Bioequivalence: An overview of statistical concepts

S. Rani, A. Pargal*

B.V. Patel Pharmaceutical Education & Research Development (PERD) Centre, Thaltej, Ahmedabad - 380 054. India
*Hindustan Lever Ltd. Mumbai, India

Received: 3.11.2003
Revised: 12.1.2004
Accepted: 2.2.2004

Correspondence to:
S. Rani
E-mail:
shubha_rani@yahoo.com

ABSTRACT

Bioequivalence (BE) means the *absence* of a greater-than-allowable difference between the systemic bioavailability of a test product and that of a reference product. Studies to test the BE of drug products, and the statistical basis for their design, analysis and interpretation, have evolved over the last two decades. A crossover design is preferred over a parallel-group design as it segregates the inter-subject variation (which is not product-dependent) from the intra-subject variation (which is product-dependent). The value of testing two one-sided null hypotheses of non-equivalence at a significance level of 0.05, and the importance of estimating a 90% confidence interval of the ratio (test/reference) of mean AUC and C_{max} values, and of the difference between mean T_{max} values, are now recognized and form the current standards for BE. The number of subjects required for a BE study with the desired power (at least 0.80) and significance level (0.05), depends on the expected deviation of the test product from the reference product and the error variance associated with the bioavailability parameters (AUC, C_{max}, T_{max} etc.) of the drug substance. At present, according to the Indian regulatory authority, the number of subjects required to conduct a BE study is 12 which is inadequate for most drug substances by the current international standards and criteria. This review expounds the foregoing principles of BE testing.

KEY WORDS: Bioequivalence study designs, comparative bioavailability, statistical analysis

Introduction

Current international regulatory authorities[1-3] require that the final quality judgment of an oral dosage form be based on its *in vitro* dissolution profile and its *in vivo* bioavailability and/or bioequivalence evaluation. The latter is based on the premise that the concentration of the drug in the systemic circulation is at equilibrium with the concentration of the drug at the site of action and that the therapeutic effect of the drug moiety is a function of its pharmacodynamic-pharmacokinetic relationship.

Bioequivalence gained increasing attention during the last 40 years after it became evident that marketed products having the same amounts of the drug may exhibit marked differences in their therapeutic responses. Generally, these differences were well correlated to dissimilar drug plasma levels caused mainly by impaired absorption. Several examples illustrating this phenomenon are available in the literature, for example, researchers correlated the bioavailability and clinical effectiveness of USP thyroid products,[4-6] digoxin,[7,8] tolbutamide,[9,10] prednisone[11,12] and phenytoin.[13] Now a considerable body of evidence has accumulated indicating that drug response is better correlated with the plasma concentration or with the amount of drug in the body than with the dose administered. Consequently, on the basis of simple pharmacokinetic concepts and parameters, bioavailability and bioequivalence studies have been established as acceptable surrogates for expensive, complicated and lengthy clinical trials, and are used extensively worldwide to establish and ensure consistent quality and a reliable, therapeutically effective performance of marketed dosage forms.

Bioavailability reflects the extent of the systemic availability of the active therapeutic moiety and is generally assessed by measuring the 'area under the concentration time curve' (AUC), the peak plasma concentration (C_{max}) and the time to reach C_{max} (T_{max}). The extent of the systemic availability is determined by the extent of drug absorbed from the site of administration, and is influenced by the drug, the dosage form and the interaction of these with the complex environment of the absorption site. For a drug that obeys linear pharmacokinetics, the AUC and C_{max} values increase proportionately with the dose.[14] Consequently, if two formulations / dosage forms of

the same drug exhibit comparative AUC values, they are considered to have similar systemic availability. The bioavailability of an oral dosage form or a drug is generally compared with an intravenous solution (100% standard), to determine the absolute bioavailability.

In case of drugs which obey non-linear kinetics, the changes in AUC and C_{max} values are not proportional to the dose administered.[14] This is because either one or more of the processes which handle the drug i.e. absorption, distribution, metabolism and excretion are saturated i.e. their capacity has been exceeded within the therapeutic concentration range of the drug (substrate). In this situation, the plasma-concentration-time profile cannot be used as an indicator of absolute bioavailability. The latter has to be then assessed by measuring the extent of the drug and its metabolites excreted in the urine.

The comparative bioavailability assessment of two or more formulations of the same active ingredient to be administered by the same route is termed bioequivalence. Bioequivalence studies compare both the rate and extent of absorption of various multisource drug formulations with the innovator (reference) product, on the basis that if two formulations exhibit similar drug concentration-time profiles in the blood/plasma, they should exhibit similar therapeutic effects. For an unapproved generic dosage form to be marketed and accepted as therapeutically equieffective to the innovator product, it must establish bioequivalence with the innovator product, *in vivo*. Bioequivalence studies provide a quality control tool to monitor production and manufacturing changes.

Three situations have thus been defined in which bioequivalence studies are required (i) when the proposed marketed dosage form is different from that used in pivotal clinical trials, (ii) when significant changes are made in the manufacture of the marketed formulation and (iii) when a new generic formulation is tested against the innovator's marketed product.

The design, performance and evaluation of bioequivalence studies have received major attention from academia, the pharmaceutical industry and health authorities over the past two decades. In this article we would like to provide an overview and walk you through the evolution and synthesis of the key statistical analysis concepts, as applied to bioequivalence studies.

General concepts of study design

As recommended by the US FDA (1992),[1] in most bioequivalence trials, a "test" formulation is compared with the standard/innovator "reference" formulation, in a group of normal, healthy subjects (18-55 yr), each of whom receive both the treatments alternately, in a crossover fashion (two-period, two-treatment crossover design), with the two phases of treatment separated by a "washout period" of generally a week's duration, but may be longer (a minimum time equivalent to 5 half-lives) if the elimination half-life of the drug is very long. The treatment is assigned to each subject, randomly, but an equal number of subjects receive each treatment in each phase, as depicted in Table 1. Thus, in case of two treatments A and B, one group gets the treatment in the order AB, and the second group in the reverse order. This is done to avoid the occurrence of possible sequence or period effects.[15] A similar allocation is done in case of a three-treatment crossover design (three-period, three-treatment crossover design).

For several drugs a great inter-subject variability in clearance is observed. The intra-subject coefficient of variation (approximately 15%) is usually substantially smaller than that between subjects (approximately 30%), and therefore, crossover designs are generally recommended for bioequivalence studies.[16,17]

The primary advantage of the crossover design is that since the treatments are compared on the same subject, the inter-subject variability does not contribute to the error variability (discussed in detail in the next section). If the drug under investigation and/or its metabolites have an extremely long half-life, a parallel group design may be indicated. In a parallel group design, subjects are divided randomly into groups, each group receiving one treatment only. Thus, each subject receives only one treatment (Table 1). In a parallel design, although one does not have to worry about sequence, period or carry over effects, or dropouts during the study, the inter-subject variability being very high, the sensitivity of the test is considerably reduced, thus requiring a larger number of subjects compared to a crossover design, to attain the same sensitivity.

Inherent in both the crossover and parallel designs are the three fundamental statistical concepts of study design, namely randomization, replication and error control.[18,19] Randomization implies allocation of treatments to the subjects without selection bias. Consequently, randomization is essential to determine an unbiased estimate of the treatment effects. Replication implies that a treatment is applied to more than one experimental unit (subject) to obtain more reliable estimates than is possible from a single observation and hence provides a more precise measurement of treatment effects. The number of replicates (sample size) required will depend upon the degree of differences to be detected and inherent variability of

Table 1

Crossover design and parallel group design with 12 subjects

Crossover design with 12 subjects			*Parallel group design with 12 subjects*	
Vol. No.	*Period 1*	*Period 2*	*Treatment A**	*Treatment B**
1	B*	A*	1	2
2	A	B	3	4
3	B	A	7	5
4	B	A	9	6
5	A	B	10	8
6	A	B	11	12
7	B	A		
8	A	B		
9	B	A		
10	B	A		
11	A	B		
12	A	B		

* Treatment allocated to subjects

the data. Replication is used concomitantly with error control to reduce the experimental error or error variability.

Analysis of variance (ANOVA)

The various pharmacokinetic parameters (AUC, C_{max}) derived from the plasma concentration-time curve are subjected to ANOVA in which the variance is partitioned into components due to subjects, periods and treatments. The classical null hypothesis test is the hypothesis of equal means, $H_0: \mu_T=\mu_R$ (i.e. products are bioequivalent), where μ_T and μ_R represent the expected mean bioavailabilities of the test and reference formulations, respectively. The alternate hypothesis therefore is $H_1: \mu_T \neq \mu_R$ (i.e. products are bioinequivalent). For a crossover trial with n subjects and t treatments, the ANOVA takes the form as shown in Table 2.

Table 3 illustrates the dependence of error variability in ANOVA on the study design. Suppose two treatments T_1 and T_2 are to be compared using a group of subjects. There are two ways of designing the experiment: (i) Design 1 (parallel group design): divide the subjects into two groups and assign one treatment to each group and; (ii) Design 2 (crossover design): consider each subject as a block and then apply both the treatments to each block (subject) on two different occasions. In a parallel group design, only the variability due to the treatment is separated out, whereas in the crossover design, variability due to treatment, block (subject) and period are separated out from error variability. Consequently, the error sum of squares (SSE) is greater in the parallel design for a specific sample size (Error sum of squares in Design 1 (SSE_1) = Error sum of squares in Design 2 (SSE_2) + (subjects sum of squares) SSS_2 + (periods sum of squares) PSS_2. As degrees of freedom for SSE are the same in both the designs (for two-treatment case), the error mean sum of square for Design 1 (MSE_1) will be greater than the error mean sum of square for Design 2 (MSE_2),[18] i.e. error variability is greater in the parallel group design compared to the crossover design.

In ANOVA, the mean sum of squares due to a factor is compared with the mean sum of squares due to error (e.g. F= MST / MSE), and if these are comparable, no difference between the levels of a factor is concluded, otherwise a difference is concluded. Suppose there is a difference in treatments i.e. the treatment mean sum of squares is larger than the error mean sum of squares. Then the chances of the treatment mean sum of squares being larger than the error mean sum of squares are more in Design 2 compared to Design 1, since $MSE_2 < MSE_1$. Therefore, chances of showing a statistically significant difference (when actually there is a difference) are higher in Design 2 compared to Design 1. This is equivalent to saying that Design 2 is more powerful than Design 1. This reveals how the power of a test is influenced by the design of the experiment.

In ANOVA, the ratio of the formulations' mean sum of squares to the error mean sum of squares gives an F-statistic to test the null hypothesis $H_0: \mu_T=\mu_R$. This provides a test of whether the mean amount of drug absorbed from the test formulation is identical to the mean amount of drug absorbed from the reference. The test of this simple null hypothesis of identity is of little interest in bioequivalence studies, since the

Table 2

Analysis of variance (ANOVA) table for t-period, t-treatment crossover design

Sources of variation	*Degree of freedom (DF)*	*Sum of squares (SS)*	*Mean sum of squares (MS)*	*F Statistic*
Treatment	t[a]-1	SST	MST	MST/MSE
Subject	n[b]-1	SSS	MSS	MSS/MSE
Period	t-1	SSP	MSP	MSP/MSE
Error	(t-1)(n-2)	SSE	MSE	
Total	tn-1			

[a]t is number of treatments
[b]n is number of subjects

SST-Sum of squares due to treatments; SSS-Sum of squares due to subjects; SSP-Sum of squares due to period; SSE-Sum of squares due to error; MST-Mean sum of squares due to treatments; MSS-Mean sum of squares due to subjects; MSP-Mean sum of squares due to period; MSE-Mean sum of squares due to error

Table 3

A comparison of ANOVA for parallel group design and 2-treatment, 2-period crossover design with n subjects

Design 1					*Design 2*				
Sources of variation	*Sum of squares (SS)*	*Degree of freedom (DF)*	*Mean sum of squares (MS)*	*F Statistic*	*Sources of variation*	*Sum of squares (SS)*	*Degree of freedom (DF)*	*Mean sum of squares (MS)*	*F Statistic*
Between treatments	SST_1	1	MST_1	MST_1/MSE_1	**Between treatments**	SST_2	1	MST_2	MST_2/MSE_2
					Between blocks (subjects)	SSS_2	n^a-1	MSS_2	
					Between periods	SSP_2	1	MSP_2	
Error	SSE_1	n^a - 2	MSE_1		**Error**	SSE_2	n-2	MSE_2	
Total		n - 1			**Total**		2n-1		

[a]n is number of subjects

answer is always negative. This is because we cannot expect the mean amounts of drug absorbed from two different formulations or two different batches of the same formulation to be identical. They may be very nearly equal, but not identical. Also, if the trial is run under tightly controlled conditions (resulting in a small error mean sum of squares in the analysis) and if the number of subjects is large enough, no matter how small the difference between the formulations, it will be detected as significant.

Thus the detection of the difference (which as indicated above, will always exist) becomes simply a function of sample size, and since the probable magnitude of the difference is the critical factor, this gives rise to two anomalies:

1. A large difference between two formulations which is nevertheless not statistically significant if error variability is high and/or sample size not large enough.
2. A small difference, probably of no therapeutic importance whatsoever, that is shown to be statistically significant if error variability is minimal and/or sample size adequately large.

The first case suggests a lack of sensitivity in the analysis, and the second an excess of it. Consequently, any practice that increases the variability of the study (sloppy designs, assay variability and within formulation variability) would reduce the chances of finding a significant difference and hence improve the chances of concluding bioequivalence.

The FDA[20] therefore, recognized that a finding of no statistical significance in the first case was not necessarily evidence of bioequivalence and consequently asked for a retrospective examination of the power of the test of null hypothesis. Specifically, it was mandated that the test of equivalence have at least an 80% power of detecting a 20% difference between μ_T and μ_R (the 80/20 rule), where 20% was apparently arbitrarily chosen to represent the minimum difference that could be regarded as of therapeutic importance. However, there was no such criterion for the second case, and consequently if a very small difference was shown to be statistically significant, the conclusion that the difference was negligible and that the formulations could be considered bioequivalent was based solely on clinical judgment. It was therefore realized that the testing of the simple null hypothesis was inadequate and inappropriate and what was needed was not a test of whether the two formulations were identical but some degree of assurance that the mean amount of drug absorbed using the test formulation was close to the mean amount absorbed in case of the reference. The test hypothesis therefore, needed to be reformulated.

Another argument which favored an alternate approach to ANOVA for bioequivalence determination was the magnitude of the manufacturer's risk versus the consumer's risk. Manufacturer's risk is defined as the probability of rejecting a formulation which is in fact bioequivalent. In other words, the manufacturer's risk is the probability (α) of rejecting H_0 when H_0 is true (Type I error), and this risk was fixed at $\alpha=0.05$ by the FDA (1977).[20] Similarly, the consumer's risk is defined as the probability (β) of accepting a formulation which is bioinequivalent, i.e. accepting H_0 when H_0 is false (Type II error). By introducing the requirement that the power (1-β) of the test should be 80%, the FDA sought to restrict the consumer's risk "β" to 20%. This, however, was not a satisfactory solution for either the consumer or the regulatory agencies. In a regulatory environment, it makes sense that the regulatory authorities control the consumer's risk and let the pharmaceutical company decide how much manufacturer's risk they are willing to accept. As indicated, neither of these risks are formally identified or controlled when using the ANOVA F-test for treatments, even with the 80/20 rule.

Also the FDA guidelines[1] for bioavailability studies state that "Products whose rate and extent of absorption differ by 20% or less are generally bioequivalent". This implies that in the case of bioequivalence studies the interest is not in testing the null hypothesis of equality but in assessing the difference in two treatments. Bioequivalence is concluded if this difference is within 20% of the reference mean.

To overcome these issues, the bioequivalence problem was dealt with two different ways. One approach was the testing of the hypothesis and another was estimation i.e. confidence interval approach.

Hypothesis testing approach

In applying the statistical hypothesis testing, the hypothesis to be tested must be stated properly. The hypothesis testing paradigm requires that the hypothesis one desires to prove must be stated as the alternative hypothesis; one rejects the null hypothesis in favor of the alternative hypothesis if the evidence is sufficiently strong against the null hypothesis. As suggested by Hauck and Anderson[21] (1984), the objective of the bioequivalence trials might be incorporated into the following interval hypotheses:

$$H_0: \mu_T - \mu_R \le -\delta \quad \text{or} \quad H_0: \mu_T - \mu_R \ge +\delta$$

$\Leftrightarrow$ H_0: Products are bioinequivalent

$$Vs \qquad (1)$$

$$H_1: -\delta < \mu_T - \mu_R < +\delta$$

$\Leftrightarrow$ H_1: Products are bioequivalent,

where $\pm\delta[-0.20\mu_R(= -\delta), +0.20\mu_R(= +\delta)]$ is the allowable range for bioequivalence. They gave a test statistic for this hypothesis. It was also shown that if degrees of freedom are small, the true level of significance is always greater than the nominal level α i.e. consumer's risk is more than 5%. Therefore, this method could not get appreciation in the statistical methodology of bioequivalence testing.

In 1987, Schuirmann[22] proposed the two one-sided t-tests procedure for bioequivalence. It consists of decomposing the interval hypotheses H_0 into two sets of one-sided hypotheses and applying two separate t-tests as follows:

$$H_{01}: \mu_T - \mu_R \le -\delta$$
$$H_{11}: \mu_T - \mu_R > -\delta$$
$$\text{and} \qquad (2)$$
$$H_{02}: \mu_T - \mu_R \ge +\delta$$
$$H_{12}: \mu_T - \mu_R < +\delta$$

The null hypotheses (H_{01} and H_{02}) of bioinequivalence will be rejected i.e. test and reference formulations will be con-

cluded as bioequivalent or μ_T and μ_R are equivalent (for a balanced study) if

$$t_1 = \frac{\hat{\mu}_T - \hat{\mu}_R - (-\delta)}{s.\sqrt{2/n}} \geq t_{1-\alpha(v)}$$

or (3)

$$t_2 = \frac{\delta - (\hat{\mu}_T - \hat{\mu}_R)}{s.\sqrt{2/n}} \geq t_{1-\alpha(v)}$$

where s is the square root of the MSE from the crossover design ANOVA, n is number of subjects per period, $t_{1-\alpha(v)}$ is the critical value of t at $\alpha = 0.05$ in the upper tail of the Student's t-distribution with degrees of freedom ν where ν = the number of degrees of freedom associated with the MSE.

Surprisingly, now in FDA's guidelines, we do not have a statement regarding power.

As is clear from Equations 1 and 2, the hypothesis is now reformulated as:

H_0 : Products are bioinequivalent

H_1 : Products are bioequivalent

Hence,

α = Probability [Reject H_0 when H_0 is true]

= Probability [Conclude bioequivalence when products are bioinequivalent]

= Consumer's risk

β = Probability [Accept H_0 when H_0 is false]

= Probability [Conclude bioinequivalence when products are bioequivalent]

= Manufacturer's risk

Power = 1- β

The roles of α and β have been interchanged. Therefore, once α is fixed at 0.05 i.e. the consumer's risk is restricted to 5%, the agency leaves the pharmaceutical industry to determine the extent of the manufacturer's risk. Consequently, in the FDA's guidelines, there is no mention of power. However, the manufacturer's risk can be minimized by an adequate sample size.[23]

Confidence interval approach

Westlake[24] was the first to suggest the use of confidence intervals as a bioequivalence test to evaluate whether the mean amount of drug absorbed using the test formulation was close to the mean amount absorbed in case of the reference. The first question that arises then is what is a confidence interval (C.I.) and what does it signify? This is exemplified by the 95% [(1-α)%] C.I. for the difference in *AUC*'s of the test and reference formulations, which is given by the equation $(\hat{AUC}_T - \hat{AUC}_R) \pm SE_D t_{0.05(2),v}$, where $\hat{AUC}_T$ and $\hat{AUC}_R$ are the estimates of the mean *AUC* for the test and reference formulations, respectively; SE_D is the standard error of $(\hat{AUC}_T - \hat{AUC}_R)$, $t_{0.05(2),v}$ is the critical value of t at $\alpha = 0.05$ (for the two sided t - test), with *v* degrees of freedom. This C.I. implies that we are 95% confident that these two limits, referred to as the lower and upper confidence limits, will cover the true value of the difference. The quantity (1-α) which is equal to 0.95, since $\alpha = 0.05$ in this example, is referred to as the confidence level or confidence coefficient. A C.I. thus expresses the precision of the sample statistic of interest (in this example the statistic is the mean difference between AUCs), and as the precision increases (by a decrease in the variability or standard error) the C.I. becomes narrower.

Westlake[24] (1972), Metzler[25] (1974) and Kirkwood[26] (1981) proposed the construction of a (1-α) C.I. (k_1,k_2) for the difference in population mean parameters of bioavailability $(\mu_T-\mu_R)$ where α is typically 0.05. The decision rule was to accept bioequivalence with the inclusion of this C.I. into the bioequivalence range where the bioequivalence range is given by $[-0.20\mu_R(=-\delta), +0.20\mu_R(=+\delta)]$. This C.I. is symmetric about zero. This method, however, had some inaccuracies, and a refinement of this procedure was described in 1976,[27] but this rather unorthodox procedure had the effect of increasing the manufacturer's chances of demonstrating the equivalence of the test formulation.

In 1981, Westlake[28] attempted to modify his C.I. approach and decided to put the proposed rule for the acceptance of bioequivalence on the same logical basis as the FDA's well-established rule for the acceptance of efficacy in the clinical trial of a new drug entity. He argued that in efficacy trials one is generally attempting to demonstrate the efficacy of a new drug by testing against a placebo, and traditionally this is done at a significance level of $\alpha = 0.05$. This indicates that the regulatory agency is ensuring that if the drug is really the same as the placebo, then there is only a low probability, 0.05, that it will be approved. Also, since the drug would never be approved for being less efficacious than the placebo, the test and its critical region is one-sided. If a similar policy is adopted for bioequivalence, then if the difference in the means of the two formulations is δ, where ±δ is the allowable range for bioequivalence, then the probability that the (1-α) confidence interval falls within ±δ should be acceptably small (say 0.05). In other words, in a borderline case, the probability of accepting the new formulation as bioequivalent to the reference should be small. However, if a (1-α) C.I. is constructed, this probability is very low (less than α/2). This results in increasing the corresponding manufacturer's risk (β), i.e. the chances of declaring bioequivalent products as bioinequivalent increase (this is because statistically if one error (α) decreases, then the other (β) increases and vice versa). Consequently, to maintain the parallelism with the FDA's requirement for efficacy testing, he proposed that one should use the (1-2α) or 90% C.I. instead of the 95% confidence interval.

The inclusion of the 90% C.I. in the bioequivalence range turns out to be same as the rejection of both null hypotheses at nominal α (=0.05) level by two one-sided t-tests. Thus even though they both overlap in meaning, both the approaches are usually adopted.

Logarithmic transformation of bioequivalence parameters

Bioequivalence studies measure and compare statistically AUC, C_{max} and T_{max} of the formulations. In case of AUC and C_{max}, the regulatory authorities[1-3] recommend that they should

be logarithmically transformed before further statistical analysis. The use of log transformed values for AUC and C_{max} is recommended for several reasons:

(i) Clinical rationale: In a meeting in September 1991, the Generic Drugs Advisory Committee (GDAC)[29] concluded that the primary comparison of interest in a bioequivalence study was the ratio rather than the difference between average parameter data from the test and reference formulations. This is achieved statistically by using log transformation;

(ii) Pharmacokinetic rationale: In the crossover design, the usual assumption is that the observation is a function of additive effects due to subject, period and treatment. But pharmacokinetic equations are of multiplicative character, for example, AUC = Clearance^{-1}.f.dose, where $0 < f < 1$ denotes the fraction absorbed. The multiplicative term "clearance" can be regarded as a function of the subject. Consequently, Westlake contended that the subject effect is not additive if the data is analyzed on the original scale of measurement. Taking logarithms transforms this pharmacokinetic equation into an additive model equation: ln AUC = -ln Clearance + ln f + ln dose, where ln denotes the natural logarithm. Similar arguments are given for C_{max}. Log transformation of C_{max} data results in additive treatment of the Volume of distribution "V";

(iii) Statistical rationale: Many biological data correspond more closely to a log normal distribution. AUC and C_{max} tend to be skewed and their variances increase with the means. Log transformation makes the variances independent of the mean and the frequency distribution is made more symmetrical.

The third measure of bioavailability, $T_{max,}$ poses a somewhat different problem. What makes T_{max} different is the discreteness of its measurement, as well as the fact that it is measured with an error that will depend on the study's sampling times. While Hauck and Anderson contended that none of the statistical procedures available for use in the analysis of bioequivalence studies appear to be appropriate for T_{max}, Westlake has suggested that a C.I. on the difference of the T_{max} for standard and test formulations following an ANOVA on the untransformed data seems appropriate. However, the regulatory agencies[1-3] recommend that differences in T_{max} be evaluated by non-parametric tests (Wilcoxon signed rank test, Wilcoxon rank sum test) on the untransformed values (additive model).

Bioequivalence range

For a broad range of drugs, the US FDA has used a range of 80-120% for the 90% C.I. of the ratio of the product averages as the standard equivalence criterion, when the study data are analyzed on the original scale. This corresponds to a range of ± 20% for the relative difference between product averages. This argument did presume that one of the two products being compared is clearly identified as the reference and is therefore the denominator of the ratio. This is important, because with this (0.80-1.20) criterion, reversing which formulation is the reference and which is the test can change the conclusion as to whether the two formulations are equivalent.

When log-transformed data are used in the analysis of AUC and C_{max}, it is recommended to use 80-125% for the 90% C.I. of the ratio of the product averages as the standard equivalence criterion. Using a range of 80-125% for the 90% C.I. of the ratio of averages has an advantage over the 80-120% criterion, in that for the analysis of log-transformed data the probability of concluding bioequivalence is at a maximum if the ratio of averages is in fact 1, i.e. exact equality. For the analysis of log-transformed data with a criterion of 80-120%, the maximum probability of concluding equivalence occurs when the ratio of product averages equals approximately 0.98. Also, in contrast to the '0.80-1.20' range the '0.80-1.25' range is a multiplicative symmetric in the sense that $1.25 = (0.80)^{-1}$, and the conclusion will not depend on the choice of reference for the denominator. Consequently, regulatory agencies prefer the equivalence criterion of 80-125% for the 90% C.I. of the ratio of product averages. For marketing approval of a generic, however, this criterion does allow the test to be 25% greater than the reference.

The ninth draft of the CPMP guidance[3] on bioequivalence studies adopted an equivalence range of (0.70-1.43) for the C.I. of C_{max} ratio, where $(0.7)^{-1} = 1.43$. However, the final version of the CPMP guidance states that for the C.I. of the C_{max} ratio a wider acceptance range may be necessary than for the C.I. of the AUC ratio. This recommendation reflects the experience that single concentrations, especially extreme concentrations like C_{max}, generally have larger variation than integrated characteristics like AUC. It has also been suggested that the choice of the appropriate bioequivalence range should be made on clinical grounds; thus for a drug with a narrow therapeutic range, tighter limits may have to be considered, e.g. '0.9-1.11' for the C.I. of the AUC ratio and '0.8-1.25' for the C.I. of the C_{max} ratio. With regard to C_{max}, the US FDA however maintains the '0.8-1.25' range for the C.I. of the ratio (for log-transformed data), while the Canadian Health Protection Branch considers the range '0.8-1.25' for the point estimate.

Current regulatory criterion of bioequivalence

At present, the regulatory authorities[1-3] recommend analysis of the data after logarithmic transformation for C_{max} and AUC, and bioequivalence is concluded if either 90% C.I. for the ratios of the bioavailabilities of the two formulations lies in the bioequivalence range $[\delta_1=0.80,\ \delta_2=1.25]$, where 90% C.I. for the ratio is given by Exp $(\mu_T - \mu_R \pm s\sqrt{2/n}t_{0.05(1),v})$ where s is the square root of the MSE from the crossover design ANOVA, n is number of subjects per period, $t_{0.05(1),v}$ is the critical value of t at $\alpha = 0.05$, and ν the number of degrees of freedom associated with the MSE; or when the left and right sides of the Schuirmann's t - test are both statistically significant ($P<0.05$).

Sample size

According to the CPMP guidance,[3] "the number of subjects required is determined by the error variance associated with the primary characteristic to be studied (as estimated from a

pilot experiment, from previous studies, or from published data), by the significance level desired, by the expected deviation from the reference product and by the required power. It should be calculated by appropriate methods and should not be less than 12". The equations for the approximate sample size calculation for the two one-sided '*t*' tests, by Liu and Chow,[30] for the additive model (untransformed data) are given below (as power curves are symmetric, the formulae are given only for $\nabla \geq 0$):

$$n \geq [t_{\alpha,2n-2} + t_{\beta/2,2n-2}]^2 \, [CV/\delta]^2 \qquad \text{for } \nabla = 0,$$

$$n \geq [t_{\alpha,2n-2} + t_{\beta,2n-2}]^2 \, [CV/(\delta - \nabla)]^2 \qquad \text{for } \nabla > 0,$$

where n = number of subjects required per sequence; CV = co-efficient of variation

δ = the bioequivalence limit

$$\nabla = \frac{\mu_T - \mu_R}{\mu_R} \cdot 100$$

$$CV = \frac{\sqrt{MSE}}{\mu_R} \cdot 100$$

This equation can be interpreted as follows:

(i) ∇, the minimum detectable difference between population means: To detect a very small difference requires a larger sample size, than if a large difference is to be detected; (ii) CV, population variance: If the variability within samples is great, then a larger sample size is required to achieve a given ability of the test to detect differences between means; (iii) the significance level α : If the test is performed at α low a, then the critical value, $t_{\alpha,\nu}$, will be large and a large sample size will be required to achieve a given ability to detect a difference between means. That is, if a low probability is desired of committing a Type I error (i.e. falsely rejecting H_0), then large sample sizes are needed. Similarly, if we desire a low probability of committing a Type II error, a large sample size is required.

The sample size calculations for the multiplicative (log-transformed) model have been provided by Hauschke *et al* [31,32] who gave the values of n for various values of %CV and μ_T/μ_R. Those values reveal that even for 15% CV (very rare phenomena, usually %CVs are greater than 15%), the sample size needed to show bioequivalence for a ratio of 0.95 or 1.05 to attain 90% power is 16. However, according to the Indian regulatory requirements, a bioequivalence trial can be conducted in 12 subjects. On conducting the trial in 12 subjects, suppose CV is 20%, the power of the test is less than 80% even when $\mu_T/\mu_R=1$, i.e. manufacturer's risk is greater than 20% for similar products. It indicates that the sample size of 12 is not adequate to show bioequivalence even if products are bioequivalent. Unquestionably, it is the manufacturer's concern, but it is a waste of time and resources. Therefore, the Indian regulatory requirement for the minimum number of subjects should be modified.

Illustration

To illustrate, the statistical analyses of data from a two-period two-treatment crossover bioequivalence trial were carried out. The data and the results are presented in Tables 4, 5 and 6.

On using the formula given in the section "Current Regulatory Criterion of Bioequivalence", the 90% C.I. is given by (73.97, 111.40) for C_{max} and (88.61, 120.77) for $AUC_{0-\infty}$.

Conclusion

There is far-reaching international consensus on the design, performance and data analysis of bioequivalence studies. The consolidation and harmonization of the methodology has been reflected in the guidance of major health authorities such as the European CPMP, the Canadian HPB and the US FDA. There is, however, increasing awareness that some fundamentals of bioequivalence assessment need to be reconsidered, such as (i) the single bioequivalence criterion for all drugs, independent of the therapeutic window or the intra-subject variability of the drug, and (ii) the differentiation between

Table 4

C_{max} and $AUC_{0-\infty}$ values for reference and test formulations of a drug

	Reference		*Test*	
Vol no.	***C_{max}*** (ng/ml)	***$AUC_{0-\infty}$*** (ng.h/ml)	***C_{max}*** (ng/ml)	***$AUC_{0-\infty}$*** (ng.h/ml)
1	120.21	264.68	132.86	237.95
2	219.14	287.32	183.42	339.90
3	199.42	202.46	219.66	205.00
4	222.86	182.52	91.43	102.54
5	150.37	153.20	96.05	226.68
6	179.31	308.65	172.04	331.59
7	166.14	122.17	170.61	143.31
8	216.16	344.68	162.82	329.19
9	211.50	324.39	134.32	285.47
10	105.35	164.48	129.06	274.84
11	89.90	232.01	129.43	200.62
12	154.51	231.88	217.81	254.11

Table 5

ANOVA for C_{max} (after logarithmic transformation)

Sources of variation	***DF***	***SS***	***MS***	***F***
Treatment	1	0.0562	0.0562	0.75
Subject	11	1.1526	0.1048	1.41
Period	1	0.0121	0.0121	0.16
Error	10	0.7449	0.0745	
Total	23	1.9658		

Table 6

ANOVA for $AUC_{0-\infty}$ (after logarithmic transformation)

Sources of variation	***DF***	***SS***	***MS***	***F***
Treatment	1	0.0069	0.0069	0.16
Subject	11	2.1624	0.1966	4.62
Period	1	0.0011	0.0011	0.03
Error	10	0.4259	0.0426	
Total	23	2.5963		

switchability and prescribability when average bioequivalence is measured. While prescribability implies prescribing a generic formulation to a first-time user (this is adequately addressed by existing bioequivalence determinations), switchability is concerned with the equivalence issues of switching individual patients already on the reference formulation to a generic formulation.

In the future, there will be continued efforts toward achieving clinically and therapeutically relevant, cost-effective bioequivalence assessment procedures, on a case-by-case basis for each drug or classes of drugs. With the promise of expanding opportunities for generic formulations in the immediate future, there is also a need for all national regulatory agencies, especially in the emerging markets, to align themselves as well as constantly update their regulatory approval processes, in accordance with the current international thinking on the subject.

References

1. Food and Drug Administration (FDA), Guidance for Industry: Statistical approaches to establishing bioequivalence 2001.
2. Canadian Health Protection Branch (HPB), Drugs Directorate Policy: Cl standard for comparative bioavailability 1991.
3. Committee for Proprietary Medicinal Products (CPMP), Working Party on Efficacy of Medicinal Products. Note for guidance: Investigation of bioavailability and bioequivalence 1998.
4. Catz B, Ginsburg E, Salenger S. Clinically inactive Thyroid USP. A preliminary report. New Engl J Med 1962;266:136-7.
5. Braverman LE, Ingbar SH. Anomalous effects of certain preparations of desiccated thyroid on serum protein-based iodine. *New Engl J Med* 1964;270:439-42.
6. Ramos-Gabatin A, Jacobson JM, Young RL. *In vivo* comparison of levothyroxine preparations. J Am Med Assoc 1982;247:203-5.
7. Whitting B, Rodger J, Summer D. New formulation of digoxin. Lancet 1972;2:922-3.
8. Stewart MJ, Simpson E. New formulation of lanoxin: Expected plasma levels of digoxin. Lancet 1972;2:541.
9. Levy G. Effect of dosage form properties on therapeutic efficacy of tolbutamide tablets. Can Med Assoc J 1964;90:978-9.
10. Lu FC, Rice WB, Mainville CW. A comparative study of some brands of tolbutamide in Canada, part II, pharmaceutical aspects. Can Med Assoc J 1965;92:1166-9.
11. Campagna FA, Cureton G, Mirigian RA, Nelson E. Inactive prednisone tablets, USP XVI. Pharm Technol 1963;52:605-6.
12. Sullivan TJ, Sakmar E, Albert KS, Blair DC, Wagner JG. *In vitro* and *in vivo* availability of commercial prednisone tablets. J Pharm Sci 1975;64:1723-5.
13. Albert KS, Sakmar E, Hallmark MR, Weidler DJ, Wagner JG. Bioavailability of diphenylhydantoin. Clin Pharmacol Ther 1974;16:727-35.
14. Gibaldi M, Perrier D. Pharmacokinetics. 2nd Ed. New York: Marcel Dekker 1982.
15. Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 1965;21:467-80.
16. Zar JH. Biostatistical Analysis. 2nd Ed. New Jersey: Prentice-Hall, Inc., Englewood Cliffs 1984.
17. Armitage P. Statistical methods in medical research. New York: Wiley and Sons 1973.
18. Cochran WG, Cox GM. Experimental designs. 2nd Ed. New York: Wiley and Sons 1957.
19. Fisher RA. The design of experiments. 8th Ed. New York: Hafner Publishing Company 1966.
20. Food and drug administration (FDA), Division of Biopharmaceutics, Bioavailability protocol guidelines for ANDA and NDA Submission, 1977.
21. Hauck WW, Anderson S. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. J Pharmacokinet Biopharm 1984; 12:83-91.
22. Schuirmann DJ. A comparison of two one-sided tests procedure and the power approach for assessing the bioequivalence of average bioavailability. J Pharmacokinet Biopharm 1987;15:657-80.
23. Westlake WJ. Bioavailability and bioequivalence of pharmaceutical formulations. In: Peace KE, editor. Biopharmaceutical statistics for drug development, 1st Ed. New York: Marcel Dekker 1988. p. 329-52.
24. Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. *J Pharm Sci* 1972;61:1340-1.
25. Metzler CM. Bioavailability: A problem in bioequivalence. Biometrics 1974;30:309-17.
26. Kirkwood TBL. Bioequivalence testing - A need to rethink. Biometrics 1981;37:589-91.
27. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics 1976;32:741-4.
28. Westlake WJ. Response to Kirkwood TBL: Bioequivalence testing - A need to rethink. Biometrics 1981;37:589-94.
29. Meeting of generic drug advisory committee to the FDA. Washington 1991.
30. Liu JP, Chow SC. Sample size determination for the two one-sided tests procedure in bioequivalence. J Pharmacokinet Biopharm 1992;20:101-4.
31. Diletti E, Hauschke D, Steinijans VW. Sample size determination for bioequivalence assessment by means of confidence intervals. Int J Clin Pharmacol Ther Toxicol 1991;29:1-8.
32. Hauschke D, Steinijans VW, Diletti E, Burke M. Sample size determination for bioequivalence assessment using a multiplicative model. J Pharmacokinet Biopharm 1992;20:557-61.