

ATTACHMENT 3

---

# Guidance for Industry

## **Statistical Approaches to Establishing Bioequivalence**

**U.S. Department of Health and Human Services  
Food and Drug Administration  
Center for Drug Evaluation and Research (CDER)  
January 2001  
BP**

---

# Guidance for Industry

## Statistical Approaches to Establishing Bioequivalence

*Additional copies are available from:*

*Office of Training and Communications  
Division of Communications Management  
Drug Information Branch, HFD-210  
5600 Fishers Lane  
Rockville MD 20857  
(Tel) 301-827-4573*

*(Internet) <http://www.fda.gov/cder/guidance/index.htm>*

**U.S. Department of Health and Human Services  
Food and Drug Administration  
Center for Drug Evaluation and Research (CDER)  
January 2001  
BP**

## Table of Contents

<b>I</b>	<b>INTRODUCTION.....</b>	<b>1</b>
<b>II</b>	<b>BACKGROUND.....</b>	<b>1</b>
	A. GENERAL.....	1
	B. STATISTICAL.....	2
<b>III</b>	<b>STATISTICAL MODEL.....</b>	<b>3</b>
<b>IV</b>	<b>STATISTICAL APPROACHES FOR BIOEQUIVALENCE.....</b>	<b>3</b>
	A. AVERAGE BIOEQUIVALENCE.....	4
	B. POPULATION BIOEQUIVALENCE.....	5
	C. INDIVIDUAL BIOEQUIVALENCE.....	6
<b>V</b>	<b>STUDY DESIGN.....</b>	<b>7</b>
	A. EXPERIMENTAL DESIGN.....	7
	B. SAMPLE SIZE AND DROPOUTS.....	8
<b>VI</b>	<b>STATISTICAL ANALYSIS.....</b>	<b>9</b>
	A. LOGARITHMIC TRANSFORMATION.....	9
	B. DATA ANALYSIS.....	10
<b>VII</b>	<b>MISCELLANEOUS ISSUES.....</b>	<b>13</b>
	A. STUDIES IN MULTIPLE GROUPS.....	13
	B. CARRYOVER EFFECTS.....	13
	C. OUTLIER CONSIDERATIONS.....	14
	D. DISCONTINUITY.....	15
	<b>REFERENCES.....</b>	<b>16</b>
	<b>APPENDIX A.....</b>	<b>21</b>
	<b>APPENDIX B.....</b>	<b>25</b>
	<b>APPENDIX C.....</b>	<b>28</b>
	<b>APPENDIX D.....</b>	<b>32</b>
	<b>APPENDIX E.....</b>	<b>34</b>
	<b>APPENDIX F.....</b>	<b>35</b>
	<b>APPENDIX G.....</b>	<b>40</b>
	<b>APPENDIX H.....</b>	<b>45</b>

# GUIDANCE FOR INDUSTRY<sup>1</sup>

## Statistical Approaches to Establishing Bioequivalence

This guidance represents the Food and Drug Administration's current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. An alternative approach may be used if such approach satisfies the requirements of the applicable statutes and regulations.

### I. INTRODUCTION

This guidance provides recommendations to sponsors and applicants who intend, either before or after approval, to use equivalence criteria in analyzing in vivo or in vitro bioequivalence (BE) studies for investigational new drug applications (INDs), new drug applications (NDAs), abbreviated new drug applications (ANDAs) and supplements to these applications. This guidance discusses three approaches for BE comparisons: average, population, and individual. The guidance focuses on how to use each approach once a specific approach has been chosen. This guidance replaces a prior FDA guidance entitled *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*, which was issued in July 1992.

### II. BACKGROUND

#### A. General

Requirements for submitting bioavailability (BA) and BE data in NDAs, ANDAs, and supplements, the definitions of BA and BE, and the types of in vivo studies that are appropriate to measure BA and establish BE are set forth in 21 CFR part 320. This guidance provides recommendations on how to meet provisions of part 320 for all drug products.

Defined as *relative BA*, BE involves comparison between a test (T) and reference (R) drug product, where T and R can vary, depending on the comparison to be performed (e.g., to-be-marketed dosage form versus clinical trial material, generic drug versus reference listed drug,

---

<sup>1</sup> This guidance has been prepared by the Population and Individual Bioequivalence Working Group of the Biopharmaceutics Coordinating Committee in the Office of Pharmaceutical Science, Center for Drug Evaluation and Research (CDER) at the Food and Drug Administration (FDA).

drug product changed after approval versus drug product before the change). Although BA and BE are closely related, BE comparisons normally rely on (1) a criterion, (2) a confidence interval for the criterion, and (3) a predetermined BE limit. BE comparisons could also be used in certain pharmaceutical product line extensions, such as additional strengths, new dosage forms (e.g., changes from immediate release to extended release), and new routes of administration. In these settings, the approaches described in this guidance can be used to determine BE. The general approaches discussed in this guidance may also be useful when assessing pharmaceutical equivalence or performing equivalence comparisons in clinical pharmacology studies and other areas.

## B. Statistical

In the July 1992 guidance on *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design* (the 1992 guidance), CDER recommended that a standard in vivo BE study design be based on the administration of either single or multiple doses of the T and R products to healthy subjects on separate occasions, with random assignment to the two possible sequences of drug product administration. The 1992 guidance further recommended that statistical analysis for pharmacokinetic measures, such as area under the curve (AUC) and peak concentration (C<sub>max</sub>), be based on the *two one-sided tests procedure* to determine whether the average values for the pharmacokinetic measures determined after administration of the T and R products were comparable. This approach is termed *average bioequivalence* and involves the calculation of a 90% confidence interval for the ratio of the averages (population geometric means) of the measures for the T and R products. To establish BE, the calculated confidence interval should fall within a BE limit, usually 80-125% for the ratio of the product averages.<sup>2</sup> In addition to this general approach, the 1992 guidance provided specific recommendations for (1) logarithmic transformation of pharmacokinetic data, (2) methods to evaluate sequence effects, and (3) methods to evaluate outlier data.

Although average BE is recommended for a comparison of BA measures in most BE studies, this guidance describes two new approaches, termed *population* and *individual bioequivalence*. These new approaches may be useful, in some instances, for analyzing in vitro and in vivo BE studies.<sup>3</sup> The average BE approach focuses only on the comparison of population averages of a BE measure of interest and not on the variances of the measure for the

---

<sup>2</sup> For a broad range of drugs, a BE limit of 80 to 125% for the ratio of the product averages has been adopted for use of an average BE criterion. Generally, the BE limit of 80 to 125% is based on a clinical judgment that a test product with BA measures outside this range should be denied market access.

<sup>3</sup> For additional recommendations on in vivo studies, see the FDA guidance for industry on *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations*. Additional recommendations on in vitro studies will be provided in an FDA guidance for industry on *Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*, when finalized.

T and R products. The average BE method does not assess a subject-by-formulation interaction variance, that is, the variation in the average T and R difference among individuals. In contrast, population and individual BE approaches include comparisons of both averages and variances of the measure. The population BE approach assesses total variability of the measure in the population. The individual BE approach assesses within-subject variability for the T and R products, as well as the subject-by-formulation interaction.

### III. STATISTICAL MODEL

Statistical analyses of BE data are typically based on a statistical model for the logarithm of the BA measures (e.g., AUC and Cmax). The model is a mixed-effects or two-stage linear model. Each subject,  $j$ , theoretically provides a mean for the log-transformed BA measure for each formulation,  $\mu_{Tj}$  and  $\mu_{Rj}$  for the T and R formulations, respectively. The model assumes that these subject-specific means come from a distribution with population means  $\mu_T$  and  $\mu_R$ , and between-subject variances  $\sigma_{BT}^2$  and  $\sigma_{BR}^2$ , respectively. The model allows for a correlation,  $\rho$ , between  $\mu_{Tj}$  and  $\mu_{Rj}$ . The subject-by-formulation interaction variance component (Schall and Luus 1993),  $\sigma_D^2$ , is related to these parameters as follows:

$$\begin{aligned}\sigma_D^2 &= \text{variance of } (\mu_{Tj} - \mu_{Rj}) \\ &= (\sigma_{BT} - \sigma_{BR})^2 + 2(1-\rho)\sigma_{BT}\sigma_{BR}\end{aligned}\tag{Equation 1}$$

For a given subject, the observed data for the log-transformed BA measure are assumed to be independent observations from distributions with means  $\mu_{Tj}$  and  $\mu_{Rj}$ , and within-subject variances  $\sigma_{WT}^2$  and  $\sigma_{WR}^2$ . The total variances for each formulation are defined as the sum of the within- and between-subject components (i.e.,  $\sigma_{TT}^2 = \sigma_{WT}^2 + \sigma_{BT}^2$  and  $\sigma_{TR}^2 = \sigma_{WR}^2 + \sigma_{BR}^2$ ). For analysis of crossover studies, the means are given additional structure by the inclusion of period and sequence effect terms.

### IV. STATISTICAL APPROACHES FOR BIOEQUIVALENCE

The general structure of a BE criterion is that a function ( $\Theta$ ) of population measures should be demonstrated to be no greater than a specified value ( $\theta$ ). Using the terminology of statistical hypothesis testing, this is accomplished by testing the hypothesis  $H_0: \Theta > \theta$  versus  $H_A: \Theta \leq \theta$  at a desired level of significance, often 5%. Rejection of the null hypothesis  $H_0$  (i.e., demonstrating that the estimate of  $\Theta$  is statistically significantly less than  $\theta$ ) results in a conclusion of BE. The choice of  $\Theta$  and  $\theta$  differs in average, population, and individual BE approaches.

A general objective in assessing BE is to compare the log-transformed BA measure after administration of the T and R products. As detailed in Appendix A, population and individual approaches are based on the comparison of an expected squared distance between the T and R formulations to the expected

squared distance between two administrations of the R formulation. An acceptable T formulation is one where the T-R distance is not substantially greater than the R-R distance. In both population and individual BE approaches, this comparison appears as a comparison to the reference variance, which is referred to as *scaling to the reference variability*.

Population and individual BE approaches, but not the average BE approach, allow two types of scaling: reference-scaling and constant-scaling. Reference-scaling means that the criterion used is scaled to the variability of the R product, which effectively widens the BE limit for more variable reference products. Although generally sufficient, use of reference-scaling alone could unnecessarily narrow the BE limit for drugs and/or drug products that have low variability but a wide therapeutic range. This guidance, therefore, recommends mixed-scaling for the population and individual BE approaches (section IV.B and C). With mixed scaling, the reference-scaled form of the criterion should be used if the reference product is highly variable; otherwise, the constant-scaled form should be used.

**A. Average Bioequivalence**

The following criterion is recommended for average BE:

$$(\mu_T - \mu_R)^2 \leq \theta_A^2 \quad \text{Equation 2}$$

where

$\mu_T$  = population average response of the log-transformed measure for the T formulation

$\mu_R$  = population average response of the log-transformed measure for the R formulation

as defined in section III above.

This criterion is equivalent to:

$$-\theta_A \leq (\mu_T - \mu_R) \leq \theta_A \quad \text{Equation 3}$$

and, usually,  $\theta_A = \ln(1.25)$ .

## B. Population Bioequivalence

The following mixed-scaling approach is recommended for population BE (i.e., use the reference-scaled method if the estimate of  $\sigma_{TR} > \sigma_{T0}$  and the constant-scaled method if the estimate of  $\sigma_{TR} \leq \sigma_{T0}$ ).

The recommended criteria are:

- Reference-Scaled:

$$\frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma_{TR}^2} \leq \theta_p \quad \text{Equation 4}$$

or

- Constant-Scaled:

$$\frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma_{T0}^2} \leq \theta_p \quad \text{Equation 5}$$

where:

- $\mu_T$  = population average response of the log-transformed measure for the T formulation
- $\mu_R$  = population average response of the log-transformed measure for the R formulation
- $\sigma_{TT}^2$  = total variance (i.e., sum of within- and between-subject variances) of the T formulation
- $\sigma_{TR}^2$  = total variance (i.e., sum of within- and between-subject variances) of the R formulation
- $\sigma_{T0}^2$  = specified constant total variance
- $\theta_p$  = BE limit

Equations 4 and 5 represent an aggregate approach where a single criterion on the left-hand side of the equation encompasses two major components: (1) the difference between the T and R population averages ( $\mu_T - \mu_R$ ), and (2) the difference between the T and R total variances ( $\sigma_{TT}^2 - \sigma_{TR}^2$ ). This aggregate measure is scaled to the total variance of the R product or to a constant value ( $\sigma_{T0}^2$ , a standard that relates to a limit for the total variance), whichever is greater.

The specification of both  $\sigma_{T0}$  and  $\theta_P$  relies on the establishment of standards. The generation of these standards is discussed in Appendix A. When the population BE approach is used, in addition to meeting the BE limit based on confidence bounds, the point estimate of the geometric test/reference mean should fall within 80-125%.

### C. Individual Bioequivalence

The following mixed-scaling approach is one approach for individual BE (i.e., use the reference-scaled method if the estimate of  $\sigma_{WR} > \sigma_{W0}$ , and the constant-scaled method if the estimate of  $\sigma_{WR} \leq \sigma_{W0}$ ). Also see section VII.D, Discontinuity, for further discussion.

The recommended criteria are:

- Reference-Scaled:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\sigma_{WR}^2} \leq \theta_t \quad \text{Equation 6}$$

or

- Constant-Scaled:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\sigma_{W0}^2} \leq \theta_t \quad \text{Equation 7}$$

where:

- $\mu_T$  = population average response of the log-transformed measure for the T formulation
- $\mu_R$  = population average response of the log-transformed measure for the R formulation
- $\sigma_D^2$  = subject-by-formulation interaction variance component

- $\sigma_{WT}^2$  = within-subject variance of the T formulation
- $\sigma_{WR}^2$  = within-subject variance of the R formulation
- $\sigma_{W0}^2$  = specified constant within-subject variance
- $\theta_1$  = BE limit

Equations 6 and 7 represent an aggregate approach where a single criterion on the left-hand side of the equation encompasses three major components: (1) the difference between the T and R population averages ( $\mu_T - \mu_R$ ), (2) subject-by-formulation interaction ( $\sigma_D^2$ ), and (3) the difference between the T and R within-subject variances ( $\sigma_{WT}^2 - \sigma_{WR}^2$ ). This aggregate measure is scaled to the within-subject variance of the R product or to a constant value ( $\sigma_{W0}^2$ , a standard that relates to a limit for the within-subject variance), whichever is greater.

The specification of both  $\sigma_{W0}$  and  $\theta_1$  relies on the establishment of standards. The generation of these standards is discussed in Appendix A. When the individual BE approach is used, in addition to meeting the BE limit based on confidence bounds, the point estimate of the geometric test/reference mean ratio should fall within 80-125%.

## V. STUDY DESIGN

### A. Experimental Design

#### 1. *Nonreplicated Designs*

A conventional nonreplicated design, such as the standard two-formulation, two-period, two-sequence crossover design, can be used to generate data where an average or population approach is chosen for BE comparisons. Under certain circumstances, parallel designs can also be used.

#### 2. *Replicated Crossover Designs*

Replicated crossover designs can be used irrespective of which approach is selected to establish BE, although they are not necessary when an average or population approach is used. Replicated crossover designs are critical when an individual BE approach is used to allow estimation of within-subject variances for the T and R measures and the subject-by-formulation interaction variance component. The following four-period, two-sequence, two-formulation design is recommended for replicated BE studies (see Appendix B for further discussion of replicated crossover designs).

		Period			
		<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
Sequence	1	<b>T</b>	<b>R</b>	<b>T</b>	<b>R</b>
	2	<b>R</b>	<b>T</b>	<b>R</b>	<b>T</b>

For this design, the same lots of the T and R formulations should be used for the replicated administration. Each period should be separated by an adequate washout period.

Other replicated crossover designs are possible. For example, a three-period design, as shown below, could be used.

		Period		
		<u>1</u>	<u>2</u>	<u>3</u>
Sequence	1	<b>T</b>	<b>R</b>	<b>T</b>
	2	<b>R</b>	<b>T</b>	<b>R</b>

A greater number of subjects would be encouraged for the three-period design compared to the recommended four-period design to achieve the same statistical power to conclude BE (see Appendix C).

#### **B. Sample Size and Dropouts**

A minimum number of 12 evaluable subjects should be included in any BE study. When an average BE approach is selected using either nonreplicated or replicated designs, methods appropriate to the study design should be used to estimate sample sizes. The number of subjects for BE studies based on either the population or individual BE approach can be estimated by simulation if analytical approaches for estimation are not available. Further information on sample size is provided in Appendix C.

Sponsors should enter a sufficient number of subjects in the study to allow for dropouts. Because replacement of subjects during the study could complicate the statistical model and analysis, dropouts generally should not be replaced. Sponsors who wish to replace dropouts during the study should indicate this intention in the protocol. The protocol should also state

whether samples from replacement subjects, if not used, will be assayed. If the dropout rate is high and sponsors wish to add more subjects, a modification of the statistical analysis may be recommended. Additional subjects should not be included after data analysis unless the trial was designed from the beginning as a sequential or group sequential design.

## **VI. STATISTICAL ANALYSIS**

The following sections provide recommendations on statistical methodology for assessment of average, population, and individual BE.

### **A. Logarithmic Transformation**

#### *1. General Procedures*

This guidance recommends that BE measures (e.g., AUC and C<sub>max</sub>) be log-transformed using either common logarithms to the base 10 or natural logarithms (see Appendix D). The choice of common or natural logs should be consistent and should be stated in the study report. The limited sample size in a typical BE study precludes a reliable determination of the distribution of the data set. Sponsors and/or applicants are not encouraged to test for normality of error distribution after log-transformation, nor should they use normality of error distribution as a reason for carrying out the statistical analysis on the original scale. Justification should be provided if sponsors or applicants believe that their BE study data should be statistically analyzed on the original rather than on the log scale.

#### *2. Presentation of Data*

The drug concentration in biological fluid determined at each sampling time point should be furnished on the original scale for each subject participating in the study. The pharmacokinetic measures of systemic exposure should also be furnished on the original scale. The mean, standard deviation, and coefficient of variation for each variable should be computed and tabulated in the final report.

In addition to the arithmetic mean and associated standard deviation (or coefficient of variation) for the T and R products, geometric means (antilog of the means of the logs) should be calculated for selected BE measures. To facilitate BE comparisons, the measures for each individual should be displayed in parallel for the formulations tested. In particular, for each BE measure the ratio of the individual geometric mean of the T product to the individual geometric mean of the R product should be tabulated side by side for each subject. The summary tables should indicate in which sequence each

subject received the product.

## B. Data Analysis

### 1. Average Bioequivalence

#### a. Overview

Parametric (normal-theory) methods are recommended for the analysis of log-transformed BE measures. For average BE using the criterion stated in equations 2 or 3 (section III.A), the general approach is to construct a 90% confidence interval for the quantity  $\mu_T - \mu_R$  and to reach a conclusion of average BE if this confidence interval is contained in the interval  $[-\theta_A, \theta_A]$ . Due to the nature of normal-theory confidence intervals, this is equivalent to carrying out two one-sided tests of hypothesis at the 5% level of significance (Schuirmann 1987).

The 90% confidence interval for the difference in the means of the log-transformed data should be calculated using methods appropriate to the experimental design. The antilogs of the confidence limits obtained constitute the 90% confidence interval for the ratio of the geometric means between the T and R products.

#### b. Nonreplicated Crossover Designs

For nonreplicated crossover designs, this guidance recommends parametric (normal-theory) procedures to analyze log-transformed BA measures. General linear model procedures available in PROC GLM in SAS or equivalent software are preferred, although linear mixed-effects model procedures can also be indicated for analysis of nonreplicated crossover studies.

For example, for a conventional two-treatment, two-period, two-sequence (2 x 2) randomized crossover design, the statistical model typically includes factors accounting for the following sources of variation: sequence, subjects nested in sequences, period, and treatment. The *Estimate* statement in SAS PROC GLM, or equivalent statement in other software, should be used to obtain estimates for the adjusted differences between treatment means and the standard error associated with these differences.

#### c. Replicated Crossover Designs

Linear mixed-effects model procedures, available in PROC MIXED in SAS or equivalent software, should be used for the analysis of replicated crossover studies for average BE. Appendix E includes an example of SAS program statements.

d. Parallel Designs

For parallel designs, the confidence interval for the difference of means in the log scale can be computed using the total between-subject variance. As in the analysis for replicated designs (section VI. B.1.b), equal variances should not be assumed.

2. *Population Bioequivalence*

a. Overview

Analysis of BE data using the population approach (section IV.B) should focus first on estimation of the mean difference between the T and R for the log-transformed BA measure and estimation of the total variance for each of the two formulations. This can be done using relatively simple unbiased estimators such as the method of moments (MM) (Chinchilli 1996, and Chinchilli and Esinhart 1996). After the estimation of the mean difference and the variances has been completed, a 95% upper confidence bound for the population BE criterion can be obtained, or equivalently a 95% upper confidence bound for a linearized form of the population BE criterion can be obtained. Population BE should be considered to be established for a particular log-transformed BA measure if the 95% upper confidence bound for the criterion is less than or equal to the BE limit,  $\theta_p$ , or equivalently if the 95% upper confidence bound for the linearized criterion is less than or equal to 0.

To obtain the 95% upper confidence bound of the criterion, intervals based on validated approaches can be used. Validation approaches should be reviewed with appropriate staff in CDER. Appendix F includes an example of upper confidence bound determination using a population BE approach.

b. Nonreplicated Crossover Designs

For nonreplicated crossover studies, any available method (e.g., SAS PROC GLM or equivalent software) can be used to obtain an unbiased estimate of the mean difference in log-transformed BA measures between the T and R products. The total variance for each formulation should be estimated by the

usual sample variance, computed separately in each sequence and then pooled across sequences.

c. Replicated Crossover Designs

For replicated crossover studies, the approach should be the same as for nonreplicated crossover designs, but care should be taken to obtain proper estimates of the total variances. One approach is to estimate the within- and between-subject components separately, as for individual BE (see section VI.B.3), and then sum them to obtain the total variance. The method for the upper confidence bound should be consistent with the method used for estimating the variances.

d. Parallel Designs

The estimate of the means and variances from parallel designs should be the same as for nonreplicated crossover designs. The method for the upper confidence bound should be modified to reflect independent rather than paired samples and to allow for unequal variances.

3. *Individual Bioequivalence*

Analysis of BE data using an individual BE approach (section IV.C) should focus on estimation of the mean difference between T and R for the log-transformed BA measure, the subject-by-formulation interaction variance, and the within-subject variance for each of the two formulations. For this purpose, we recommend the MM approach.

To obtain the 95% upper confidence bound of a linearized form of the individual BE criterion, intervals based on validated approaches can be used. An example is described in Appendix G. After the estimation of the mean difference and the variances has been completed, a 95% upper confidence bound for the individual BE criterion can be obtained, or equivalently a 95% upper confidence bound for a linearized form of the individual BE criterion can be obtained. Individual BE should be considered to be established for a particular log-transformed BA measure if the 95% upper confidence bound for the criterion is less than or equal to the BE limit,  $\theta_1$ , or equivalently if the 95% upper confidence bound for the linearized criterion is less than or equal to 0.

The restricted maximum likelihood (REML) method may be useful to estimate mean differences and variances when subjects with some missing data are included in the statistical analysis. A key distinction between the REML and MM methods relates to

differences in estimating variance terms and is further discussed in Appendix H. Sponsors considering alternative methods to REML or MM are encouraged to discuss their approaches with appropriate CDER review staff prior to submitting their applications.

## VII. MISCELLANEOUS ISSUES

### A. Studies in Multiple Groups

If a crossover study is carried out in two or more groups of subjects (e.g., if for logistical reasons only a limited number of subjects can be studied at one time), the statistical model should be modified to reflect the multigroup nature of the study. In particular, the model should reflect the fact that the periods for the first group are different from the periods for the second group. This applies to all of the approaches (average, population, and individual BE) described in this guidance.

If the study is carried out in two or more groups and those groups are studied at different clinical sites, or at the same site but greatly separated in time (months apart, for example), questions may arise as to whether the results from the several groups should be combined in a single analysis. Such cases should be discussed with the appropriate CDER review division.

A *sequential* design, in which the decision to study a second group of subjects is based on the results from the first group, calls for different statistical methods and is outside the scope of this guidance. Those wishing to use a sequential design should consult the appropriate CDER review division.

### B. Carryover Effects

Use of crossover designs for BE studies allows each subject to serve as his or her own control to improve the precision of the comparison. One of the assumptions underlying this principle is that *carryover effects* (also called *residual effects*) are either absent (the response to a formulation administered in a particular period of the design is unaffected by formulations administered in earlier periods) or equal for each formulation and preceding formulation. If carryover effects are present in a crossover study and are not equal, the usual crossover estimate of  $\mu_T - \mu_R$  could be biased. One limitation of a conventional two-formulation, two-period, two-sequence crossover design is that the only statistical test available for the presence of unequal carryover effects is the sequence test in the analysis of variance (ANOVA) for the crossover design. This is a between-subject test, which would be expected to have poor discriminating power in a typical BE study. Furthermore, if the possibility of unequal carryover effects cannot be ruled out, no unbiased estimate of  $\mu_T - \mu_R$  based on within-subject

comparisons can be obtained with this design.

For replicated crossover studies, a within-subject test for unequal carryover effects can be obtained under certain assumptions. Typically only first-order carryover effects are considered of concern (i.e., the carryover effects, if they occur, only affect the response to the formulation administered in the next period of the design). Under this assumption, consideration of carryover effects could be more complicated for replicated crossover studies than for nonreplicated studies. The carryover effect could depend not only on the formulation that preceded the current period, but also on the formulation that is administered in the current period. This is called a *direct-by-carryover* interaction. The need to consider more than just *simple* first-order carryover effects has been emphasized (Fleiss 1989). With a replicated crossover design, a within-subject estimate of  $\mu_T - \mu_R$  unbiased by general first-order carryover effects can be obtained, but such an estimate could be imprecise, reducing the power of the study to conclude BE.

In most cases, for both replicated and nonreplicated crossover designs, the possibility of unequal carryover effects is considered unlikely in a BE study under the following circumstances:

- It is a single-dose study.
- The drug is not an endogenous entity.
- More than an adequate washout period has been allowed between periods of the study and in the subsequent periods the predose biological matrix samples do not exhibit a detectable drug level in any of the subjects.
- The study meets all scientific criteria (e.g., it is based on an acceptable study protocol and it contains sufficient validated assay methodology).

The possibility of unequal carryover effects can also be discounted for multiple-dose studies and/or studies in patients, provided that the drug is not an endogenous entity and the studies meet all scientific criteria as described above. Under all other circumstances, the sponsor or applicant could be asked to consider the possibility of unequal carryover effects, including a direct-by-carryover interaction. If there is evidence of carryover effects, sponsors should describe their proposed approach in the study protocol, including statistical tests for the presence of such effects and procedures to be followed. Sponsors who suspect that carryover effects might be an issue may wish to conduct a BE study with parallel designs.

### **C. Outlier Considerations**

Outlier data in BE studies are defined as subject data for one or more BA measures that are

discordant with corresponding data for that subject and/or for the rest of the subjects in a study.

Because BE studies are usually carried out as crossover studies, the most important type of subject outlier is the within-subject outlier, where one subject or a few subjects differ notably from the rest of the subjects with respect to a within-subject T-R comparison. The existence of a subject outlier with no protocol violations could indicate one of the following situations:

1. *Product Failure*

Product failure could occur, for example, when a subject exhibits an unusually high or low response to one or the other of the products because of a problem with the specific dosage unit administered. This could occur, for example, with a sustained and/or delayed-release dosage form exhibiting dose dumping or a dosage unit with a coating that inhibits dissolution.

2. *Subject-by-Formulation Interaction*

A subject-by-formulation interaction could occur when an individual is representative of subjects present in the general population in low numbers, for whom the relative BA of the two products is markedly different than for the majority of the population, and for whom the two products are not bioequivalent, even though they might be bioequivalent in the majority of the population.

In the case of product failure, the unusual response could be present for either the T or R product. However, in the case of a subpopulation, even if the unusual response is observed on the R product, there could still be concern for lack of interchangeability of the two products. For these reasons, deletion of outlier values is generally discouraged, particularly for nonreplicated designs. With replicated crossover designs, the *retest* character of these designs should indicate whether to delete an outlier value or not. Sponsors or applicants with these types of data sets may wish to review how to handle outliers with appropriate review staff.

**D. Discontinuity**

The mixed-scaling approach has a discontinuity at the changeover point,  $\sigma_{w0}$  (individual BE criterion) or  $\sigma_{T0}$  (population BE criterion), from constant- to reference-scaling. For example, if the estimate of the within-subject standard deviation of the reference is just above the changeover point, the confidence interval will be wider than just below. In this context, the confidence interval could pass the predetermined BE limit if the estimate is just below the boundary and could fail if just above. This guidance recommends that sponsors applying the individual BE approach may use either reference-scaling or constant-scaling at either side of the changeover point. With this approach, the multiple testing inflates the type I error rate slightly, to approximately 6.5%, but only over a small interval of  $\sigma_{WR}$  (about 0.18-0.20).

## REFERENCES

- Anderson, S., and W.W. Hauck, 1990, "Consideration of Individual Bioequivalence," *J. Pharmacokin. Biopharm.*, 18:259-73.
- Anderson, S., 1993, "Individual Bioequivalence: A Problem of Switchability (with discussion)," *Biopharmaceutical Reports*, 2(2):1-11.
- Anderson, S., 1995, "Current Issues of Individual Bioequivalence," *Drug Inf. J.*, 29:961-4.
- Chen, M.-L., 1997, "Individual Bioequivalence — A Regulatory Update (with discussion)," *J. Biopharm. Stat.*, 7:5-111.
- Chen, M.-L., R. Patnaik, W.W. Hauck, D.J. Schuirmann, T. Hyslop, R.L. Williams, and the FDA Population and Individual Bioequivalence Working Group, 2000, "An Individual Bioequivalence Criterion: Regulatory Considerations," *Stat. Med.*, 19:2821-42.
- Chen, M.-L., S.-C. Lee, M.-J. Ng, D.J. Schuirmann, L. J. Lesko, and R.L. Williams, 2000, "Pharmacokinetic analysis of bioequivalence trials: Implications for sex-related issues in clinical pharmacology and biopharmaceutics," *Clin. Pharmacol. Ther.*, 68(5):510-21.
- Chinchilli, V.M., 1996, "The Assessment of Individual and Population Bioequivalence," *J. Biopharm. Stat.*, 6:1-14.
- Chinchilli, V.M., and J.D. Esinhart, 1996, "Design and Analysis of Intra-Subject Variability in Cross-Over Experiments," *Stat. Med.*, 15:1619-34.
- Chow, S.-C., 1999, "Individual Bioequivalence — A Review of the FDA Draft Guidance," *Drug Inf. J.*, 33:435-44.
- Diletti E., D. Hauschke, and V.W. Steinijans, 1991, "Sample Size Determination for Bioequivalence Assessment By Means of Confidence Intervals," *Int. J. Clin. Pharmacol. Therap.*, 29:1-8.
- Efron, B., 1987, "Better Bootstrap Confidence Intervals (with discussion)," *J. Amer. Stat. Assoc.*, 82:171-201.
- Efron, B., and R.J. Tibshirani, 1993, *An Introduction to the Bootstrap*, Chapman and Hall, Ch. 14.

Ekbohm, G., and H. Melander, 1989, "The Subject-by-Formulation Interaction as a Criterion for Interchangeability of Drugs," *Biometrics*, 45:1249-54.

Ekbohm, G., and H. Melander, 1990, "On Variation, Bioequivalence and Interchangeability," Report 14, Department of Statistics, Swedish University of Agricultural Sciences.

Endrenyi, L., and M. Schulz, 1993, "Individual Variation and the Acceptance of Average Bioequivalence," *Drug Inf. J.*, 27:195-201.

Endrenyi, L., 1993, "A Procedure for the Assessment of Individual Bioequivalence," in *Bio-International: Bioavailability, Bioequivalence and Pharmacokinetics* (H.H.Blume, K.K. Midha, eds.), Medpharm Publications, 141-6.

Endrenyi, L., 1994, "A Method for the Evaluation of Individual Bioequivalence," *Int. J. Clin. Pharmacol. Therap.*, 32:497-508.

Endrenyi, L., 1995, "A Simple Approach for the Evaluation of Individual Bioequivalence," *Drug Inf. J.*, 29:847-55.

Endrenyi, L., and K.K. Midha, 1998, "Individual Bioequivalence — Has Its Time Come?," *Eur. J. Pharm. Sci.*, 6:271-8.

Endrenyi, L., G.L. Amidon, K.K. Midha, and J.P. Skelly, 1998, "Individual Bioequivalence: Attractive in Principle, Difficult in Practice," *Pharm. Res.*, 15:1321-5.

Endrenyi, L., and Y. Hao, 1998, "Asymmetry of the Mean-Variability Tradeoff Raises Questions About the Model in Investigations of Individual Bioequivalence," *Int. J. Pharmacol. Therap.*, 36:450-7.

Endrenyi, L., and L. Tothfalusi, 1999, "Subject-by-Formulation Interaction in Determination of Individual Bioequivalence: Bias and Prevalence," *Pharm. Res.*, 16:186-8.

Esinhart, J.D., and V.M. Chinchilli, 1994, "Sample Size Considerations for Assessing Individual Bioequivalence Based on the Method of Tolerance Interval," *Int. J. Clin. Pharmacol. Therap.*, 32(1):26-32.

Esinhart, J.D., and V.M. Chinchilli, 1994, "Extension to the Use of Tolerance Intervals for the Assessment of Individual Bioequivalence," *J. Biopharm. Stat.*, 4(1):39-52.

Fleiss, J.L., 1989, "A Critique of Recent Research on the Two-Treatment Crossover Design," *Controlled Clinical Trials*, 10:237-43.

- Graybill, F., and C.M. Wang, 1980, "Confidence Intervals on Nonnegative Linear Combinations of Variances," *J. Amer. Stat. Assoc.*, 75:869-73.
- Hauck, W.W., and S. Anderson, 1984, "A New Statistical Procedure for Testing Equivalence in Two-Group Comparative Bioavailability Trials," *J. Pharmacokin. Biopharm.*, 12:83-91.
- Hauck, W.W., and S. Anderson, 1992, "Types of Bioequivalence and Related Statistical Considerations," *Int. J. Clin. Pharmacol. Therap.*, 30:181-7.
- Hauck, W.W., and S. Anderson, 1994, "Measuring Switchability and Prescribability: When is Average Bioequivalence Sufficient?," *J. Pharmacokin. Biopharm.*, 22:551-64.
- Hauck, W.W., M.-L. Chen, T. Hyslop, R. Patnaik, D. Schuirmann, and R.L. Williams for the FDA Population and Individual Bioequivalence Working Group, 1996, "Mean Difference vs. Variability Reduction: Tradeoffs in Aggregate Measures for Individual Bioequivalence," *Int. J. Clin. Pharmacol. Therap.*, 34:535-41.
- Holder, D.J., and F. Hsuan, 1993, "Moment-Based Criteria for Determining Bioequivalence," *Biometrika*, 80:835-46.
- Holder, D.J., and F. Hsuan, 1995, "A Moment-Based Method for Determining Individual Bioequivalence," *Drug Inf. J.*, 29:965-79.
- Howe, W.G., 1974. "Approximate Confidence Limits on the Mean of  $X+Y$  Where  $X$  and  $Y$  are Two Tabled Independent Random Variables," *J. Amer. Stat. Assoc.*, 69:789-94.
- Hsu, J.C., J.T.G. Hwang, H.-K. Liu, and S.J. Ruberg, 1994, "Confidence Intervals Associated with Tests for Bioequivalence," *Biometrika*, 81:103-14.
- Hwang, S., P.B. Huber, M. Hesney, and K.C. Kwan, 1978, "Bioequivalence and Interchangeability," *J. Pharm. Sci.*, 67:IV "Open Forum."
- Hyslop, T., F. Hsuan, and D.J. Holder, 2000, "A Small-Sample Confidence Interval Approach to Assess Individual Bioequivalence," *Stat. Med.*, 19:2885-97.
- Kimanani, E.K., and D. Potvin, 1998, "Parametric Confidence Interval for a Moment-Based Scaled Criterion for Individual Bioequivalence," *J. Pharm. Biopharm.*, 25:595-614.
- Liu, J.-P, 1995, "Use of the Repeated Crossover Designs in Assessing Bioequivalence," *Stat. Med.*, 14:1067-78.

Patnaik, R. N., L.J. Lesko, M.-L. Chen, R.L. Williams, and the FDA Population and Individual Bioequivalence Working Group, 1997, "Individual Bioequivalence: New Concepts in the Statistical Assessment of Bioequivalence Metrics," *Clin. Pharmacokin.*, 33:1-6.

Schall, R., 1995, "Unified View of Individual, Population and Average Bioequivalence," in *Bio-International 2: Bioavailability, Bioequivalence and Pharmacokinetic Studies* (H.H.Blume, K.K.Midha, eds.), Medpharm Scientific Publishers, 91-105.

Schall, R., and H.G. Luus, 1993, "On Population and Individual Bioequivalence," *Stat. Med.*, 12:1109-24.

Schall R., 1995, "Assessment of Individual and Population Bioequivalence Using the Probability That Bioavailabilities Are Similar," *Biometrics*, 51:615-26.

Schall, R., and R.L. Williams for the FDA Individual Bioequivalence Working Group, 1996, "Towards a Practical Strategy for Assessing Individual Bioequivalence," *J. Pharmacokin. Biopharm.*, 24:133-49.

Schuirmann, D.J., 1987, "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *J. Pharmacokin. Biopharm.*, 15:657-80.

Schuirmann, D.J., 1989, "Treatment of Bioequivalence Data: Log Transformation," in *Proceedings of Bio-International '89 — Issues in the Evaluation of Bioavailability Data*, Toronto, Canada, October 1-4, 159-61.

Senn, S., and D. Lambrou, 1998, "Robust and Realistic Approaches to Carry-Over," *Stat. Med.*, 17:2849-64.

Sheiner, L. B., 1992, "Bioequivalence Revisited," *Stat. Med.*, 11:1777-88.

Ting, N., R.K. Burdick, F.A. Graybill, S. Jeyaratnam, and T.F.C. Lu, 1990, "Confidence Intervals on Linear Combinations of Variance Components That Are Unrestricted in Sign," *J. Stat. Comp. Sim.*, 35:135-43.

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.), Lea and Febiger, 149-79.

Westlake, W.J., 1979, "Statistical Aspects of Comparative Bioavailability Trials," *Biometrics*, 35:273-80.

Westlake, W.J., 1981, "Response to Kirkwood, TBL.: Bioequivalence Testing — A Need to Rethink," *Biometrics*, 37:589-94.

Westlake, W.J., 1988, "Bioavailability and Bioequivalence of Pharmaceutical Formulations," in *Biopharmaceutical Statistics for Drug Development* (K.E. Peace, ed.), Marcel Dekker, Inc., 329-52.

## APPENDIX A

### Standards

The equations in section IV call for standards to be established (i.e.,  $\sigma_{T0}$  and  $\theta_p$  for assessment of population BE,  $\sigma_{w0}$  and  $\theta_i$  for individual BE). The recommended approach to establishing these standards is described below.

#### A. $\sigma_{T0}$ and $\sigma_{w0}$

As indicated in section IV, a general objective in assessing BE should be to compare the difference in the BA log-measure of interest after the administration of the T and R formulations, T-R, with the difference in the same log-metric after two administrations of the R formulation, R-R'.

##### 1. Population Bioequivalence

For population BE, the comparisons of interest should be expressed in terms of the ratio of the expected squared difference between T and R (administered to different individuals) and the expected squared difference between R and R' (administered to different individuals), as shown below.

$$E(T - R)^2 = (\mu_T - \mu_R)^2 + \sigma_{TT}^2 + \sigma_{TR}^2 \quad \text{Equation 8}$$

$$E(R - R')^2 = 2\sigma_{TR}^2 \quad \text{Equation 9}$$

$$\frac{E(T - R)^2}{E(R - R')^2} = \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 + \sigma_{TR}^2}{2\sigma_{TR}^2} \quad \text{Equation 10}$$

The population BE criterion in equation 4 (section IV.B.) is derived from equation 10, such that the criterion equals zero for two identical formulations. The square root of equation 10 yields the "population difference ratio" (PDR):

$$\text{PDR} = \left[ \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 + \sigma_{TR}^2}{2\sigma_{TR}^2} \right]^{1/2} \quad \text{Equation 11}$$

The PDR is the square root of the ratio of the expected squared T-R difference compared to the expected squared R-R' difference in the population. It should be noted that the PDR is monotonically related to the population BE criterion (PBC) described in equation 4 as follows:

$$\text{PDR} = (\text{PBC}/2 + 1)^{1/2} \quad \text{Equation 12}$$

Sponsors or applicants wishing to use the population BE approach should contact the Agency for further information on  $\sigma_{T0}$ .

## 2. Individual Bioequivalence

For individual BE, the comparisons of interest should be expressed in terms of the ratio of the expected squared difference between T and R (administered to the same individual) and the expected squared difference between R and R' (two administrations of R to the same individual), as shown below.

$$E(T - R)^2 = (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 + \sigma_{WR}^2 \quad \text{Equation 13}$$

$$E(R - R')^2 = 2\sigma_{WR}^2 \quad \text{Equation 14}$$

$$\frac{E(T - R)^2}{E(R - R')^2} = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 + \sigma_{WR}^2}{2\sigma_{WR}^2} \quad \text{Equation 15}$$

The individual BE criterion in equation 6 (section IV.C.) is derived from equation 15, such that the criterion equals zero for two identical formulations. The square root of equation 15 is the *individual difference ratio* (IDR):

$$\text{IDR} = \left[ \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 + \sigma_{WR}^2}{2\sigma_{WR}^2} \right]^{1/2} \quad \text{Equation 16}$$

The IDR is the square root of the ratio of the expected squared T-R difference compared to the expected squared R-R' difference within an individual. The IDR is monotonically related to the individual BE criterion (IBC) described in equation 6 as follows:

$$\text{IDR} = (\text{IBC}/2 + 1)^{1/2} \quad \text{Equation 17}$$

This guidance recommends that  $\sigma_{W0} = 0.2$ , based on the consideration of the maximum allowable IDR of 1.25.<sup>4</sup>

#### B. $\theta_P$ and $\theta_I$

The determination of  $\theta_P$  and  $\theta_I$  should be based on the consideration of average BE criterion and the addition of variance terms to the population and individual BE criterion, as expressed by the formula below.

$$\theta = \frac{\text{average BE limit} + \text{variance factor}}{\text{variance}}$$

##### 1. Population Bioequivalence

$$\theta_P = \frac{(\ln 1.25)^2 + \epsilon_P}{\sigma_{T0}^2} \quad \text{Equation 18}$$

The value of  $\epsilon_P$  for population BE is guided by the consideration of the variance term ( $\sigma_{TT}^2 - \sigma_{TR}^2$ ) added to the average BE criterion. Sponsors or applicants wishing to use the population BE approach should contact the Agency for further information on  $\epsilon_P$  and  $\theta_P$ .

---

<sup>4</sup> The IDR upper bound of 1.25 is drawn from the currently used upper BE limit of 1.25 for the average BE criterion.

2. Individual Bioequivalence

$$\theta_1 = \frac{(\ln 1.25)^2 + \varepsilon_1}{\sigma_{w0}^2} \quad \text{Equation 19}$$

The value of  $\varepsilon_1$  for individual BE is guided by the consideration of the estimate of subject-by-formulation interaction ( $\sigma_D$ ) as well as the difference in within-subject variability ( $\sigma_{wT}^2 - \sigma_{wR}^2$ ) added to the average BE criterion. The recommended allowance for the variance term ( $\sigma_{wT}^2 - \sigma_{wR}^2$ ) is 0.02. In addition, this guidance recommends a  $\sigma_D^2$  allowance of 0.03. The magnitude of  $\sigma_D$  is associated with the percentage of individuals whose average T to R ratios lie outside 0.8-1.25. It is estimated that if  $\sigma_D = 0.1356$ , ~10% of the individuals would have their average ratios outside 0.8-1.25, even if  $\mu_T - \mu_R = 0$ . When  $\sigma_D = 0.1741$ , the probability is ~20%.

Accordingly, on the basis of consideration for both  $\sigma_D$  and variability ( $\sigma_{wT}^2 - \sigma_{wR}^2$ ) in the criterion, this guidance recommends that  $\varepsilon_1 = 0.05$ .

## APPENDIX B

### Choice of Specific Replicated Crossover Designs

Appendix B describes why FDA prefers replicated crossover designs with only two sequences, and why we recommend the specific designs described in section V.A of this guidance.

#### 1. Reasons Unrelated to Carryover Effects

Each unique combination of sequence and period in a replicated crossover design can be called a *cell* of the design. For example, the two-sequence, four-period design recommended in section V.A.1 has 8 cells. The four-sequence, four-period design below has 16 cells.

		Period			
		<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
Sequence	1	T	R	R	T
	2	R	T	T	R
	3	T	T	R	R
	4	R	R	T	T

The total number of degrees-of-freedom attributable to comparisons among the cells is just the number of cells minus one (unless there are cells with no observations).

The fixed effects that are usually included in the statistical analysis are sequence, period, and treatment (i.e., formulation). The number of degrees-of-freedom attributable to each fixed effect is generally equal to the number of levels of the effect, minus one. Thus, in the case of the two-sequence, four-period design recommended in section V.A.1, there would be  $2-1=1$  degree-of-freedom due to sequence,  $4-1=3$  degrees-of-freedom due to period, and  $2-1=1$  degree-of-freedom due to treatment, for a total of  $1+3+1=5$  degrees-of-freedom due to the three fixed effects. Because these 5 degrees-of-freedom do not account for all 7 degrees-of-freedom attributable to the eight cells of the design, the fixed effects model is not *saturated*. There could be some controversy as to whether a fixed effects model that accounts for more or all of the degrees-of-freedom due to cells (i.e., a more saturated fixed effects model) should be used. For example, an effect for sequence-by-treatment interaction might be included in addition to the three *main effects* — sequence, period, and treatment. Alternatively, a sequence-by-

period interaction effect might be included, which would fully saturate the fixed effects model.

If the replicated crossover design has only two sequences, use of only the three main effects (sequence, period, and treatment) in the fixed effects model or use of a more saturated model makes little difference to the results of the analysis, provided there are no missing observations and the study is carried out in one group of subjects. The least squares estimate of  $\mu_T - \mu_R$  will be the same for the main effects model and for the saturated model. Also, the method of moments (MM) estimators of the variance terms in the model used in some approaches to assessment of population and individual BE (see Appendix H), which represent within-sequence comparisons, are generally fully efficient regardless of whether the main effects model or the saturated model is used.

If the replicated crossover design has more than two sequences, these advantages are no longer present. Main effects models will generally produce different estimates of  $\mu_T - \mu_R$  than saturated models (unless the number of subjects in each sequence is equal), and there is no well-accepted basis for choosing between these different estimates. Also, MM estimators of variance terms will be fully efficient only for saturated models, while for main effects models fully efficient estimators would have to include some between-sequence components, complicating the analysis. Thus, use of designs with only two sequences minimizes or avoids certain ambiguities due to the method of estimating variances or due to specific choices of fixed effects to be included in the statistical model.

## 2. Reasons Related to Carryover Effects

One of the reasons to use the four-sequence, four-period design described above is that it is thought to be optimal if carryover effects are included in the model. Similarly, the two-sequence, three-period design

		Period		
		<u>1</u>	<u>2</u>	<u>3</u>
Sequence	1	T	R	R
	2	R	T	T

is thought to be optimal among three-period replicated crossover designs. Both of these designs are *strongly balanced for carryover effects*, meaning that each treatment is preceded by each other treatment *and itself* an equal number of times.

With these designs, no efficiency is lost by including *simple* first-order carryover effects in the statistical model. However, if the possibility of carryover effects is to be considered in the statistical analysis of BE studies, the possibility of direct-by-carryover interaction should also be considered. If direct-by-

carryover interaction is present in the statistical model, these favored designs are no longer optimal. Indeed, the TRR/RTT design does not permit an unbiased within-subject estimate of  $\mu_T - \mu_R$  in the presence of general direct-by-carryover interaction.

The issue of whether a purely main effects model or a more saturated model should be specified, as described in the previous section, also is affected by possible carryover effects. If carryover effects, including direct-by-carryover interaction, are included in the statistical model, these effects will be partially confounded with sequence-by-treatment interaction in four-sequence or six-sequence replicated crossover designs, but not in two-sequence designs.

In the case of the four-period and three-period designs recommended in section V.A.1, the estimate of  $\mu_T - \mu_R$ , adjusted for first-order carryover effects including direct-by-carryover interaction, is as efficient or more efficient than for any other two-treatment replicated crossover designs.

### **3. Two-Period Replicated Crossover Designs**

For the majority of drug products, two-period replicated crossover designs such as the Balaam design (which uses the sequences TR, RT, TT, and RR) should be avoided for individual BE because subjects in the TT or RR sequence do not provide any information on subject-by-formulation interaction. However, the Balaam design may be useful for particular drug products (e.g., a long half-life drug for which a two-period study would be feasible but a three- or more period study would not).

## APPENDIX C

### Sample Size Determination

Sample sizes for average BE should be obtained using published formulas. Sample sizes for population and individual BE should be based on simulated data. The simulations should be conducted using a default situation allowing the two formulations to vary as much as 5% in average BA with equal variances and certain magnitude of subject-by-formulation interaction. The study should have 80 or 90% power to conclude BE between these two formulations. Sample size also depends on the magnitude of variability and the design of the study. Variance estimates to determine the number of subjects for a specific drug can be obtained from the biomedical literature and/or pilot studies.

Tables 1-4 below give sample sizes for 80% and 90% power using the specified study design, given a selection of within-subject standard deviations (natural log scale), between-subject standard deviations (natural log scale), and subject-by-formulation interaction, as appropriate.

**Table 1**

**Average Bioequivalence  
Estimated Numbers of Subjects  
 $\Delta=0.05$**

$\sigma_{WT} =$	$\sigma_D$	80% Power		90% Power	
		2P	4P	2P	4P
0.15	0.01	12	6	16	8
	0.10	14	10	18	12
	0.15	16	12	22	16
0.23	0.01	24	12	32	16
	0.10	26	16	36	20
	0.15	30	18	38	24
0.30	0.01	40	20	54	28
	0.10	42	24	56	30
	0.15	44	26	60	34
0.50	0.01	108	54	144	72
	0.10	110	58	148	76
	0.15	112	60	150	80

- Note:
1. Results for two-period designs use method of Diletti et al. (Diletti 1991).
  2. Results for four-period designs use relative efficiency data of Liu (Liu 1995).

**Table 2**

**Population Bioequivalence  
Four-Period Design (RTRT/TRTR)  
Estimated Numbers of Subjects  
 $\epsilon_p = 0.02, \Delta = 0.05$**

$\sigma_{WR} = \sigma_{WT}$	$\sigma_{BR} = \sigma_{BT}$	80% Power	90% Power
0.15	0.15	18	22
	0.30	24	32
0.23	0.23	22	28
	0.46	24	32
0.30	0.30	22	28
	0.60	26	34
0.50	0.50	22	28
	1.00	26	34

Note: Results for population BE are approximate from simulation studies (1,540 simulations for each parameter combination), assuming two-sequence, four-period trials with a balanced design across sequences.

**Table 3**

**Individual Bioequivalence  
Estimated Numbers of Subjects  
 $\varepsilon_1=0.05, \Delta=0.05$**

$\sigma_{WT} =$	$\sigma_D$	80% Power		90% Power	
		3P	4P	3P	4P
0.15	0.01	14	10	18	12
	0.10	18	14	24	16
	0.15	28	22	36	26
0.23	0.01	42	22	54	30
	0.10	56	30	74	40
	0.15	76	42	100	56
0.30	0.01	52	28	70	36
	0.10	60	32	82	42
	0.15	76	42	100	56
0.50	0.01	52	28	70	36
	0.10	60	32	82	42
	0.15	76	42	100	56

Note: Results for individual BE are approximate using simulations (5,000 simulations for each parameter combination). The designs used in simulations are RTR/TRT (3P) and RTRT/TRTR (4P) assuming two-sequence trials with a balanced design across sequences.

While the above sample sizes assume equal within-subject standard deviations, simulation studies for 3-period and 4-period designs reveal that if  $\Delta = 0$  and  $\sigma_{WT}^2 - \sigma_{WR}^2 = 0.05$ , the sample sizes given will provide either 80% or 90% power for these studies.

To maintain consistency with section V.C, which suggests a minimum of 12 subjects in all BE studies, the one case where  $n = 10$  provides 80% power should be increased to  $n = 12$ .

**Table 4**  
**Individual Bioequivalence**  
**Estimated Numbers of Subjects**  
 $\varepsilon_1 = 0.05, \Delta = 0.10$   
**With Constraint on  $\Delta$  ( $0.8 \leq \exp(\Delta) \leq 1.25$ )**

$\sigma_{WT} =$	$\sigma_D$	80% Power	90% Power
		4P	4P
0.30	0.01	30	40
	0.10	36	48
	0.15	42	56
0.50	0.01	34	46
	0.10	36	48
	0.15	42	56

Note: Results for individual BE are approximate using simulations (5,000 simulations for each parameter combination). The designs used in simulations are RTRT/TRTR (4P), assuming two-sequence trials with a balanced design across sequences. When  $\Delta = 0.05$ , sample sizes remain the same as given in Table 3. This is because the studies are already powered for variance estimation and inference, and therefore, a constraint on the point estimate of  $\Delta$  has little influence on the sample size for small values of  $\Delta$ .

## APPENDIX D

### Rationale for Logarithmic Transformation of Pharmacokinetic Data

#### A. Clinical Rationale

The FDA Generic Drugs Advisory Committee recommended in 1991 that the primary comparison of interest in a BE study is the ratio, rather than the difference, between average parameter data from the T and R formulations. Using logarithmic transformation, the general linear statistical model employed in the analysis of BE data allows inferences about the difference between the two means on the log scale, which can then be retransformed into inferences about the ratio of the two averages (means or medians) on the original scale. Logarithmic transformation thus achieves a general comparison based on the ratio rather than the differences.

#### B. Pharmacokinetic Rationale

Westlake observed that a multiplicative model is postulated for pharmacokinetic measures in BA/BE studies (i.e., AUC and C<sub>max</sub>, but not T<sub>max</sub>) (Westlake 1973 and 1988). Assuming that elimination of the drug is first-order and only occurs from the central compartment, the following equation holds after an extravascular route of administration:

$$AUC_{0-\infty} = FD/CL \quad \text{Equation 20}$$

$$= FD/(VK_e) \quad \text{Equation 21}$$

where F is the fraction absorbed, D is the administered dose, and FD is the amount of drug absorbed. CL is the clearance of a given subject that is the product of the apparent volume of distribution (V) and the elimination rate constant (K<sub>e</sub>).<sup>5</sup> The use of AUC as a measure of the amount of drug absorbed

---

<sup>5</sup> Note that a more general equation can be written for any multicompartmental model as

$$AUC_{0-\infty} = FD/V_{db} \lambda_n \quad \text{Equation 22}$$

where V<sub>db</sub> is the volume of distribution relating drug concentration in plasma or blood to the amount of drug in the body during the terminal exponential phase, and λ<sub>n</sub> is the terminal slope of the concentration-time curve.

involves a multiplicative term (CL) that might be regarded as a function of the subject. For this reason, Westlake contends that the subject effect is not additive if the data are analyzed on the original scale of measurement.

Logarithmic transformation of the AUC data will bring the CL ( $VK_e$ ) term into the following equation in an additive fashion:

$$\ln AUC_{0-\infty} = \ln F + \ln D - \ln V - \ln K_e \quad \text{Equation 23}$$

Similar arguments were given for  $C_{max}$ . The following equation applies for a drug exhibiting one compartmental characteristics:

$$C_{max} = (FD/V) \times e^{-k_e \cdot T_{max}} \quad \text{Equation 24}$$

where again F, D and V are introduced into the model in a multiplicative manner. However, after logarithmic transformation, the equation becomes

$$\ln C_{max} = \ln F + \ln D - \ln V - K_e T_{max} \quad \text{Equation 25}$$

Thus, log transformation of the  $C_{max}$  data also results in the additive treatment of the V term.

## APPENDIX E

### SAS Program Statements for Average BE Analysis of Replicated Crossover Studies

The following illustrates an example of program statements to run the average BE analysis using PROC MIXED in SAS version 6.12, with SEQ, SUBJ, PER, and TRT identifying sequence, subject, period, and treatment variables, respectively, and Y denoting the response measure (e.g., log(AUC), log(Cmax)) being analyzed:

```
PROC MIXED;  
CLASSES SEQ SUBJ PER TRT;  
MODEL Y = SEQ PER TRT/ DDFM=SATTERTH;  
RANDOM TRT/TYPE=FA0(2) SUB=SUBJ G;  
REPEATED/GRP=TRT SUB=SUBJ;  
ESTIMATE 'T vs. R' TRT 1 -1/CL ALPHA=0.1;
```

The *Estimate* statement assumes that the code for the T formulation precedes the code for the R formulation in sort order (this would be the case, for example, if T were coded as 1 and R were coded as 2). If the R code precedes the T code in sort order, the coefficients in the Estimate statement would be changed to -1 1.

In the *Random* statement, TYPE=FA0(2) could possibly be replaced by TYPE=CSH. This guidance recommends that TYPE=UN not be used, as it could result in an invalid (i.e., not non-negative definite) estimated covariance matrix.

Additions and modifications to these statements can be made if the study is carried out in more than one group of subjects.

## APPENDIX F

### Method for Statistical Test of Population Bioequivalence Criterion

#### Four-Period Crossover Designs

Appendix F describes a method for using the population BE criterion (see section IV.B, equations 4 and 5). The procedure involves the computation of a test statistic that is either positive (does not conclude population BE) or negative (concludes population BE).

Consider the following statistical model which assumes a four-period design with equal replication of T and R in each of  $s$  sequences with an assumption of no (or equal) carryover effects (equal carryovers go into the period effects)

$$Y_{ijkl} = \mu_k + \gamma_{ikl} + \delta_{ijk} + \varepsilon_{ijkl}$$

where  $i = 1, \dots, s$  indicates sequence,  $j = 1, \dots, n_i$  indicates subject within sequence  $i$ ,  $k = R, T$  indicates treatment,  $l = 1, 2$  indicates replicate on treatment  $k$  for subjects within sequence  $i$ .  $Y_{ijkl}$  is the response of replicate  $l$  on treatment  $k$  for subject  $j$  in sequence  $i$ ,  $\gamma_{ikl}$  represents the fixed effect of replicate  $l$  on treatment  $k$  in sequence  $i$ ,  $\delta_{ijk}$  is the random subject effect for subject  $j$  in sequence  $i$  on treatment  $k$ , and  $\varepsilon_{ijkl}$  is the random error for subject  $j$  within sequence  $i$  on replicate  $l$  of treatment  $k$ . The  $\varepsilon_{ijkl}$ 's are assumed to be mutually independent and identically distributed as

$$\varepsilon_{ijkl} \sim N(0, \sigma_{wk}^2)$$

for  $i = 1, \dots, s$ ,  $j = 1, \dots, n_i$ ,  $k = R, T$ , and  $l = 1, 2$ . Also, the random subject effects

$\delta_{ij} = (\mu_R + \delta_{ijR}, \mu_T + \delta_{ijT})'$  are assumed to be mutually independent and distributed as

$$\delta_{ij} \sim N_2 \left[ \begin{pmatrix} \mu_R \\ \mu_T \end{pmatrix}, \begin{pmatrix} \sigma_{BR}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BT}^2 \end{pmatrix} \right].$$

The following constraint is applied to the nuisance parameters to avoid overparameterization of the model for  $k=R, T$ :

$$\sum_{i=1}^s \sum_{l=1}^2 \gamma_{ikl} = 0$$

This statistical model proposed by Chinchilli and Esinhart assumes  $s \cdot p$  location parameters (where  $p$  is the number of periods) that can be partitioned into  $t$  treatment parameters and  $sp-t$  nuisance parameters (Chinchilli and Esinhart 1996). This produces a saturated model. The various *nuisance* parameters are estimated in this model, but the focus is on the parameters needed for population BE. In some designs, the sequence and period effects can be estimated through a reparametrization of the nuisance effects.

This model definition can be extended to other crossover designs.

**Linearized Criteria (from section IV. B, equations 4 and 5):**

- Reference-Scaled:

$$\eta_h = (\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2) - \theta_p \cdot \sigma_{TR}^2 < 0$$

- Constant-Scaled:

$$\eta_b = (\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2) - \theta_p \cdot \sigma_{T0}^2 < 0$$

**Estimating the Linearized Criteria:**

The estimation of the linearized criteria depends on study designs. The remaining estimation and confidence interval procedures assume a four-period design with equal replication of T and R in each of  $s$  sequences. The reparametrizations are defined as:

$$U_{Ty} = \frac{1}{2} * (Y_{yT1} + Y_{yT2})$$

$$U_{Ry} = \frac{1}{2} * (Y_{yR1} + Y_{yR2})$$

$$V_{Ty} = \frac{1}{\sqrt{2}} * (Y_{yT1} - Y_{yT2})$$

$$V_{Ry} = \frac{1}{\sqrt{2}} * (Y_{yR1} - Y_{yR2})$$

$$I_{y} = Y_{yT.} - Y_{yR.},$$

for  $i = 1, \dots, s$  and  $j = 1, \dots, n_i$ , where

$$Y_{yT\Box} = \frac{1}{2}(Y_{yT1} + Y_{yT2}) \quad \text{and} \quad Y_{yR\Box} = \frac{1}{2}(Y_{yR1} + Y_{yR2}) .$$

Compute the formulation means pooling across sequences:

$$\hat{\mu}_k = \frac{1}{S} \sum_{i=1}^s \bar{Y}_{ik}, \quad k = R, T \quad \text{and} \quad \hat{\Delta} = \hat{\mu}_T - \hat{\mu}_R$$

where

$$\bar{Y}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{2} \sum_{l=1}^2 Y_{ykl} .$$

Compute the variances of  $U_{Tij}, U_{Rij}, V_{Tij}, V_{Rij}$ , pooling across sequences, and denote these variance estimates by  $MU_T, MU_R, MV_T, MV_R$ , respectively. Specifically,

$$MU_T = \frac{1}{n_{U_T}} \sum_{i=1}^s \sum_{j=1}^{n_i} (U_{Tij} - \bar{U}_{Ti})^2$$

$$MV_T = \frac{1}{n_{V_T}} \sum_{i=1}^s \sum_{j=1}^{n_i} (V_{Tij} - \bar{V}_{Ti})^2$$

$$MU_R = \frac{1}{n_{U_R}} \sum_{i=1}^s \sum_{j=1}^{n_i} (U_{Rij} - \bar{U}_{Ri})^2$$

$$MV_R = \frac{1}{n_{V_R}} \sum_{i=1}^s \sum_{j=1}^{n_i} (V_{Rij} - \bar{V}_{Ri})^2$$

$$n_i = n_{U_T} = n_{U_R} = n_{V_T} = n_{V_R} = \left( \sum_{i=1}^s n_i \right) - s$$

Then, the linearized criteria are estimated by:

- Reference-Scaled:

$$\hat{\eta}_1 = \hat{\Delta} + MU_T + 0.5 \cdot MV_T - (1 + \theta_p) \cdot [MU_R + 0.5 \cdot MV_R]$$

- Constant-Scaled:

$$\hat{\eta}_2 = \hat{\Delta}^2 + MU_T + 0.5 \cdot MV_T - (1) \cdot [MU_R + 0.5 \cdot MV_R] - \theta_p \cdot \sigma_{T0}$$

### 95% Upper Confidence Bounds for Criteria:

The table below illustrates the construction of a  $(1 - \alpha)$  level upper confidence bound based on the two-sequence, four-period design, for the reference-scaled criterion,  $\hat{\eta}_1$ . Use  $\alpha=0.05$  for a 95% upper confidence bound.

$H_q = \text{Confidence Bound}$	$E_q = \text{Point Estimate}$	$U_q = (H_q - E_q)^2$
$H_D = \left( \left  \hat{\Delta} \right  + t_{1-\alpha, n-s} \left( \frac{1}{s^2} \sum_{i=1}^s n_i^{-1} M_i \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	$U_D$
$H1 = \frac{(n-s) \cdot E1}{\chi^2_{n-s, \alpha}}$	$MU_T = E1$	$U1$
$H2 = \frac{(n-s) \cdot E2}{\chi^2_{n-s, \alpha}}$	$0.5 \cdot MV_T = E2$	$U2$
$H3rs = \frac{(n-s) \cdot E3rs}{\chi^2_{n-s, 1-\alpha}}$	$-(1 + \theta_p) MU_R = E3rs$	$U3rs$
$H4rs = \frac{(n-s) \cdot E4rs}{\chi^2_{n-s, 1-\alpha}}$	$-(1 + \theta_p) \cdot 0.5 \cdot MV_R = E4rs$	$U4rs$
$H_{\eta_1} = \sum E_q + \left( \sum U_q \right)^{1/2}$		

$H_{\eta_1} = \sum E_q + \left( \sum U_q \right)^{1/2}$  is the upper 95% confidence bound for  $\hat{\eta}_1$ . Note  $n = \sum_{i=1}^s n_i$ , where  $s$  is the number of sequences,  $n_i$  is the number of subjects per sequence, and  $\chi^2_{\alpha, n-s}$  is from the cumulative distribution function of the chi-square distribution with  $n - s$  degrees of freedom, i.e.

$\Pr(\chi^2_{n-s} \leq \chi^2_{\alpha, n-s}) = \alpha$ . The confidence bound for  $\hat{\eta}_2$  is computed similarly, adjusting the constants associated with the variance components where appropriate (in particular, the constant associated with  $MU_R$  and  $MV_R$ ).

$H_q = \text{Confidence Bound}$	$E_q = \text{Point Estimate}$	$U_q = (H_q - E_q)^2$
$H_D = \left( \left  \hat{\Delta} \right  + t_{1-\alpha, n-s} \left( \frac{1}{s^2} \sum_{i=1}^s n_i^{-1} M_i \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	$U_D$
$H1 = \frac{(n-s) \cdot E1}{\chi^2_{n-s, \alpha}}$	$MU_T = E1$	$U1$
$H2 = \frac{(n-s) \cdot E2}{\chi^2_{n-s, \alpha}}$	$0.5 \cdot MV_T = E2$	$U2$
$H3cs = \frac{(n-s) E3cs}{\chi^2_{n-s, 1-\alpha}}$	$-1 \cdot MU_R = E3cs$	$U3cs$
$H4cs = \frac{(n-s) E4cs}{\chi^2_{n-s, 1-\alpha}}$	$-0.5 \cdot MV_R = E4cs$	$U4cs$
$H_{\eta_2} = \sum E_q - \theta_p \cdot \sigma_{T0}^2 + \left( \sum U_q \right)^{1/2}$		

Using the mixed-scaling approach, to test for population BE, compute the 95% upper confidence bound of either the reference-scaled or constant-scaled linearized criterion. The selection of either reference-scaled or constant-scaled approach depends on the study estimate of total standard deviation of the reference product (estimated by  $[MU_R + 0.5 \cdot MV_R]^{1/2}$  in the four-period design). If the study estimate of standard deviation is  $\leq \sigma_{T0}$ , the constant-scaled criterion and its associated confidence interval should be computed. Otherwise, the reference-scaled criterion and its confidence interval should be computed. The procedure for computing each of the confidence bounds is described above. If the upper confidence bound for the appropriate criterion is negative or zero, conclude population BE. If the upper bound is positive, do not conclude population BE.

## APPENDIX G

### Method for Statistical Test of Individual Bioequivalence Criterion

Appendix G describes a method for using the individual BE criterion (see section IV.C, equations 6 and 7). The procedure (Hyslop, Hsuan, and Holder 2000) involves the computation of a test statistic that is either positive (does not conclude individual BE) or negative (concludes individual BE).

Consider the following statistical model that assumes a four-period design with equal replication of T and R in each of  $s$  sequences with an assumption of no (or equal) carryover effects (equal carryovers go into the period effects)

$$Y_{ijkl} = \mu_k + \gamma_{ikl} + \delta_{ijk} + \varepsilon_{ijkl}$$

where  $i = 1, \dots, s$  indicates sequence,  $j = 1, \dots, n_i$  indicates subject within sequence  $i$ ,  $k = R, T$  indicates treatment,  $l = 1, 2$  indicates replicate on treatment  $k$  for subjects within sequence  $i$ .  $Y_{ijkl}$  is the response of replicate  $l$  on treatment  $k$  for subject  $j$  in sequence  $i$ ,  $\gamma_{ikl}$  represents the fixed effect of replicate  $l$  on treatment  $k$  in sequence  $i$ ,  $\delta_{ijk}$  is the random subject effect for subject  $j$  in sequence  $i$  on treatment  $k$ , and  $\varepsilon_{ijkl}$  is the random error for subject  $j$  within sequence  $i$  on replicate  $l$  of treatment  $k$ . The  $\varepsilon_{ijkl}$ 's are assumed to be mutually independent and identically distributed as

$$\varepsilon_{ijkl} \sim N(0, \sigma_{wk}^2)$$

for  $i = 1, \dots, s$ ,  $j = 1, \dots, n_i$ ,  $k = R, T$ , and  $l = 1, 2$ . Also, the random subject effects

$\delta_{ij} = (\mu_R + \delta_{ijR}, \mu_T + \delta_{ijT})'$  are assumed to be mutually independent and distributed as

$$\delta_{ij} \sim N_2 \left[ \begin{pmatrix} \mu_R \\ \mu_T \end{pmatrix}, \begin{pmatrix} \sigma_{BR}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BT}^2 \end{pmatrix} \right].$$

The following constraint is applied to the nuisance parameters to avoid overparameterization of the model for  $k = R, T$ :

$$\sum_{i=1}^s \sum_{l=1}^2 \gamma_{ikl} = 0$$

This statistical model proposed by Chinchilli and Esinhart assumes  $s \cdot p$  location parameters (where  $p$  is the number of periods) that can be partitioned into  $t$  treatment parameters and  $s \cdot t$  nuisance parameters (Chinchilli and Esinhart, 1996). This produces a saturated model. The various *nuisance* parameters are estimated in this model, but the focus is on the parameters needed for individual BE. In some designs, the sequence and period effects can be estimated through a reparametrization of the nuisance effects.

This model definition can be extended to other crossover designs.

**Linearized Criteria (from section IV. C, equations 6 and 7) :**

- Reference-Scaled:

$$\eta_h = (\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2) - \theta_l \cdot \sigma_{WR}^2 < 0$$

- Constant-Scaled:

$$\eta_b = (\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2) - \theta_l \cdot \sigma_{W0}^2 < 0$$

**Estimating the Linearized Criteria:**

The estimation of the linearized criteria depends on study designs. The remaining estimation and confidence interval procedures assume a four-period design with equal replication of T and R in each of  $s$  sequences. The reparametrizations are defined as:

$$I_{ij} = Y_{ijT} - Y_{ijR}$$

$$T_{ij} = Y_{ijT1} - Y_{ijT2}$$

$$R_{ij} = Y_{ijR1} - Y_{ijR2}$$

for  $i = 1, \dots, s$  and  $j = 1, \dots, n_i$ , where

$$Y_{yT\Box} = \frac{1}{2}(Y_{yT1} + Y_{yT2}) \quad \text{and} \quad Y_{yR\Box} = \frac{1}{2}(Y_{yR1} + Y_{yR2})$$

Compute the formulation means, and the variances of  $I_y$ ,  $T_y$ , and  $R_y$ , pooling across sequences, and denote these variance estimates by  $M_I$ ,  $M_T$ , and  $M_R$ , respectively, where

$$\hat{\mu}_k = 1/S \sum_{i=1}^s \bar{Y}_{i,k}, \quad k=R, T \quad \text{and} \quad \hat{\Delta} = \hat{\mu}_T - \hat{\mu}_R$$

$$\bar{Y}_{i,k} = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{2} \sum_{l=1}^2 Y_{yjl}$$

$$M_I = \hat{\sigma}_I^2 = \frac{1}{n_I} \sum_{i=1}^s \sum_{j=1}^{n_i} (I_y - \bar{I}_i)^2$$

$$n_I = n_T = n_R = \left( \sum_{i=1}^s n_i \right) - s$$

$$M_T = \hat{\sigma}_{WT}^2 = \frac{1}{2n_T} \sum_{i=1}^s \sum_{j=1}^{n_i} (T_y - \bar{T}_i)^2$$

$$M_R = \hat{\sigma}_{WR}^2 = \frac{1}{2n_R} \sum_{i=1}^s \sum_{j=1}^{n_i} (R_y - \bar{R}_i)^2 .$$

Then, the linearized criteria are estimated by:

- Reference-Scaled:

$$\hat{\eta}_1 = \hat{\Delta}^2 + M_I + 0.5 \cdot M_T - (1.5 + \theta_I) \cdot M_R$$

- Constant-Scaled:

$$\hat{\eta}_2 = \hat{\Delta}^2 + M_I + 0.5 \cdot M_T - 1.5 \cdot M_R - \theta_I \cdot \sigma_{W0}^2$$

and the subject-by-formulation interaction variance component can be estimated by:

$$\hat{\sigma}_D^2 = \hat{\sigma}_I^2 - \frac{1}{2}(\hat{\sigma}_{WT}^2 + \hat{\sigma}_{WR}^2)$$

**95% Upper Confidence Bounds for Criteria:**

The table below illustrates the construction of a  $(1-\alpha)$  level upper confidence bound based on the two-sequence, four-period design, for the reference-scaled criterion,  $\hat{\eta}_1$ . Use  $\alpha=0.05$  for a 95% upper confidence bound.

$H_q$ = Confidence Bound	$E_q$ = Point Estimate	$U_q=(H_q- E_q)^2$
$H_D = \left( \left  \hat{\Delta} \right  + t_{1-\alpha, n-s} \left( \frac{1}{s^2} \sum_{i=1}^s n_i^{-1} M_I \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	$U_D$
$H_I = \frac{(n-s) \cdot M_I}{\chi^2_{\alpha, n-s}}$	$E_I = M_I$	$U_I$
$H_T = \frac{0.5 \cdot (n-s) \cdot M_T}{\chi^2_{\alpha, n-s}}$	$E_T = 0.5 \cdot M_T$	$U_T$
$H_R = \frac{-(1.5 + \theta_I) \cdot (n-s) \cdot M_R}{\chi^2_{1-\alpha, n-s}}$	$E_R = -(1.5 + \theta_I) \cdot M_R$	$U_R$
$H_{\eta_1} = \sum E_q + \left( \sum U_q \right)^{1/2}$		

where  $n = \sum_{i=1}^s n_i$ ,  $s$  is the number of sequences, and  $\chi^2_{\alpha, n-s}$  is from the cumulative distribution function of the chi-square distribution with  $n-s$  degrees of freedom, i.e.  $\Pr(\chi^2_{n-s} \leq \chi^2_{\alpha, n-s}) = \alpha$ . Then,  $H_{\eta_1} = \sum E_q + \left( \sum U_q \right)^{1/2}$  is the upper 95% confidence bound for  $\hat{\eta}_1$ . The confidence bound for  $\hat{\eta}_2$  is computed similarly, adjusting the constants associated with the variance components where appropriate (in particular, the constant associated with  $M_R$ ).

$H_q$ = Confidence Bound	$E_q$ = Point Estimate	$U_q=(H_q- E_q)^2$
$H_D = \left( \left  \hat{\Delta} \right  + t_{1-\alpha, n-s} \left( \frac{1}{s^2} \sum_{i=1}^s n_i^{-1} M_i \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	$U_D$
$H_I = \frac{(n-s) \cdot M_I}{\chi^2_{\alpha, n-s}}$	$E_I = M_I$	$U_I$
$H_T = \frac{0.5 \cdot (n-s) \cdot M_T}{\chi^2_{\alpha, n-s}}$	$E_T = 0.5 \cdot M_T$	$U_T$
$H_R = \frac{-(1.5) \cdot (n-s) \cdot M_R}{\chi^2_{1-\alpha, n-s}}$	$E_R = -(1.5) \cdot M_R$	$U_R$
$H_{\eta_2} = \sum E_q - \theta_l \cdot \sigma_{w_0}^2 + \left( \sum U_q \right)^{1/2}$		

Using the mixed-scaling approach, to test for individual BE, compute the 95% upper confidence bound of either the reference-scaled or constant-scaled linearized criterion. The selection of either reference-scaled or constant-scaled criterion depends on the study estimate of within-subject standard deviation of the reference product. If the study estimate of standard deviation is  $\leq \sigma_{w_0}$ , the constant-scaled criterion and its associated confidence interval should be computed. Otherwise, the reference-scaled criterion and its confidence interval should be computed. The procedure for computing each of the confidence bounds is described above. If the upper confidence bound for the appropriate criterion is negative or zero, conclude individual BE. If the upper bound is positive, do not conclude individual BE.

This guidance recommends that sponsors use either reference-scaling or constant-scaling at the changeover point (see section VII.D, Discontinuity). To test for individual BE, compute the 95% upper confidence bounds of both reference-scaled and constant-scaled linearized criteria. The procedure for computing these confidence bounds is described above. If the upper bound of either criterion is negative or zero (either  $H_{\eta_1}$  or  $H_{\eta_2}$ ), conclude individual BE. If the upper bounds of both criteria are positive, do not conclude individual BE.

## APPENDIX H

### Variance Estimation

Relatively simple unbiased estimators, the method of moments (MM) or the restricted maximum likelihood (REML) method, can be used to estimate the mean and variance parameters in the individual BE approach. A key distinction between the REML and MM methods relates to differences in estimating variance terms. The REML method estimates each of the three variances,  $\sigma_D^2$ ,  $\sigma_{WR}^2$ ,  $\sigma_{WT}^2$ , separately and then combines them in the individual BE criterion. The REML estimate of  $\sigma_D^2$  is found from estimates of  $\sigma_{BR}^2$ ,  $\sigma_{BT}^2$ , and the correlation,  $\rho$ . The MM approach is to estimate the sum of the variance terms in the numerator of the criterion,  $\sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2$ , and does not necessarily estimate each component separately. One consequence of this difference is that the MM estimator of  $\sigma_D^2$  is unbiased but could be negative. The REML approach can also lead to negative estimates, but if the covariance matrix of the random effects is forced to be a proper covariance matrix, the estimate of  $\sigma_D^2$  can be made to be non-negative. This forced non-negativity has the effect of making the estimate positively biased and introduces a small amount of conservatism to the confidence bound. The REML method can be used in special cases (e.g., when substantial missing data are present). In addition, the MM approaches have not yet been adapted to models that allow assessment of carryover effects.