applied to log-transformed data (7,11), and the bioequivalence ranges for M_{mu} and M_{ms} were taken to be $\Delta^2_{mu} = [\log(1.25)]^2$, and $\Delta^2_{ms} = \Delta^2_{mu}/\sigma^2_{W0}$ with $\sigma_{W0} = 0.15$.

For convenience, the *CI* for M_{mu} and M_{ms} are reported on the percentage scale which can be motivated as follows: Using the unscaled approach bioequivalence is declared if the upper bound UB_{mu} of a one-sided 95% confidence interval for M_{mu} is less than or equal to $\Delta^2_{mu} = [\log(1.25)]^2$. This is equivalent to

$$UB'_{mu} = 100\% \cdot \exp(\sqrt{UB_{mu}}) \leq 125\%$$

which suggests reporting UB'_{mu} instead of UB_{mu}, and declaring bioequivalence if UB'_{mu} is less than or equal to 125%. In this way the results of the statistical analysis are reported on a familiar scale. Similarly, using the scaled approach bioequivalence is declared if the upper bound UB_{ms} of a one-sided 95% confidence interval for M_{ms} is less than or equal to $\Delta^2_{ms} = [\log(1.25)]^2/0.15^2$. This is equivalent to

$$UB'_{ms} = 100\% \cdot \exp(0.15 \cdot \sqrt{UB_{ms}}) \leq 125\%$$

Thus we report UB'_{ms} instead of UB_{ms} and declare bioequivalence if UB'_{ms} is less than or equal to 125%.

For Drug 1, an analysis of the variable *AUC* is reported (see Table II). We note that the estimate of the test/reference mean ratio is close to 1, and the two formulations satisfy the conventional criterion for average bioequivalence. Furthermore, $\hat{\sigma}_D$ is zero, and $\hat{\sigma}_{WT}$ is smaller than $\hat{\sigma}_{WR}$ which should favor the test formulation. Thus one should expect that the two formulations satisfy the criterion for individual bioequivalence, which is indeed the case, using either the scaled or unscaled approach.

To illustrate what would happen if the within-test variability were greater than the within-reference variability, the labels of test and reference were exchanged for this data. This leaves the results for average bioequivalence essentially unchanged, but the two formulations no longer satisfy the criteria for individual bioequivalence. The test formulation is penalized for being more variable than the reference formulation.

For Drug 2, an analysis of the variable C_{max} is reported. Again the estimate of the test/reference mean ratio is close to 1, the two formulations satisfy the criterion for average bioequivalence, $\hat{\sigma}_D$ is zero, and $\hat{\sigma}_{WT}$ is smaller than $\hat{\sigma}_{WR}$. C_{max} for this drug is highly variable, so that one should expect that it is easier to show individual bioequivalence using the scaled approach than using the unscaled approach. This is indeed the case; the two formulations can be declared bioequivalent using the scaled approach, but not using the unscaled approach. If this is a drug from Class 3 (high variability/wide therapeutic range) one would be justified to accept individual bioequivalence according to the "mixed strategy."

For Drug 3, an analysis of the variable AUC is reported. The estimate of the test/reference mean ratio is fairly close to 1, and the two formulations satisfy the criterion for average bioequivalence. However, $\hat{\sigma}_D$ is large, greater than either $\hat{\sigma}_{WT}$ or $\hat{\sigma}_{WR}$, which are similar. This is an example of a test formulation that satisfies the criterion for average, but not individual bioequivalence; the test formulation fails because of the large subject by formulation interaction.

DISCUSSION: PROGRESS MADE AND SOME OPEN QUESTIONS

Progress Made

In recent years, several steps have been achieved in moving towards a more appropriate assessment of bioequivalence.

First, the various potential shortcomings of average bioequivalence are now understood (3–5). Switchability, and thus individual bioequivalence, has become a reasonable expectation when changing from one pharmaceutically equivalent drug product to another.

Second, progress has been made in developing new criteria for individual bioequivalence (see Table I).

Third, an overview and classification of most of the different approaches to the assessment of individual and population bioequivalence has been achieved (13). As a consequence of this classification, the different character of scaled and unscaled bioequivalence measures has been recognized and, in turn, leads to the proposal, made in this paper, of using both scaled and unscaled criteria for bioequivalence assessment of different classes of drugs, depending on their within-subject variability and therapeutic range.

Finally, and encouragingly, practical application of the mixed strategy proposed in this paper suggests that it has some satisfactory properties. Perhaps most important, the examples show that it is indeed possible to demonstrate individual bioequivalence with sample sizes between 24 and 38. The scaled approach in particular seems suitable for the analysis of highly variable drugs. The proposed approach clearly rewards a test formulation for being less variable than the reference formulation, and it penalizes a test formulation for being more variable than the reference formulation. Examples exist where a test and reference formulation satisfy the criterion for average bioequivalence, but not for individual bioequivalence due to large formulation by subject interaction.

Open Question 1: Bioequivalence Ranges

The regulatory authorities must determine bioequivalence ranges for M_{mu} and M_{ms}, which involves fixing Δ^2_{mu} and σ_{W0}. There should be little controversy about choosing $\Delta^2_{mu} = \Delta^2_{av} = [\log(1.25)]^2$. There may be more discussion about the value of σ_{W0}. In this paper we suggest that σ_{W0} should be smaller than 0.3; it might be chosen to lie between 0.15 and 0.25.

Open Question 2: Statistical Implementation of Mixed Strategy

For drugs in Classes 2 and 3 [that is, low intrasubject variability/usual toxicity (wide therapeutic range), and high intrasubject variability/low toxicity (wide therapeutic range)], which comprise the majority of drugs, one can demonstrate bioequivalence statistically using either the unscaled or the scaled criterion. This is statistically acceptable if the choice of criterion is specified a priori (in the study protocol). The choice could be made based on prior information about the within-subject variability of the drug under investigation: One would choose the unscaled criterion for drugs known to have low variability, and the scaled criterion for drugs known to have high variability.

A more liberal approach would be to use both criteria, and declare bioequivalence if either criterion is satisfied. This approach might be slightly anticonservative statistically, but should for all practical purposes be acceptable. Research into this question is needed. Even the former approach, which is statistically conservative and thus somewhat stricter than the latter, allows one to show bioequivalence with reasonable sample sizes as suggested by simulation studies currently in progress.

Consequences

A consequence of adopting a regulatory requirement of individual bioequivalence is that three- or four-period designs will be required for

bioequivalence studies. Preliminary simulation studies (11) suggest that samples of between 20 and 40 subjects may be needed to show bioequivalence with reasonable power (more than 70 to 80%). Thus the samples will generally not be significantly larger than those used at present in bioequivalence studies. For highly variable drugs the sample sizes might actually be lower than at present. However, the study size (number of subject times number of drug applications) will increase because of the requirement for three- or four-period studies.

ACKNOWLEDGMENTS

The authors thank Walter Hauck for serving as an expert consultant to the working group, and Walter Hauck and two referees for comments on this paper. The views expressed in this article are those of the authors and do not necessarily represent the official view of the U.S. Food and Drug Administration.

REFERENCES

1. Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design, Guidance, Division of Bioequivalence, Office of Generic Drugs, Food and Drug Administration (revision date: 1 July 1992).
2. Investigation of bioavailability and bioequivalence. In *The Rules Governing Medicinal Products in the European Community*, Vol. 2, Addendum No. 2, Office for Official Publications of the European Communities, Luxemburg, May 1992, pp. 149–168.
3. S. Hwang, P. B. Huber, M. Hesney and K. C. Kwan. Bioequivalence and interchangeability (letter). *J. Pharma. Sci.* **67**:IV (1978).
4. S. Anderson and W. W. Hauck. Consideration of individual bioequivalence. *J. Pharmacokin. Biopharma.* **18**:259–273 (1990).
5. G. Ekbohm and H. Melander. *On Variation, Bioequivalence and Interchangeability*, Report 14, Department of Statistics, Swedish University of Agricultural Science, 1990.
6. L. B. Sheiner. Bioequivalence revisited. *Statist. Med.* **11**:1777–1788 (1992).
7. R. Schall and H. G. Luus. On population and individual bioequivalence. *Statis. Med.* **12**:1109–1124 (1993).
8. L. Endrenyi. A procedure for the assessment of individual bioequivalence. In K. K. Midha and H. H. Blume (eds.), *Bio-International. Bioavailability, Bioequivalence and Pharmacokinetics*, Medpharm, Stuttgart, 1993, pp. 141–146.
9. D. J. Holder and F. Hsuan. Moment-based criteria for determining bioequivalence. *Biometrika* **80**:835–846 (1993).
10. J. D. Esinhart and V. M. Chinchilli. Extension to the use of tolerance intervals for the assessment of individual bioequivalence. *J. Biopharm. Statist.* **4**:39–52 (1994).
11. R. Schall. Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics* **51**:615–626 (1995).
12. W. W. Hauck and S. Anderson. Measuring switchability and prescribability: when is average bioequivalence sufficient? *J. Pharmacokin. Biopharma.* **22**:551–564 (1994).
13. R. Schall. A unified view of individual, population, and average bioequivalence. In H. H. Blume and K. K. Midha (eds.), *Bio-International 2. Bioavailability, Bioequivalence and Pharmacokinetic Studies*, Medpharm, Stuttgart, 1995, pp. 91–106.

14. W. J. Westlake. Bioavailability and bioequivalence of pharmaceutical formulations. In K. E. Peace (ed.), *Pharmaceutical Statistics for Drug Development*, Marcel Dekker, New York, 1988, pp. 329–352.
15. G. Ekbohm and H. Melander. The subject-by-formulation interaction as a criterion of interchangeability of drugs. *Biometrics* **45**:1249–1254 (1989).
16. L. Z. Benet. Bioavailability and bioequivalence: Definitions and difficulties in acceptance criteria. In K. K. Midha and H. H. Blume (eds.), *Bio-International. Bioavailability, Bioequivalence and Pharmacokinetics*, Medpharm, Stuttgart, 1993, pp. 27–35.