Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Chow and Liu, J Biom Biostat 2019, 10:5
Research Article
Open Access

Practical Statistical Issues in Evaluation of Average Bioequivalence

Shein-Chung Chow[1]* and Mo Liu[2]

[1]Duke University School of Medicine, Durham, North Carolina, USA
[2]Beijing Friendship Hospital, Capital Medical University, Beijing, China

Abstract

For approval of generic drug products, the United States Food and Drug Administration (FDA) has published several regulatory guidance to assist the sponsors in preparing documents, which provide substantial evidence for demonstration of bioequivalence between a generic (test) product and its innovative (reference) product (e.g., FDA, 1992, 2003) through the conduct of bioavailability and bioequivalence studies. Bioavailability and bioequivalence studies are usually conducted under crossover designs such as a standard 2x2 crossover design or a higher-order crossover design. Under a crossover design, bioequivalence is commonly evaluated using a two one-sided tests procedure (each at a 5% level of significance) or a 90% confidence interval approach. Bioequivalence is claimed if the constructed 90% confidence interval for the geometric mean ratio falls entirely within the bioequivalence limit of (80%, 125%). Statistical methods for bioequivalence evaluation are well established and widely accepted in the pharmaceutical industry since the publication of the FDA guidance in 2003. However, several practical issues are commonly encountered during the review of regulatory submissions of generic drug products. In this article, these issues are described. In addition, some recommendations for possible clarification and/or resolutions are made.

Keywords: Two one-sided tests procedure (TOST); Confidence interval approach; Outlier detection; Missing values; Binary response

Introduction

The current concept for the assessment of bioequivalence (BE) is based on the Fundamental Bioequivalence Assumption that when two formulations of the same drug product or two drug products (e.g., an innovative or brand name drug and its generic copy) are equivalent in the rate and the extent of drug absorption, it is assumed that they will reach the same therapeutic effect or that they are therapeutically equivalent [1]. Pharmacokinetic (PK) responses, such as area under the plasma or blood concentration–time curve (AUC) and maximum concentration (Cmax), are usually considered to assess the rate and the extent of drug absorption. The United States Food and Drug Administration (FDA) requires that evidence of BE in average bioavailabilities in terms of some primary PK responses such as AUC and Cmax between the two formulations of the same drug product or the two drug products be provided (FDA, 1992, 2003). This type of BE is referred to as average bioequivalence (ABE).

In practice, bioavailability and bioequivalence studies are usually conducted under a crossover design such as a standard 2x2 crossover design or a higher-order crossover design. Under a crossover design, bioequivalence is commonly evaluated using a two one-sided tests (TOST) procedure (each test is performed at a 5% level of significance) or a 90% confidence interval (CI) approach. Bioequivalence is claimed if the constructed 90% confidence interval for the geometric mean ratio (GMR) of the primary PK parameters (e.g., AUC and Cmax) of the two drug products (i.e., the test product and the reference product) falls entirely within the bioequivalence limit of (80%, 125%). Statistical methods for bioequivalence evaluation are well established and widely accepted in the pharmaceutical industry since the publication of the FDA guidance in 2003.

However, several practical issues are commonly encountered during the review of regulatory submissions of generic drug products. These practical issues include, but are not limited to, (i) the mixed use of the concepts of interval hypotheses and the confidence interval approach for bioequivalence evaluation, (ii) mis-interpretation of the power of TOST and the probability of claiming bioequivalence based on the confidence interval approach, (iii) sample size requirement under higher-order crossover designs, (iv) inconsistency between test statistics under a 2x2 crossover design and a 2x2m replicated crossover design, (v) justification for log-transformation, (vi) statistical methods for detection of outlying subjects, (vii) missing values at later dosing periods, (viii) the relationship between bioequivalence criteria and variability, (ix) bioequivalence assessment based on binary responses, and (x) post-approval equivalence in manufacturing process. In this article, these issues are described. In addition, some recommendations for possible clarification and/or resolutions are made whenever possible.

TOST versus CI Approach

As indicated in the 2003 FDA guidance on bioequivalence, the FDA recommends the following interval hypotheses for testing bioequivalence between a test product and a reference product in terms of pharmacokinetics (PK) responses such as area under the blood concentration time curve (AUC) or maximum concentration (Cmax) based on log-transformed PK data:

$$H_0: \theta \le \delta_L \text{ or } \theta \ge \delta_U \theta_U \text{ vs. } H_a: \delta_L < \theta < \delta_U, \quad (1)$$

Where $\theta=\mu_T/\mu_R$, μ_T and μ_R are mean PK response for the test product and the reference product, respectively, δ_L=0.8 and $\delta_U = \frac{1}{\delta_L} = 1.25$.

Under the interval hypotheses (1), Schuirmann [2] suggested a two one-sided tests (TOST) procedure be used. In many cases, under certain conditions, TOST (each side is tested at the α level

***Corresponding author:** Shein-Chung Chow, Duke University School of Medicine, Durham, North Carolina, USA, Tel: +1 919-681-1400; E-mail: sheinchung.chow@duke.edu

Received November 20, 2019; **Accepted** November 25, 2019; **Published** December 02, 2019

Citation: Chow SC, Liu M (2019) Practical Statistical Issues in Evaluation of Average Bioequivalence. J Biom Biostat 10: 435.

of significance) is operationally equivalent to the (1-2α) × 100% confidence interval (CI) approach for the evaluation of bioequivalence. In other words, we may claim that the test product is bioequivalent to the reference product if the constructed (1-2α) × 100% CI falls entirely within the bioequivalence limits (δ_L, δ_U). As a result, in practice, for convenience sake, the (1-2α) × 100% CI is often used for evaluation of bioequivalence. This CI approach, however, has created the following confusion that the use of 90% CI (if we choose α=5%) may have inflated the overall type I error rate from 5% to 10%. This confusion has been challenged by many authors for adopting different standards for regulatory approval of generic drug products (i.e., α=10%) and new drugs (i.e., α=5%). To address this issue, Chow and Shao [3] showed that Schuirmann's TOST is a size-α test. In addition, it should be noted that (i) the concept of interval hypotheses is different from that of the CI approach, (ii) TOST is the official test procedure recommended by the agency (see, e.g., FDA 1992, 2003), (iii) TOST, each side is tested at the α level of significance, is not generally equivalent to the (1 - 2α) × 100% CI approach for evaluation of bioequivalence. For example, for bioequivalence studies with binary responses, TOST, each side is tested at α level of significance, is not equivalent to the (1 - 2α) × 100% CI approach for evaluation of bioequivalence.

Thus, for evaluation of bioequivalence between a test product and a reference product, it is then suggested that which method is recommended by the agency should be clarified in the future revision of the guidance.

Power and the Probability of Claiming Bioequivalence

Based on the discussion above, it is clear that the concept of interval hypotheses testing (i.e., TOST) is very different from that of the 90% CI approach. Under a standard 2x2 crossover design and the assumption of log-normality, assuming that $-\delta_L=\delta_{U=}\delta$, and denote by $\varepsilon=\mu_T-\mu_R$, the following interval hypotheses is often tested for equivalence

$$H_0: |\varepsilon| \geq \delta \text{ vs. } H_a: |\varepsilon| < \delta, \tag{2}$$

Where δ is bioequivalence limit (margin). The test drug is then concluded to be equivalent to the reference product in average if the null hypothesis is rejected at significance level α. Let $\overline{x}$ and $\overline{x}_2$ be the sample mean for the test product and the reference product, respectively. Also let n_1 and n_2 be the sample size for the test product and the reference product, respectively. When σ^2 is known, the null hypothesis H_0 of (2) is rejected at the α level of significance if

$$\frac{\overline{x}_1-\overline{x}_2-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -z_\alpha \text{ and } \frac{\overline{x}_1-\overline{x}_2-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > z_\alpha. \tag{3}$$

Under the alternative hypothesis that $|\varepsilon| < \delta$, the power of this test is given by

$$\begin{aligned}&\Phi\left(\frac{\delta-\varepsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-z_\alpha\right)+\Phi\left(\frac{\delta+\varepsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-z_\alpha\right)-1\\ &\approx 2\Phi\left(\frac{\delta-|\varepsilon|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-z_\alpha\right)-1.\end{aligned} \tag{4}$$

Based on (4), an appropriate sample size is often chosen for achieving a desired power for establishment of bioequivalence (i.e., the probability of correctly claiming bioequivalence between a test product and a reference product when in fact the test product is bioequivalent to the reference product).

On the other hand, if the 90% CI approach is used, we may choose an appropriate sample size in order to have the following desired probability of claiming bioequivalence

$$p_{90\%CI} = P\{90\%CI \sqsubset (\delta_L, \delta_U)\} \tag{5}$$

It should be noted that $p_{90\%CI}$ given in (5) is not the same as the power of the TOST given in (4). In practice, however, power analysis for sample size calculation is usually performed based on interval hypotheses (2) but the bioequivalence assessment is often done based on the 90% CI approach. Thus, it is suggested that (i) the agency should clarify what is the official method (either TOST or the 90% CI approach) for bioequivalence evaluation and (ii) sample size determination should be made under the official method for consistency.

Sample size

Under a standard 2x2 crossover design and interval hypotheses (2), the power function of TOST when σ^2 is known is given in (4). Thus, the sample size needed to achieve power 1− β can be obtained by solving the following equation

$$\frac{\delta-|\varepsilon|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-z_\alpha = z_{\beta/2}.$$

This leads to

$$n_1 = kn_2 \text{ and } n_2 = \frac{(z_\alpha+z_{\beta/2})^2\sigma^2(1+1/k)}{(\delta-|\varepsilon|)^2} \tag{6}$$

When σ^2 is unknown, it can be replaced by s^2 in (3). The null hypothesis H_0 is rejected at the α level of significance if

$$\frac{\overline{x}_1-\overline{x}_2-\delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -t_{\alpha,n_1+n_2-2} \text{ and } \frac{\overline{x}_1-\overline{x}_2-\delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha,n_1+n_2-2}. \tag{7}$$

Under the alternative hypothesis that $|\varepsilon| < \delta$, the power of this TOST test is given by

$$\begin{aligned}&1-T_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta-\varepsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right)\\ &-T_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta+\varepsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right).\end{aligned}$$

Hence, with $n_1=\kappa n_2$, the sample size n_2 needed to achieve power 1 − β can be obtained by setting the power to 1 − β. Since the power is larger than,

$$1-2T_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta-|\varepsilon|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right),$$

a conservative approximation to the sample size n_2 can be obtained by solving,

$$T_{(1+\kappa)n_2-2}\left(t_{\alpha,(1-\kappa)n_2-2}\left|\frac{\sqrt{n_2}(\delta-|\varepsilon|)}{\sigma\sqrt{1+1/\kappa}}\right.\right)=\frac{\beta}{2}. \tag{8}$$

Thus, under a standard 2x2 crossover design, sample size can be obtained either using (6) when σ^2 is known or solving equation (8) when σ^2 is unknown. In practice, however, power analysis for sample size calculation for bioequivalence studies is often performed under a 2 x 2 crossover design regardless a higher-order crossover design such as 4 x 2 Balaam's crossover design, 2 x 3 dual crossover design, or 2 x 4 crossover design (or duplicated 2 x 2 crossover design) is used. Table 1 summarizes four commonly used higher-order crossover designs in bioequivalence evaluation.

k	pxq Crossover Design	Description
1	4 x 2	Balaam's design
2	2 x 3	Two-sequence dual design
3	2 x 4	Four-period design with two sequences
4	4 x 4	Four-period design with four sequences

Table 1: Four commonly used crossover designs in bioequivalence studies.

For higher-order crossover designs comparing two formulations of the same drug products or two drug products, similar formulae can be derived [4]. Because the power curves of Schuirmann's two one-sided tests procedure are symmetric about zero, we present only the equations for the case where $q \geq 0$. Let n_i be the number of subjects in each sequence i, have the same value n, and F_ν denote the cumulative distribution function of the t distribution with ν degrees of freedom. Then the power function, $P_k(\theta)$ of Schuirmann's two one-sided tests at the a level of significance for design (k) is given by

$$P_k(\theta) = F_{\nu_k}([(\Delta-\theta)/(CV\sqrt{b_k/n})] - t(\alpha,\nu_k)) - F_{\nu_k}(t(\alpha,\nu_k) - [(\Delta+\theta)/(CV\sqrt{b_k/n})]) \quad for\ k = 1, 2, 3, 4 \tag{9}$$

Where ν_1=4n - 3, ν_2=4n - 4, ν_3=6n - 5, ν_4=12n - 5, b_1=2, b_2=3/4, b_3=11/20, and b_4=1/4.

Hence, the exact equation for determination of n required to achieve a $1 - b$ power at the a nominal level for each design (k) when q=0 is the following:

$$n \geq b_k[t(\alpha,\nu_k) + t(\beta/2,\nu_k)]^2[CV/\Delta]^2 \qquad for\ k = 1, 2, 3, 4 \tag{10}$$

And if $q > 0$ the approximate formula for n is

$$n \geq b_k[t(\alpha,\nu_k) + t(\beta,\nu_k)]^2[CV/(\Delta-\theta)]^2 \qquad for\ k = 1, 2, 3, 4 \tag{11}$$

For the multiplicative model, we consider the (0.8, 1.25) bioequivalence range of m_T/m_R, denoted by d, where m_T and m_R denote the median bioavailabilities of the test and reference formulations, and let 1n denote the natural logarithm. Similarly, the sample size n required to achieve a $1 - b$ power at the a nominal level for each corresponding design (k) after the logarithmic transformation is determined by the following equations:

$$n \geq b_k[t(\alpha,\nu_k) + t(\beta/2,\nu_k)]^2[CV_m/\ln(1.25)]^2 \qquad \text{if } \delta = 1$$

$$n \geq b_k[t(\alpha,\nu_k) + t(\beta,\nu_k)]^2[CV_m/(\ln(1.25) - \ln\delta)]^2 \qquad \text{if } 1 < \delta < 1.25$$

And

$$n \geq b_k[t(\alpha,\nu_k) + t(\beta,\nu_k)]^2[CV/(\Delta-\theta)]^2 \qquad for\ k = 1, 2, 3, 4 \tag{12}$$

In the above equations, b is the probability of a type II error concluding bioinequivalence when, in fact, the two formulations are bioequivalent, d, $CV_m = \sqrt{\exp(\sigma^2) - 1}$, the coefficient of variation in the multiplicative model, and s^2, the residual (within-subject) variance of the logarithmically transformed characteristics, can usually be obtained from previous studies. However, because the degrees of freedom are usually unknown, an easy way to find the sample size is to enumerate n.

Inconsistency between test statistics under a 2 x 2 crossover design and a 2 x 2 m replicated crossover design

A commonly asked question in the assessment of average bioequivalence is that there is an inconsistency between test statistics given in this book (second edition) and the one as described in the FDA draft guidance. It should be noted that test statistic for assessment of ABE given in this book was derived under a 2 x 2 crossover design and the test statistic as described in the FDA guidance was derived under a replicated 2 x 2 m crossover design. However, the test statistics derived under a replicated 2 x 2 m crossover design is reduced to the test statistics derived under a 2 x 2 crossover design given in our book if m=1. In addition, the current regulatory requirement for approval of generic drug is still average bioequivalence. As a result, the 2003 FDA guidance for general considerations recommends non-replicate 2 x 2 crossover design for bioequivalence studies of immediate-release and modified-release dosage forms (p. 7 of the guidance). It follows that the test statistics under the standard 2 x 2 crossover design should be used for evaluation of average bioequivalence.

To address the inconsistency, first we would like to point out that the assessment of ABE is usually done under a 2 x 2 crossover design under certain assumptions (e.g., $\sigma_{BT}=\sigma_{BR}=\sigma_S$, where σ_{BT} and σ_{BR} are between subject variability for the test product and the reference product, respectively, and $\sigma_{BT}=\sigma_{BR}=\sigma$, where σ_{WT} and σ_{WR} are within subject variability for the test product and the reference product, respectively.). For convenience's sake, we will refer to the statistical model under the 2 x 2 crossover design with these assumptions as the classical model. However, in practice, these assumptions may not hold. If there are replicates, we will be able to provide independent estimates for σ_{BT}, σ_{BR}, σ_{WT}, and σ_{WR}. In this case, the FDA suggests a mixed effects model be used. I will refer to the statistical model with the assumption that (i) σ_{BT} and σ_{BR} are not necessarily the same and (ii) σ_{WT} and σ_{WR} are not necessarily the same as the FDA's model. The difference between the classical model and the FDA's model is summarized below.

FDA's model – As an example, consider a 2x6 crossover design, i.e., (ABABAB, BABABA), the following mixed effect model is considered:

$$y_{ijkl} = \mu_k + \gamma_{ik} + s_{ik} + e_{ijkl},$$

where y_{ijkl} is the pharmacokinetics (PK) response from the jth (j=1,⋯,n) subject in the ith (i=1,2) sequence under the lth (l=1,2,3) replicate of treatment k (k=1:*test*,2: *reference*), μ_k is the k th formulation effect such that $\mu_1 - \mu_2 = \delta$, γ_{ik}, is the fixed effect of the ith sequence under treatment k, S_{ik} is the random effect of the ith subject under treatment k, (S_{i1}, S_{i2}), i=1,⋯,n are assumed to be independent and identically distributed (i.i.d.) as bivariate normal random variable with mean 0 and covariance matrix.

$$\begin{pmatrix} \sigma_{BT}^2, \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR}, \sigma_{BR}^2 \end{pmatrix}.$$

e_{ij1l}'s are assumed to be i.i.d normal random variables with mean 0 and variance, σ_{WT}^2 and e_{ij2l}'s are assumed to be i.i.d normal random variables with mean 0 and variance σ_{WR}^2.

Classical model under a 2x2 design – Classical model is essentially the same as FDA's model under the assumption that $S_{i1}=S_{i2}$, i=1, ⋯, n, which implies that $\sigma_{BT}=\sigma_{BR}$ and ρ=1. Consequently, the variability due to formulation-by-subject interaction

$$\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} = 0.$$

This may not be true under the FDA's model. As a result, under different models with different assumptions, test statistics for assessment of ABE could be different.

For example, under the 2x2 design, the unbiased estimates for δ and σ_{WT}^2, σ_{BR}^2 are given by $\hat{\delta} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n}(y_{ij11} - y_{ij21})$, which follows a normal distribution $\hat{\delta} \sim N(\delta, \frac{\sigma_{WT}^2, \sigma_{WR}^2}{2n})$. On the other hand, under

the FDA's model, we have $\hat{\delta} \sim N(\delta, \frac{\sigma_D^2 + \sigma_{WT}^2, \sigma_{WR}^2}{2n})$ Now, under a 2x*m* design (e.g., *m*=4,6), let $\bar{y}_{ijk\bullet} = \frac{1}{m}(y_{ijk1} + \cdots + y_{ijkm})$. Then the unbiased estimates for δ is given by $\hat{\delta} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n}(\bar{y}_{ij1\bullet} - \bar{y}_{ij2\bullet})$. Under the classical model, we have $\hat{\delta} \sim N(\delta, \frac{\sigma_{WT}^2, \sigma_{WR}^2 / m}{2n})$. On the other hand, under the FDA's model, $\hat{\delta} \sim N(\delta, \frac{\sigma_D^2 + (\sigma_{WT}^2, \sigma_{WR}^2)/m}{2n})$.

As discussed above, we have the following observations. First, the ABE is established based on $\hat{\delta}$. According to the above discussion, it is clear that based on the FDA's model, increasing the number of replicates does not decrease the variability due to the subject-by-formulation interaction. Especially in our simulation study, we choose ρ=0.75 and, σ_{BT}^2 and σ_{BR}^2 are not necessarily equal to each other. Therefore, $\sigma_D^2 \neq 0$, which prevent the further improvement of ABE. Second, for assessment of ABE under a 2x2 crossover design, the methods described in Chow and Liu (2008), which assumes $\sigma_D^2 = 0$, has been widely used and accepted in practice. The reason of such observations comes can be explained as follows. Under the 2x2 crossover design, ρ, σ_{BT}^2 and σ_{BR}^2 are confounded with σ_{WR}^2 thus cannot be separated. As a result, the assumption that $\sigma_D^2 = 0$ is necessarily made for a valid statistical assessment of ABE. However, as indicated in the 2003 FDA guidance, the assumption of $\sigma_D^2 = 0$ may not hold. In addition, replicated crossover designs provide independent estimates of ρ and all variance components.

Log-transformation

Both the 1992 and 2003 FDA guidance provide the pharmacokinetic rationale as the clinical rationale for use of logarithmic transformation of exposure measures. In addition, both guidance do not encourage the sponsors to test for normality of error distribution after log-transformation, nor to use normality of error distribution as a reason for carrying out the statistical analysis on the original scale.

With respect to the pharmacokinetic rationale, deterministic multiplicative pharmacokinetic models are used to justify the routine use of logarithmic transformation for AUC and C_{max}. However, the deterministic PK models are theoretical derivations of AUC and C_{max} for a single object. Both guidance suggest that AUC be calculated from the observed plasma–blood concentration–time curve using the trapezoidal rule, and that C_{max} be obtained directly from the curve, without interpolation. It is not known whether the observed AUC and C_{max} can provide good approximations to those under the theoretical models if the models are incorrect.

On the other hand, assessment of bioequivalence requires statistical models that take into consideration the design features and the random components caused by inter-subject and intrasubject variations. The validity of the statistical inferences, such as confidence intervals and hypotheses testing, relies on the normality assumption of the random components in the statistical models. Consequently, determination of a scale of the exposure responses for assessment of bioequivalence also should be based solely on whether the random components in the statistical models satisfy the normality assumption.

The AUC and C_{max} are calculated from the observed plasma–blood concentrations. Therefore, the distributions of the observed AUC and C_{max} depend on the distributions of plasma–blood concentrations. Liu and Weng [5] showed that the log-transformed AUC and C_{max} do not generally follow a normal distribution, even when either the plasma concentrations or log-plasma concentrations are normally distributed. This argues against the routine use of the logarithmic transformation in assessment of bioequivalence. Moreover, Patel [6] also pointed out that performing a routine log-transformation of data and then applying normal theory-based methods is not a scientific approach. In addition, the sample size of a typical BE study is general too small to allow an adequate large-sample normal approximation.

Because current statistical methods for evaluation of bioequivalence are based on the normality assumption on the inter-subject and intrasubject variabilities, the examination of the normal probability plots for the studentized inter-subject and intrasubject residuals should always be carried out for the scale intended to be used in the analysis. In addition, formal statistical tests for normality of the inter-subject and intrasubject variabilities can also be carried out through Shapiro-Wilk's method. Contrary to the misconception of many people, Shapiro-Wilk's method is an exact method for small samples, such as bioequivalence studies. It is then scientifically imperative that tests for normality be routinely performed for the sale used in analysis, such as log-scale, is suggested in the guidance. If normality cannot be satisfied by both original-scale and log-scale, nonparametric methods should be employed.

Other issues concerning the routine use of the logarithmic transformation of exposure responses are the equivalence limits and presentation of the results on the original scale. The guidance recommends that the bioequivalence limits of (80%, 125%) on the original scale for assessment of average bioequivalence be used. On the log-scale, they are (log(0.8), log(1.25))=(-0.2231, 0.2231), where log denotes the natural logarithm. This set of limits is symmetrical about zero on the log-scale but it is not symmetrical about on the original scale. It should be noted that the rejection region of Schuirmann's two one-sided tests procedure associated with the new limits of (80%, 125%) is larger than that with the limits of (80%, 120%). As a result, a 90% confidence interval of (82%, 122%), for the ratio of averages of AUC between the test and reference formulations, will pass the bioequivalence test by the new limits, but not by the old limits. The new bioequivalence limits are 12.5% wider and 25% more liberal in the upper limit than the old limits. A new, wider upper bioequivalence limit may have an influence on the safety of the test formulation, which should be carefully examined if the new bioequivalence limits are adopted.

The FDA guidance requires that the results of analyses be presented on the log-scale as well as on the original scale, which can be obtained by taking the inverse transformation. Because the logarithmic transformation is not linear, the inverse transformation of the results to the original scale is not straightforward [7]. For example, the point estimator of the ratio of averages on the original scale obtained from the antilog of the estimator of difference in averages on the log-scale is biased and is always overestimated. Furthermore, the antilog of the standard deviation of the difference in averages on the log-scale is not the standard deviation for the point estimator of the ratio of the averages on the original scale. Further research is needed for the presentation of the results on the original scale, especially the estimation of variability after the analyses are performed on the log-scale.

Current regulation does not encourage the verification of the assumption of log-normality for the primary PK parameters such as AUC and Cmax [8]. The requirement of log-transformation needs to be scientifically or statistically justifiable. Also, "What if the distribution is still skewed after log-transformation?" and "Can non-parametric

method be used for bioequivalence assessment?" are questions of particular interest to the sponsors which need to be addressed.

Outlier detection

The 1992 FDA guidance provides a pharmacokinetic definition of subject outliers and provides possible causes for their occurrence. The FDA guidance suggests that Lund's method [9] be used for outlier detection. Although Lund's method is useful in a linear regression setting that requires statistical independence of all PK responses, this method may not be appropriate for a crossover design in which the PK responses from the same subject are correlated. Although Lund's method may be applied to the difference of the PK responses between the test and the reference formulations from the same subject in a standard two-sequence, two-period crossover design, it does not account for the feature of the study design. Moreover, it does not eliminate other nuisance effects; hence, it cannot be applied to other crossover designs.

As an alternative, Chow and Tse [10] first proposed two formal statistical test procedures for detection of a subject outlier in any crossover designs for assessment of bioequivalence. Their methods are the extension of Cook's likelihood distance [11]. Their methods are valid for large samples. For small samples, Liu and Weng proposed the use of Hotelling T^2 for detection of multiple subject outliers. Wang and Chow [12,13] proposed a procedure for detection of outliers under a mean-shift model. Although Liu and Weng's method is an exact method for small samples, it requires some special tables for critical values. Frequently, a subject may be considered an outlier based on the difference between the test and reference formulations. This subject, however, may not be considered an outlier based on the ratio. The reason for this conflict is that the design structure, statistical model, and scale for analysis are not taken into account for outlier detection. As a result, it is recommended that the detection of subject outliers should be carried out with a scale (original scale or the log-scale) intended for analysis under an appropriate statistical model for the design employed. In summary, Lund's method cannot take all of these factors into account for detecting subject outliers. Chow and Tse's method, Liu and Weng's procedures, and Wang and Chow test can accommodate the design, scale, and statistical models for detection of subject outliers. Ramsay and Elkum [14] conducted a simulation study to compare the above four methods.

The 2003 FDA guidance suggest that product failure and subject-by-formulation interaction are the two causes of outliers in a BE study. In addition, it discourages any deletion of outliers. Although statistical procedures for detection of outliers are available, these methods are derived from the model for assessment of average bioequivalence. Consequently, they are inadequate for identification of outliers in assessment of either PBE or ABE. More research on this topic is urgently needed.

In the current guidance, Lund's method was recommended for outlier detection. Lund's method is a valid method under a parallel-group design. It is not a valid statistical method for outlier detection in bioequivalence trials.

Missing data

In bioequivalence trials comparing a test product with a reference product, the dataset is often incomplete for various reasons (protocol violations, failure of assay methods, missed visits, etc.) if there are more than two dosing periods. For example, for bioequivalence studies using a two-sequence, three-period crossover design or a two-sequence, four-period crossover design, subjects are likely to drop out at the third period because they are required to return for tests more often than a standard two-sequence, two-period crossover design. Also, due to cost or other administrative reasons, sometimes not all of the subjects receive treatments beyond the second dosing period. In this case, one may not apply directly standard statistical methods for a crossover design to an incomplete or unbalanced dataset. Current regulation does not address the issue of missing data in depth.

A simple and naive way to analyze an incomplete dataset from a two-sequence three-period crossover design is to exclude the data from subjects who do not receive all three treatments so that one can treat the dataset as if it is from a two-sequence, three-period crossover design with smaller sample sizes. This, however, may result in a substantial loss in efficiency when the dropout rate is appreciable. Alternatively, it is suggested that subjects with missing data be replaced for achieving the desired power for establishment of bioequivalence. In this case, the intended bioequivalence study become an add-on bioequivalence study. It is a concern that the subjects for replacement of subjects with missing data may come from a similar but different target population. Statistical methods for bioequivalence evaluation of an add-on bioequivalence study are not well established. More research is needed.

Under a two-sequence, three-period crossover design, for inference on the treatment and carry-over effects, Chow and Shao [15] proposed a method based on differences of the observations that eliminates the random subject effects and thus does not require any distributional condition on the random subject effects. When no data is missing, Chow and Shao's method provides the same results as the ordinary least squares method. When there are missing data, Chow and Shao's method still provides exact confidence intervals for the treatment and carry-over effects, as long as the dropout is independent of the measurement errors.

Bioequivalence Criteria and Variability

Chow [16] studied the relationship between bioequivalence limit and variability (or the impact of variability on the bioequivalence limit under a parallel design for assessment of biosimilarity between a proposed biosimilar product and a reference product. The idea can be similarly carried out under a crossover design. Let's denote independent samples of T_i and R_j be the observations of T and R with $i=1,\cdots,n_T$ and $j=1,\cdots,n_R$. Without loss of generality, assume T_i and R_j are independent samples from $N(\mu_T, V_T)$ and $N(\mu_R, V_R)$ respectively. Then the 100(1-2α)% confidence interval based on the parallel design for $\mu_T - \mu_R$ can be expressed as

$$\left[\left(\bar{T}-\bar{R}\right)-Z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}}\,,\left(\bar{T}-\bar{R}\right)+Z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}}\right]$$

where $\bar{T}$ and $\bar{R}$ are the unbiased estimators of μ_T and μ_R, and $Z_{1-\alpha}$ is the (1-α) percentile of standard normal distribution. The ABE of the test product and reference product will be concluded at significance level of α if the above confidence interval lies entirely within (δ_L, δ_U). Thus, the probability of concluding ABE can be expressed as

$$P(\delta_L \le \left(\bar{T}-\bar{R}\right)-z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}} \text{ and } \left(\bar{T}-\bar{R}\right)+z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}} \le \delta_U)$$

$$= P(\delta_L + z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}} \le \left(\bar{T}-\bar{R}\right) \le \delta_U - z_{1-\alpha}\sqrt{\frac{V_T}{n_T}+\frac{V_R}{n_R}}.$$

In particular, if $\delta=\delta_U=-\delta_L$), $n_T=an_R$, $V_T=bV_R$, and denote

$C = \sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right)} * |\mu_R|$, the above equation can be expressed as

$$P\left\{-\left[\delta - z_{1-\alpha}\sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right) * V_R}\right] \le (\bar{T} - \bar{R}) \le \delta - z_{1-\alpha}\sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right) * V_R}\right\}$$

$$= P\left\{-\left[\delta - z_{1-\alpha}\sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right) * |\mu_R| * CV_R}\right] \le (\bar{T} - \bar{R}) \le \delta - z_{1-\alpha}\sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right) * |\mu_R| * CV_R}\right\}$$

$$= \Phi\left\{\frac{\delta - z_{1-\alpha} * C * CV_R - (\mu_T - \mu_R)}{C * CV_R}\right\} - \Phi\left\{\frac{-[\delta - z_{1-\alpha} * C * CV_R] - (\mu_T - \mu_R)}{C * CV_R}\right\}$$

Since the power of the test is defined as correctly concluding average bioequivalence when $\mu_T - \mu_R$ is 0 or close to 0 (within the bioequivalence limit), we can obtain the required bioequivalence limit (δ) to achieve desired power and type I error given the variability as measured by coefficient of variation (CV) by solving the equation of

$$\Phi\left\{\frac{\delta - z_{1-\alpha} * C * CV_R - (\mu_T - \mu_R)}{C * CV_R}\right\} - \Phi\left\{\frac{-[\delta - z_{1-\alpha} * C * CV_R] - (\mu_T - \mu_R)}{C * CV_R}\right\} = 1 - \beta \quad (13)$$

In (13) using the first order of Taylor expansion around $\mu_T - \mu_R$, we obtain

$$2 * \Phi\left\{\frac{\delta - z_{1-\alpha} * C * CV_R - (\mu_T - \mu_R)}{C * CV_R}\right\} - 1 + o(\mu_T - \mu_R) = 1 - \beta.$$

Solving the equation, we get

$$\delta = (Z_{1-\alpha} + Z_{1-\beta/2}) * C * CV_R + (\mu_T - \mu_R) + o(\mu_T - \mu_R)$$

Therefore, when $\mu_T - \mu_R$ is close 0, the closed form of relationship between δ and CV can be approximated as

$$\delta = (Z_{1-\alpha} + Z_{1-\beta/2}) * C * CV_R \quad (14)$$

When $\mu_T - \mu_R$ is largely deviated from 0 (outside of the bioequivalence limit for example), the probability of concluding bioequivalence in expression (13) will be mainly obtained by one side of the interval.

$$P\left\{-\left[\delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}\right] \le (\bar{T} - \bar{R}) \le \delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}\right\}$$

$$\approx \Phi\left(\frac{\delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}} - (\mu_T - \mu_R)}{\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}}\right) \text{ if } \mu_T - \mu_R \gg 0$$

or

$$P\left\{-\left[\delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}\right] \le (\bar{T} - \bar{R}) \le \delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}\right\}$$

$$\approx 1 - \Phi\left(\frac{-\left[\delta - z_{1-\alpha}\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}\right] - (\mu_T - \mu_R)}{\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}}}\right) \text{ if } \mu_T - \mu_R \ll 0$$

Under this situation, we get

$$\delta = (Z_{1-\alpha} + Z_{1-\beta})\sqrt{\frac{V_T}{n_T} + \frac{V_R}{n_R}} + |\mu_T - \mu_R| = (Z_{1-\alpha} + Z_{1-\beta})\sqrt{\left(\frac{1}{n_R} + \frac{b}{an_R}\right)} * |\mu_R| * CV_R + |\mu_T - \mu_R| \quad (15)$$

Expression (14) and (15) above provide us some closed form of relationship between bioequivalence limit (δ) and variability as measured by CV, approximately. The precise numerical solution and the approximation from those close forms are off slightly as CV goes beyond 1. But the difference decreases as the sample size increase. The relationship provided in closed form (14) and (15) motivates the use of scaled bioequivalence limits.

Another observation from expression (14) and (15) is that the required margin is linearly related with the coefficient of variation (CV) given the fixed choice of type I error, desired power, and sample size. In a traditional PK bioequivalence study where sample size is generally less (e.g. n is from 18-24), the margin of 20%± 20%*μ_R will provide sufficient power for $CV \le 30\%$, which is consistent with current generally accepted criteria. For highly variable drug products, the sample size per group could go up to fifty or hundreds per group. With the larger sample sizes, a fixed margin of ± 20%*μ_R can provide sufficient power for CV up to 40%. However, when CV is even larger than 40% which is commonly seen in biological products, scaled margin need to be applied to account for the large variability of the reference drug itself.

Apparently, bioequivalence limit depends upon the variability associated with the reference product. Table 2 summarizes current bioequivalence criteria for *in vitro* and *in vivo* bioequivalence testing recommended by the FDA. As it can be seen from Table 2, for *in vitro* bioequivalence testing, FDA recommends the use of (90%, 111%) as the bioequivalence limit, while for *in vivo* bioequivalence testing, bioequivalence limit of (80%, 125%) is suggested. In most cases, we expect to have a less variability (say less than 6%) in *in vitro* bioequivalence testing, while a moderate variability (say 20% to 30%) in *in vivo* bioequivalence testing. As the variability increases, a wider bioequivalence limit is expected or justified. For example, for highly variable drug products, FDA suggested a scaled average bioequivalence (SABE) criterion be used (Haidar et al., 2008). SABE is a criterion based on (80%, 125%) adjusted for the variability associated with the reference. Thus, in the revised guidance, it is suggested the relationship between the bioequivalence limit and variability associated with the reference product be established (Table 2).

Continuous Endpoint versus Binary Response

Current regulation for bioequivalence evaluation only focuses on continuous endpoints (i.e., AUC and Cmax). However, in many bioequivalence studies, binary responses are considered as the primary study endpoints (e.g., for locally acting drug products such as nasal spray drug products). In this case, standard methods (e.g., the 90% confidence interval approach) based on continuous endpoint cannot be applied directly for bioequivalence evaluation because TOST is not operationally equivalent to the 90% confidence interval approach for binary responses. Besides, it is not clear that (i) should log-transformation be performed before data analysis? (ii) whether the bioequivalence should be assessed based on difference in proportion between the two products or ratio of proportions or odds ratio of the two products, and (iii) what bioequivalence criteria should be used for difference in proportion, ratio of proportions, and/or odds ratio of proportions between the test product and the reference product.

However, there is little or no mentions of statistical approaches for bioequivalence evaluation based on study endpoint of binary response

Variability	BE Criterion	Application
<10%	(90%, 111%)	*In vitro* BE testing
10%-20%	(85%, 118%)[1] or SABE[2]	*In vivo* BE testing
20%-30%	(80%, 125%)	*In vivo* BE testing
>30%	(70%, 143%)[1] or SABE[2]	Highly variable drugs

Note: [1]suggested BE criterion. [2]Scaled average bioequivalence (SABE) criterion.

Table 2: Bioequivalence criteria and variability.

in the current guidance on bioequivalence. Thus, for bioequivalence studies with binary endpoints, bioequivalence criteria, study endpoints, and the corresponding statistical methods need to be developed under the study design for a valid, accurate, and reliable assessment of bioequivalence. Along the same line, criteria and statistical methods need to developed for bioequivalence studies with other types of study endpoints such as time-to-event data.

Post-Approval BE in Manufacturing Process

Since bioequivalence evaluation is usually done based on few small-scaled laboratory batches, it is important to evaluate the performance of the manufacturing process post-approval. For this purpose, the manufacturing process is necessarily validated according regulations as described in the current Good Manufacturing Practices (cGMP). The validation of a manufacturing process assures not only that the process does what it purports to do but also that the drug product will conform to United States Pharmacopeia and National Formulary (USP/NF) specifications. In practice, however, although the manufacturing process is validated, it is still a concern whether the manufacturing process will produce approved generic drug products which will possess good drug characteristics such as identity, strength, quality, purity, and stability as compared to the brand-name drugs. Thus, it may be a good idea to establish/document equivalence in manufacturing process between the test product and the reference product post-approval.

Conclusion

In the past several decades, criteria, statistical designs, and analysis methods for bioequivalence evaluation of generic drug products are well established (FDA, 1992, 2003). However, some practical issues are inevitably encountered during the review and approval of regulatory submissions. These practical issues may have an impact on the review and approval process. Thus, these issues need to be either clarified or resolved.

Practical issues that need clarification or resolution include (i) what is the official method (either TOST or the confidence interval approach) for bioequivalence evaluation? (ii) power analysis for sample size calculation should be performed based on the official method for bioequivalence evaluation under study design of the intended bioequivalence trial, (iii) inconsistency between the FDA's recommended method (based on mixed effects model) and the classical method for bioequivalence under a higher-order crossover design or a replicated crossover design, (iv) why log-transformation? (v) the validity of Lund's method for outlier detection under a crossover design is questionable, (vi) appropriate statistical methods for bioequivalence evaluation based on incomplete dataset should be developed when there are missing data, and (vii) the relationship between bioequivalence criteria and the variability of the reference product.

In addition, current regulatory guidance does not cover bioequivalence studies with binary response or time-to-event data as the primary study endpoint. Thus, bioequivalence criteria and corresponding statistical methods are necessarily developed under a valid study design for an accurate and reliable assessment of bioequivalence. Furthermore, bioequivalence evaluation is usually done based on few small-scaled laboratory batches. It is a concern whether the approved generic drug products will possess similar good drug characteristics such as identity, strength, quality, purity, and stability post-approval. Thus, it is suggested that a provision on the establishment of equivalence in manufacturing process between the test product and the reference product be documented.

References

1. Chow SC, Liu JP (2008) Design and Analysis of Bioavailability and Bioequivalence Studies. Third Edition, Taylor & Francis, New York.
2. Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioequivalence. Journal of Pharmacokinetics and Bio pharmaceutics 15: 657-680.
3. Chow SC, Shao J (2002) A note on statistical methods for assessing therapeutic equivalence. Controlled Clinical Trials 23: 515-520.
4. Chen KW, Chow SC, Li G (1997) A note on sample size determination for bioequivalence studies with higher-order crossover designs. Journal of Pharmacokinetics and Bio pharmaceutics 25: 753-765.
5. Liu JP, Weng CS (1994) Estimation of log-transformation in assessing bioequivalence. Communications in Statistics – Theory and Methods 23: 421-434.
6. Patel HI (1994) Dose-response in pharmacokinetics. Communications in Statistics – Theory and Methods 23: 451-465.
7. Liu JP, Weng CS (1992) Estimation of direct formulation effect under log-normal distribution in bioavailability/bioequivalence studies. Statistics in Medicine 11: 881-896.
8. FDA (2003) Guidance on Bioavailability and Bioequivalence Studies for Orally Administrated Drug Products – General Consideration. CDER pp: 1-27.
9. Lund RE (1975) Tables for an approximate test for outliers in linear models. Technometrics 17: 473-476.
10. Chow SC, Tse SK (1990) Outlier detection in bioavailability/bioequivalence studies. Statistics in Medicine 9: 549-558.
11. Cook RD, Weisberg S (1982) Residuals and Influence in Regression. Chapman and Hall New York and London
12. Wang W, Chow SC (2003) Examining outlying subjects and outlying records in bioequivalence trials. Journal of Biopharmaceutical Statistics 13: 43-56.
13. Liu JP, Weng CS (1991) Detection of outlying data in bioavailability/bioequivalence studies. Statistics in Medicine 10: 1375-1389.
14. Ramsay T, Elkum N (2005) A comparison of four different methods for outlier detection in bioequivalence studies. Journal of Biopharmaceutical Statistics 15: 43-52.
15. Chow SC, Shao J (1997) Statistical methods for two-sequence dual crossover designs with incomplete data. Statistics in Medicine 16: 1031-1039.
16. Chow SC (2014) Biosimilars: Design and Analysis of Follow-on Biologics. Chapman and Hall/CRC Press, New York, p: 444.