Statistical Approaches to Establishing Bioequivalence

Guidance for Industry

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

May 2026
Biopharmaceutics
Generic Drugs

Statistical Approaches to Establishing Bioequivalence

Guidance for Industry

Additional copies are available from:
Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration

Phone: 855-543-3784 or 301-796-3400
Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs

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

May 2026
Biopharmaceutics
Generic Drugs

TABLE OF CONTENTS

Contains Nonbinding Recommendations

Statistical Approaches to Establishing Bioequivalence Guidance for Industry[1]

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA office responsible for this guidance as listed on the title page.

I. INTRODUCTION

This guidance provides recommendations on how to meet provisions of part 320 (21 CFR part 320) for all drug products. This guidance replaces prior FDA guidance for industry of the same name issued in February 2001 (2001 guidance). Part 320 sets forth requirements for (1) submitting bioavailability (BA) and bioequivalence (BE) data in investigational new drug applications (INDs), new drug applications (NDAs), and abbreviated new drug applications (ANDAs), as well as amendments and supplements to these applications; (2) the definitions of BA and BE; and (3) the types of in vitro and in vivo studies that are appropriate to measure BA and establish BE.[2]

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

A. Overview

This guidance provides recommendations to applicants who intend to use equivalence criteria in analyzing in vivo or in vitro BE studies for INDs, NDAs, and ANDAs as well as amendments and supplements to these applications.[3] This guidance discusses statistical approaches for BE comparisons and focuses on how to use these approaches both generally and in specific situations. This guidance provides recommendations on the topics covered in the 2001 guidance, including average BE and population BE, study design, logarithmic transformation, studies in multiple groups, carryover effects, and outliers, as well as recommendations on additional topics,

[1] This guidance has been prepared by the Office of Generic Drugs in the Center for Drug Evaluation and Research in cooperation with the Center for Drug Evaluation and Research's Office of Translational Sciences and Office of Pharmaceutical Quality at the Food and Drug Administration.

[2] The definitions of BA and BE are set forth in 21 CFR 314.3(b) and cross-referenced in 21 CFR 320.1.

[3] The term "applicant" as used in this guidance refers to both sponsors of INDs and applicants of NDAs and ANDAs.

including missing data and intercurrent events, adaptive design, and specific situations, such as narrow therapeutic index (NTI) drugs and highly variable drugs.

The evaluation of *relative BA* involves the BE comparison between a test product (test or T) and a reference material (reference or R), where T and R can vary depending on the comparison to be performed (e.g., to-be-marketed formulation versus clinical trial formulation, generic drug versus reference listed drug (RLD), originally approved formulation versus postapproval change(s) formulation). 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.[4] In these contexts, 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 equivalents (i.e., drug products in identical dosage forms and route(s) of administration that contain identical amounts of the identical active drug ingredient)[5] or performing equivalence comparisons in clinical pharmacology studies and other areas.

This guidance is intended to encourage the use of science-based approaches to making statistical BE assessments. Given the evolving nature of statistical approaches and technologies, FDA encourages generic and new drug applicants to propose and discuss novel methodologies, such as model-based BE or novel adaptive designs, with the Agency through appropriate regulatory meetings or controlled correspondences, as described below.

B. Statistical Guidance Background

In the July 1992 guidance for industry *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design* (the 1992 guidance), the Center for Drug Evaluation and Research (CDER) recommended that a standard in vivo BE study design be based on the administration of either single or multiple doses of T and R 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 (PK) measures, such as area under the curve (AUC) and peak concentration (C_{max}), be based on the *two one-sided tests procedure* to determine whether the average values for the PK measures determined after administration of T and R were comparable. This approach is termed *average BE* and involves the calculation of a 90 percent confidence interval for the ratio of the averages (population geometric means) of the PK measures for T and R. To establish BE, the calculated confidence interval should fall within a BE limit, usually 80.00

[4] For example, to submit an ANDA that is not the same as its RLD because it has a different strength, dosage form, or route of administration than that of the RLD, an applicant first must obtain permission from FDA through the citizen petition process. See section 505(j)(2)(C) of the Federal Food, Drug and Cosmetic Act (21 U.S.C. 355(j)(2)(C)); 21 CFR 314.93(b). Such petitions are referred to as *suitability petitions*.

[5] 21 CFR 314.3(b).

percent to 125.00 percent for the ratio of the product averages.[6] In addition to this general approach, the 1992 guidance provided specific recommendations for (1) logarithmic transformation of PK data, (2) methods to evaluate sequence effects, and (3) methods to evaluate outlier data.

In addition to reiterating the key points from the 1992 guidance and replacing that guidance, the 2001 guidance introduced two additional approaches to assessing BE: *population BE* and *individual BE*. Both of these approaches, unlike the *average BE* approach, include a comparison of the variabilities of the PK metrics of the two products being compared, as well as the average responses. However, the individual BE approach is not currently used in the regulatory setting, while the population BE approach is mainly used for certain in vitro BE studies when variability as well as means need to be compared for clinical reasons. The 2001 guidance also included discussion of *replicate crossover designs*, which involve the administration of at least one of the products to at least some of the subjects more than once. The discussion of these designs in that guidance included their implications for possible carryover effects and their use in screening for outliers.

This guidance provides recommendations on the topics covered by the 1992 guidance and the 2001 guidance, as well as recommendations on some additional topics. As noted in the Introduction section above, this guidance replaces the 2001 guidance.

II. GENERAL CONSIDERATIONS

A. Study Design

1. Experimental Design

Various experimental designs can be used to establish BE. For most BE studies, the typical experimental design is a nonreplicate design (see section II.A.1.a, Nonreplicate designs). For highly variable drugs (see section III.C, Statistical Methods for Highly Variable Drugs) or NTI drug products (see section III.B, Statistical Methods for Narrow Therapeutic Index Drugs), a replicate crossover design (see section II.A.1.b, Replicate crossover designs) is often used. When applicants are uncertain about study design parameters, such as the variability of the PK parameters or comparative clinical endpoint, an adaptive design (see section II.A.1.c, Adaptive design) may be appropriate. In studies where sampling can only be done at a single time point or limited number of time points for each subject, a sparse design (see section II.A.1.d, Design with

[6] For a broad range of drugs, a BE limit of 80.00 percent to 125.00 percent for the ratio of the product averages has been adopted for use of an average BE criterion. Generally, the BE limit of 80.00 percent to 125.00 percent is based on a clinical judgment that a test product with BA measures outside this range should be denied market access. To pass a confidence interval limit of 80 percent to 125 percent, the rounded confidence interval value should be at least 80.00 percent and not more than 125.00 percent. See the guidance for industry *Bioequivalence Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (May 2026). FDA updates guidances periodically. For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

sparse sampling) may be appropriate. In these last two cases, the underlying experimental design can be either a crossover or parallel group design.

a. Nonreplicate designs

A conventional nonreplicate design, such as the two-formulation, two-period, two-sequence crossover design, commonly referred to as a two-way crossover design, can be used to generate data when an average or population approach is chosen for BE comparisons. Under certain circumstances, such as products with apparent, long half-lives, where crossover designs are impractical, parallel designs can be used.

b. Replicate crossover designs

A replicate crossover study design (either partial or fully replicate) is appropriate for drugs regardless of whether the reference is a highly variable drug or not. A replicate design can have the advantage of using fewer subjects compared to a nonreplicate design, although each subject in a replicate design would receive more treatments.

Further, a replicate design is recommended to be used under the following scenarios:

- A replicate design is advantageous over a nonreplicate design for non-NTI drugs with a high within-subject variability in at least one of the PK measures. Either a partial or fully replicate design can be used in this scenario, but the reference-scaled average BE analysis approach should only be applied to specific PK metrics that exhibit a high within-subject variability for the reference in the pivotal BE study.
- A fully replicate design is recommended for NTI drugs, where within-subject variability for both the reference material and test product can be computed and a reference scaled-average BE analysis can be conducted to properly adjust the BE acceptance criteria.
- A fully replicate design may be recommended for certain non-NTI drugs when the BE approach includes a comparison of the within-subject variability of the test product and reference material.

Replicate crossover designs can be used irrespective of which BE approach is selected to establish BE, although they are not necessary when an average or population BE approach is used. When a reference-scaled average BE approach is used, replicate crossover designs are critical to allow estimation of within-subject variances of each PK parameter for R (and T if a fully replicate study is used). For BE studies with fully replicate designs, the four-period, two-sequence, two-formulation design, commonly referred to as four-way, full-replicate design, shown below, is recommended.

		Period			
		1	***2***	***3***	***4***
Sequence	***1***	*T*	*R*	*T*	*R*
	2	*R*	*T*	*R*	*T*

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

Other fully replicate crossover designs are also possible. For example, a three-period design, as shown below, could be used. A fully replicate design can estimate the subject-by-formulation interaction variance components.

		Period		
		1	***2***	***3***
Sequence	***1***	*T*	*R*	*T*
	2	*R*	*T*	*R*

The following three-period, three-sequence, two-formulation, partially replicate design can also be used for assessing reference-scaled BE, though it cannot fully estimate the subject-by-formulation interaction variance component (as a fully replicate design can).

		Period		
		1	***2***	***3***
Sequence	***1***	*T*	*R*	*R*
	2	*R*	*T*	*R*
	3	*R*	*R*	*T*

For these replicate designs, a greater number of subjects would be needed for the three-period designs compared to the recommended four-period design to achieve the same statistical power to conclude BE. For further discussion on replicate crossover designs see Appendix A.

c. Adaptive design

An adaptive design is a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial. An adaptive design can be a group sequential design, or other design with one or more adaptive features.[7] For example, for PK BE studies, Potvin's methods (Potvin et al. 2008, Xu et al. 2016),[8] are a combination of a group sequential design and an adaptive design with sample size re-estimation; Maurer et al.'s proposed adaptive design allows control of the Type I error rate analytically (Maurer et al. 2019).

[7] See the guidance for industry *Adaptive Designs for Clinical Trials of Drugs and Biologics* (November 2019).

[8] See the References section for the complete citations for all published literature referenced in this guidance.

Adaptive design can provide ethical advantages[9] and statistical efficiency. When appropriately implemented, adaptive designs can reduce resources used, decrease time to study completion, and increase the chance of study success, especially when the prior information needed for the study design is limited. However, use of adaptive designs can also have limitations. For example, adaptive designs may call for certain statistical methods to avoid increasing the chance of erroneous conclusions and introducing bias in estimates, and for complex adaptive designs, such methods may not be readily available.[10] The decision to use or not use an adaptive design is at the applicant's discretion.

In general, the design, conduct, and analysis of a proposed adaptive study design should satisfy the following recommendations:

- The details of the adaptive design should be completely specified prior to initiation of the study and documented accordingly. For example, prospective planning should include prespecification of the anticipated number and timing of interim analyses, the type of adaptation, the statistical inference methods to be used and the specific algorithm or criteria governing the adaptive decision. If a study could be stopped early (e.g., for futility or for success in demonstrating BE), detailed stopping criteria should be prespecified and scientifically justified.

- The applicant should establish that estimation of treatment effect will be sufficiently reliable, and the chance of erroneous conclusions will be adequately controlled. The Agency will accept appropriately designed BE studies that are scientifically justified. Support might include published literature in peer-reviewed journals in which the applicant's proposed approach is validated or simulation results that meet desired criteria (e.g., the Type I error probability of the proposed approach is controlled at a nominal level of 0.05 for a BE test). Appropriate details (e.g., literature references, proofs, simulation codes/results) for the methodology should be submitted.

- The applicant should ensure that study integrity will be appropriately maintained. A comprehensive written data access plan defining how study integrity will be maintained in the presence of the planned adaptation should be included in the protocol or statistical analysis plan (SAP). This applies to both comparative clinical endpoint BE studies and PK BE studies, whether blinded or unblinded by design.

For details, refer to the guidance for industry *Adaptive Design for Clinical Trials of Drugs and Biologics* (November 2019).

Due to the increased complexity of adaptive studies and uncertainties regarding their operating characteristics, applicants are encouraged to contact the Agency early to discuss their proposed

[9] See footnote 7. For example, the ability to stop a trial early if it becomes clear that the trial is unlikely to demonstrate equivalence can reduce the number of patients exposed to the unnecessary risk of an ineffective investigational treatment and allow subjects the opportunity to explore more promising therapeutic alternatives.

[10] See footnote 7, and Potvin et al. 2008, Xu et al. 2016, and Maurer et al. 2019.

adaptive study designs and statistical methods via the controlled correspondence,[11] pre-ANDA meeting,[12] pre-IND meeting, or pre-NDA meeting pathway, as appropriate.[13]

d. Design with sparse sampling

For certain generic products, a sparse BE design is used, where the sampling for each subject is done at a single or very limited number of time points rather than the number of time points needed to get a full concentration profile. For example, some ophthalmic products are studied using a sparse BE design, where only a single sample is collected from a single eye of each subject at one assigned sampling time point for that subject. More generally, a sparse BE study design can be a parallel design where each subject should receive only one treatment, T or R, but not both. Alternatively, a crossover sparse study design can be used where each subject receives both T and R (e.g., in subjects undergoing indicated cataract surgery for both eyes).

For a sparse BE study design, the mean concentration for each product at each time point of measurement is calculated by using the mean concentration of the subjects measured at each time point to derive the mean profile for each product. Based on the trapezoid rule, the AUC_{0-t} for each product is computed as a weighted linear combination of these mean concentrations at each time point through time t. The AUC_{0-t} is the area under the concentration – time curve from zero to the time t. C_{max} and time to maximum observed concentration (T_{max}) can be determined accordingly. The ratios of AUC_{0-t} and C_{max} between the test product and the reference material are used to assess BE. Estimation of the standard deviation and confidence interval for the ratio of AUC_{0-t} may be done by bootstrap or parametric methods (e.g., Bailer's methods (Bailer 1988) for a parallel study design), and estimation of the standard deviation and confidence interval for the ratio of C_{max} may be done by bootstrap methods. BE is supported if the 90 percent confidence interval for the ratio of AUC_t, at each time point of interest, between the test product and the reference material lies within the BE margin (80.00 percent to 125.00 percent). Model-based approaches using nonlinear mixed effects models can be considered when they can reliably control the error rate of concluding BE for bioinequivalent products (Type I error) (Zhao et al. 2019; Loingeville et al. 2020; Gong et al. 2023).

For complicated issues, such as other forms of sparse design or alternative statistical methods, applicants are encouraged to contact the Agency early to discuss their proposed study design and statistical methods via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway, as appropriate.[14]

---

[11] See the guidance for industry *Controlled Correspondence Related to Generic Drug Development* (March 2024).

[12] See the guidance for industry *Formal Meetings Between FDA and ANDA Applicants of Complex Products Under GDUFA* (October 2022).

[13] See the draft guidance for industry *Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products* (September 2023). When final, this guidance will represent FDA's current thinking on this topic.

[14] See footnotes 11, 12, and 13.

2. Estimands and Intercurrent Events

The guidance for industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (May 2021) introduces the concept of an estimand, which is a precise description of the treatment effect reflecting the clinical question posed by a particular study objective. The trial protocol of a BE study should include the following five components of an estimand:

(1) the treatment of interest and alternative treatment(s) to which comparison will be made (e.g., test product compared with reference material);
(2) the population;
(3) the variable (or endpoint) to be measured for each subject (e.g., AUC and C_{max});
(4) strategies for handling intercurrent events in assessing the scientific question of interest. Subjects may have intercurrent (post-randomization) events, such as noncompliance for various reasons or use of rescue medication due to lack of efficacy, that affect either the interpretation or the existence of the measurements associated with the question of interest. For example, in a comparative clinical endpoint BE study with a binary endpoint, it may be reasonable to consider including subjects who discontinue study treatment early due to lack of treatment effect as treatment failures;
(5) the population-level summary for the variable to compare between treatment conditions (for example, the geometric mean ratio of the test product to reference material in a PK BE study).

3. Sample Size Determination

It is an applicant's responsibility to ensure a proposed BE study is designed with adequate power. FDA recommends that applicants enroll enough subjects to power the study at a level of 0.8 or higher for a BE test to be carried out with a target effect size (e.g., geometric mean ratio 0.95) at a Type I error rate of 0.05 (see section II.C.1.a, Logarithmic transformation for PK measures for more details). When determining the sample size, rates of attrition and noncompliance should be taken into consideration. Applicants should enroll enough subjects in the study to account for possible dropouts. Data from all subjects who were dosed should be submitted. Eligibility criteria for inclusion in the analysis should be prespecified in the protocol, SAP, or both. Enough subjects should be recruited, randomized, and dosed at the beginning of the study to ensure that the desired number of evaluable subjects will be available for analysis.

For BE studies, dropouts generally should not be replaced because replacement of subjects during the study could impact the statistical model, analysis, and inference. In certain situations, e.g., if the number of evaluable subjects falls below the calculated sample size, and applicants want to add more subjects with justification that the addition of subjects will not bias the study conclusion, additional cohort(s) of subjects may be added to the study prior to bioanalysis. However, this should be prespecified in the study protocol, and a prespecified modification of the statistical analysis is recommended. Additional subjects can be added after bioanalysis only in a prespecified adaptive design with a prespecified adaptation to add subjects and statistical methods to control the Type I error rate under the nominal level.

The number of subjects to be included in a study should be based on an appropriate sample size calculation for the proposed study design (Chow and Liu 2008; Patterson and Jones 2017).[15] For example, the sample size calculation for the standard two-way crossover study should not be used for a different study design. For sample size re-estimation in an adaptive study design, refer to section II.A.1.c. Adaptive design.

Sample size and power calculation should be supported by established scientific practice. For complex study designs with no analytical solutions for sample size calculation, simulation(s) can be used to estimate the needed sample size in order to reach a desired power. The method by which the sample size is determined should be specified in the protocol, together with the estimates of any quantities used in the calculations (e.g., variances, mean values, response rates, the assumed effect size for both the test product and reference material, and the assumed rates of attrition). The basis for these estimates should also be specified. For example, variance estimates can be obtained from the biomedical literature and/or pilot studies. It is important to investigate the sensitivity of the sample size calculated to a variety of deviations from the assumed estimates. This may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from the assumptions or alternative approaches supported by published peer-reviewed literature.

In general, for PK BE or in vitro BE studies, sample size calculation should be based on BE metrics (e.g., AUC, C_{max}) after log-transformation; for comparative clinical endpoint BE studies, sample size calculation should be based on the untransformed comparative clinical endpoints unless otherwise noted in the relevant FDA product-specific guidance (PSG) or prespecified and justified in the protocol.[16] The number of evaluable subjects in a PK BE study should not be less than 12. For highly variable drug products, a minimum of 24 subjects are recommended for BE assessment (Davit and Conner 2010).

B. Data Preparation

The drug concentration in biological fluid determined at each sampling time point as well as the PK measures of systemic exposure should be provided on the original scale for each subject participating in the study. The variables for a comparative clinical endpoint BE study should also be provided on the original scale. The mean, standard deviation, and coefficient of variation for each variable should be computed and tabulated in the final report.

1. Log-Transformation

A general approach to assessing BE is to compare the log-transformed BA measures after administration of T and R.

[15] Guidance for industry *Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (May 2026).

[16] For the most recent version of a product-specific guidance, check the product-specific guidances web page at https://www.accessdata.fda.gov/scripts/cder/psg/index.cfm.

a. Logarithmic transformation for PK measures

The limited sample size in a typical BE study precludes a reliable determination of the distribution of the data set. This guidance thus recommends that PK BE measures (e.g., AUC and C_{max}) be log-transformed (see Appendix B). The choice of common or natural logs should be consistent and should be stated in the study report. Applicants should not 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 applicants believe that their BE study data should be statistically analyzed on the original rather than on the log scale. Note that the hypotheses after anti-log transformation should be consistent with that noted in the PSG.

b. Data transformation for comparative pharmacodynamic and clinical endpoint BE study

The decision on whether and how to transform a variable for a comparative pharmacodynamic (PD) or comparative clinical endpoint BE study should be specified in the protocol, especially for the primary variable(s). The basis for the variables should also be given in the protocol. For example, these variables can be obtained from FDA guidance, the biomedical literature, or pilot studies. Similar considerations apply to other derived variables, such as the use of change from baseline, percentage change from baseline, the AUC of repeated measures, or the ratio of two different variables. Subsequent clinical interpretation should be carefully considered. Regarding comparative clinical endpoint studies, in general, the log-transformation is not used. For example, in the case of the Fieller's confidence interval for the ratio of two means, the raw (untransformed) data are used for the confidence interval derivation (Fieller 1954). If an applicant believes that their comparative clinical endpoint study data should be analyzed on the log-transformed scale, justification should be provided in the protocol or SAP.

c. Negative values for baseline corrected PK or PD endpoints

Because data transformation and scales might affect BE conclusions, they should be chosen carefully and appropriately justified in the protocol.[17] If a baseline correction results in a negative plasma concentration value, the value should be set equal to 0 before calculating the baseline-corrected AUC. Refer to the relevant PSG for the baseline correction method.

2. Missing Data

Subjects may have missing data in the study for various reasons (e.g., subject's refusal to continue in the study, attrition due to worsening of conditions or emergence of adverse events, subject's failure to meet scheduled appointments for evaluation). Missing data is distinct from intercurrent events; however, both can introduce problems such as bias, misleading inference, loss of precision, and loss of power, any of which make it hard to interpret the trial outcome.

[17] For example, see Sun, W, S Grosser, and Y Tsong, 2017, Ratio of Means vs. Difference of Means as Measures of Superiority, Noninferiority, and Average Bioequivalence, Journal Biopharmaceutical Statistics, 27(2): 338-355.

The protocol or the protocol and the SAP should include plans to minimize missing data. The trial protocol or SAP should prospectively define anticipated causes of missing data, the corresponding statistical assumptions about reasons for the missing data, and how missing data will be treated in the statistical analysis. The treatment of missing data in the statistical analysis should be justified such that valid statistical inferences can be made under the assumptions about the missing data mechanism.

Statistical methods for handling missing data include complete case analysis, available case analysis, weighting methods, imputation, and model-based approaches. For example, in a two-way crossover study, a complete case analysis could be a general linear model as implemented in SAS PROC GLM, which removes all subjects with any missing observations for any variables included in the GLM model (i.e., removes subjects missing one or both periods). An available case analysis could be done using SAS PROC MIXED, which uses all observed data (e.g., in a two-way crossover study, uses all subjects with one or two complete periods of data).

Approaches for handling missing data and the statistical methods for the primary BE analysis (e.g., GLM vs. MIXED) should be prespecified in the study protocol or SAP. Depending on the nature of the assumed or likely missing data mechanism, statistical methods from any of these categories may be appropriate. The validity of a statistical approach to handle missing data depends on a variety of factors, including, but not limited to, the mechanism for missingness, the fraction of incomplete cases, the values that are missing, specifics of the analysis, and definition of the estimand. Sensitivity analyses should be prespecified in the trial protocol to evaluate the robustness of conclusions to deviations from the assumptions about the missing data mechanism. The applicant should provide detailed information about reasons for missing data and any observed intercurrent events.

For a particular drug product, if the PSG recommends certain approaches to handling missing data, the applicants should refer to that PSG. Applicants may choose to contact the Agency via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway, as appropriate, to discuss their proposed approach to handling missing data if such an approach is different from what is recommended in the PSG or if the applicants have further questions.[18]

3. Outlier Detection

Outlier data, or extreme values, in BE studies are 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, when 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 and for which there are not bioanalytical errors could indicate one of the following situations:

[18] See footnotes 11, 12, and 13.

- 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.

- 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 from that for most of the population, and for whom the two products are not bioequivalent, even though they might be bioequivalent in most of the population. In the case of product failure, the unusual response could be present for either T or R. However, in the case of a subpopulation, even if the unusual response is observed on R, there could still be concern about lack of BE of the two products. For these reasons, applicants should not remove data from the statistical analysis of BE studies solely because those data are identified as statistical outliers.

In general, outlier data (whether due to product failure, subject-by-formulation interaction, or another cause) may only be removed from the BE statistical analysis if there is real-time documentation demonstrating a protocol violation during the clinical and/or analytical/experimental phase of the BE study. Applicants should include a prospective plan with scientific rationale in the BE study protocol for handling outliers related to protocol deviation or violation in the BE statistical analysis. Note that all subject data should be submitted, and potential extreme values flagged with appropriate documentation as part of the application. However, for a replicate PK BE study, if reference-scaled average BE is used, the applicant should ensure that the calculated intrasubject variability is not inflated due to extreme values or situations. To illustrate the impact of such values on the assessment of BE, the applicant is encouraged to submit the statistical analysis with and without the statistical outliers.

To characterize aberrant observations for exploratory or quality control purposes, the choice of the appropriate technique depends on whether there are outlying subjects or outlying observations, as well as on the study design.

C. Statistical Models

1. General Statistical Criteria for BE

The general structure of a BE criterion is that a function (Θ) of population measures should be demonstrated to be no greater than a specified value (θ). Using the terminology of statistical hypothesis testing, this is accomplished by testing the hypothesis H_0: $\Theta \geq \theta$ versus H_a: $\Theta < \theta$ at a desired level of significance, often 5 percent. Rejection of the null hypothesis H_0 (i.e., demonstrating that the estimate of Θ is statistically significantly less than θ) results in a conclusion of BE.

a. Use of confidence intervals to perform two one-sided tests

In BE assessment we are frequently interested in testing whether a parameter (e.g., the difference of means for T and R for a specific endpoint) is contained within a defined interval, called $[\theta_1, \theta_2]$. The recommended method for doing such a test is the *Two One-Sided Tests Procedure* (Schuirmann 1987). A one-sided statistical test is carried out to determine whether the parameter is $\geq \theta_1$, and a second one-sided test is carried out to determine whether the parameter is $\leq \theta_2$; both tests are carried out at a level of significance α, which is usually 0.05. If both tests are successful (i.e., we reject the null hypothesis in both cases), we conclude that the parameter is contained in $[\theta_1, \theta_2]$.

These two one-sided tests are sometimes carried out by calculating a 100 (1-2α) percent confidence interval for the parameter and determining whether this confidence interval is completely contained in the interval $[\theta_1, \theta_2]$. For this confidence interval method of carrying out the tests to be valid, the confidence interval should be an *equal tails* confidence interval. If the lower and upper confidence limits of the 100 (1-2α) percent confidence interval are L_1 and L_2, respectively, then the confidence interval is *equal tails* if L_1, by itself, is at least a 100 (1-α) percent lower confidence bound for the parameter and L_2, by itself, is at least a 100 (1-α) percent upper confidence bound for the parameter.

In some cases, there may not be general agreement as to the best choice of a particular statistical testing methodology for carrying out the two one-sided tests (for example, if the parameter of interest is the difference between the success probabilities for T and R for a binary endpoint). In such cases, careful consideration should be given to the choice of statistical methods for doing the two one-sided tests, which may or may not correspond to a confidence interval method.

2. Statistical Information and Implementation of Criteria for PK Measures (AUC_{0-t}, $AUC_{0-\infty}$, and C_{max})

FDA recommends that applicants provide the following statistical information for AUC_{0-t}, $AUC_{0-\infty}$, and C_{max}:

- Geometric means for the formulations tested
- Arithmetic means for the formulations tested
- Geometric mean ratios of T vs. R and their corresponding 90 percent confidence intervals or 95 percent upper confidence bounds (e.g., for highly variable drugs or NTI drugs)

Recommended statistical information for other types of outcome measures is discussed in section III: Specific Situations.

To facilitate BE comparisons, for crossover studies, the measures for each individual should be displayed in parallel for the formulations tested. For each BE measure, the ratio of the individual geometric mean of the test product to the individual geometric mean of the reference material should be tabulated side by side. The summary tables should indicate in which sequence each subject received the product.

Statistical analyses of BE data are typically based on a statistical model for the logarithm of the BA measures (e.g., AUC and C_{max}). 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, μ_{Tj} and μ_{Rj} for the T and R formulations, respectively. The model assumes that these subject-specific means come from a distribution with population means μ_T and μ_R, and between-subject variances σ_{BT}^2 and σ_{BR}^2, respectively. The model allows for a correlation, ρ, between μ_{Tj} and μ_{Rj}. The subject-by-formulation interaction variance component, σ_D^2, is related to these parameters as follows:

$$\sigma_D^2 = variance\ of\ \left(\mu_{T_i} - \mu_{TR_i}\right)$$

$$= (\sigma_{BT} - \sigma_{BR})^2 + 2(1-\rho)\sigma_{BT}\sigma_{BR}\ ^{[19]}$$

For a given subject, the observed data for the log-transformed BA measure are assumed to be independent observations from distributions with means μ_{Tj} and $\mu_{Rj}$and within-subject variances σ_{WT}^2 and σ_{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.

The applicant may also consider prespecifying inclusion of important demographic and baseline prognostic covariates in the statistical model for parallel studies. This sort of adjustment can increase the precision and power of the statistical analysis and compensate for any lack of balance between treatment groups with no inflation of Type I error.

III. SPECIFIC SITUATIONS[20]

A. In Vitro BE and Population BE

This section discusses statistical methods for assessment of in vitro BE, including population BE (PBE), a similarity index (f_2), statistical approaches respectively for in vitro release tests (IVRT), in vitro permeation tests (IVPT) and in vitro abuse-deterrent formulations (ADF) comparative studies, and a profile comparison approach based on Earth Mover's Distance (EMD).

1. PBE

One of the recommended statistical approaches for evaluating in vitro BE is PBE. To test for PBE, the null and alternative hypotheses are given as follows:

---

[19] Schall, R., and H. G. Luus, 1993, On Population and Individual Bioequivalence, Statistics in Medicine, 12(12): 1109-1124.

[20] Some specific situations are addressed in the following subsections with specified choices of BE criteria. Further discussion regarding these specified choices can be found in the guidances cited in those subsections.

$$H_0: \theta \geq \theta_P \quad \text{vs.} \quad H_a: \theta < \theta_P$$

where $\theta = \frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\sigma_R^2}$ if the estimated $\sigma_R > \sigma_0$ or $\theta = \frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\sigma_0^2}$ if the estimated $\sigma_R \leq \sigma_0$.

Here, μ_T and μ_R are the population means, and σ_T^2 and σ_R^2 are the population variances of the log-transformed measure for T and R, respectively; σ_0^2 is a regulatory constant for variance; and θ_P is the PBE limit. The concept of PBE is to compare the difference of T and R with that of the R versus R itself. This comparison can be denoted in terms of the population difference ratio as follows:

$$\sqrt{\frac{E(Y_T - Y_R)^2}{E(Y_R - Y_R')^2}} = \sqrt{\frac{(\mu_T - \mu_R)^2 + \sigma_R^2 + \sigma_T^2}{2\sigma_R^2}}$$

The regulatory constant variance, σ_0^2, is set based on the following considerations. Due to the low variability of in vitro measurements, this guidance recommends that the ratio of geometric means should fall within 0.90 and 1.11. As a result, an upper BE limit of 1.11 is recommended for the average BE limit for in vitro data. Assuming $\sigma_R^2 = \sigma_T^2 = \sigma_0^2$, $\mu_T - \mu_R = \ln 1.11$ and the maximum allowable limit for population difference ratio is 1.25, this leads to the recommended choice of $\sigma_0^2 = 0.01$.

The determination of PBE limit, θ_P, is based on the consideration of average BE criterion and the addition of variance terms to PBE criterion as the following form:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\max\{\sigma_0^2, \sigma_R^2\}} = \frac{\text{Average BE limit} + \text{Variance term}}{\text{Scaled variance term}}$$

The FDA-recommended allowance for the variance term is 0.01. This value may be adjusted depending on the average BE limit for in vitro data based on further communication with the Agency. Accordingly, the PBE limit, θ_P, is recommended as follows:

$$\theta_P = \frac{(\ln 1.11)^2 + 0.01}{0.01} = 2.089$$

A linearized form is recommended to use to test $H_0: \theta \geq \theta_P$. That is, testing $H_0: \theta \geq \theta_P$ is equivalent to testing $H_0: \gamma \geq 0$ where $\gamma = (\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2) - \theta_P \sigma_R^2$ if the estimated $\sigma_R > \sigma_0$ or $\gamma = (\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2) - \theta_P \sigma_0^2$ if the estimated $\sigma_R \leq \sigma_0$. Here, $\gamma_1 = (\mu_T - \mu_R)^2$, $\gamma_2 = \sigma_T^2$ and $\gamma_3 = \sigma_R^2 + \theta_P \sigma_R^2$ if the estimated $\sigma_R > \sigma_0$ or $\gamma_3 = \sigma_R^2 + \theta_P \sigma_0^2$ if the estimated $\sigma_R \leq \sigma_0$.

Suppose $\hat{\gamma}_U$ is a 95 percent upper confidence bound for γ. Then, PBE is supported if and only if $\hat{\gamma}_U \leq 0$. Based on the work of Howe (1974) and Ting et al. (1990), an approximate 95 percent upper confidence bound for γ is given as follows:

$$\hat{\gamma}_U = \hat{\gamma}_1 + \hat{\gamma}_2 - \hat{\gamma}_3 + \sqrt{(\tilde{\gamma}_1 - \hat{\gamma}_1)^2 + (\tilde{\gamma}_2 - \hat{\gamma}_2)^2 + (\tilde{\gamma}_3 - \hat{\gamma}_3)^2}$$

where $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$ are point estimators of γ_1, γ_2, and γ_3, respectively; $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are 95 percent upper confidence bounds for γ_1 and γ_2 and $\tilde{\gamma}_3$ is a 95 percent lower confidence bound for γ_3. For further detail, see Appendix D, Statistical Analysis Using Population Bioequivalence, and Appendix E, Statistical Analysis Using Modified Population BE.

2. Similarity Index (f_2)

For a comparison of dissolution profiles, similarity is assessed using the similarity index, f_2 (Shah et al., 1998), as described in detail in the guidance for industry *Immediate Release Solid Oral Dosage Forms Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation* (November 1995). In particular, given that all profiles are conducted on a minimum of 12 individual dosage units, 2 profiles are similar if the value of their similarity factor f_2 is between 50 and 100. If applicants believe that f_2 is not suitable for their comparison of dissolution profiles, applicants are encouraged to contact the Agency early to discuss their proposed methods via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway, as appropriate.[21]

3. In-Vitro Release Test

When an IVRT is used to support a demonstration of BE for topical dermatological drug products as part of an in vitro characterization-based BE approach, a two-stage, nonparametric statistical approach is recommended and described in the draft guidance for industry *In Vitro Release Test Studies for Topical Drug Products Submitted in ANDAs* (October 2022).[22] The statistical approach is the same as that used to assess the equivalence of drug release rates for nonsterile semisolid dosage forms evaluated by a comparative IVRT study in the context of certain postapproval changes; this is shown in detail in the guidance for industry *Nonsterile Semisolid Dosage Forms — Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls; In Vitro Release Testing and In Vivo Bioequivalence Documentation* (May 1997).

The assessment of equivalence by an IVRT involves a comparison of the median in vitro drug release rates of two formulations using a nonparametric statistical test which is resistant to outliers that are expected to occur under the testing conditions.

4. In-Vitro Permeation Test

When an IVPT is used to support a demonstration of BE for topical dermatological drug products as part of an in vitro characterization-based BE approach, a mixed scaled criterion is recommended, and described in detail in the draft guidance for industry *In Vitro Permeation Test*

[21] See footnotes 11, 12, and 13.

[22] When final, this guidance will represent FDA's current thinking on this topic.

Studies for Topical Drug Products Submitted in ANDAs (October 2022).[23] According to that methodology, a confidence interval is calculated for each of the endpoints, log-transformed maximum flux (J_{max}) and log-transformed total (cumulative) amount (AMT) permeated. The permeation test is performed with excised skin sections from patients undergoing a surgical procedure or from cadaver donors and the statistical test uses the within-reference standard deviation, s_{WR}, as the threshold that prompts use of either the unscaled or scaled confidence interval.

The mixed-scaled criterion uses the within-reference standard deviation as a threshold, independently, for each endpoint. Specifically, for J_{max} or log-transformed total (cumulative) amount permeated, the reference-scaled average BE approach is used for the endpoint only if it has a $s_{WR} > 0.294$. The regular average BE approach (refer to Schuirmann, 1987) is used for the endpoint with $s_{WR} \leq 0.294$.

In the reference-scaled average BE approach, the hypotheses to be tested are:

$$H_0 : \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} \geq \theta$$

$$H_a : \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} < \theta$$

Here we determine the $100(1-\alpha)$ percent upper confidence bound for $(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$ where:

- $\mu_T - \mu_R$ = mean difference of T and R
- σ_{WR}^2 = within-subject variance of R
- $\theta = \frac{(\ln(m))^2}{(\sigma_{W0})^2}$, $m = 1.25$, and $\sigma_{W0} = 0.25$ (regulatory constant)

For T to be bioequivalent to R, both of the following conditions must be satisfied for each endpoint tested:

a. The 95 percent upper confidence bound for $(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$ must be less than or equal to zero (numbers should be kept to a minimum of four significant figures for comparison).

b. The point estimate for the mean difference of T and R must fall within the prespecified limits $[-\ln(m), \ln(m)]$, where m = 1.25.

In the case of the nonscaled approach, we calculate the 100(1-2α) percent confidence interval for $\mu_T - \mu_R$ as

[23] When final, this guidance will represent FDA's current thinking on this topic.

$$\bar{I}_{\cdot} \pm t_{(1-\alpha),(n-1)} * \sqrt{\frac{S_I^2}{n}}$$

where:

- $\bar{I}_{\cdot}$ is the point estimate for the mean difference of T and R
- S_I^2 is the estimate of interdonor variability
- $t_{(1-\alpha),(n-1)}$ is the 100 $(1-\alpha)$ percentile of the student's t-distribution with $(n-1)$ degrees of freedom
- n is the number of donors
- the value of α is usually set at 0.05

For T to be bioequivalent to R, the 100(1-2α) percent confidence interval for $\mu_T - \mu_R$ must be contained within the limits $\left[ln\left(\frac{1}{m}\right), \ln\ (m)\right]$ for each endpoint tested, where m = 1.25.

5. Abuse-Deterrent Formulation Comparative Studies

An ADF is a formulation that has abuse-deterrent properties, which are defined as drug product properties that are expected to meaningfully deter certain types of abuse, even if they do not fully prevent abuse.[24] The general BE statistical considerations for in vitro ADF comparative studies presented in this guidance align with the guidance for industry *Abuse-Deterrent Opioids — Evaluation and Labeling*[25] and the guidance for industry *General Principles for Evaluating the Abuse Deterrence of Generic Solid Oral Opioid Drug Products* (November 2017). The potential route of abuse (i.e., ingestion (oral), injection (parenteral), insufflation (nasal), or smoking (inhalation)) and its relevance to ADF design feature(s) will determine how an applicant should evaluate the abuse deterrence of the product utilizing a tier-based approach. To support in vitro ADF comparative studies, the Agency recommends applicants provide justification for the sample size, statistical test, and number of batches to assess the abuse-deterrent properties and demonstrate consistency of abuse-deterrent performance throughout the drug product shelf-life and lifecycle (i.e., postapproval changes). Applicants should consider a standardized accept/reject criterion based on the noninferiority margin delta or confidence interval relevant to the abuse-deterrent outcome. The Agency recommends the use of relevant statistics (e.g., sampling plans) to support evaluation of abuse-deterrent properties.

For ANDA submissions, a noninferiority approach should be taken when comparing T with R to conclude that T is no less abuse deterrent than R.[26] FDA recommends inferential analyses to evaluate the abuse deterrence of T versus R. In these analyses, a hierarchical set of null hypotheses serves as a gatekeeper for subsequent null hypotheses, evaluating the abuse deterrence of T and R under progressively more challenging conditions. A hierarchical

[24] See the guidance for industry *Abuse-Deterrent Opioids - Evaluation and Labeling* (April 2015).

[25] Ibid.

[26] Guidance for industry *General Principles for Evaluating the Abuse Deterrence of Generic Solid Oral Opioid Drug Products* (November 2017).

inferential approach is used to maintain a fixed, family-wise experiment Type I error rate. Typically, the acceptable Type I error probability (α) will be set at 5 percent.

6. Earth Mover's Distance Based Profile Comparison Approach

EMD is a statistical metric that measures the discrepancy (distance) between distributions without a prior assumption of the distribution (Rubner et al. 2000). The EMD has been recommended in a profile comparison approach to assess equivalence of particle size distribution profile, where the profile exhibits complex distribution (i.e., multiple peaks) that cannot be accurately described by some conventional descriptors (e.g., the D50 and SPAN).[27] The EMD-based profile comparison approach is briefly described as follows. To assess equivalence between the T and R formulations in the particle size distribution shape, R center, an average profile of all R samples is calculated and serves as the reference profile to compute the distance between an R or a T sample to the R center using the EMD algorithm. After obtaining the profile distances between each R sample and the R average (R – R center distance), and the profile distances between each T sample and the R average (T – R center distance), a statistical equivalence method, e.g., the PBE, is then applied to the two groups of distances to indicate whether T and R are statistically equivalent in the particle size distribution shape. For details, refer to Rubner et al. (2000).

Importantly, considering the emerging technologies and methods for in vitro BE studies, applicants are encouraged to contact the Agency early to discuss their proposed study designs and statistical methods via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway, as appropriate.[28]

B. Statistical Method for Narrow Therapeutic Index Drugs

NTI drugs are those drugs where small differences in dose or blood concentration may lead to serious therapeutic failures and/or adverse drug reactions that are life-threatening or result in persistent or significant disability or incapacity. For NTI drugs, a fully replicate crossover design should be used. The statistical analysis should be carried out using both the average BE and the reference-scaled average BE tests for both AUC and C_{max}.

The reference-scaled average BE is evaluated by testing the null hypothesis:

$$H_0 \ : \ \frac{(\mu_T-\mu_R)^2}{\sigma_{WR}^2} \geq \theta$$

versus the alternative hypothesis:

$$H_a \ : \ \frac{(\mu_T-\mu_R)^2}{\sigma_{WR}^2} < \theta$$

[27] For example, see the draft PSG for industry on Cyclosporine emulsion (October 2016). When final, this guidance will represent FDA's current thinking on this topic.

[28] See footnotes 11, 12, and 13.

where:

- μ_T is the population average response of the log-transformed measure for the Test formulation.
- μ_R is the population average response of the log-transformed measure for the Reference formulation.
- σ^2_{WR} is the population within subject variance of the Reference formulation.
- $\theta = \frac{[\ln(\Delta)]^2}{\sigma^2_{W0}}$ is the BE limit.
- Δ and σ^2_{W0} are predetermined constants. For NTI products, $\Delta = \frac{1}{0.9}$ and $\sigma_{w0} = 0.1$.

Testing is usually done at α = 0.05, and rejection of the null hypothesis supports the conclusion of BE.

NTI BE studies should pass both the reference-scaled average BE criteria and the unscaled average BE limits of 80.00 percent to 125.00 percent.

In addition, the test/reference ratio of the within-subject standard deviation should be evaluated. The within-subject variability comparison of T and R is carried out by a one-sided F test. The null hypothesis for this test is the following:

$$H_0 : \frac{\sigma_{WT}}{\sigma_{WR}} \geq \delta$$

And the alternative hypothesis is:

$$H_a : \frac{\sigma_{WT}}{\sigma_{WR}} < \delta$$

where σ_{WT} is the within-subject standard deviation for the test product, σ_{WR} is the within-subject standard deviation for the reference material, and δ is the limit to declare the within-subject standard deviation of the test product is not greater than or equal to δ times the within-subject standard deviation of the reference material (refer to Appendix F, Statistical Analysis for Reference-scaled Average BE for Narrow Therapeutic Index Drugs, where δ is set to 2.5).

- The $100(1-\alpha)$ percent confidence interval for $\frac{\sigma_{WT}}{\sigma_{WR}}$ is given by
- $$\left(\frac{\frac{s_{WT}}{s_{WR}}}{\sqrt{F_{\frac{\alpha}{2}}(\nu_1,\nu_2)}}, \frac{\frac{s_{WT}}{s_{WR}}}{\sqrt{F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2)}} \right)$$

Here, $\alpha = 0.1$, $F_{\frac{\alpha}{2}}(v_1, v_2)$ and $F_{1-\frac{\alpha}{2}}(v_1, v_2)$ are the values of the F-distribution with v_1 (numerator) and v_2 (denominator) degrees of freedom that has probability of $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ to its right, respectively.

Refer to Appendix F, Statistical Analysis for Reference-scaled Average BE for Narrow Therapeutic Index Drugs, for the steps that can be followed to carry out the statistical analysis for the reference-scaled average BE for NTI drugs.

C. Statistical Method for Highly Variable Drugs

Highly variable drugs are drugs for which within subject variability (%CV) in BE measures 30 percent or greater and that are not considered NTI drugs. In order to use the reference-scaled average BE approach for highly variable drugs, a partial or fully replicate crossover design should be used. The statistical analysis should be carried out using the mixed scaling approach below for both AUC and C_{max}.

If the estimated within-subject standard deviation of the reference is < 0.294, the average BE two one-sided test procedure should be used to determine BE for the individual PK parameter. Otherwise, the reference-scaled average BE procedure should be used to determine BE for the individual PK parameter together with a point estimate constraint for the estimated test/reference geometric mean ratio, which should be bounded by 80.00 percent to 125.00 percent.

For the reference-scaled average BE approach for highly variable drugs, refer to Appendix G, Statistical Analysis for Reference-scaled Average BE for Highly Variable Drugs, where $\Delta = \frac{1}{0.8}$ and $\sigma_{w0} = 0.25$.

Refer to Appendix G, Statistical Analysis for Reference-scaled Average BE for Highly Variable Drugs, for the steps that can be followed to carry out the statistical analysis for the reference-scaled average BE for highly variable drugs.

D. Comparative Clinical Endpoint BE Studies

For some products, the PSG may recommend an appropriately designed comparative clinical endpoint BE study. In particular, a comparative clinical endpoint BE study is an option to be considered for measuring BA or demonstrating BE of dosage forms intended to deliver the active ingredient or active moiety locally, e.g., topical preparations for the skin, eye, and mucous membranes; oral dosage forms not intended to be systemically absorbed, such as an antacid; and bronchodilators administered by oral inhalation.

In general, these studies will have a randomized, parallel group design, with three arms: T, R, and placebo/vehicle.

A placebo/vehicle arm is recommended to demonstrate that T and R are active, and to establish that the study is sufficiently sensitive to detect differences between products at the lower end of the dose/response curve.

To establish BE, it is recommended that the following compound hypotheses (continuous endpoint or dichotomous endpoint) be tested. Rejection of the null hypothesis supports the conclusion of equivalence of the two products.

For a continuous endpoint, if the ratio of means is used to assess BE, the null hypothesis for this test is:

$$H_0 \ : \ \frac{\mu_T}{\mu_R} \leq \ \theta_1 \ or \ \frac{\mu_T}{\mu_R} \geq \ \theta_2$$

versus the alternative hypothesis:

$$H_a \ : \ \theta_1 < \frac{\mu_T}{\mu_R} < \ \theta_2$$

where:

- μ_T = mean of the primary endpoint for the test group, and
- μ_R = mean of the primary endpoint for the reference group.

The null hypothesis, H_0, is rejected with a Type I error (α) of 0.05 (two one-sided tests) if the 90 percent confidence interval for the ratio of the means between T and R $\frac{\mu_T}{\mu_R}$ is contained within the interval $[\theta_1, \theta_2]$.

Or, if the difference in means is used to assess BE for a continuous endpoint, the null and alternative hypotheses are as follows:

$$H_0 \ : \ \mu_T - \mu_R \leq \ \theta_1 \ or \ \mu_T - \mu_R \geq \ \theta_2$$

$$H_a \ : \ \theta_1 < \ \mu_T - \mu_R < \ \theta_2$$

The null hypothesis, H_0, should be rejected at the Type I error rate as recommended in the PSG for a particular product.

For a dichotomous endpoint, the risk difference is usually used to assess BE. The null hypothesis for this test is:

$$H_0 \ : \ \pi_T - \pi_R \leq \ \Delta_1 \ or \ \pi_T - \pi_R \geq \ \Delta_2$$

versus the alternative hypothesis:

$$H_a \ : \ \Delta_1 < \ \pi_T - \pi_R < \ \Delta_2$$

where:

- π_T = the success rate of the primary endpoint for the treatment group, and
- π_R = the success rate of the primary endpoint for the reference group.

The null hypothesis, H_0, is rejected with a Type I error (α) of 0.05 (two one-sided tests) if the estimated 90 percent confidence interval for the difference of the success rates between T and R $(\pi_T - \pi_R)$is contained within the interval $[\Delta_1, \Delta_2]$.

- For continuous and binary endpoints, to demonstrate adequate study sensitivity, T and R should both be statistically superior to placebo (p-value < 0.05) with regard to the primary endpoint.

- Refer to PSGs for comparative clinical endpoint BE study designs, definitions of study populations, regulatory constant (e.g., equivalence interval limit), and analyses specific to a given product.

E. Studies in Multiple Groups

There can be multiple sources of group[29] effects in BE studies. Sometimes, groups reflect factors arising from study design and conduct. For example, a PK BE study can be carried out in two or more clinical centers and the study may be considered a multigroup BE study. The combination of multiple factors arising from study design and conduct may complicate the designation of group. Therefore, applicants should minimize the group effect in a PK BE study as recommended below:

- Dose all groups at the same clinic unless multiple clinics are needed to enroll a sufficient number of subjects.

- Recruit subjects from the same enrollment pool to achieve similar demographics among groups.

- Recruit all subjects, and randomly assign them to an arm of the BE study at study outset.

- Follow the same protocol and SAP criteria and procedures for all groups.

- When feasible (e.g., when healthy volunteers are enrolled), assign an equal sample size to each group (Alosh et al. 2015).

BE should be determined based on the overall treatment effect among the analysis population for BE in the whole study population. In general, the assessment of BE in the whole study population should be done without including the treatment and group interaction(s) term in the

[29] In literature, the term *group* is sometimes referred to as *subgroup*.

model, but applicants may also use other prespecified models, as appropriate (Fleiss 1986; Permutt 2003; Tsiatis et al. 2008). FDA recommends that the applicant assess the interaction between the treatment and group(s), especially if any of the first four study design criteria recommended above are not met and the PK BE data are considered pivotal information for drug approval. If the interaction term of group and treatment is significant (Sun et al. 2023; Alosh et al. 2015; Grizzle 1965), heterogeneity of treatment effect across groups should be carefully examined and interpreted with care. If the observed treatment effect of the products varies greatly among the groups, vigorous attempts should be made to find an explanation for the heterogeneity in terms of other features of trial management or subject characteristics, which may suggest appropriate further analysis and interpretation.

It is important that statistical methods and models for the primary BE analysis are fully prespecified in the protocol or SAP (e.g., in an ANDA study, the applicant should prespecify detailed statistical criteria and models to be used if the interaction term of group and treatment is applicable). In addition, the statistical model should reflect the multigroup nature of the study. For example, if subjects are dosed in two groups in a crossover BE study, the model should reflect the fact that the periods for the first group are different from the periods for the second group, i.e., the period effect should be nested within the group effect.

When there are multiple centers with very few subjects in some centers and applicants want to combine centers in the analysis, any rules for combination should be prespecified in the protocol or SAP and a sensitivity analysis is recommended. More complicated scenarios may be discussed with the appropriate CDER review division before submission.

F. BE Statistics for Adhesion and Irritation Studies

For information about the statistical method used in irritation, sensitization, and adhesion studies for Transdermal and Topical Delivery Systems, refer to the Statistical Consideration section in the draft guidance for industry *Assessing the Irritation and Sensitization Potential of Transdermal and Topical Delivery Systems for ANDAs* (April 2023) and the Considerations for Statistical Analysis section in the draft guidance for industry *Assessing Adhesion With Transdermal and Topical Delivery Systems for ANDAs* (April 2023).[30]

G. Dose Scale for BE Assessment

In this method, the BE assessment is based on relative BA of the test and reference formulations at the site(s) of action. The relative BA, F, is the ratio of the doses of test and reference formulations that produce an equivalent PD response.

Generally, the F is estimated by fitting an E_{max} model that describes the within-study dose-response relationship. Among available statistical methods for E_{max} model fitting, nonlinear mixed effect (NLME) modeling is recommended, because the NLME modeling is capable of

[30] See also the draft guidance for industry *Assessment of Adhesion for Topical and Transdermal Systems Submitted in New Drug Applications* (July 2021). When final, these guidances will represent FDA's current thinking on these topics.

characterizing between-subject variability and residual unexplained variability and is less sensitive to aberrant observation and missing values.

Relative BA of the test product and reference material can be determined by simultaneously fitting the within-study pooled individual dose response data of both the test product and reference material to the following model:

$$y = E_0 + \frac{E_{max} * Dose * F^i}{ED_{50} + Dose * F^i}$$

where y is the response, $Dose$ is the administered dose, E_0 is the baseline response in the absence of the drug, E_{max} is the fitted maximum drug effect, ED_{50} is the dose required to produce 50 percent of the fitted maximum effect, and i is the treatment indicator (0 = Ref, 1 = Test), with the understanding that F^0 equals 1 and that F^1 is the relative potency used to evaluate BE.

This model is based on the assumption that both E_0 and E_{max} are the same for the test and reference. ED_{50} for the reference material is ED_{50} itself, while ED_{50} for the test product is ED_{50}/F^1. When applying NLME modeling, the fixed effects are E_0, E_{max}, ED_{50}, and F, the between-subject random effects should be specified for parameters such as E_0 and E_{max}, and the residual error random effect should be included. Appropriate justification may be submitted to support the final selected model.

To determine BE, the 90 percent confidence interval for F can be estimated by a bootstrap procedure. Each bootstrap estimation includes the calculation of F by fitting the selected model to a sample dose-response data set, which is generated by resampling with replacement. To maintain the correlation of observations within subject, resampling by subject (remaining observations from all T and R treatment arms) is recommended rather than resampling by observations. The Agency has also recommended using Efron's bias corrected and accelerated method to compute a 90 percent confidence interval for F.[31] Alternatively, the 90 percent confidence interval for F can be estimated without a bootstrap procedure, directly from the point estimate of logF and its standard error calculated using NLME modeling.

Given the complexity of dose scale analyses for comparative PD BE studies, applicants are encouraged to contact FDA early to discuss their proposed study designs and statistical methods (e.g., alternative modeling approaches, impact of missing data, and the handling strategy) via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway, as appropriate.[32]

[31] See draft PSG for industry on Orlistat oral capsule (August 2021). When final, this guidance will represent FDA's current thinking.

[32] See footnotes 11, 12, and 13.

H. BE Studies Using Multiple References

In BE studies with more than two reference treatment arms (e.g., a three-period study including two references, one from the European Union (EU) and another from the United States, or a four-period study including test and reference in fed and fasted states), the BE determination should be based on the comparison between the relevant test product and reference material, using only the data from those products. The BE analysis for this comparison should be conducted excluding the data from the non-relevant treatment(s). For example, in a BE study with a test, an EU reference, and a U.S. reference, the comparison of the test to the U.S. reference should be based on an analysis excluding the data from the EU reference. However, full data from the BE studies, including data comparing the test that is the subject of the application with the non-U.S. reference, should be submitted in the application for completeness. The applicant may discuss the study design and statistical approach with the appropriate CDER review division before study conduct.

IV. REFERENCES

Alosh, M, K Fritsch, M Huque, K Mahjoob, G Pennello, M Rothmann, E Russek-Cohen, F Smith, S Wilson, and L Yue, 2015, Statistical Considerations on Subgroup Analysis in Clinical Trials, Stat Biopharm Res, 7(4):286-303.

Bailer, AJ, 1988, Testing for the Equality of Area Under the Curves When Using Destructive Measurement Techniques, J Pharmacokinetics Biopharmaceutics, 16(3):303-309.

Chow, S-C and J-P Liu, 2008, Design and Analysis of Bioavailability and Bioequivalence Studies, 3rd Edition, New York: Chapman and Hall/CRC.

Davit, B and D Conner, 2010, Reference-Scaled Average Bioequivalence Approach. In: I Kanfer and L Shargel, editors. Generic Drug Product Development — International Regulatory Requirements for Bioequivalence, New York, NY: Informa Healthcare, 271-272; Food and Drug Administration, Advisory Committee for Pharmaceutical Science, October 5-6, 2006.

Fieller, E, 1954, Some Problems in Interval Estimation, J Royal Stat Soc, 16(2):175-185.

Fleiss, JL, 1986, Analysis of Data from Multiclinic Trials, Controlled Clin Trials, 7(4):267-275.

Gong, Y, P Zhang, M Yoon, H Zhu, A Kohojkar, AC Hooker, MP Ducharme, J Gobburu, G Cellière, P Gajjar, BV Li, R Velagapudi, YC Tsang, A Schwendeman, J Polli, L Fang, R Lionberger, and L Zhao, 2023, Establishing the Suitability of Model-integrated Evidence to Demonstrate Bioequivalence for Long-acting Injectable and Implantable Drug Products: Summary of Workshop. CPT Pharmacometrics Syst Pharmacol, 12(5):624-630.

Grizzle, JE, 1965, The Two-Period Change-Over Design and Its Use in Clinical Trials, Biometrics, 21(2):467-480.

Howe, WG, 1974, Approximate Confidence Limits of the Mean of X+Y Where X and Y are Two Tabled Independent Random Variables, J Amer Stat Assoc, 69(347):789-794.

Loingeville, F, J Bertrand, TT Nguyen, S Sharan, K Feng, W Sun, J Han, S Grosser, L Zhao, L Fang, K Möllenhoff, H Dette, and F Mentré, 2020, New Model-Based Bioequivalence Statistical Approaches for Pharmacokinetic Studies with Sparse Sampling, AAPS J, 22(6):141.

Maurer, W, B Jones, and Y Chen, 2018, Controlling the Type I Error Rate in Two-stage Sequential Adaptive Designs when Testing for Average Bioequivalence, Stat Med, 37(10):1587-1607.

Patterson, SD and B Jones, 2017, Bioequivalence and Statistics in Clinical Pharmacology, 2nd Edition, New York: Chapman and Hall/CRC.

Permutt, T, 2003, Probability Models and Computational Models for ANOVA in Multicenter Clinical Trials, J Biopharmaceutical Stat, 13(3):495-505.

Potvin, D, CE DiLiberti, WW Hauck, AF Parr, DJ Schuirmann, and RA Smith, 2008, Sequential Design Approaches for Bioequivalence Studies With Crossover Designs, Pharm Stat, 7(4):245-262.

Rubner, Y, C Tomasi, and LJ Guibas, 2000, The Earth Mover's Distance as a Metric for Image Retrieval, Int J Computer Vision, 40(2):99-121.

Schuirmann, DJ, 1987, A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability, J Pharmacokinetics Biopharmaceutics, 15(6): 657-680.

Shah, VP, Y Tsong, P Sathe, and JP Liu, 1998, In Vitro Dissolution Profile Comparison—Statistics and Analysis of the Similarity Factor, f_2, Pharm Res, 15(6):889-896.

Sun, W, D Schuirmann, and S Grosser, 2023, Qualitative versus Quantitative Treatment-by-Subgroup Interaction in Equivalence Studies with Multiple Subgroups, Stat Biopharm Res, 15(4):737-747.

Ting, N, RK Burdick, F Graybill, S Jeyaratnam, and TFC Lu, 1990, Confidence Intervals on Linear Combinations of Variance Components That Are Unrestricted in Sign, J Stat Computation Simul, 35:135-143.

Tsiatis, AA, M Davidian, M Zhang, and X Lu, 2008, Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: A Principled Yet Flexible Approach, Stat Med, 27(23):4658-4677.

Xu, J, C Audet, CE DiLiberti, WW Hauck, TH Montague, AF Parr, D Potvin, and DJ Schuirmann, 2016, Optimal Adaptive Sequential Designs for Crossover Bioequivalence Studies, Pharm Stat, 1(15):15-27.

Zhao, L, M-J Kim, L Zhang, and R Lionberger, 2019, Generating Model Integrated Evidence for Generic Drug Development and Assessment, Clin Pharmacol Therapeutics, 105(2):338-349.

V. APPENDICES

A. Choice of Specific Replicate Crossover Designs

This Appendix (Appendix A) describes why FDA prefers replicate crossover designs with only two sequences, and why the Agency recommends the specific designs described in section II.A.1.b, Replicate crossover designs.

1. Reasons Unrelated to Carryover Effects

Each unique combination of sequence and period in a replicate crossover design can be called a cell of the design. For example, the two-sequence, four-period design recommended in section II.A.1.b, Replicate crossover designs, has eight cells. The four-sequence, four-period design below has 16 cells.

		Period			
		1	2	3	4
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 II.A.1, Experimental Design, 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 five degrees-of-freedom do not account for all seven 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, a sequence-by-period-by-treatment interaction effect might be included, which would fully saturate the fixed-effects model.

If the replicate 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 point estimate of $\mu_T - \mu_R$ will be the same for the main-effects model and for the saturated model.

If the replicate crossover design has more than two sequences, these advantages are no longer present. Main-effects models will generally produce different point 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 (though $\mu_T - \mu_R$ from the saturated model was determined to be appropriate for use in the reference-scaled average bioequivalence (BE) assessment). Thus, use of designs with only two sequences minimizes or avoids certain ambiguities 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 is thought to be optimal among three-period replicate 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.

		Period		
		1	2	3
Sequence	1	**T**	**R**	**R**
	2	**R**	**T**	**T**

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, where the direct effect of a treatment is the effect that that treatment has in the period in which it is administered (i.e., the treatment effect). 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 replicate crossover designs, but not in two-sequence designs.

In the case of the four-period and three-period designs recommended in section II.A.1.b, Replicate crossover designs, 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 replicate crossover designs.

3. Two-Period Replicate Crossover Designs

For most drug products, two-period replicate crossover designs such as the Balaam design (which uses the sequences TR, RT, TT, and RR) should be avoided. However, the modified Balaam design (TR, RT, RR) 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) when reference-scaled average BE is needed.

B. Rationale for Logarithmic Transformation of Pharmacokinetic Data

1. 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 pharmacokinetic (PK) 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 (geometric means) on the original scale. Logarithmic transformation thus achieves a general comparison based on the ratio rather than the differences.

2. Pharmacokinetic Rationale

Westlake observed that a multiplicative model is postulated for PK measures in bioavailability (BA)/BE studies (i.e., area under the curve (AUC) and peak concentration (C_{max}), but not time to maximum observed concentration (T_{max})).[1,2] 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:

$$AUC_{0-\infty} = F*D/CL$$

$$= F*D/(V*Ke)$$

where F is the fraction absorbed, D is the administered dose, and F*D 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 (Ke). The use of AUC as a measure of the amount of drug absorbed

[1] Westlake, WJ, 1973, The Design and Analysis of Comparative Blood-Level Trials, J Swarbick, editor, Current Concepts in the Pharmaceutical Sciences, Dosage Form Design and Bioavailability, Philadelphia:Lea and Febiger, 149-179.

[2] Westlake, WJ, 1988, Bioavailability and Bioequivalence of Pharmaceutical Formulations, In: Biopharmaceutical Statistics for Drug Development, K.E. Peace, editor, Marcel Dekker, Inc., 329-352.

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 (i.e., V*Ke) term into the following equation in an additive fashion:

$$\ln AUC_{0-\infty} = \ln F + \ln D - \ln V - \ln Ke$$

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

$$C_{max} = (F*D/V) * \exp(-Ke*T_{max})$$

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

$$\ln C_{max} = \ln F + \ln D - \ln V - Ke*T_{max}$$

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

C. SAS Program Statements for Average BE Analysis of Replicate Crossover Studies

The following illustrates an example of program statements to run the unscaled average BE analysis using PROC MIXED in SAS version 9, 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 (C_{max})) 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 test formulation precedes the code for the reference 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 or UNR.

In the *Model* statement, DDFM=SATTERTH could possibly be replaced by DDFM=KR2. However, the detailed model specification should be prespecified in the protocol or SAP and data driven post hoc selection of the model is not allowed.

Additions and modifications to these statements can be made if the study is carried out in more than one group of subjects or other complicated scenarios. Alternative software could also be used if same results are generated as in PROC MIXED in SAS.

D. Statistical Analysis Using Population BE

The following steps can be followed to carry out the statistical analysis for using population bioequivalence **(PBE)**:

Step 1. Establish PBE criterion:

$$\frac{(\mu_T-\mu_T)^2+(\sigma_T^2-\sigma_R^2)}{\sigma_R^2} \leq \theta \textbf{ or } \frac{(\mu_t-\mu_R)^2+(\sigma_T^2-\sigma_R^2)}{\sigma_{TO}^2} \leq \theta$$

Linearized criteria:

$\eta_1 = (\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2) - \theta_P \cdot \sigma_R^2 < 0$ **for** $\sigma_R > \sigma_{TO}$

$\eta_2 = (\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2) - \theta_P \cdot \sigma_{TO}^2 < 0$ **for** $\sigma_R \leq \sigma_{TO}$

where:

- $\mu_T - \mu_R$ is the mean difference of T (log scale) and R (log scale)
- σ_T^2, σ_R^2 is Total variance of T and R
- σ_{TO} is 0.1 (regulatory constant)
- θ_p is 2.0891 (regulatory constant) calculated as follows:

$$\frac{[\ln(1.11)]^2 + 0.01}{0.1^2} = 2.089$$

Estimating the linearized criteria:

$\hat{\eta}_1 = \hat{\Delta}^2 + \frac{MSB_T}{m} + \frac{(m-1)MSW_T}{m} - (1+\theta_p)\frac{MSB_R}{m} - (1+\theta_p)\frac{(m-1)MSW_R}{m}$ **for** $\sigma_R > \sigma_{T0}$

$\hat{\eta}_2 = \hat{\Delta}^2 + \frac{MSB_T}{m} + \frac{(m-1)MSW_T}{m} - \frac{MSB_R}{m} - \frac{(m-1)MSW_R}{m} - \theta_p\sigma_{T0}^2$ **for** $\sigma_R \leq \sigma_{T0}$

where:

- $\hat{\Delta} = \bar{X}_{..T.} - \bar{X}_{..R.}$
- m is the number of life stages
- MSW_T is the within-bottle variability for test product
- MSW_R is the within-bottle variability for reference material
- $\frac{MSB_T - MSW_T}{m}$ is the between-bottle variability for test product
- $\frac{MSB_R - MSW_R}{m}$ is the between-bottle variability for reference material

Step 2. Calculate MSB and MSW:

Calculation for MSW_T, MSW_R, MSB_T, and MSB_R can be conducted as follows:

$MSB_k = \frac{m \cdot \sum_{j=1}^{\ell_k} \sum_{i=1}^{n_k} (\bar{X}_{ijk\cdot} - \bar{X}_{\cdot\cdot k\cdot})^2}{n_k \cdot \ell_k - 1}$ k refers to either test product or reference material

$$MSW_k = \frac{\sum_{j=1}^{\ell_k} \sum_{i=1}^{n_k} \sum_{s=1}^{m} (X_{ijks} - \bar{X}_{ijk\cdot})^2}{n_k \cdot \ell_k \cdot (m-1)}$$

$$\bar{X}_{ijk\cdot} = \frac{\sum_{s=1}^{m} X_{ijks}}{m}; \; \bar{X}_{\cdot\cdot k\cdot} = \frac{\sum_{i=1}^{\ell_k} \sum_{j=1}^{n_k} \bar{X}_{ijk\cdot}}{n_k \cdot \ell_k}$$

where:

- n_T, n_R is the number of canisters or bottles per batch, for test product and reference material
- ℓ_T, ℓ_R is the number of batches of test product and reference material
- X_{ijks} is the i[th] bottle in batch # j at life stage s for test product or reference material
- $\bar{X}_{ijk\cdot}$ is the average m life stages for i[th] bottle in batch # j
- $\bar{X}_{\cdot\cdot k\cdot}$ is the population mean for the test products or reference material

Step 3. Calculate σ_R and σ_T

1. σ_R can be conducted as follows:

$$\sigma_R = \sqrt{\frac{MSB_R}{m} + \frac{(m-1)MSW_R}{m}}$$

   a. If $\sigma_R > \sigma_{TO}$ (regulatory constant, 0.1), using the reference-scaled procedure to determine BE for the measured parameter(s).

   b. If $\sigma_R \leq \sigma_{TO}$ (regulatory constant, 0.1), using the constant-scaled procedure to determine BE for the measured parameter(s).

2. σ_T can be conducted as follows:

$$\sigma_T = \sqrt{\frac{MSB_T}{m} + \frac{(m-1)MSW_T}{m}}$$

Step 4. Calculate linearized point estimate and 95% upper confidence bound:

1. Reference-scaled criterion ($\hat{\eta}_1$): Use $\alpha = 0.05$ for a 95% upper confidence bound:

Equation for linearized point estimate:

$Eq = E_D + E1 + E2 + E3s + E4s$

95% Upper Confidence Bound (Hη1):

$H\eta_1 = (E_D + E1 + E2 + E3s + E4s) + (U_D + U1 + U2 + U3s + U4s)^{½}$

The following are the equations to compute each component:

E_q = Point Estimate	**H_q = Confidence Bound**	**$U_p = (H_q - E_q)^2$**
$E_D = \hat{\Delta}^2$	$H_D = \left(\lvert\hat{\Delta}\rvert + t_{1-a,n_T\cdot\ell_T+n_R\cdot\ell_R-2}\left(\frac{MSB_T}{n_T\cdot\ell_T\cdot m}+\frac{MSB_R}{n_R\cdot\ell_R\cdot m}\right)^{\frac{1}{2}}\right)^2$	U_D
$E1 = \frac{MSB_T}{m}$	$H1 = \frac{(\ell_T\cdot n_T - 1)\cdot E1}{\chi^2_{\ell_T\cdot n_T-1,\alpha}}$	$U1$
$E2 = \frac{(m-1)\cdot MSW_T}{m}$	$H2 = \frac{\ell_T\cdot n_T\cdot(m-1)\cdot E2}{\chi^2_{\ell_T\cdot n_T\cdot(m-1),\alpha}}$	$U2$
$E3s = -(1+\theta_P)\frac{MSB_R}{m}$	$H3s = \frac{(\ell_R\cdot n_R - 1)\cdot E3s}{\chi^2_{\ell_R\cdot n_R-1,1-\alpha}}$	$U3s$
$E4s = -(1+\theta_P)\frac{(m-1)MSW_R}{m}$	$H4s = \frac{(\ell_R\cdot n_R)\cdot(m-1)\cdot E4s}{\chi^2_{\ell_R\cdot n_R\cdot(m-1),1-\alpha}}$	$U4s$

where $\chi^2_{\ell_T\cdot n_T-1,\alpha}$ is from the cumulative distribution function of the chi-square distribution with $\ell_T\cdot n_T - 1$ degrees of freedom, i.e.,
$\Pr(\chi^2_{\ell_T\cdot n_T-1} \le \chi^2_{\ell_T\cdot n_T-1,\alpha}) = \alpha$

For data collected on one life stage (m=1), ignore E2 and E4s and their corresponding H and U terms in the calculation. For data collected on more than one stage (m≥2), use the equations listed above.

2. Constant-scaled criterion ($\hat{\eta}_2$): Use $\alpha = 0.05$ for a 95% upper confidence bound:

Equation for Linearized Point Estimate:

$E_q = E_D + E1 + E2 + E3c + E4c - \theta_p\sigma_{T0}^2$

95% Upper Confidence Bound ($H\eta_2$):

$H\eta_2 = (E_D + E1 + E2 + E3c + E4c - \theta_p\sigma_{T0}^2) + (U_D + U1 + U2 + U3c + U4c)^{1/2}$

The following are the equations to compute each component:

E_q = Point Estimate	**H_q = Confidence Bound**	**$U_p = (H_q - E_q)^2$**
$E_D = \hat{\Delta}^2$	$H_D = \left(\lvert\hat{\Delta}\rvert + t_{1-a,n_T\cdot\ell_T+n_R\cdot\ell_R-2}\left(\frac{MSB_T}{n_T\cdot\ell_T\cdot m}+\frac{MSB_R}{n_R\cdot\ell_R\cdot m}\right)^{\frac{1}{2}}\right)^2$	U_D
$E1 = \frac{MSB_T}{m}$	$H1 = \frac{(\ell_T\cdot n_T - 1)\cdot E1}{\chi^2_{\ell_T\cdot n_T-1,\alpha}}$	$U1$
$E2 = \frac{(m-1)\cdot MSW_T}{m}$	$H2 = \frac{\ell_T\cdot n_T\cdot(m-1)\cdot E2}{\chi^2_{\ell_T\cdot n_T\cdot(m-1),\alpha}}$	$U2$
$E3c = -\frac{MSB_R}{m}$	$H3c = \frac{(\ell_R\cdot n_R - 1)\cdot E3c}{\chi^2_{\ell_R\cdot n_R-1,1-\alpha}}$	$U3c$
$E4c = -\frac{(m-1)MSW_R}{m}$	$H4c = \frac{\ell_R\cdot n_R\cdot(m-1)\cdot E4rc}{\chi^2_{\ell_R\cdot n_R\cdot(m-1),1-\alpha}}$	$U4c$

For data collected on one life stage (m =1), ignore E2 and E4c and their corresponding H and U terms in the calculation. For data collected on more than one stage (m≥2), use the equations listed above.

Step 5. For the test product to be bioequivalent to the reference material, the following condition must be satisfied:

The 95% upper confidence bound for linearized criteria Hη must be ≤ 0.

E. Statistical Analysis Using Modified PBE

The following steps can be followed to carry out the statistical analysis for a modified PBE criterion for drugs in small particles or droplets (e.g., fluticasone propionate nasal spray):

Step 1. Establish modified PBE criterion.

Modified PBE criterion:

$$\frac{(\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2)}{\sigma_R^2} = \theta p \text{ ; If } \mu_T \geq \mu_R$$

$$\frac{(\sigma_T^2 - \sigma_R^2)}{\sigma_R^2} = \theta p \text{ ; If } \mu_T < \mu_R$$

where:

- $\mu_T - \mu_R$ is the mean difference of T (log scale) and R (log scale)
- $\sigma_T^2 - \sigma_R^2$ is the total variance of T and R
- σ_{TO} is 0.1 (regulatory constant)
- θp is 2.0891 (regulatory constant)

Step 2. When $\mu_T \geq \mu_R$, use traditional PBE analysis.

When $\mu_T \geq \mu_R$, proceed 95% upper bound calculation, as described in the traditional PBE in Appendix D.

Step 3. When $\mu_T < \mu_R$, follow Step 3A to Step 3E.

Step 3A. Estimate the linearized criteria:

$$\hat{\eta}_1 = \frac{MSB_T}{m} + \frac{(m-1)MSW_T}{m} - (1+\theta_p)\frac{MSB_R}{m} - (1+\theta_p)\frac{(m-1)MSW_R}{m} \quad \textbf{for } \sigma_R > \sigma_{T0}$$

$$\hat{\eta}_2 = \frac{MSB_T}{m} + \frac{(m-1)MSW_T}{m} - \frac{MSB_R}{m} - \frac{(m-1)MSW_R}{m} - \theta_p\sigma_{T0}^2 \quad \textbf{for } \sigma_R \leq \sigma_{T0}$$

where:

- m is the number of life stages

- MSW_T is the within-bottle variability for test product
- MSW_R is the within-bottle variability for reference material
- $\frac{(MSB_T - MSW_T)}{m}$ is the between-bottle variability for test product
- $\frac{MSB_R - MSW_R}{m}$ is the between-bottle variability for reference material

Step 3B. Calculate MSB and MSW

Calculation for MSW_T, MSW_R, MSB_T, and MSB_R can be conducted as follows:

$$MSB_k = \frac{m \cdot \sum_{j=1}^{\ell_k} \sum_{i=1}^{n_k} (\bar{X}_{ijk\cdot} - \bar{X}_{\cdot\cdot k\cdot})^2}{n_k \cdot \ell_k - 1}$$ k refers to either test product or reference material

$$MSW_k = \frac{\sum_{j=1}^{\ell_k} \sum_{i=1}^{n_k} \sum_{s=1}^{m} (X_{ijks} - \bar{X}_{ijk\cdot})^2}{n_k \cdot \ell_k \cdot (m-1)}$$

$$\bar{X}_{ijk\cdot} = \frac{\sum_{s=1}^{m} X_{ijks}}{m};\ \bar{X}_{\cdot\cdot k\cdot} = \frac{\sum_{i=1}^{\ell_k} \sum_{j=1}^{n_k} \bar{X}_{ijk\cdot}}{n_k \cdot \ell_k}$$

where:

- n_T, n_R is the number of canisters or bottles per batch, for test product and reference material
- ℓ_T, ℓ_R is the number of batches of test product and reference material
- X_{ijks} is the i[th] bottle in batch # j at life stage s for test product or reference material
- $\bar{X}_{ijk\cdot}$ is the average m life stages for i[th] bottle in batch # j
- $\bar{X}_{\cdot\cdot k\cdot}$ is the population mean for the test products or reference material

Step 3C. Calculate σ_R:

σ_R can be calculated as follows:

$$\sigma_R = \sqrt{\frac{MSB_R}{m} + \frac{(m-1)MSW_R}{m}}$$

If $\sigma_R > \sigma_{TO}$ (regulatory constant, 0.1), using the reference-scaled procedure to determine BE for the measured parameter(s).

If $\sigma_R \leq \sigma_{TO}$ (regulatory constant, 0.1), using the constant-scaled procedure to determine BE for the measured parameter(s).

Step 3D. Calculate linearized point estimate and 95% upper confidence bound:

1. Reference-scaled Criterion ($\hat{\eta}_1$): Use $\alpha = 0.05$ for a 95% upper confidence bound:

Equation for Linearized Point Estimate:

Eq = E1+ E2+ E3s+ E4s

95% Upper Confidence Bound ($H\eta_1$):

$H\eta_1$ = (E1+ E2+ E3s+ E4s) + (U1+ U2+U3s+U4s)^{½}

The following are the equations to compute each component:

E_q = Point Estimate	**H_q = Confidence Bound**	**$U_p = (H_q - E_q)^2$**
$E1 = \frac{MSB_T}{m}$	$H1 = \frac{(\ell_T \cdot n_T - 1) \cdot E1}{\chi^2_{\ell_T \cdot n_T - 1, \alpha}}$	*U*1
$E2 = \frac{(m-1) \cdot MSB_T}{m}$	$H2 = \frac{\ell_T \cdot n_T \cdot (m-1) \cdot E2}{\chi^2_{\ell_T \cdot n_T \cdot (m-1), \alpha}}$	*U*2
$E3 = -(1+\theta_P)\frac{(MSB_R}{m}$	$H3 = \frac{(\ell_R \cdot n_R - 1) \cdot E3s}{\chi^2_{\ell_R \cdot n_R - 1, 1-\alpha}}$	*U*3*s*
$E4s = -(1+\theta_P)\frac{(m-1)MSW_R}{m}$	$H4s = \frac{(\ell_R \cdot n_R) \cdot (m-1) \cdot E4s}{\chi^2_{\ell_R \cdot n_R \cdot (m-1), 1-\alpha}}$	*U*4*s*

where $\chi^2_{\ell_T \cdot n_T - 1, \alpha}$ is from the cumulative distribution function of the chi-square distribution with $\ell_T \cdot n_T - 1$ degrees of freedom, i.e.,
Pr ($\chi^2_{\ell_T \cdot n_T - 1} \leq \chi^2_{\ell_T \cdot n_T - 1, \alpha}$) = α

For data collected on one life stage (m =1), ignore E2 and E4s and their corresponding H and U terms in the calculation. For data collected on more than one stage (m≥2), use the equations listed above.

2. Constant-scaled criterion ($\hat{\eta}_2$): Use $\alpha = 0.05$ for a 95% upper confidence bound:

Equation for Linearized Point Estimate:

$E_q = E1 + E2 + E3c + E4c - \theta_p\sigma_{T0}^2$

95% Upper Confidence Bound ($H\eta_2$):

$H\eta_2 = (E1 + E2 + E3c + E4c - \theta_p\sigma_{T0}^2) + (U1 + U2 + U3c + U4c)^{1/2}$

The following are the equations to compute each component:

E_q = Point Estimate	**H_q = Confidence Bound**	**$U_p = (H_q - E_q)^2$**
$E1 = \frac{MSB_T}{m}$	$H1 = \frac{(\ell_T \cdot n_T - 1) \cdot E1}{\chi^2_{\ell_T \cdot n_T - 1, \alpha}}$	*U*1
$E2 = \frac{(m-1) \cdot MSW_T}{m}$	$H2 = \frac{\ell_T \cdot n_T \cdot (m-1) \cdot E2}{\chi^2_{\ell_T \cdot n_T \cdot (m-1), \alpha}}$	*U*2
$E3c = -\frac{MSB_R}{m}$	$H3c = \frac{(\ell_R \cdot n_R - 1) \cdot E3c}{\chi^2_{\ell_R \cdot n_R - 1, 1-\alpha}}$	*U*3c
$E4c = -\frac{(m-1)MSW_R}{m}$	$H4c = \frac{\ell_R \cdot n_R \cdot (m-1) \cdot E4rc}{\chi^2_{\ell_R \cdot n_R \cdot (m-1), 1-\alpha}}$	*U*4c

For data collected on one life stage (m =1), ignore E2 and E4c and their corresponding H and U terms in the calculation. For data collected on more than one stage (m≥2), use the equations listed above.

Step 3E. For the test product to be bioequivalent to the reference material, the following conditions must be satisfied. The 95% upper confidence bound for linearized criteria Hη must be ≤ 0

F. Statistical Analysis for Reference-scaled Average BE for Narrow Therapeutic Index Drugs

The following steps can be followed to carry out the statistical analysis for the reference-scaled average BE for narrow therapeutic index drugs:

Step 1. Determine s_{WR}, the estimate of within-subject standard deviation of the reference material (R or reference), for the PK parameters including AUC and C_{max}.

Calculation for s_{WR} can be conducted as follows:

$$s_{WR}^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(D_{ij} - \bar{D}_{i.}\right)^2}{2(n-m)}$$

where:

- i = number of sequences m used in the study

  [$m = 2$ for fully replicate design: TRTR and RTRT]

- j = number of subjects within each sequence
- T = Test product
- R = Reference
- $D_{ij} = R_{ij1} - R_{ij2}$ (where 1 and 2 represent replicate reference treatments)
- $\bar{D}_{i.} = \frac{\sum_{j=1}^{n_i} D_{ij}}{n_i}$
- $n = \sum_{i=1}^{m} n_i$ (i.e., total number of subjects used in the study, while n_i is number of subjects used in sequence i)

Step 2. Use the reference-scaled procedure to determine BE for individual PK parameter(s).

Determine the 95% upper confidence bound[3] for:

$$(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$$

[3] The method of obtaining the upper confidence bound is based on Howe's Approximation I, which is described in Howe, WG, 1974, Approximate Confidence Limits on the Mean of *X*+*Y* Where *X* and *Y* are Two Tabled Independent Random Variables, J Am Stat Assoc, 69(347):789–794.

where:

- μ_T and μ_R are the means of the log-transformed PK endpoint (AUC and/or C_{max}) for T and R, respectively
- σ^2_{WR} is the within-subject variance of the reference material
- $\theta = \left(\frac{\ln(\Delta)}{\sigma_{W0}}\right)^2$ (scaled average BE limit)
- and $\sigma_{W0} = 0.10$ (regulatory constant), $\Delta = 1/0.9$ (approximately=1.11111)

Step 3. Use the unscaled average BE procedure to determine BE for individual PK parameter(s).

Step 4. Calculate the 90% confidence interval of the ratio of the within-subject standard deviation of test product to reference material σ_{WT}/σ_{WR}. The upper limit of the 90% confidence interval for σ_{WT}/σ_{WR} will be evaluated to determine if σ_{WT} and σ_{WR} are comparable.

The $100(1-\alpha)\%$ confidence interval for $\frac{\sigma_{WT}}{\sigma_{WR}}$ is given by

$$\left(\frac{s_{WT}/s_{WR}}{\sqrt{F_{\frac{\alpha}{2}}(v_1, v_2)}}, \frac{s_{WT}/s_{WR}}{\sqrt{F_{1-\frac{\alpha}{2}}(v_1, v_2)}}\right)$$

where:

- s_{WT} is the estimate of σ_{WT} with v_1 as the degree of freedom
- s_{WR} is the estimate of σ_{WR} with v_2 as the degree of freedom
- $F_{\alpha/2}(v_1, v_2)$ is the value of the F-distribution with v_1 (numerator) and v_2 (denominator) degrees of freedom that has probability of $\alpha/2$ to its right
- $F_{1-\alpha/2}(v_1, v_2)$ is the value of the F-distribution with v_1 (numerator) and v_2 (denominator) degrees of freedom that has probability of $1-\alpha/2$ to its right
- here, $\alpha = 0.1$

Step 5. For T to be bioequivalent to R, the following conditions must be satisfied for each PK parameter tested:

a. The 95% upper confidence bound for $(\mu_T - \mu_R)^2 - \theta\sigma^2_{WR}$ must be ≤ 0 (numbers should be kept to a minimum of four significant figures for comparison).

b. Regular unscaled BE limits of 80.00%-125.00% should be passed.

c. The proposed requirement for the upper limit of the 90% equal-tails confidence interval for $\sigma_{WT}/\sigma_{WR}$is less than or equal to 2.500.

Example SAS codes for a fully replicate four-period, two-sequence, four-way crossover design are presented below. It is not necessary to use SAS if other software accomplish the same objectives.

If SAS® is used for statistical analysis, PROC MIXED should be used for fully replicate four-way crossover BE studies.

The following codes are an example of the determination of reference-scaled average BE for LAUCT. Assume that the datasets TEST and REF have already been created, with TEST having all the test observations and REF having all the reference observations.

Dataset containing TEST 1 observations:

```
data test1;
 set test;
 if (seq=1 and per=1) or (seq=2 and per=2);
 lat1t=lauct;
run;
```

Dataset containing TEST 2 observations:

```
data test2;
  set test;
  if (seq=1 and per=3) or (seq=2 and per=4);
  lat2t=lauct;
run;
```

Dataset containing REFERENCE 1 observations:

```
data ref1;
  set ref;
  if (seq=1 and per=2) or (seq=2 and per=1);
  lat1r=lauct;
run;
```

Dataset containing REFERENCE 2 observations:

```
data ref2;
  set ref;
  if (seq=1 and per=4) or (seq=2 and per=3);
  lat2r=lauct;
run;
```

The number of subjects in each sequence is n1 and n2 for sequences 1 and 2, respectively.

Define the following quantities:

$T_{ijk} = k^{th}$ observation (k = 1 or 2) on T for subject j within sequence i

$R_{ijk} = k^{th}$ observation (k = 1 or 2) on R for subject j within sequence i

$$I_{ij} = \frac{T_{ij1} + T_{ij2}}{2} - \frac{R_{ij1} + R_{ij2}}{2}$$

and

$$D_{ij} = R_{ij1} - R_{ij2}$$

I_{ij} is the difference between the mean of a subject's (specifically subject j within sequence i) two observations on T and the mean of the subject's two observations on R, while D_{ij}Dij is the difference between a subject's two observations on R.

Determine I_{ij} and D_{ij}

```
data scavbe;
  merge test1 test2 ref1 ref2;
  by seq subj;
  ilat=0.5*(lat1t+lat2t-lat1r-lat2r);
  dlat=lat1r-lat2r;
run;
```

Intermediate analysis - ilat

```
proc mixed data=scavbe;
  class seq;
  model ilat =seq/ddfm=satterth;
  estimate 'average' intercept 1 seq 0.5 0.5/e cl alpha=0.1;
  ods output CovParms=iout1;
  ods output Estimates=iout2;
  ods output NObs=iout3;
  title1 'scaled average BE';
  title2 'intermediate analysis - ilat, mixed';
run;
```

From the dataset IOUT2, calculate the following:
IOUT2:

```
pointest=exp(estimate);
x=estimate**2–stderr**2;
boundx=(max((abs(lower)),(abs(upper))))**2;
```

Intermediate analysis - dlat

```
proc mixed data=scavbe;
  class seq;
  model dlat=seq/ddfm=satterth;
  estimate 'average' intercept 1 seq 0.5 0.5/e cl alpha=0.1;
  ods output CovParms=dout1;
  ods output Estimates=dout2;
  ods output NObs=dout3;
  title1 'scaled average BE';
  title2 'intermediate analysis - dlat, mixed';
run;
```

From the dataset DOUT1, calculate the following:
DOUT1:

```
s2wr=estimate/2;
```

From the dataset DOUT2, calculate the following:
DOUT2:

```
dfd=df;
```

From the above parameters, calculate the final 95% upper confidence bound:

```
theta=((log(1.11111))/0.1)**2;
y=-theta*s2wr;
boundy=y*dfd/cinv(0.95,dfd);
sWR=sqrt(s2wr);
critbound=(x+y)+sqrt(((boundx-x)**2)+((boundy-y)**2));
```

G. Statistical Analysis for Reference-scaled Average BE for Highly Variable Drugs

The following steps can be followed to carry out the statistical analysis for the reference-scaled average BE assessment for highly variable drugs:

Step 1. Determine s_{WR}, the within-subject standard deviation of the reference material (R or reference), for PK parameters including AUC and C_{max}.

a. If $s_{WR} < 0.294$, use the two one-sided tests procedure to determine BE for the individual PK parameter(s).

b. If $s_{WR} \geq 0.294$, use the reference-scaled procedure to determine BE for the individual PK parameter(s).

Calculation for s_{WR} can be conducted as follows:

$$s_{WR}^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(D_{ij} - \overline{D}_{i.}\right)^2}{2(n-m)}$$

where:

- I = number of sequences m used in the study

  [$m = 3$for partially replicate design: TRR, RTR, and RRT;
  $m = 2$ for fully replicate design: TRTR and RTRT]

- j = number of subjects within each sequence

- T = Test product

- R = Reference material
- $D_{ij} = R_{ij1} - R_{ij2}$ (where 1 and 2 represent replicate reference treatments)
- $\overline{D}_{i.} = \frac{\sum_{j=1}^{n_i} D_{ij}}{n_i}$
- $n = \sum_{i=1}^{m} n_i$((i.e., total number of subjects used in the study, while $n_i$is number of subjects used in sequence *i*)

Continue with steps 2 and 3 for PK parameters that have a $s_{WR} \geq 0.294$.

Step 2. Determine the 95% upper confidence bound[4] for:

$$(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$$

where:

- μ_T and μ_R are the means of the log-transformed PK endpoint (AUC and/or C_{max}) for T and R, respectively.
- σ_{WR}^2 is the within-subject variance of the reference material
- $\theta = \left(\frac{\ln(1.25)}{\sigma_{W0}}\right)^2$ (scaled average BE limit).
- $\sigma_{W0} = 0.25$ (regulatory constant).

Step 3. For T to be bioequivalent to R, *both* of the following conditions must be satisfied for each PK parameter tested:

a. The 95% upper confidence bound for $(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$ must be ≤ 0 (numbers should be kept to a minimum of four significant figures for comparison).

b. The point estimate of the T/R geometric mean ratio must fall within [0.8000, 1.2500].

Example SAS® codes are presented below. It is not necessary to use SAS® if other software accomplish the same objectives.

If SAS® is used for statistical analysis, note the following:

[4] The method for obtaining the upper confidence bound is based on *Howe's Approximation I*, which is described in Howe, WG, 1974, Approximate Confidence Limits on the Mean of *X*+*Y* Where *X* and *Y* are Two Tabled Independent Random Variables, J Am Stat Assoc, 69(347):789–794.

- PROC GLM should be used for partially replicate (three-way) BE studies
- PROC MIXED should be used for fully replicate (four-way) BE studies

1. Example SAS Codes: Partially Replicate Three-Way Design

For PK parameters with a $s_{WR} \geq 0.294$, use the reference-scaled procedure to determine BE.

The following codes are an example of the determination of reference-scaled average BE for LAUCT with a partially replicate three-way BE design:

Dataset containing TEST observations:

```
data test;
   set pk;
   if trt='T';
   latt=lauct;
run;
```

Dataset containing REFERENCE 1 observations:

```
data ref1;
   set ref;
   if (seq=1 and per=2) or (seq=2 and per=1) or (seq=3 and per=1); lat1r=lauct;
run;
```

Dataset containing REFERENCE 2 observations:

```
data ref2;
   set ref;
   if (seq=1 and per=3) or (seq=2 and per=3) or (seq=3 and per=2); lat2r=lauct;
run;
```

Define the following quantities:

T_{ij} = the observation on T for subject j within sequence i

R_{ijk} = k^{th} observation (k = 1 or 2) on R for subject j within

$$I_{ij} = T_{ij} - \frac{R_{ij1} + R_{ij2}}{2}$$

$$D_{ij} = R_{ij1} - R_{ij2}$$

I_{ij} is the difference between a subject's (specifically, subject j within sequence i) observation on T and the mean of the subject's two observations on R, while D_{ij} is the difference between a subject's two observations on R.

Determine I_{ij} and D_{ij}

```
data scavbe;
   merge test ref1 ref2;
```

```
    by seq subj;
    ilat=latt - 0.5*(lat1r+lat2r));
    dlat=lat1r-lat2r;
run;
```

Intermediate analysis - ilat

```
proc glm data=scavbe;
    class seq;
    model ilat=seq/clparm alpha=0.1;
    estimate 'average' intercept 1 seq 0.3333333333 0.3333333333 0.3333333333;
    ods output overallanova=iglm1;
    ods output Estimates=iglm2;
    ods output NObs=iglm3;
    title1 'scaled average BE';
run;
```

From the dataset IGLM2, calculate the following:

IGLM2:

```
    pointest=exp(estimate);
    x=estimate**2–stderr**2;
    boundx=(max((abs(LowerCL)),(abs(UpperCL))))**2;
```

Intermediate analysis - dlat

```
proc glm data=scavbe;
    class seq;
    model dlat=seq;
    ods output overallanova=dglm1;
    ods output NObs=dglm3;
    title1 'scaled average BE';
run;
```

From the dataset DGLM1, calculate the following:

DGLM1:

```
  dfd=df;
s2wr=ms/2;
```

From the above parameters, calculate the final 95% upper confidence bound:

```
theta=((log(1.25))/0.25)**2;
y=-theta*s2wr;
boundy=y*dfd/cinv(0.95,dfd);
sWR=sqrt(s2wr);
critbound=(x+y)+sqrt(((boundx-x)**2)+((boundy-y)**2));
```

2. Example SAS Codes: Fully Replicate Four-Period, Two-Sequence, Four-Way Crossover Design

For PK parameters with a $s_{WR} \geq 0.294$, use the reference-scaled procedure to determine BE.

The following codes are an example of the determination of reference-scaled average BE for LAUCT with a fully replicate four-way BE design:

- **Dataset containing TEST 1 observations:**

```
data test1;
  set test;
  if (seq=1 and per=1) or (seq=2 and per=2);
  lat1t=lauct;
run;
```

- **Dataset containing TEST 2 observations:**

```
data test2;
    set test;
    if (seq=1 and per=3) or (seq=2 and per=4);
  lat2t=lauct;
run;
```

- **Dataset containing REFERENCE 1 observations:**

```
data ref1;
  set ref;
  if (seq=1 and per=2) or (seq=2 and per=1);
  lat1r=lauct;
run;
```

- **Dataset containing REFERENCE 2 observations:**

```
data ref2;
  set ref;
  if (seq=1 and per=4) or (seq=2 and per=3);
  lat2r=lauct;
run;
```

The number of subjects in each sequence is n_1 and n_2 for sequences 1 and 2, respectively.

Define the following quantities:

$T_{ijk} = k^{th}$ observation (k = 1 or 2) on T for subject j within sequence i

$R_{ijk} = k^{th}$ observation (k = 1 or 2) on R for subject j within sequence i

$$I_{ij} = \frac{T_{ij1} + T_{ij2}}{2} - \frac{R_{ij1} + R_{ij2}}{2}$$

$$D_{ij} = R_{ij1} - R_{ij2}$$

I_{ij} is the difference between the mean of two observations of a subject (specifically, subject j within sequence i) on T and the mean of the subject's two observations on R, while D_{ij} is the difference between a subject's two observations on R.

Determine I_{ij} and D_{ij}

```
data scavbe;
  merge test1 test2 ref1 ref2;
  by seq subj;
  ilat=0.5*(lat1t+lat2t-lat1r-lat2r);
  dlat=lat1r-lat2r;
run;
```

Intermediate analysis – ilat

```
proc mixed data=scavbe;
  class seq;
  model ilat =seq/ddfm=satterth;
  estimate 'average' intercept 1 seq 0.5 0.5/e cl alpha=0.1;
  ods output CovParms=iout1;
  ods output Estimates=iout2;
  ods output NObs=iout3;
  title1 'scaled average BE';
  title2 'intermediate analysis - ilat, mixed';
run;
```

From the dataset IOUT2, calculate the following:

IOUT2:

```
  pointest=exp(estimate);
  x=estimate**2–stderr**2;
  boundx=(max((abs(lower)),(abs(upper))))**2;
```

Intermediate analysis – dlat

```
proc mixed data=scavbe;
  class seq;
  model dlat=seq/ddfm=satterth;
  estimate 'average' intercept 1 seq 0.5 0.5/e cl alpha=0.1;
  ods output CovParms=dout1;
  ods output Estimates=dout2;
  ods output NObs=dout3;
  title1 'scaled average BE';
  title2 'intermediate analysis - dlat, mixed';
run;
```

From the dataset DOUT1, calculate the following:

DOUT1:

s2wr=estimate/**2;**

From the dataset DOUT2, calculate the following:

DOUT2:

dfd=df;

From the above parameters, calculate the final 95% upper confidence bound:

```
theta=((log(1.25))/0.25)**2;
y=-theta*s2wr;
boundy=y*dfd/cinv(0.95,dfd);
sWR=sqrt(s2wr);
critbound=(x+y)+sqrt(((boundx-x)**2)+((boundy-y)**2));
```