Statistical Approaches to Establishing Bioequivalence

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) David Coppersmith at 301-796-9193.

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

December 2022
Biopharmaceutics

Revision 1

Statistical Approaches to Establishing Bioequivalence

Guidance for Industry

Additional copies are available from:
Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
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)

December 2022
Biopharmaceutics

Revision 1

TABLE OF CONTENTS

Statistical Approaches to Establishing Bioequivalence Guidance for Industry[1]

This draft guidance, when finalized, will represent the current thinking of the Food and Drug Administration (FDA or Agency) on this topic.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 staff responsible for this guidance as listed on the title page.

I. INTRODUCTION

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

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 sponsors and applicants who intend to use equivalence criteria in analyzing in vivo or in vitro BE studies for INDs, NDAs, ANDAs, and supplements to these applications. This guidance discusses statistical approaches for BE comparisons and focuses on how to use these approaches both generally and in specific situations. When finalized, this guidance will replace the guidance for industry *Statistical Approaches to Establishing Bioequivalence*, which was issued in February 2001 (2001 guidance). This guidance provides recommendations on the topics covered in the 2001 guidance as well as recommendations on additional topics, including missing data and intercurrent events, adaptive design, and specific situations, such as narrow therapeutic index drugs and highly variable drugs.

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

Defined as *relative BA*, the assessment of BE involves comparison between a test (T) and reference (R) drug product, where T and R can vary depending on the comparison to be performed (e.g., to-be-marketed formulation versus clinical trial formulation, generic drug versus reference listed drug (RLD), originally approved formulation versus postapproval formulation changes). 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.[2] 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 equivalence (i.e., the identical dosage form and route(s) of administration that contain identical amounts of the identical active drug ingredient) 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 (e.g., model-based BE and novel adaptive designs for comparative clinical endpoint BE studies) with the Agency through appropriate regulatory meetings, as described below.

B. Statistical Guidance Background

In the July 1992 guidance on *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 the T and R products to healthy subjects on separate occasions, with random assignment to the two possible sequences of drug product administration. The 1992 guidance further recommended that statistical analysis for pharmacokinetic (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 the T and R products were comparable. This approach is termed *average BE* (ABE) and involves the calculation of a 90% confidence interval for the ratio of the averages (population geometric means) of the measures for the T and R products. To establish BE, the calculated confidence interval should fall within a BE limit, usually 80 to 125% for the ratio of the product averages.[3] 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.

[2] 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.

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

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. The 2001 guidance also includes discussion of *replicated crossover designs* — crossover designs in which at least some of the subjects receive at least one of the products 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 Overview section above, when finalized, this guidance will replace the 2001 guidance.

II. GENERAL CONSIDERATIONS

A. Study Design

1. Experimental Design

a. Nonreplicated designs

A conventional nonreplicated design, such as the standard two-formulation, two-period, two-sequence 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 studies are impractical, parallel designs can be used.

b. Replicated crossover designs

Replicated 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 BE approach is used, replicated crossover designs are critical to allow estimation of within-subject variances for the R (and T if a fully replicated study is used) measures. In particular, the following four-period, two-sequence, two-formulation design is recommended for fully replicated BE studies (see Appendix A for further discussion of replicated crossover designs).

		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 replicated administration. Each period should be separated by an adequate washout period.

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

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

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.

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.[4] For example, Potvin's methods (Potvin et al. 2008, Xu et al. 2016)[5] are a combination of a group sequential design and an adaptive design with sample size re-estimation.

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

[5] Potvin, D., C.E. DiLiberti, W.W. Hauck, A.F. Parr, D.J. Schuirmann, and R.A. Smith, 2008, Sequential Design Approaches for Bioequivalence Studies With Crossover Designs, Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 7, no. 4: 245-262; Xu, J., C. Audet, C.E. DiLiberti, W.W. Hauck, T.H. Montague, A.F. Parr, D. Potvin, and D.J. Schuirmann, 2016, Optimal Adaptive Sequential Designs for Crossover Bioequivalence Studies, Pharmaceutical Statistics (15) 1:15-27.

Adaptive design can provide ethical advantages[6] 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.[7] 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 governing the adaptive decision. If a study should be stopped early (e.g., for futility or for success in demonstrating BE), detailed stopping criteria should be pre-specified 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 meeting 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 adaption should be included in the protocol or statistical analysis plan (SAP). This applies to both adaptive 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).

[6] See footnote 4. 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.

[7] See footnotes 4 and 5.

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 adaptive study designs and statistical methods via the controlled correspondence,[8] pre-ANDA meeting,[9] pre-IND meeting, or pre-NDA meeting pathway.[10]

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 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 test and reference treatments (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 T_{max} (time to maximum observed concentration) can be determined accordingly. The ratios of AUC_{0-t} and C_{max} between the test and the reference product 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)[11] for a parallel study design), and that for the ratio of C_{max} may be done by bootstrap methods. BE is supported if the 90% confidence interval for the ratio of AUC_t between the test and the reference product lies within the BE margin (80.00%, 125.00%). Model-based approaches can be considered when they can reliably control the error rate of concluding BE for bio inequivalent products (Type I error).[12]

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.[13]

[8] See the guidance for industry *Controlled Correspondence Related to Generic Drug Development* (December 2020).

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

[10] See the draft guidance for industry *Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products* (December 2017). When final, this guidance will represent FDA's current thinking on this topic.

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

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

[13] See footnotes 8, 9, and 10.

2. Sample Size Determination

It is an applicant's responsibility to design an adequately powered BE study for the proposed study. We recommend 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 type 1 error rate of 0.05 (see section III.C.1.a for more details). When determining the sample size, rates of attrition and noncompliance (e.g., protocol violation) should be taken into consideration. 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. All eligible subjects who were dosed should be included in the analysis. For BE studies, add-on subjects after the pre-specified number of subjects have been reached are generally not encouraged except in an adaptive study design with a pre-specified 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.[14,15,16] For example, the standard 2×2 cross-over study will use a particular calculation while studies with a different design or set of endpoints will use different calculations. 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 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 given in the protocol, together with the estimates of any quantities used in the calculations (such as variances, mean values, response rates, the assumed effect size). The basis for these estimates should also be given. 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.

Applicants should enter a sufficient number of subjects in the study to allow for dropouts. Dropouts generally should not be replaced because replacement of subjects during the study could complicate the statistical model and analysis. Applicants who wish to replace dropouts during the study should indicate this intention in the protocol. The protocol should also state whether samples from replacement subjects, if not used, will be assayed. If the dropout rate is high and applicants wish to add more subjects, a modification of the statistical analysis may be

[14] Chow, S.-C. and J.-P. Liu, 2008, Design and Analysis of Bioavailability and Bioequivalence Studies, 3rd Edition, New York: Chapman and Hall/CRC.
[15] Draft guidance for industry *Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (August 2021). When final, this guidance will represent FDA's current thinking on this topic.
[16] Patterson, S.D. and B. Jones, 2017, Bioequivalence and Statistics in Clinical Pharmacology, 2nd Edition, New York: Chapman and Hall/CRC.

recommended. Additional subjects should not be included after data analysis unless the study was designed from the beginning as an adaptive design.

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 un-transformed comparative clinical endpoints unless otherwise noted in the relevant FDA product-specific guidance (PSG).[17] 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.[18]

B. Data Preparation

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

1. Log-Transformation

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

a. Logarithmic transformation for PK measures

This guidance 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. The limited sample size in a typical BE study precludes a reliable determination of the distribution of the data set. Sponsors and/or applicants are not encouraged to test for normality of error distribution after log-transformation, nor should they use normality of error distribution as a reason for carrying out the statistical analysis on the original scale. Justification should be provided if sponsors or applicants believe that their BE study data should be statistically analyzed on the original rather than on the log scale.

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

[18] 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.

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 the biomedical literature and/or pilot studies. Similar considerations apply to other derived variables, such as the use of change from baseline, percentage change from baseline, the area under the curve 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.[19]

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.[20] 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.

2. Missing Data and Intercurrent Events

Subjects may have missing data in the study for various reasons (e.g., subject's refusal to continue in the study, worsening of conditions or emergence of adverse events, subject's failure to meet scheduled appointments for evaluation). Subjects may also have intercurrent (post-randomization) events that affect either the interpretation or the existence of the measurements associated with the question of interest (e.g., noncompliance with the protocol for various reasons, use of rescue medication due to lack of efficacy, death). Missing data and intercurrent events can introduce problems such as bias, misleading inference, loss of precision and loss of power, which make it hard to interpret the trial outcome.

The ICH (Internal Council for Harmonization) E9(R1) Addendum 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.[21] The trial protocol of a BE study should include the following components of an estimand: (1) the treatment of interest and alternative treatment(s) to which comparison will be made: e.g., test drug compared with reference drug; (2) the analysis population for BE assessment; (3) the variable (or endpoint) to be measured for each subject (e.g., AUC or C_{max}); (4) the specification of how to account for intercurrent events in assessing the scientific question of interest (for example, in a comparative clinical endpoint BE study with

[19] Fieller, E., Some Problems in Interval Estimation, 1954, Journal of the Royal Statistical Society, 16(2): 175-185.

[20] 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.

[21] Guidance for industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials*, Revision 1 (May 2021).

a binary endpoint, subjects who discontinue study treatment early due to lack of treatment effect should be included as treatment failures); and (5) the population-level summary for the variable to compare between treatment conditions, e.g., the geometric mean ratio of the test to reference drug in a PK BE study.

The protocol should include plans to minimize missing data. The trial protocol 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 pre-specified 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 using alternative approaches may also be used in the statistical analysis to address missing data. Sensitivity analyses should be pre-specified 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 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.

3. *Outlier Detection*

Outlier data in BE studies are defined as subject data for one or more BA measures that are discordant with corresponding data for that subject and/or for the rest of the subjects in a study. Because BE studies are usually carried out as crossover studies, the most important type of subject outlier is the within-subject outlier, 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:

a. Product failure

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

b. Subject-by-formulation interaction

A subject-by-formulation interaction could occur when an individual is representative of subjects present in the general population in low numbers, for whom the relative BA of the two products is markedly different 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 the T or R product. However, in the case of a subpopulation, even if the unusual response is observed on the R product, there could still be concern about lack of bioequivalence 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 in the BE study protocol for handling subjects (experimental outliers) in the BE statistical analysis. Data from redosing studies are not considered valid evidence to support removal of outlier data from the statistical analysis. All subject data should be submitted, with potential outliers flagged with appropriate documentation as part of the submission. However, for a replicated PK BE study, if reference-scaled average BE is used, the applicant should ensure that the calculated intra-subject variability is not inflated due to extreme values or situations.

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 Bioequivalence

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%. 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 do two one-sided tests

In BE assessment we are frequently interested in testing whether a parameter (for example, the difference of means for a T and R product for a specific endpoint) is contained within a defined interval, call it $[\theta_1, \theta_2]$. The recommended method for doing such a test is the *Two One-Sided Tests Procedure*.[22] 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 (that is, 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) % 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) % 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) % lower confidence bound for the parameter and L_2, by itself, is at least a 100 (1-α) % 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 a T and R product 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 ($AUC_{0\text{-}t}$, $AUC_{0\text{-}\infty}$, and C_{max})

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

- Geometric means for the formulations tested
- Arithmetic means for the formulations tested
- Geometric mean ratios of Test vs. Reference and their corresponding 90% confidence intervals or 95% upper confidence bounds (e.g., for highly variable drugs or narrow therapeutic index drugs)

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

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 T product to the individual geometric mean of the R product 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:

$$\sigma_D^2 = \text{variance of } (\mu_{Tj} - \mu_{Rj})$$

$$= (\sigma_{BT} - \sigma_{BR})^2 + 2\,(1-\rho)\sigma_{BT}\sigma_{BR}$$ [23]

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 μ_{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 1 error.

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

III. SPECIFIC SITUATIONS[24]

A. In Vitro Bioequivalence and Population Bioequivalence

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. Population Bioequivalence

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

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

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, σ_T^2 and σ_R^2 are the population variances of the log-transformed measure for T and R products, 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 the T and R products with that of the reference versus reference itself. This comparison can be denoted in terms of the population difference ratio as follows:

$$\sqrt{\frac{E(Y_T-Y_R)^2}{E(Y_R-Y_R')^2}} = \sqrt{\frac{(\mu_T-\mu_R)^2+\sigma_R^2+\sigma_T^2}{2\sigma_R^2}} = \sqrt{\frac{\theta}{2}+1}.$$

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:

[24] 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.

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

A linearized form is recommended to use to test $H_0: \theta \geq \theta_P$. That is, testing $\mathrm{H}_0: \theta \geq \theta_P$ is equivalent to testing $\mathrm{H}_0: \gamma \geq 0$ where $\gamma = (\mu_T - \mu_R)^2 + (\sigma_\mathrm{T}^2 - \sigma_\mathrm{R}^2) - \theta_P \sigma_R^2$ if the estimated $\sigma_R > \sigma_0$ or $\gamma = (\mu_T - \mu_R)^2 + (\sigma_\mathrm{T}^2 - \sigma_\mathrm{R}^2) - \theta_P \sigma_0^2$ if the estimated $\sigma_R \leq \sigma_0$. Here, $\gamma_1 = (\mu_T - \mu_R)^2$, $\gamma_2 = \sigma_T^2$ and $\gamma_3 = \sigma_R^2 + \theta_P \sigma_\mathrm{R}^2$ 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% 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)[25] and Ting et al. (1990)[26], an approximate 95% 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% upper confidence bounds for γ_1 and γ_2 and $\tilde{\gamma}_3$ is a 95% lower confidence bound for γ_3. For further detail, see, e.g., the draft PSGs for Budesonide suspension (September 2012) and Fluticasone Propionate metered spray (June 2020).[27]

2. *Similarity Index (f_2)*

For a comparison of dissolution profiles, similarity is assessed using the similarity index, f_2 (Shah et al., 1998),[28] 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.

3. *In-Vitro Release Test*

When an in-vitro release test (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).[29] The statistical approach is the same as that used to assess the equivalence of drug release rates for non-sterile 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

[25] Howe, W.G., 1974, Approximate Confidence Limits of the Mean of X+Y Where X and Y are Two Tabled Independent Random Variables, Journal of the American Statistical Association, 69:789-794.

[26] Ting, N., R.K. Burdick, F. Graybill, S. Jeyaratnam, and T.F.C. Lu, 1990, Confidence Intervals on Linear Combinations of Variance Components That Are Unrestricted in Sign, Journal of Statistical Computation and Simulation, 35:135-143.

[27] When final, these guidances will represent FDA's current thinking on these topics.

[28] Shah, V.P., Y. Tsong, P. Sathe, and J.P. Liu, 1998, In Vitro Dissolution Profile Comparison—Statistics and Analysis of the Similarity Factor, f2, Pharmaceutical Research, 15(6):889-896.

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

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 non-parametric statistical test which is resistant to outliers that are expected to occur under the particular testing conditions.

4. *In-Vitro Permeation Test*

When an in-vitro permeation test (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 Studies for Topical Drug Products Submitted in ANDAs* (October 2022).[30] 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 ABE approach (refer to Schuirmann, 1987)[31] 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-α)% upper confidence bound for $(\mu_T - \mu_R)^2 - \theta\sigma_{WR}^2$ where:

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

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

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

[31] See footnote 22.

a. The 95% 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 of the T/R geometric mean ratio must fall within the pre-specified limits $\left[\frac{1}{m},\ m\right]$, where *m* = 1.25.

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

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

where:

- $\bar{I}_{.}$ is the point estimate for the mean difference of T and R products
- S_I^2 stimate of inter-donor 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 the T product to be bioequivalent to the R product, the 100(1-2α)% confidence interval for $\mu_T - \mu_R$ must be contained within the limits $\left[\frac{1}{m},\ m\right]$ in the original scale 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.[32] 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*[33] 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 route), injection (parenteral route), insufflation (nasal route), or smoking (inhalation route)) 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

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

[33] Ibid.

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 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 non-inferiority approach should be taken when comparing T product with R product to conclude that T product is no less abuse deterrent than R product.[34] The Agency recommends inferential analyses to evaluate the abuse deterrence of T product versus R product. In the analyses, a hierarchical set of null hypotheses serves as a gatekeeper for subsequent null hypotheses, evaluating the abuse deterrence of T and R products under progressively more challenging conditions. A hierarchical 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%.

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.[35] The EMD has been recommended in a profile comparison approach to assess equivalence of particle size distribution profile,[36] 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). The EMD-based profile comparison approach is briefly described as follows. To assess equivalence between the T and R product formulations in the particle size distribution shape, an average profile of all R product samples (i.e., R center) is calculated and serves as the reference profile to compute the distance between an R or a T product sample to the R center using the EMD algorithm. After obtaining the profile distances between each R product sample and the R product average (R – R center distance), and the profile distances between each T product sample and the R product 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 the T and R products are statistically equivalent in the particle size distribution shape. For details, refer to Rubner et al. (2000).[37]

Importantly, considering the increasingly 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.[38]

[34] Guidance for Industry *Evaluating the Abuse Deterrence of Generic Solid Oral Opioid Drug Products* (November 2017).

[35] Rubner, Y., C. Tomasi, and L.J. Guibas, 2000, The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision, 40(2):99-121.

[36] Draft PSG for industry on Cyclosporine emulsion (October 2016). When final, this guidance will represent the FDA's current thinking on this topic.

[37] See footnote 35.

[38] See footnotes 8, 9, and 10.

B. Statistical Methods for Narrow Therapeutic Index and Highly Variable Drug Products

1. Statistical Method for Narrow Therapeutic Index Drugs

If a drug is a narrow therapeutic index drug, a fully replicated cross-over design should be used. The statistical analysis should be carried out using both the ABE 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:

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

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.
- $\sigma_{WR}^2$is the population within subject variance of the Reference formulation.
- $\theta = \frac{[\ln(\Delta)]^2}{\sigma_{W0}^2}$ is the BE limit.
- Δ and σ_{W0}^2 are predetermined constants. Refer to the draft guidance for industry *Bioequivalence Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (August 2021) for the values of Δ and σ_{W0}^2.[39]

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

Narrow therapeutic index BE studies should pass both the reference-scaled approach and the unscaled average BE limits of 80.00 to 125.00%.

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

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

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

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 product and δ is the limit to declare the within-subject variability of the test product is not greater than that of the reference product (refer to the draft guidance for industry *Bioequivalence Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (August 2021) where δ was set to 2.5).[40]

- The 100(1-α)% CI for σ_{WT}/σ_{WR} is given by
- $$\left(\frac{s_{wt}/s_{wR}}{\sqrt{F_{\frac{\alpha}{2}}(v_1,v_2)}}, \frac{s_{wt}/s_{wR}}{\sqrt{F_{1-\frac{\alpha}{2}}(v_1,v_2)}} \right)$$

  Here, α=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 α/2 and 1-α/2 to its right, respectively.

2. *Statistical Method for Highly Variable Drugs*

If a drug is a high variable drug, a partial or fully replicated cross-over design should be used. The statistical analysis should be carried out using the mixed scaling approach below for both AUC and C_{max}.

The mixed scaling approach:

If the estimated within-subject standard deviation of the RLD is < 0.294, the two one-sided test procedure should be used to determine BE for the individual PK parameter. Otherwise, the reference-scaled 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.

For the reference-scaled approach the upper BE limit for Test/Reference ratio of geometric means is $\Delta = \frac{1}{0.8}$, the regulatory constant is $\sigma_{w0} = 0.25$ and the point estimate constraint is 80.00 to 125.00%.

Refer to the draft guidance for industry *Bioequivalence Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA* (August 2021) for further details.[41]

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

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

C. Comparative Clinical Endpoint Bioequivalence 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 moiety locally, e.g., topical preparations for the skin, eye, and mucous membranes; oral dosage forms not intended to be systemically absorbed, e.g., an antacid; bronchodilators administered by oral inhalation.

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

- A placebo/vehicle arm is recommended to demonstrate that the T product and R product 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:
The null hypothesis for this test is:

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

versus the alternative hypothesis:
H_a: $\theta_1 < \mu_T / \mu_R < \theta_2$

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% confidence interval for the ratio of the means between T and R products (μ_T / μ_R) is contained within the interval [θ_1, θ_2].

For a dichotomous endpoint:
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% confidence interval for the difference of the success rates between T and R products ($\pi_T – \pi_R$) is contained within the interval [Δ_1, Δ_2].

- For continuous and binary endpoints, in order to demonstrate adequate study sensitivity, the test product and reference product should both be statistically superior to placebo ($p<0.05$) 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.

D. Studies in Multiple Groups

There can be multiple sources of group[42] 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 multi-group BE study. The combination of multiple factors may complicate the designation of group. Therefore, sponsors should minimize the group effect in a PK BE study as recommended below:

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

(2) Recruit subjects from the same enrollment pool to achieve similar demographics among groups.

(3) Recruit all subjects, and randomly assign them to group and treatment arm, at study outset.

(4) Follow the same protocol criteria and procedures for all groups.

(5) When feasible (e.g., when healthy volunteers are enrolled), assign an equal sample size to each group.

Bioequivalence should be determined based on the overall treatment effect 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 model, but applicants may also use other pre-specified models, as appropriate (Fleiss 1986, Permutt 2003, Tsiatis et al.

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

2008).[43] The assessment of interaction between the treatment and group(s) is important, 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 (Alosh et al. 2015, Grizzle 1965),[44] 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 pre-specified in the protocol or SAP (e.g., in an ANDA study, the applicant should pre-specify 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 sponsors want to combine centers in the analysis, any rules for combination should be pre-specified 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.

E. Bioequivalence Statistics for Adhesion and Irritation Studies

In terms of the statistical method used in irritation, sensitization or/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* (October 2018) and the Considerations for Statistical Analysis section in the draft guidance for industry *Assessing Adhesion With Transdermal and Topical Delivery Systems for ANDAs* (October 2018).[45]

[43] Fleiss, J.L., 1986, Analysis of Data from Multiclinic Trials, Controlled Clinical Trials, 7(4):267-275; Permutt, T., 2003, Probability Models and Computational Models for ANOVA in Multicenter Clinical Trials, Journal of Biopharmaceutical Statistics, 13(3):495-505; Tsiatis, A.A., M. Davidian, M. Zhang, and X. Lu, 2008, Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: A Principled Yet Flexible Approach, Statistics in Medicine, 27(23):4658-4677.

[44] 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, Statistics in Biopharmaceutical Research, 7(4):286-303; Grizzle, J.E., 1965, The Two-Period Change-Over Design and Its Use in Clinical Trials, Biometrics, 21(2):467-480.

[45] 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.

F. Dose Scale for Bioequivalence Assessment

In this method, the BE assessment is based on relative bioavailability of the test and reference formulations at the site(s) of action. The relative bioavailability, 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 Emax model that describes the within-study dose-response relationship. Among available statistical methods for Emax model fitting, nonlinear mixed effect (NLME) modeling is recommended, because the NLME modeling is capable of characterizing between-subject variability and residual unexplained variability, and less sensitive to aberrant observation and missing values.

For model fitting details, refer to the PSG on Orlistat oral capsule.[46]

To determine BE, the 90% 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% confidence interval for F.[47] Alternatively, the 90% 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 analysis for comparative PD BE studies, applicants are encouraged to contact the Agency early to discuss their proposed study designs and statistical methods (e.g., alternative modeling approaches, impact of the missing data and the handling strategy) via the controlled correspondence, pre-ANDA meeting, pre-IND meeting, or pre-NDA meeting pathway.[48]

G. Bioequivalence 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 and reference products, 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 T product, an EU reference product, and a U.S. reference product, the comparison of the T product to the U.S. reference product should be based on an analysis excluding the data from the EU reference. However, full data from the BE studies, including data comparing the T product that

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

[47] Ibid.

[48] See footnotes 8, 9, and 10.

is the subject of the application with non-U.S. reference products, 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.

V. APPENDICES

A. Choice of Specific Replicated Crossover Designs

Appendix A describes why FDA prefers replicated crossover designs with only two sequences, and why the Agency recommends the specific designs described in section II.A.1.b of this guidance.

1. Reasons Unrelated to Carryover Effects

Each unique combination of sequence and period in a replicated crossover design can be called a cell of the design. For example, the two-sequence, four-period design recommended in section II.A.1.b 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 V.A.1, there would be 2-1=1 degree-of-freedom due to sequence, 4-1=3 degrees-of-freedom due to period, and 2-1=1 degree-of-freedom due to treatment, for a total of 1+3+1=5 degrees-of-freedom due to the three fixed effects. Because these 5 degrees-of-freedom do not account for all 7 degrees-of-freedom attributable to the eight cells of the design, the fixed-effects model is not saturated. There could be some controversy as to whether a fixed-effects model that accounts for more or all of the degrees-of-freedom due to cells (i.e., a more saturated fixed-effects model) should be used. For example, a sequence-by-period-by-treatment interaction effect might be included, which would fully saturate the fixed-effects model.

If the replicated crossover design has only two sequences, use of only the three main effects (sequence, period, and treatment) in the fixed-effects model or use of a more saturated model makes little difference to the results of the analysis, provided there are no missing observations, and the study is carried out in one group of subjects. The least squares point estimate of $\mu_T - \mu_R$ will be the same for the main-effects model and for the saturated model.

If the replicated 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 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 replicated crossover designs. Both of these designs are strongly balanced for carryover effects, meaning that each treatment is preceded by each other treatment and itself an equal number of times.

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

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

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

3. *Two-Period Replicated Crossover Designs*

For most drug products, two-period replicated 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 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 BA/BE studies (i.e., AUC and $C_{max,}$ but not T_{max}) (Westlake 1973 and 1988).[49,50] 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:

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

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

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 involves a multiplicative term (CL) that might be regarded as a function of the subject. For this reason,

[49] Westlake, W. J., 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.

[50] Westlake, W. J., 1988, Bioavailability and Bioequivalence of Pharmaceutical Formulations, Biopharmaceutical Statistics for Drug Development, 329-352.

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:

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

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 Bioequivalence Analysis of Replicated 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 pre-specified 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.