Journal of
Bioequivalence & Bioavailability
ISSN: 0975-0851

Lu et al., J Bioequiv Availab 2014, 6:2
DOI: 10.4172/jbb.1000182

Research Article Open Access

In vivo and *In vitro* Bioequivalence Testing

Ying Lu[1]*, Shein-Chung Chow[2,3] and Shichen Zhu[3,4]

[1]*Beijing University of Technology, Beijing, China*
[2]*Duke University School of Medicine, Durham, North Carolina, USA*
[3]*Civil Aviation University of China, Tianjin, China*
[4]*Statistics Department, North Carolina State University, Raleigh, North Carolina, USA*

Abstract

For approval of generic drug products, the FDA requires that evidence of average bioequivalence in drug absorption be provided through the conduct of bioequivalence studies. As indicated in 21CFR320.24, bioequivalence may be established by *in vivo* (e.g., pharmacokinetic, pharmacodynamic, or clinical) and *in vitro* studies or with suitable justification by *in vitro* studies alone. In this presentation, an overview of statistical considerations including study design, criteria, and statistical methods for assessment of bioequivalence will be provided. For *in vivo* bioequivalence testing, in addition to average bioequivalence, the concept of population bioequivalence and individual bioequivalence for addressing drug interchangeability will also be discussed. For *in vitro* bioequivalence testing, an overview regarding some *in vitro* tests such as dose or spray content uniformity through container's life, droplet and drug particle size distribution, spray pattern, plume geometry, priming and repriming, and tail off profile that are commonly employed for local action drug products such as nasal aerosols and nasal sprays products will be provided. Recent development and future research topics will also be discussed.

Keywords: Fundamental bioequivalence assumption; Individual bioequivalence; Population bioequivalence; Highly variable drugs; *In vitro-in vivo* correlation

Introduction

For approval of generic drug products, bioequivalence testing is considered as a surrogate for clinical evaluation of the therapeutic equivalence of drug products based on the Fundamental Bioequivalence Assumption that when two drug products (e.g., a brand-name drug and its generic copy) are equivalent in bioavailability, they will reach the same therapeutic effect. Although bioavailability for *in vivo* bioequivalence studies is usually assessed through the measures of the rate and extent to which the drug product is absorbed into the bloodstream of human subjects, for some locally acting drug products such as nasal aerosols (e.g., etered-dose inhalers) and nasal sprays (e.g., metered-dose spray pumps) that are not intended to be absorbed into the bloodstream, bioavailability may be assessed by measurements intended to reflect the rate and extent to which the active ingredient or active moiety becomes available at the site of action. For those local delivery drug products, the United States Food and Drug Administration (FDA) indicates that bioequivalence may be assessed, with suitable justification, by *in vitro* bioequivalence studies alone (e.g., Part 21 Codes of Federal Regulations Section 320.24).

In practice, although it is recognized that *in vitro* methods are less variable, easier to control, and more likely to detect differences between products if they exist, the clinical relevance of the *in vitro* tests or the magnitude of the differences in the tests are not clearly established until a draft guidance on bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action [1] and a draft guidance on Nasal Spray and Inhalation Solution, Suspension and Spray Drug Product [1] were issued by the FDA. The 1999 FDA draft guidance on bioavailability and bioequivalence was subsequently revised and issued in [1]. The 2003 FDA [2] draft guidance indicates that *in vitro* bioequivalence can be established through seven *in vitro* tests. These *in vitro* tests include tests for (i) single actuation content through container life, (ii) droplet size distribution by laser diffraction, (iii) drug in small particles/droplets, or particle/droplet size distribution by cascade impactor, (iv) drug particle size distribution by microscopy, (v) spray pattern, (vi) plume geometry, and (vii) priming and re-priming. For bioequivalence assessment of the seven *in vitro* tests, the FDA classifies statistical methods as either the non-profile analysis or the profile analysis.

In the next two sections, an overview regarding design and analysis of *in vivo* and *in vitro* bioequivalence testing for generic development are provided. Current issues are discussed in Section 4. Recent developments are discussed in the last section of this article.

In vivo bioequvalence testing

The process of *in vivo* bioequivalence testing starts with Fundamental Bioequivalence Assumption followed by conducting a bioequivalence study under a valid study design, appropriate statistical methods for assessment of average bioequivalence, and regulatory submission, review, and approval.

Fundamental bioequivalence assumptions

As indicated [2], bioequivalence studies are necessarily conducted under the Fundamental Bioequivalence Assumption, which constitutes legal basis (from the Hatch-Waxman Act) for regulatory review and approval of small molecule generic drug products. The Fundamental Bioequivalence Assumption states that "If two drug products are shown to be bioequivalent, it is assumed that they will reach the same therapeutic effect or they are therapeutically equivalent."

In practice, bioequivalence in drug absorption has been interpreted

***Corresponding author:** Ying Lu, Beijing University of Technology, Beijing, 121004, China, Tel: 86 1342 6181 945; E-mail: Lvying@emails.bjut.edu.cn

Received November 06, 2013; **Accepted** March 24, 2014; **Published** March 31, 2014

Citation: Lu Y, Chow SC, Zhu S (2014) *In vivo* and *In vitro* Bioequivalence Testing. J Bioequiv Availab 6: 067-074. doi:10.4172/jbb.1000182

that the confidence interval for the ratio of means (of drug absorption) is within bioequivalence limits. An alternative would be to show that the tolerance intervals (or a distribution free model) overlap sufficiently. On the basis of the Fundamental Bioequivalence Assumption, many practitioners interpret that generic drug products and the innovative drug product can be used interchangeably because they are therapeutically equivalent. The FDA, however, does not indicate that approved generic drug products and the innovative drug products can be used interchangeably. The FDA only indicates that an approved generic drug product can be used as a substitute for the innovative drug product.

Study design

As indicated in the Federal Register [Vol. 42, No. 5, Sec. 320.26(b) and Sec. 320.27(b), 1977], a bioavailability study (single-dose or multi-dose) should be crossover in design, unless a parallel or other design is more appropriate for valid scientific reasons. Thus, in practice, a standard two-sequence, two-period (or 2×2) crossover design is often considered for a bioavailability or bioequivalence study. Denote by T and R the test product and the reference product, respectively. Thus, a 2x2 crossover design can be expressed as (TR, RT), where TR is the first sequence of treatments and RT denotes the second sequence of treatments. Under the (TR, RT) design, qualified subjects who are randomly assigned to sequence 1 (TR) will receive the test product (T) first and then cross-over to receive the reference product (R) after a sufficient length of wash-out period. Similarly, subjects who are randomly assigned to sequence 2 (RT) will receive the reference product (R) first and then cross-over to receive the test product (T) after a sufficient length of wash-out period.

One of the limitations of the standard 2×2 crossover design is that it does not provide independent estimates of intra-subject variabilities since each subject receives the same treatment only once. In the interest of assessing intra-subject variabilities, the following alternative crossover designs for comparing two drug products are often considered:

Design 1: Balaam's design – e.g., (TT, RR, RT, TR);

Design 2: Two-sequence, three-period dual design – e.g., (TRR, RTT);

Design 3: Four-period design with two sequences – e.g., (TRRT, RTTR);

Design 4: Four-period design with four sequences – e.g., (TTRR, RRTT, TRTR, RTTR).

The above study designs are also referred to as higher-order crossover designs. A higher-order crossover design is defined as a design with the number of sequences or the number of periods greater than the number of treatments to be compared.

For comparing more than two drug products, a Williams' design is often considered. For example, for comparing three drug products, a six-sequence, three-period (6×3) Williams' design is usually considered, while a 4×4 Williams' design is employed for comparing 4 drug products. Williams' design is a variance stabilizing design. More information regarding the construction and good design characteristics of Williams' designs can be found in [3].

Statistical methods

Average bioequivalence (ABE) is claimed if the geometric means ratio (GMR) of average bioavailabilities between test and reference products is within the bioequivalence limit of 80%-125% with 90% assurance based on log-transformed data. Along this line, commonly employed statistical methods are the confidence interval approach and the method of interval hypotheses testing. For the confidence interval approach, a 90% confidence interval for the ratio of means of the primary pharmacokinetic response such as AUC or C_{max} is obtained under an analysis of variance model. We claim bioequivalence if the obtained 90% confidence interval is totally within the bioequivalence limit of 80%-125%.

For the method of interval hypotheses testing, the interval hypotheses that:

H_0: Bioinequivalence vs. H_a: Bioequivalence (1)

Note that the above hypotheses are usually decomposed into two sets of one-sided hypotheses. For the first set of hypotheses is to verify that the average bioavailability of the test product is not too low, whereas the second set of hypotheses is to verify that average bioavailability of the test product is not too high. Under the two one-sided hypotheses, Schuirmann's [4] two one-sided tests procedure is commonly employed for testing ABE.

In practice, other statistical methods such as Westlake's symmetric confidence interval approach [5], exact confidence interval based on Fieller's theorem [6], Chow and Shao's joint confidence region approach [7], Bayesian methods, and non-parametric methods such as Wilcoxon-Mann-Whitney [8] one-sided tests procedure, distribution-free confidence interval based on the Hodges-Lehmann [9] estimator, and bootstrap confidence interval are sometimes considered (Chow et al. [3]).

Issue of drug interchangeability

Basically, drug interchangeability can be classified either as drug prescribability or drug switchability (Chow et al. [3]). Drug prescribability is defined as the physician's choice for prescribing an appropriate drug product for his/her new patients between a brand-name drug product and a number of generic drug products that have been shown to be bioequivalent/biosimilar to the brand-name drug product. The underlying assumption of drug prescribability is that the brand-name drug product and its generic copies can be used alternatively in terms of the efficacy and safety of the drug product. Drug prescribability, therefore, is the interchangeability for a new patient. Drug switchability, on the other hand, is related to the switch from a drug product (e.g., a brand-name drug product) to an alternative drug product (e.g., a generic copy of the brand-name drug product) within the same subject, whose concentration of the drug product has been titrated to a steady, efficacious, and safe level. As a result, drug switchability is considered more critical than drug prescribability in the study of drug interchangeability for patients who have been on medication for a while. Drug switchability, therefore, is exchangeability within the same subject.

Note that in practice, many use the terms interchangeability and switchability synonymously. (Another term used in this context, is substitutability.) These terms are meant to replace, in a given patient, the administration of one drug product by another. Thus, these usages refer to subjects to whom the drug has already been administered and who are not naïve to it. Also noteworthy is the definition of interchangeability in the Biologics Price Competition and Innovation (BPCI) [10] Act of 2010, Section 7002: "(3) The term 'interchangeable'

or 'interchangeability', in reference to a biological product that is shown to meet the standards described in subsection (4), means that the biological product may be substituted for the reference product without the intervention of the health care provider who prescribed the reference product."

Population bioequivalence for drug prescribability: As indicated in [1], average bioequivalence can guarantee neither drug prescribability nor drug switchability. Therefore, it is suggested that the assessment of bioequivalence should take into consideration of drug prescribability and drug switchability. To address drug interchangeability, it is recommended that population bioequivalence (PBE) and individual bioequivalence (IBE) be considered for testing drug prescribability and drug switchability, respectively. More specifically, the FDA recommends that PBE be applied to new formulations, additional strengths, or new dosage forms in NDAs, while IBE should be considered for ANDA or AADA (abbreviated antibiotic drug application) for generic drugs.

To address drug prescribability, FDA [11] proposed the following aggregated, scaled, moment-based, one-sided criterion:

$$PBC = \frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\max(\sigma_{TR}^2 - \sigma_{T0}^2)} \le \theta_p \quad (2)$$

where μ_T and μ_R are the mean of the test drug product and the reference drug product, respectively, σ_{TT}^2 and σ_{TR}^2 are the total variance of the test drug product and the reference drug product, respectively, σ_{T0}^2 is a constant that can be adjusted to control the probability of passing PBE, and θ_p is the bioequivalence limit for PBE. The numerator on the left-hand side of the criterion is the sum of the squared difference of the population averages and the difference in total variance between the test and reference drug products which measure the similarity for the marginal population distribution between the test and reference drug products. The denominator on the left-hand side of the criterion is a scaling factor that depends upon the variability of the drug class of the reference drug product. The FDA guidance suggests that θ_p be chosen as

$$\theta_p = \frac{(\log 1.25)^2 \varepsilon_p}{\sigma_{T0}^2} \quad (3)$$

Where ε_P is guided by the consideration of the variability term $\sigma_{TT}^2 - \sigma_{TR}^2$ added to the ABE criterion. As suggested by the FDA guidance, it may be appropriate that ε_P chosen to be 0.02. For the determination of σ_{T0}^2, the guidance suggests the use of so-called population difference ratio (PDR), which is defined as

$$PDR = \left[\frac{E(T-R)^2}{E(T-R')^2}\right]^{1/2} = \left[\frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 + \sigma_{TR}^2}{2\sigma_{TR}^2}\right]^{1/2} = \left[\frac{PBC}{2} + 1\right]^{1/2} \quad (4)$$

Therefore, assuming that the maximum allowable PDR is 1.25, substitution of $(\log 1.25)^2 / \sigma_{T0}^2$ for PBC without adjustment of the variance term approximately yield σ_{T0}^2.

Individual bioequivalence for drug switch ability: Similarly, to address drug switch ability, the FDA recommended the following aggregated, scaled, moment-based, one-sided criterion:

$$IBC = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\max(\sigma_{WT}^2, \sigma_{W0}^2)} \le \theta_I \quad (5)$$

where σ_{WT}^2 and σ_{WR}^2 are the within-subject variances of the test drug product and the reference drug product, respectively, σ_D^2 is the variance component due to subject-by-drug interaction, σ_{W0}^2 is a constant that can be adjusted to control the probability of passing IBE, and θ is the bioequivalence limit for IBE. The FDA guidance suggests that θ_I be chosen

$$\theta_I = \frac{(\log 1.25)^2 + \varepsilon_I}{\sigma_{W0}^2} \quad (6)$$

Where ε_I is the variance allowance factor, which can be adjusted for sample size control. Note that the FDA guidance suggests $\varepsilon_I = 0.05$.

For the determination of σ_{T0}^2, the guidance suggests the use of individual difference ratio (IDR), which is defined as:

$$IDR = \left[\frac{E(T-R)^2}{E(T-R')^2}\right]^{1/2} = \left[\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 + \sigma_{WR}^2)}{2\sigma_{WR}^2}\right]^{1/2} = \left[\frac{IBC}{2} + 1\right]^{1/2} \quad (7)$$

Therefore, assuming that the maximum allowable IDR is 1.25, substitution of $(\log 1.25)^2 / \sigma_{W0}^2$ for IBC without adjustment of the variance term approximately yield $\sigma_{W0} = 0.2$. It should be noted that although the FDA guidance recommends $\sigma_{W0} = 0.2$, FDA uses (in a different context) $\sigma_{W0} = 0.2$.

In vitro Bioequivalence Testing

For the assessment of *in vitro* bioequivalence, the FDA [1] guidance requires that seven *in vitro* testing of single actuation content uniformity through container life, droplet/particle size distribution, spray pattern, plume geometry, and priming/re-priming be done to demonstrate comparable delivery characteristics between two drug products. In this section, a brief description of the recommended study design and each of the seven *in vitro* tests are given.

Study design

According to the FDA, three products from each lots/sub-lots are required to be tested for *in vitro* emitted dose uniformity, droplet size distribution, spray pattern, plume geometry, priming/re-priming, and tail-off profile. For each *in vitro* test, ten samples are randomly drawn from each lot. Samples are randomized for *in vitro* tests. The analysts will not have access to the randomization codes. An automated actuation station with a fixed setting (actuation force, dose time, return time, and hold time) is usually used for the *in vitro* tests.

Emitted dose uniformity, priming, priming/re-priming, and tail-off profile: Following the FDA's recommendations, the priming, emitted dose uniformity, priming/re-priming, and tail-off tests may be tested in the following setting. Three individual lots of test product and reference product are evaluated. For each lot, ten samples are then tested for pump priming, unit spray content through life, and tail-off studies. Then, additional samples for each lot are evaluated for the prime hold study (re-prime study).

For each sample unit, spray samples are collected for sprays 1-8 and analyzed in order to determine the minimum number of actuations required before the pump delivers the labeled dose of drug (sprays 1-8). To characterize emitted dose uniformity at the beginning of unit life, spray 9 is collected. Sprays 10-15 are wasted by the automatic actuation station. Spray 16 is collected in the middle of unit life. Sprays 17-20 are wasted. Sprays 21-23 are collected at the end of the unit life. Additional sprays after spray 23 are collected and analyzed to determine the tail-off profile.

Ten additional samples are drawn randomly from each lot of drug product for the pump prime hold study. For each unit, the first

12 sprays (sprays 1-12) are wasted. Sprays 13 and 14 are collected as fully primed sprays. The unit is then stored undisturbed for 24 hours. Within each lot, five samples are placed in the upright position, while the other five samples are placed in a side position. After that, sprays 15-17 are collected. The unit is then stored undisturbed in its former position for another 24 hours. After that, the doses emitted by sprays 18-20 are collected. All spray samples are weighted in order to obtain re-priming characteristics.

Spray pattern: A spray pattern produced by a nasal spray pump evaluates in part the integrity and the performance of the orifice and pump mechanism in delivering a dose to its intended site of deposition. Measurements can be made on the diameter of the horizontal intersection of the spray plume at different distances from the actuator tip. Spray patterns are usually measured at three distances (e.g., 1, 2, and 4 cm) at both the beginning (sprays 8-10) and the end (sprays 17-19) of unit life. As a result, a total of six pray patterns is collected for each sample unit. For each spray pattern image, the diameters (the longest and shortest diameters) and the ovality (which is defined by the ratio of the longest to the shortest diameters) are measured.

Droplet size distribution: For a test of droplet size distribution, methods of laser diffraction and cascade impaction are commonly used. These methods are briefly described below.

Laser diffraction: For a test of droplet size distribution using laser diffraction particle analyzer, each sample unit is first primed by actuating the pump eight times using an automatic actuation station. Droplet size distribution is then determined at three distances (e.g., 3, 5, and 7 cm) from the laser beam and at the beginning, the middle, and the end of unit life. At each distance, three measurements of delay times (plume, formation, start of dissipation, and intermediate measurements) and overall evaluation are used to characterize the droplet size. As a result, a total of 36 measurements are recorded for each sample unit.

Cascade impaction: When the spray pump is actuated in the nasal cavity, a fine mist of droplets is generated. Droplets that are greater than 9 in diameter are considered non-respirable and are therefore useful for nasal deposition. As recommended in the 1999 FDA [1] draft guidance, the data should be reported as follows:

Group 1: Adaptor (expansion chamber, i.e., 5-L flask), rubber gasket, throat, and Stage 0

Group 2: Stage 1

Group 3: Stage 2 to filter

Each sample unit is first primed by actuating the pump seven times using an automatic actuation station. Droplet size distribution is then determined at the beginning and the end of the life of the sample. Thus, a total of six groups of results are reported for each spray unit.

Plume geometry: Plume geometry is performed on the nasal spray plume that is allowed to develop into an unconstrained space that far exceeds the volume of nasal cavity. It represents a frozen moment in spray plume development that is viewed from two axes perpendicular to the axis of plume development. The samples should be actuated vertically. Prime the pump with 10 actuations until a steady fine mist is produced from the pump. A fast-speed video camera is placed in front of the sample bottle and starts recording. Repeat the test by rotating the actuator 90 degree to the previous actuator placement so that two side views are at 90 degrees to each other (two perpendicular planes) and, relative to the axis of the plume of the spray, are captured when actuated into space. Spray plumes are characterized at three stages: early upon formation, as the plume starts dissipate, and at some intermediate time. Longest vertical distance (LVD), widest horizontal distance (WHD), and plume angle (ANG) are recorded and analyzed.

Methods for data analysis

For assessment of bioequivalence for the six *in vitro* tests, in addition to so-called non-comparative analysis, the FDA classifies statistical methods as either non-profile analysis or profile analysis [12,13], which are briefly described below.

Non-comparative analysis: For each *in vitro* test, the FDA requires that a non-comparative analysis be performed. Non-comparative analysis refers to the statistical summarization of the bioavailability data by descriptive statistics. As a result, means, standard deviations, and coefficients of variation (CVs) in percentage of the six *in vitro* tests should be documented. More specifically, the overall sample means for a given formulation should be averaged over all samples (e.g., bottle/canisters), life stages (except for priming and re-priming evaluations), and lots or batches. In addition to the overall means, means at each life stage for each batch averaged over all bottles/canisters and for each life stage averaged over all lots (or batches) should be presented. For profile data, means, standard deviations, and percent CVs should be reported for each stage. The between-lot (or batch), within-lot (or batch) between-sample (e.g., bottle or canister), and within-sample (e.g., bottle or canister) between-life stage variability should be evaluated through appropriate statistical models.

Non-profile analysis: The FDA classifies statistical methods for assessment of the six *in vitro* bioequivalence tests for nasal aerosols and sprays as either the non-profile analysis or the profile analysis. In this paper we focus on the non-profile analysis, which applies to tests for dose or spray content uniformity through container life, droplet size distribution, spray pattern, and priming and re-priming. Non-profile analysis applied to emitted dose or sprays content uniformity, through container life, droplet size distribution, spray pattern, and priming/re-priming. Suppose that m_T and m_R canisters from respectively the test and the reference products are randomly selected for *in vitro* bioequivalence testing and one observation from each canister is obtained. The data can be described by the following model:

$$y_{jk} = \mu_k + \varepsilon_{jk} \quad (8)$$

$$j=1,\ldots, m_k$$

Where $k=T$ for the test product, $k=R$ for the reference product, μ_T and μ_R are fixed product effects, ε_{jk}'s are independent random measurement errors distributed as $N\left(0,\sigma_k^2\right)$, $k=T,R$. Under model (8), the parameter θ is given by

$$\theta = \frac{\left(\mu_T - \mu_R\right)+\sigma_T^2 - \sigma_R^2}{\max\left(\sigma_0^2,\sigma_R^2\right)} \quad (9)$$

and $\theta < \theta_{BE}$ if and only if $\varsigma < 0$, where

$$\varsigma = \left(\mu_T - \mu_R\right)^2 + \sigma_T^2 - \sigma_R^2 - \theta_{PE}\max\left(\sigma_0^2,\sigma_R^2\right) \quad (10)$$

To test bioequivalence at level 5%, it suffices to construct a 95% upper confidence bound for ς. Under model (8), the best unbiased estimator of $\delta = \mu_T - \mu_R$ is

$$\hat{\delta} = \bar{y}_T - \bar{y}_R \sim N\left(0,\ \frac{\sigma_T^2}{m_T} + \frac{\sigma_R^2}{m_R}\right) \quad (11)$$

where $\bar{y}_k$ is the average of y_{jk} over j for a fixed k. The best unbiased estimator of σ_k^2 is

$$s_k^2 = \frac{1}{m_k - 1}\sum_{j=1}^{m_k}\left(y_{jk} - \bar{y}_k\right)^2 \sim \frac{\sigma_k^2 \chi_t^2 \sigma_{m_k-1}^2}{m_k - 1},$$

where $k = T, R$ and χ_t^2 denotes the central chi-square distribution with t degrees of freedom. Using the method in [14] for individual bioequivalence testing, an approximate 95% upper confidence bound for ζ in (10) is

$$\tilde{\zeta}_U = \hat{\delta}^2 + s_T^2 - s_R^2 - \theta_{BE}\max\left\{\sigma_0^2, s_R^2\right\} + \sqrt{U} \tag{12}$$

where U is the sum of the following three quantities:

$$\left[\left(|\delta| + z_{0.95}\sqrt{\frac{s_T^2}{m_T} + \frac{s_R^2}{m_R}}\right)^2 - \delta^2\right],$$

$$s_T^4\left(\frac{m_T - 1}{\chi^2_{0.05;m_{T-1}}} - 1\right)^2 \text{ and}$$

$$\left(1 + c\theta_{BE}\right)^2 s_R^4\left(\frac{m_R - 1}{\chi^2_{0.95;m_{R-1}}} - 1\right)^2$$

$c = 1$ if $s_R^2 \geq \sigma_0^2$, $c = 0$ if $s_R^2 < \sigma_0^2$, z_a is the ath quantile of the standard normal distribution, and $\chi^2_{t;\alpha}$ is the ath quantile of the central chi-square distribution with t degrees of freedom. *In vitro* bioequivalence can be claimed if $\tilde{\zeta}_U < 0$. This procedure is recommended by the FDA guidance [1].

As indicated in the FDA draft guidance, the FDA requires that m_k be at least 30. However, m_k = 30 may not be enough to achieve a desired power of the bioequivalence test in some situations. Increasing m_k can certainly increase the power, but in some situations, obtaining replicates from each canister may be more practical, and/or cost-effective. With replicates from each canister, however, the previously described test procedure is necessarily modified in order to address the between- and within-canister variabilities.

Profile analysis: As indicated in the FDA draft guidance, profile analysis using a confidence interval approach should be applied to cascade impactor or multistage liquid impringer (MSLI) for particle size distribution. Equivalence may be assessed based on chi-square differences. The idea is to compare the profile difference between test product and reference product samples to the profile variation between reference product samples. More specifically, let y_{ijk} denote the observation from the jth subject's ith stage of the kth treatment. Given a sample $_{(j_0, j_1)}$ from test product and two samples (j_0, j_1) from reference products and assuming that there are a total of S stages, the profile distance between test and reference is given by

$$d_{TR} = \sum_{i=1}^{S}\frac{\left(y_{ij_0T} - 0.5\left(y_{ij_1R} + y_{ij_2R}\right)\right)^2}{\left(y_{ij_0T} + 0.5\left(y_{ij_1R} + y_{ij_2R}\right)\right)} \tag{13}$$

Similarly, the profile variability within reference is defined as

$$d_{RR} = \sum_{i=1}^{S}\frac{\left(y_{ij_1R} - y_{ij_2R}\right)^2}{0.5\left(y_{ij_1R} + y_{ij_2R}\right)} \tag{14}$$

For a given triplet sample of (Test, Reference 1, Reference 2), the ratio of d_{TR} and d_{RR}, i.e.,

$$rd = \frac{d_{TR}}{d_{RR}} \tag{15}$$

can then be used as a bioequivalence measure for the triplet samples between the two drug products. For a selected sample, the 95% upper confidence bound of $E(rd)=E\left(d_{TR}/d_{RR}\right)$ is then used as a bioequivalence measure for the determination of bioequivalence. In other words, if the 95% upper confidence bound is less than the bioequivalence limit, then we claim that the two products are bioequivalent. The 1999 FDA [1] draft guidance recommends a bootstrap procedure to construct the 95% upper bound for *E(rd)*.The procedure is described below;

Assume that the samples are obtained in a two-stage sampling manner. In other words, for each treatment (test or reference), three lots are randomly sampled. Within each lot, ten samples (e.g., bottles or canisters) are sampled. The following is quoted from the 1999 FDA [1] draft guidance regarding the bootstrap procedure to establish profile bioequivalence.

For an experiment consisting of three lots each of test and reference products, and with 10 canisters per lot, the lots can be matched into six different combinations of triplets with two different reference lots in each triplet. The 10 canister of a test lot can be paired with the 10 canister of each of the two reference lots in (10 factorial)^2=3,628,800^2 combinations in each of the lot triplets. Hence a random sample of the N canister pairing of the six Test-Reference 1-Reference 2 lot triplets is needed. rd is estimated by the sample mean of the rds calculated for the triplets in 10 selected samples of N. Note that the FDA recommends that N=500 be considered.

Current Issues

Fundamental assumption

For *in vivo* bioequivalence testing, the Fundamental Bioequivalence Assumption states that: If two drug products are shown to be bioequivalent, it is assumed that they will reach the same therapeutic effect or they are therapeutically equivalent and hence can be used interchangeably. For *in vitro* bioequivalence testing, the fundamental assumption is that *in vitro* testing (for drug release or delivery) is predictive of *in vivo* testing (for drug absorption). Under the Fundamental Bioequivalence Assumption, one of the controversial issues is that bioequivalence may not necessarily imply therapeutic equivalence and therapeutic equivalence does not guarantee bioequivalence either. The assessment of average bioequivalence for generic approval has been criticized that it is based on legal/political deliberations rather than scientific considerations. In the past several decades, many sponsors/researchers have made an attempt to challenge this assumption with no success.

Note that the Fundamental Bioequivalence Assumption is also applied to drug products with local action such as nasal spray products via the assessment of *in vitro* bioequivalence testing. In either *in vivo* or *in vitro* bioequivalence testing, the verification of the Fundamental Bioequivalence Assumption is often difficult, if not impossible, without the conduct of clinical trials. It should be noted that the Fundamental Bioequivalence Assumption is for drug products with identical active ingredient (s). Note that for two products to be bioequivalent they must have, by general understanding, the same active ingredients.

In practice, the verification of the Fundamental Bioequivalence Assumption is often difficult, if not impossible, without the conduct of clinical trials. In practice, there are following four possible scenarios:

(1) Drug absorption profiles are similar and they are therapeutic equivalent;

(2) Drug absorption profiles are not similar but they are therapeutic equivalent;

(3) Drug absorption profiles are similar but they are not therapeutic equivalent;

(4) Drug absorption profiles are not similar and they are not therapeutic equivalent.

The Fundamental Bioequivalence Assumption is nothing but scenario (1). Scenario (1) works if the drug absorption (in terms of the rate and extent of absorption) is predictive of clinical outcome. In this case, PK responses such as AUC (area under the blood or plasma concentration-time curve for measurement of the extent of drug absorption) and C_{max} (maximum concentration for measurement of the rate of drug absorption) serve as surrogate endpoints for clinical endpoints for assessment of efficacy and safety of the test product under investigation. Scenario (2) is the case where generic companies use to argue for generic approval of their drug products especially when their products fail to meet regulatory requirement for bioequivalence. In this case, it is doubtful that there is a relationship between PK responses and clinical endpoints. The innovator companies usually argue with the regulatory agency to against generic approval with scenario (3). However, more studies are necessarily conducted in order to verify scenario (3). There are no arguments with respect to scenario (4).

In practice, the Fundamental Bioequivalence Assumption is applied to all drug products across therapeutic areas without convincing scientific justification. In the past several decades, however, no significant safety incidences were reported for the generic drug products approved under the Fundamental Bioequivalence Assumption. One of the convincing explanations is that the Fundamental Bioequivalence Assumption is for drug products with identical active ingredient(s). Whether the Fundamental Bioequivalence Assumption is applicable to drug products with similar but different active ingredient(s) as in the case of follow-on products becomes an interesting but controversial question.

One-size-fits-all criteria

For the assessment of bioequivalence both *in vivo* and *in vitro*, FDA adopted a one size-fits-all criterion. That is, for *in vivo* (*in vitro*), a test drug product is said to be bioequivalent to a reference drug product if the estimated 90% confidence interval for the ratio of geometric means of the primary PK parameters (AUC and C_{max}) is totally within the bioequivalence limits of 80% to 125% (90% to 111%). The one size-fits-all criterion does not take into consideration the therapeutic window and intra-subject variability of a drug which have been identified to have non-negligible impact on the safety and efficacy of generic drug products as compared to the innovative drug products.

In the past several decades, this one size-fits-all criterion has been challenged and criticized by many researchers. It was suggested flexible criteria in terms of safety (upper bioequivalence limit) and efficacy (lower bioequivalence limit) should be developed based on the characteristics of the drug, its therapeutic window (TW) and intra-subject variability (ISV) (Table 1).

The approach of one size-fits-all has begun to dissipate in recent years. For instance, in some jurisdictions such as Europe and Canada, narrower BE limits have been proposed for drugs with narrow therapeutic windows (Health Canada. 2006, [15]). However, FDA has maintained its usual requirement for these drugs with BE limits to be between 80% and 125%.

Class	TW	ISV	Example
A	Narrow	High	Cyclosporine
B	Narrow	Low	Theophylline
C	Wide	Low to moderate	Most drugs
D	Wide	High	Chlopromazine or topical corticosteroids

TW: Therapeutic Window; ISV: Intra-Subject Variability

Table 1: Classification of drugs.

On the other hand, for orally administered drugs with high within-subject variability and wide therapeutic window (Class D, highly variable drugs, see Table 1), the regulatory expectation has become, in some cases, more relaxed. For these drugs, the approach of scaled average bioequivalence has been proposed [16,17]. This method is closely related to, and is a simplification of, the procedure recommended earlier for individual BE when the within-subject variation is high ($\sigma^2_{WR} > \sigma^2_{W0}$). While the current FDA guidance does not contain special provisions for this class of drugs, the agency actually entertains submissions based on the criteria described in an 'informal' publication [16] which recommends the approach of scaled average bioequivalence. Europe has recently also suggested the application of a variant of this procedure. However, some other agencies still apply the one size-fits-all approach and require the usual BE limits of 80% to 125% also for this class of drugs.

Profile analysis for *in vitro* bioequivalence testing

The bootstrap procedure described in the FDA guidance [1] has received much attention and criticisms. Major criticisms are described below;

First, the statistical properties of this procedure are unknown. It includes two aspects. One is that the statistical model, which should be used to describe the profile data, is not clearly defined in the FDA draft guidances. In addition, even under an appropriate statistical model, the statistical properties of the bootstrap procedure are still unknown. More specifically, it is not clear whether the bootstrap sample mean a consistent estimator for E(rd). As a result, the 95% percentile of the bootstrap samples may not be an appropriate 95% upper bound for E(rd). These questions are not addressed in the FDA draft guidances.

Second, no criteria are given regarding the passage or failure of the bioequivalence study. This is the issue that confuses most researchers/scientists in practice. After the conduct of a valid trial and an appropriate statistical analysis following the FDA draft guidance, the sponsor still cannot tell if its product has passed or failed the bioequivalence test. This is a direct consequence of our first point (i.e., the statistical properties of the recommended bootstrap procedure are unknown).

Third, the simulation study using different random number generation schemes may produce contradictory results. It is possible for a good product to fail the bioequivalence test simply because of bad luck. It is also possible for a bad product to pass the bioequivalence test with an "appropriate" choice of random number generation scheme. As a result, researchers/scientists tend to rely more on the descriptive statistics of the two products in order to assess their bioequivalence instead of the bootstrap procedure. The proposed bootstrap procedure recommended by the FDA is not as reliable as it should be.

As a result, further research of profile analysis becomes a problem of interest in practice. More specifically, the questions of interest include (i) what statistical model should be used to describe the profile data?

(ii) is E(rd) defined by the FDA a good parameter for characterizing the bioequivalence between test and reference products? (Can we define the test-to-reference distance and reference-to-reference variability differently?) (iii) what bioequivalence limit should we use to evaluate the *in vitro* bioequivalence between two products based on appropriate model, parameter, and bioequivalence criterion?

Recent Development

Highly variable drug products

As indicated earlier, the assessment of ABE focuses on average bioavailability but ignores the variability associated with the PK responses. Thus, two drug products may fail the evaluation of ABE if the variability associated with the PK responses is large even though they have identical means. A drug with large variability is considered highly variable. FDA defines a highly variable drug (HVD) as a drug whose within-subject (or intra-subject) variation is greater than or equal to 30%.This definition based on intra-subject variation, however, rather arbitrary. One of problematic aspects of this definition is that the estimated within-subject variability depends on the metrics of pharmacokinetic responses such as AUC and C_{max}. In practice, the observed C_{max} is usually more variable than AUC. As indicated by [18], among the 212 bioequivalence studies submitted to the FDA, 33 studies were considered highly variable. In 28 of the 33 studies, only the C_{max} but not the AUC had a variation higher than 30%. Among the 33 studies, no cases indicate that the AUC but not the C_{max} is highly variable. [17] pointed out that HVDs show variable pharmacokinetics as a result of their inherent properties (e.g. distribution, systemic metabolism and elimination). A drug may have low variability if it is administered intravenously, whereas it can be highly variable after oral administration.

In practice, HVDs often fail to meet current regulatory acceptance criteria for ABE. In the past decade, the topic for evaluation of bioequivalence for HVDs has received much attention. This topic has been discussed several times at regulatory forums and international conferences, but academics, representatives of pharmaceutical industries and regulatory agencies failed to reach a consensus until recently that the approach of scaled average bioequivalence (SABE) is proposed by Haidar et al. [17] and Tothfalusi et al. [18] provided an excellent review for evaluation of bioequivalence for HVDs with SABE. The approach of SABE is briefly described below.

Scaled Average Bioequivalence (SABE): To introduce SABE, we first consider the criterion for ABE. As indicated earlier, the PK response is a logarithmically transformed metric, e.g., log (AUC) or log(C_{max}). The two one-sided tests (TOST) procedure is usually applied to assess bioequivalence [3]. Accordingly, the average logarithmic kinetic responses of the test (T) and reference (R) formulations, denoted by μ_R and μ_R respectively, are compared. The acceptance of bioequivalence is claimed it the difference between the logarithmic means is between pre-specified regulatory limits. The limits (θ_A) are generally symmetrical on the logarithmic scale and usually equal to ± ln (1.25). Thus, the criterion for ABE can be expressed as follows:

$$-\theta_A \le (\mu_T - \mu_R) \le \theta_A \tag{16}$$

In a bioequivalence study, the individual kinetic responses are evaluated from the measured concentrations. The means of the logarithmic responses of the two formulations are calculated. These sample averages estimate the true population means. A variance is also estimated for each kinetic response. It is a measure of the intra-subject variance but not always identical to it. FDA suggests the above ABE could be scaled by a standard deviation as follows:

$$-\theta_S \le \frac{(\mu_T - \mu_R)}{\sigma_W} \le \theta_S \tag{17}$$

Where, θ_S is the SABE regulatory cutoff. Here the standard deviation (σ_W) is the within-subject standard deviation. In replicate design, σ_W is generally the within-subject standard deviation of the reference formulation (denoted by σ_{WR}). Thus, the scaling factor of SABE has similar features to the scaling factor of IBE.

Recent considerations by regulatory agencies: Between early 1990 to early 2000, the FDA considered individual bioequivalence (IBE) and population bioequivalence (PBE) as a possible solution for the problem of bioequivalence for HVDs. However, the development of this approach has been abandoned. In 2004, the FDA kicked off a Critical Path Initiative that focused on the challenges involved in the development new drugs and generics. As part of this initiative, the FDA established a working group on the bioequivalence of HVDs for development of guidance on dealing with HVDs. The group made presentations to a meeting of its advisory committee in 2004 and at an AAPS symposium in 2005. The results and conclusions of the group's work were summarized very recently by Haidar et al. [17,19]. The summary [19] then serves as a basis for consideration by the FDA of actual submissions. Consequently, SABE appears to have gained a measure of recognition and implementation.

For evaluation of bioequivalence of HVDs with SABE, as indicated by David et al. [20], the bioequivalence limits for SABE can be expressed in the form of

$$\theta_S = \frac{\ln(1.25)}{\sigma_0} \tag{18}$$

Where, σ_0 is a so-called regulatory standardized variation, which defines the proportionality factor between the logarithmic bioequivalence limits and σ_W in the highly variable region. The value of σ_0 must be defined by the regulators. The magnitude of σ_0 defines the bioequivalence limits (θ_S). For instance, when $\sigma_0 = 0.294$, then θ_S is 0.760.

Remarks

The assessment of bioequivalence *in vivo* and *in vitro* has taken more and more attention in pharmaceutical companies and biological companies. There are some statistical criterions in bioequivalence *in vivo*, but *in vitro* assessment of bioequivalence the data analysis, the designs of clinical trial, and the criterion to assess bioequivalence are still incomplete. More notice should be taken on the *in vitro* bioequivalence.

References

1. FDA (1999) Guidance for Industry Nasal Spray and Inhalation Solution, Suspension, and Spray Drug Products - Center for Drug Evaluation and Research, the US Food and Drug Administration, Rockville, Maryland, USA.
2. FDA (2003) Guidance on Bioavailability and Bioequivalence Studies for Orally Administrated Drug Products - General Considerations, Center for Drug Evaluation and Research, the US Food and Drug Administration, Rockville, Maryland, USA.
3. Chow SC, Liu JP (2008) Design and Analysis of Bioavailability and Bioequivalence Studies. (3rdedn), Chapman Hall/CRC Press, Taylor & Francis, New York, New York, USA.
4. Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15: 657-680.

5. Westlake WJ (1976) Symmetrical confidence intervals for bioequivalence trials. Biometrics 32: 741-744.

6. Fieller EC (1954) Some problems in interval estimation. Journal of the Royal Statistical Society, Series B 16: 175-185.

7. Chow SC, Shao J, Wang H (2002) Individual bioequivalence testing under 2x3 designs. Stat Med 21: 629-648.

8. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics *Bulletin*, 1: 80-83.

9. Hodges JL, Lehmann EL (1963) Estimation of location based on ranks. Annals of Mathematical Statistics, 34: 598-611.

10. FDA (2010) Biologics Price Competition and Innovation Act, Center for Drug Evaluation and Research, the US Food and Drug Administration, Rockville, Maryland, USA.

11. FDA (2012) Scientific considerations in demonstrating biosimilarity to a reference product. The United States Food and Drug Administration, Silver Spring, Maryland, USA.

12. Wang H, Zhang Y, Shao J, Chow SC (2000) *In vitro* bioequivalence testing. In Encyclopedia of Biopharmaceutical Statistics, Ed. Chow, S.C., 2nd edition, Marcel Dekker, Inc., New York, New York.

13. Chow SC, Shao J, Wang H (2003) In vitro bioequivalence testing. Stat Med 22: 55-68.

14. Hyslop T, Hsuan F, Holder DJ (2000) A small sample confidence interval approach to assess individual bioequivalence. Stat Med 19: 2885-2897.

15. HC (2006) The Safety and Effectiveness of Generic Drugs. Ottawa: Health Canada, Canada.

16. EMA (2010) Concept paper on similar biological medicinal products containing recombinant follicle stimulation hormone. A/CHMP/BMWP/94899/2010 London, United Kingdom.

17. Haidar SH, Makhlouf F, Schuirmann DJ, Hyslop T, Davit B, et al. (2008) Evaluation of a scaling approach for the bioequivalence of highly variable drugs. AAPS J 10: 450-454.

18. Tothfalusi L, Endrenyi L, Arieta AG (2009) Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. Clin Pharmacokinet 48: 725-743.

19. Haidar SH, Davit B, Chen ML, Conner D, Lee L, et al. (2008) Bioequivalence approaches for highly variable drugs and drug products. Pharm Res 25: 237-241.

20. Davit BM, Conner DP, Fabian-Fritsch B, Haidar SH, Jiang X, et al. (2008) Highly variable drugs: observations from bioequivalence data submitted to the FDA for new generic drug applications. AAPS J 10: 148-156.