REVIEW ARTICLE

Accepted 23 June 2001
Revised 2 June 2001
Received 13 December 2000
Indian J. Pharm. Sci., 2001, 63(3) 169-186

A Review of the Development of Biostatistical Design and Analysis Techniques for Assessing *In Vivo* Bioequivalence: Part Two

SCOTT PATTERSON
Clinical Pharmacology Statistics
SmithKline Beecham Pharmaceuticals
One Franklin Plaza, FP1670
P O Box 7929, Philadelphia, PA, 19101, USA
and Statistical Sciences, University College, London.

Bioequivalence studies are performed to demonstrate that two pharmaceutically equivalent products are equal in rate and extent of absorption *in vivo*. Following on from developments in the pharmaceutical industry and government mandates in the 1970's and 1980's and since the early 1990's, average bioequivalence has served as the international standard for demonstrating that two formulations of drug product will provide the same therapeutic benefit and safety profile when used in the marketplace. Population (PBE) and Individual (IBE) bioequivalence has been the subject of intense international debate since methods for their assessment were proposed in the late 1980's. Guidance has been proposed by the Food and Drug Administration of the United States Government for the implementation of these techniques in the pioneer and generic pharmaceutical industries. A previous article described the basis for bioequivalence and discussed the development of techniques for design and analysis in average bioequivalence assessment. This paper, the conclusion of this series, describes the implementation of average bioequivalence in the pharmaceutical industry, and discusses the development of population and individual bioequivalence.

The basis for bioequivalence was presented in a previous paper also describing the development of techniques for design and analysis in average bioequivalence. To summarise, bioequivalence studies evolved to meet the practical needs of the marketplace for less expensive formulations and to provide a practical means for improved or altered formulations to gain market access.

In these cases, rather than repeat clinical trials to establish the safety and efficacy of the proposed formulations, the pharmacokinetic (PK) characteristics of the plasma-concentration time curve are used to infer that two drug formulations will provide similar therapeutic benefit. The PK is expressed in terms of rate and extent of absorption as characterized by the maximum observed plasma concentration (Cmax) and the area under the concentration time curve (AUC). In turn, bioequivalence is expressed in terms of the 'similarity' of these two metrics between the two formulations. For approximately the past 10 years, international regulatory agencies in North America, Europe, and Asia have used a criterion of average bioequivalence (ABE) with regulatory limits of 20% (FDA Guidance, 1992; FDA Guidance, 2000b)[1,2] except in instances where dissolution profiles suffice (FDA Guidance, 2000a)[3].

This criterion focuses on the average PK metrics of the 2 formulations being studied. The framework for statistical inference is based on exact 90% confidence intervals for the difference in formulation means. This ABE criterion has been used for both instances of pre-market approval and post-market formulation changes described above. Average bioequivalence is assessed

For correspondence
email: scott d patterson@sbphrd.com

using a two-one sided hypothesis testing procedure (Schuirmann, 1987)[4] as follows:

$$H_{01}: \mu_T - \mu_R \leq \log_e(0.80) \quad (1)$$

or

$$H_{02}: \mu_T - \mu_R \leq \log_e(1.25) \quad (2)$$

Versus

$$H_A: \log_e(0.80) < \mu_T - \mu_R < \log_e(1.25) \quad (3)$$

where μ_i is the mean for T = test and R = reference formulations adjusted for period and sequence effects on the $\log_e$-transformed scale.

Plasma concentration-time profiles are obtained after each administration in a crossover design (Jones and Kenward, 1989; Senn, 1998)[5,6], and non-compartmental methods are used to derive summary measures AUC and Cmax. Log_e-transformed AUC and Cmax are analyzed separately by Analysis of Variance, and point estimates for the difference between the test and reference formulations and the mean squared residual error are used to derive the associated 90% confidence intervals. ABE is concluded if the 90% confidence interval for the ratio of the test to reference formulation is completely contained within the regulatory limits of 0.80 to 1.25 (transformed to the $\log_e$-scale) for both AUC and Cmax.

The two-one sided testing procedure (TOST) was easy to implement for nearly any study design and had the benefit of being easy to interpret in practice. In practical terms, the ninety percent confidence interval provides a plausible range of values within which the true difference in means can be expected to fall (Hauck and Anderson, 1986)[7]. Use of the procedure quickly became the norm in clinical pharmacology studies of pharmacokinetics. Lack of a meaningful pharmacokinetic difference when a drug product was administered with and without food or with and without a concomitantly administered medication could often be inferred based on the results of such studies under such an approach (Steinijans *et al.*, 1991)[8]. Similar techniques could also be used to infer that administration in patients with a concomitant disease state (e.g. hepatic impairment or renal insufficiency) or administration in patients not usually studied in typical pivotal efficacy studies (e.g. a paediatric population) would not result in clinically significant change in AUC or Cmax.

Extent of bioavailability as measured by AUC was judged, under this type of approach, to be a surrogate marker for efficacy in those drugs having been demonstrated to be acceptably efficacious to enter the marketplace. Comparable mean AUC following administration with or without food or a concomitantly administered medication (or in another population) were held to be indicative of efficacy in that condition. Decreases or increases would be used to adjust the dosing strategy for the drug product under study.

Rate of bioavailability as measured by Cmax was held, under this approach, to be a surrogate marker for safety for drugs in the marketplace. Comparable or decreased mean Cmax following administration with or without food or a concomitantly administered medication (or in another population) were held to be indicative of changes in safety profile. Increases in mean Cmax were potentially suggestive of a less acceptable safety profile for the drug under study.

The range of plausible values as expressed by a confidence interval was used to assess the degree of equivalence or comparability. Confidence level (Type 1 error) was termed 'consumer' or 'regulator' risk - i.e. the risk of the regulator in making an incorrect decision. Though often a prespecified goalpost interval was difficult or impossible to define prior to study initiation for some drug products, inhibiting the ability of study sponsors to adequately ensure adequate power to demonstrate equivalence, power was of less concern when assessing the results of such studies than the confidence level. This gave Regulators an easy standard under which to assess the results of such studies. Choice of whether or not to implement a change in dosing strategy under this approach was often a judgment call on the part of Regulators and was dependent upon their choice of the acceptance interval.

In contrast, bioequivalence studies were held to a higher standard under the legislation described in Part One of this series. New formulations were not admissible to market unless a successful bioequivalence study demonstrated that they met the regulatory standard under a well-controlled study using the TOST with predetermined acceptance interval of 0.80-1.25 (though some nations in Europe allowed a wider standard of 0.70-1.43 for Cmax, known to be a more variable endpoint than AUC). This *average bioequivalence approach* (so-called as it pertains to the equivalence of the means of the test and reference formulations) has safeguarded public health since its adoption (Barrett *et al.*, 2000)[9].

Equivalence for narrow therapeutic index drugs, those drugs for which a small change in dose or exposure can cause a large alteration in response to treatment, is sometimes regarded as particularly problematic under the average bioequivalence approach (Benet and Goyan, 1995)[10]. Examples of such drugs, digoxin and warfarin, (Colaizzi and Lowenthal, 1986)[11] generally exhibit low within-subject variability (i.e. within-subject coefficients of variation less than ten percent.) Under the average bioequivalence approach, it is possible (Phillips, 1990)[12] to demonstrate equivalence of means with prespecified $\delta = log_e(1.25)$; however, small average changes in means of statistically significant magnitude are possible. Such small changes in mean test to reference rate and extent of exposure are potentially clinically meaningful in a proportion of patients (Barrett *et al.*, 2000b)[13], and some have advocated (Ansbacher, 1990)[14], special equivalence definitions for narrow therapeutic index products whereby such drugs would be held to a more strict regulatory standard (e.g. equivalence limits corresponding to a ten percent range, 0.90 to 1.11).

In contrast, high variability products, defined as those products with within-subject coefficients of variation in excess of thirty percent (Blume and Midha, 1993)[15], require sample sizes in excess of thirty subjects in order to have eighty to ninety percent power to demonstrate average bioequivalence in a two period crossover design (Phillips, 1990)[12]. Some have argued (Midha *et al.*, 1997a and 1997b)[16,17] that small changes in rate and extent of exposure for such products are not clinically meaningful and have advocated allowance of a less strict regulatory standard - e.g. equivalence limits corresponding to a thirty percent equivalence range, 0.70 to 1.43, as allowed by European Regulators for Cmax. As an alternative, equivalence limits could be widened based upon the within-subject variability observed in the study (Boddy *et al.*, 1995; Schall and Williams, 1996; Midha *et al.*, 1997a and 1997b)[16-19] allowing such drug products easier market access.

HISTORY OF BIOEQUIVALENCE SINCE 1992: DEVELOPING POPULATION AND INDIVIDUAL BIOEQUIVALENCE

Structure of within-subject variability in a two-period crossover thus becomes a question of concern as it (in combination with the sample size and true mean difference between formulations) determines whether a formulation meets or fails to demonstrate average bioequivalence. The structure of this variance term can be explored in several ways.

Under a restricted maximum likelihood approach developed in Patterson and Thompson, 1971[20] and in the model notation of Laird and Ware, 1982[21], one can fit a random-intercept and random-slope model on the assumption that random slope and intercept are normally and independently distributed with null mean and variance of σ^2_W and σ^2_B, respectively, as follows (Jones and Kenward, 1989; Gaffney, 1992)[5,22]. Let X_{ij} be the response (log_e-transformed AUC or Cmax) for the j-th subject in the cross-over trial administered formulation t ($t = T, R$) and

$$X_{ij} = \mu_t + \nu_j + \varepsilon_{ij} \qquad (4)$$

Where, μ_t and ε_{ij} are independent with mean zero,

$Var\ (\nu_j) = \sigma^2_B$, the between-subject variance, and

$Var\ (\varepsilon_{ij}) = \sigma^2_W$, the within-subject variance.

Period, sequence, and carryover effects would be fitted in the model in practice in such a crossover trial (see Jones and Kenward, Chapter 4, 1989)[5] but are omitted from the description here for the sake of clarity.

Developing this idea further however, assuming that random-effects with mean zero, between-subject variance of (σ^2_{BT} and σ^2_{BR}), between-subject co-variance (σ_{BTR}), and independent within-subject variance (σ^2_{WT} and σ^2_{WR}) for test (**T**) and reference (**R**) formulations are present (though not all moments are estimable in most two-period crossover designs), the variance of $\hat{\mu}_T - \hat{\mu}_R$ is $(\sigma^2_{BT} + \sigma^2_{BR} - 2\ \sigma_{BTR} + \sigma^2_{WT} + \sigma^2_{WR})/n$ (Chinchilli and Esinhart, 1996; Vonesh and Chinchilli, Chapter 4, 1997)[23,24]. Note that under the model developed in Part One of this series, $\sigma^2_B = \sigma^2_{BT} = \sigma^2_{BR} = \sigma_{BTR}$ *and* $\sigma^2_W = \sigma^2_{WT} = \sigma^2_{WR}$. Equality in this manner is known as the Huyhn-Feldt condition (Hinkelmann and Kempthorne, 1994)[25]. Estimates for within-subject variation in a two period crossover study are thus held to be composed of measurement error (not estimable), within-subject variance components (estimable under the Huyhn-Feldt condition), and components associated with between-subject variation (estimable under the Huyhn-Feldt condition).

As an aside we note here that the component of the variance for $\hat{\mu}_T - \hat{\mu}_R$ associated with the variance of differences in between-subject variation, $(\sigma^2_{BT} + \sigma^2_{BR} - 2\sigma_{BTR}) = \sigma^2_D$, is an important consideration in the assessment of what has been termed individual

bioequivalence and will be discussed later in this article. Here we will only note that in a two-period crossover, under the Huyhn-Feldt condition, this variance is assumed to be null. Operationally, however, it should be noted that between-subject variability is known to be related to the extent of absorption (Hellriegel *et al.*, 1996)[26] complicating assessment of a meaningful difference in between-subject variance (as its magnitude is dependent on the choice of endpoint measuring extent of absorption.)

Average bioequivalence compares the distance between formulations as measured by mean rate and extent of exposure. Variation under this approach is of secondary interest and generally impacts only the choice of design (when sufficient sample size is considered to provide adequate power) and when assessing the final conclusions of bioequivalence in terms of the distance between means. Increased variation beyond that expected (consider the data presented in Part One of this series) can result in reduced power to demonstrate average bioequivalence. From a sponsor's perspective, therefore, it is preferable to increase sample size to an extent such that if unexplained increases in estimates of variation are observed (e.g. from the presence of an outlier or a group of outliers), the sample size is still sufficient to demonstrate bioequivalence in the mean rate and extent of exposure. Outliers are a frequent occurrence in bioequivalence studies and can result from a variety of factors (FDA Guidance, 1992)[1]; some may simply be indicative of random variation; however, some outliers may be indicative of subgroups in the population who absorb, distribute, metabolize, or eliminate the formulations differently than the general population.

The concept of switchability of formulations for the individual patient is not addressed by the average bioequivalence criterion (Hwang, *et al.*, 1978)[27]. Population means are compared, and variation between individual subjects (or patients) is factored out of the variation used to assess the distance between population means as described above. Peace (1986)[28], Anderson and Hauck (1990)[29], Hauck and Anderson (1992)[30], and Welleck (1993)[31] introduced the concept of individual bioequivalence. Under this approach, the question, 'Can I safely and effectively switch my patient from their current formulation to another?' is addressed using an approach similar to the 75/75 rule discussed in Part One of this series.

Under the TIER procedure (Test of Individual Equivalence Ratios) introduced by Anderson and Hauck (1990)[29], a predetermined minimum number of subjects for given sample size and Type 1 error rate must demonstrate individual ratios for test to reference rate and extent of exposure falling within a predetermined equivalence interval. This approach assumes that period and carryover effects are negligible. The hypothesis that is tested is:

$$H_0 : P_E < MINP \quad (5)$$

versus

$$H_1 : P_E \geq MINP \quad (6)$$

where *MINP* is the minimum proportion of subjects falling in the predetermined equivalence interval and P_E is the true proportion of equivalent individual ratios. The number of subjects falling in the equivalence interval (Y) is evaluated relative to the null hypothesis using a binomial probability. If the *p*-value equal to the *Prob* (Number of equivalent subjects $\geq$ Y given $P_E = MINP$ and the sample size in the data set) is less than the pre-set Type 1 error rate, then bioequivalence under the TIER is demonstrated.

When the TIER is applied to the AUC and Cmax data in Tables 3/4 of Part One, assuming an equivalence interval for individual test to reference formulations of 0.80 to 1.20, it is observed that sixteen of forty-five subjects and fourteen of forty-seven subjects have individual ratios of test to reference within the equivalence interval for AUC and Cmax, respectively. Assuming a *MINP* of 0.75, neither AUC nor Cmax demonstrates bioequivalence under the TIER (*p*-values are approximately unity in both cases).

TIER based assessment of bioequivalence was discussed in Hwang and Wang (1997)[32]. Sensitivity to normal and distributional assumptions was demonstrated; however, as discussed in of Part One, these assumptions are not held to be pivotal in the assessment of bioequivalence. Period effects however, (Schuirmann, 1990)[33], are held to be a frequent occurrence in crossover studies and are a confounding factor in the assessment of individual ratios (Welleck, 1997)[34].

Esinhart and Chinchilli (1994)[35] developed a method for assessment of an extension of the TIER using tolerance intervals for the ratio of individual responses, which accounts for period effects in a two-period crossover study. A tolerance interval is derived for the ratio of individual ratios using a model accounting for period effects, and should the tolerance interval fall within

predetermined acceptance limits, bioequivalence is demonstrated. Assessment under higher order designs was also discussed in Esinhart and Chinchilli (1994)[35] and is developed in more detail in Chinchilli (1996)[36] and Brown *et al.* (1997)[37].

Sample size determination is described in Esinhart and Chinchilli (1994)[38]. While this method was intuitively attractive, it is evident that sample size requirements for many drug products (those with within-subject coefficient of variation greater than twenty percent) are too great to be addressed in a small, well-controlled two-period clinical pharmacology trial and are still too large to be practically implemented in a higher order design (Esinhart and Chinchilli, 1994)[38].

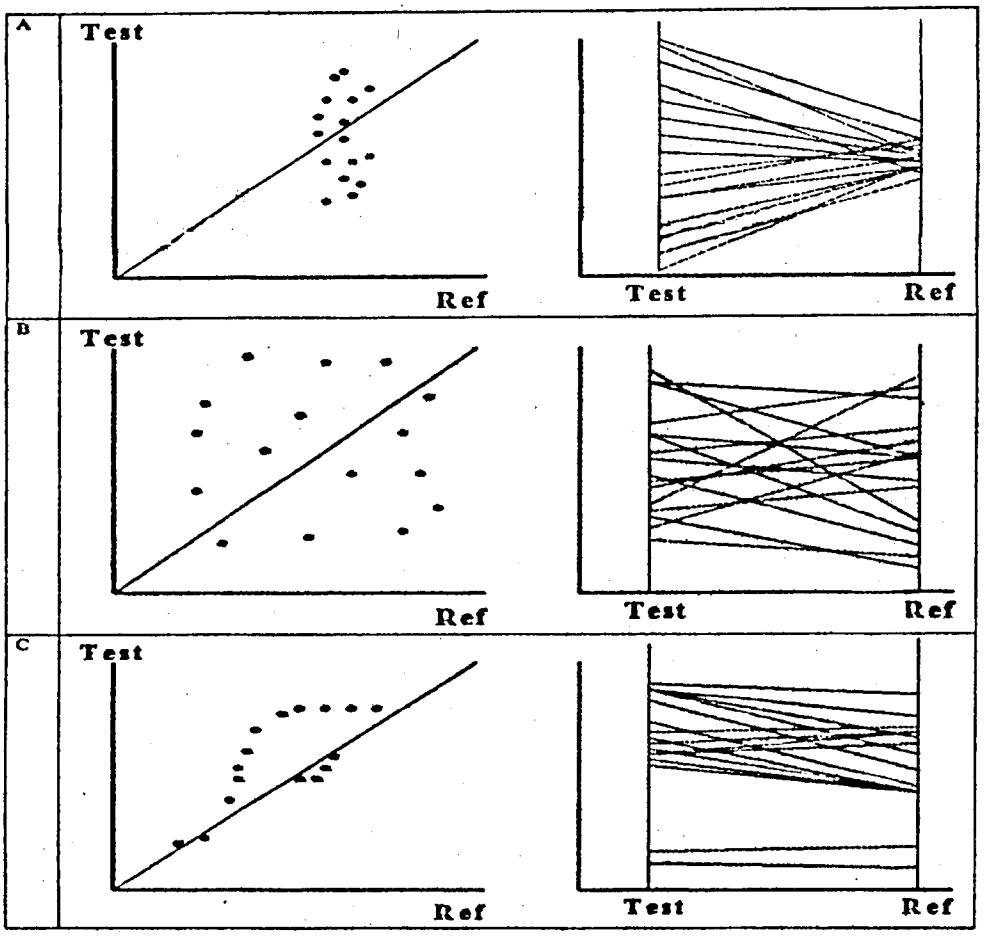
Average bioequivalence is a special case of what Hauck and Anderson (1992)[30] have termed population bioequivalence. This type of bioequivalence addresses the question, 'Can I safely and effectively start my patient on the currently approved formulation or another?' Differences in variation between formulations should also be considered when determining whether a formulation will be equally effective and safe when administering the commercial formulation of a new drug product relative to that used in clinical trials in Phase 3. It is not clear in this context whether comparison of within-subject variances or total variances (so termed as the sum of between- and within-subject variance for a given formulation) is the appropriate variance for comparison between formulations, and arguments (Hauck *et al.*, 1997; Grahnen *et al.*, 1984)[39,40] have been offered for both in this context.

Techniques for comparing within-subject variances in a two period crossover (under the assumption that between-subject variances across formulations are homogeneous) had been developed by Pitman (1939)[41] and Morgan (1939)[42]. Alternatively the total variances between formulations (between- plus within-subject variance) can be compared using a similar procedure. Most techniques for assessment of the equality of variances assume that variance components are independent (Brown and Forsythe, 1974; Balakrishnan and Ma, 1989)[43,44], a condition not met in the correlated data encountered in crossover trials. Bristol (1991a)[45] developed practical maximum likelihood techniques for comparing within-subject variances in this context based on techniques discussed in Mallet (1986)[46]. Cornell (1991)[47] derived nonparametric tests of dispersion for the two-period crossover design. Chow and Liu (Chapter 7, 2000)[48] described similar procedures, and Wang (1997)[49] and Guilbaud (1993, 1999)[50,51] described similar procedures in subsequent publications. These techniques reduce to different transformations to assess unequal marginal scales in a bivariate normal population (Kepner and Randles, 1982)[52], and such comparisons were also addressed in work by Bhoj (1979)[53], Ekbohm (1981)[54], McCulloch (1987)[55], and Bauer and Bauer (1994)[56].

Comparisons of total- or within-subject variance between formulations can be accomplished using such procedures; however, it is known (Zariffa *et al.*, 2000)[57] that variance components are ill characterized in crossover studies of the size usually performed in bioequivalence studies. Increasing sample size (Zariffa and Patterson, 2000)[58] can improve the precision of estimated variance components; however, it is unusual for such studies to be performed except in the case of highly variable drug products (Zariffa *et al.*, 2000)[57]. Moreover, while such procedures are theoretically and statistically viable, they are highly dependent (Vonesh and Chinchilli, Chapter 2, 1997)[24] on the choice of estimation procedure. Estimates for between-subject variance can be negative under a method-of-moments based procedure or maximum- likelihood procedure (Bristol, 1991 and 1991)[45,59]. Such estimates are positively biased (Endrenyi and Tothfalusi, 1999)[60] when using restricted-maximum-likelihood based estimation procedure as would be expected in a procedure constrained in the likelihood to only permit estimates greater than or equal to zero for between-subject variances and correlation constrained to lie in the range [-1, 1] (Patterson and Thompson, 1971; Jones and Kenward, Chapter 7, 1989; Davidian and Giltinan, 1995; Vonesh and Chinchilli, Chapter 4, 1997.)[5,20,24,61]

Regardless of the poor quality and high dependence of variance component estimation on choice of estimation procedure, such estimates continued to be of interest in the assessment of switchability for bioequivalence (Ekbohm and Melander, 1989)[62]. Under this approach, subject-by-formulation interaction, $\sigma^2_D = (\sigma^2_{BT}+\sigma^2_{BR}-2\sigma_{BTR}) \leq 0$, is termed a measure of individual switchability. Such an estimate is estimable in what is termed a replicate design (Gaffney, 1992)[22].

A number of scenarios can give rise to a quantitatively large subject-by-formulation interaction. These are presented in fig. 1. A classical subject-by-formulation interaction (Ekbohm and Melander, 1989)[62] occurs when



A: Between-Subject Variance Inequivalence contributing to a Subject-by-formulation Interaction
B: Unpredictability (low correlation) contributing to a Subject-by-formation Interaction
C: Subgroup-by-formulation Interaction contributing to a Subject-by-formulation Interaction

Fig. 1: Sources of Subject-by-Formulation Interaction Variation

subjects experience a more variable response when receiving one formulation relative to the other as illustrated in fig.1-A. In this example, variation was greater for the test product relative to the reference product. Subject-by- formulation interaction can also occur when unpredictable responses are observed between regimens, as illustrated in fig. 1-B. This is essentially the case when low correlation (where correlation $\rho = \sigma_{BTR}/\sqrt{\sigma^2_{BT}\sigma^2_{BR}}$) is observed. Subject-by-formulation interaction can also be generated from sub-groups having differential reactions to drug products as illustrated in fig. 1-C.

In a replicate crossover design (Patterson, 1950)[63], each subject receives each formulation twice as follows. Eligible subjects are randomized to one of two treatment sequences, e.g. TRTR or RTRT (where T denotes the test and R the reference formulations, respectively, see Jones and Kenward, Chapter 4, 1989)[5]. Thus, each subject is studied in four periods and receives each formulation twice over the course of the study. Similar to the two period crossover described in Part One, a washout period adequate to the drug under study (a least five half lives) separates each treatment periods. In each period, the formulation is administered following an overnight fast. Such a design (described in greater detail in Jones and Kenward, Chapter 4, 1989)[5] allows for the estimation of this subject-by-formulation interaction component as it is only partially confounded (Chinchilli and Esinhart, 1996)[23] with within-subject variation for each formulation.

Method-of-moment based, maximum likelihood based, or restricted maximum likelihood based estimation procedures (Harville, 1977)[64] can be used to compute the variance components. Within-subject variance estimates can be computed in a straightforward manner based on these procedures (Chinchilli and Esinhart, 1996)[23].

In bioequivalence studies, using a replicate design with sequences RTRT and TRTR, the following mixed model for log_e-transformed observations is commonly accepted (Jones and Kenward, 1989)[5]. Let X_{tjk} be the k-th response (k = 1, 2, . . .) for the j-th subject in the cross-over trial administered formulation t (t = T, R) and

$$X_{tjk} = \xi_{tj} + \varepsilon_{tjk} = \mu_t + \nu_{tj} + \varepsilon_{tjk} \qquad (7)$$

ν_{tj} and ε_{tjk} are independent with mean zero

$Var(\nu_{tj}) = \sigma^2_{BT}$, the between-subject variance,

$Var(\nu_{Tj} - \nu_{Rj}) = \sigma^2_{D}$, the subject-by-formulation interaction variance,

$Cov(\nu_{Tj}, \nu_{Rj}) = \rho\sigma_{BT}\sigma_{BR}$,

$Var(\varepsilon_{tjk}) = \sigma_{Wt}$, the within-subject variance,

$Cov(\varepsilon_{tjk}, \varepsilon_{tjk'}) = 0$, for $k = k'$.

Note that nuisance effects (period, sequence, and carryover effects) are fit in practice (Jones and Kenward, Chapter 4, 1989)[5] but are omitted from the above description for the sake of clarity.

Under approach to analysis of the replicate design, it can be shown (Vonesh and Chinchilli, Chapter 4, 1997)[24] that the variance for $\hat{\mu}_T - \hat{\mu}_R$ is equal to $(\sigma^2_{BT}+\sigma^2_{BR}-2\sigma_{BTR}+((\sigma^2_{WT}+\sigma^2_{WR})/2))/n = (\sigma^2_{D}+((\sigma^2_{WT}+\sigma^2_{WR})/2))/n$ in a balanced design with n subjects. Under the assumption that σ^2_D is zero and that within-subject variances are equal, this design is approximately twice as efficient as the two period crossover (in terms of the sample size required to demonstrate bioequivalence with equal power). It should be noted that the replicate design with sequences RTTR and TRRT is more efficient in those situations where first-order carryover cannot be assumed to be negligible (or equal) between formulations (Jones and Kenward, 1989)[5].

Model estimates for differences in means and variances can be used in other ways. Sheiner (1992)[65], Schall and Luus (1993)[66], and Schall (1995)[67] introduced an alternative method for individual bioequivalence assessment based on models of dose-response (Sheiner, *et al.* 1989)[68], risk assessment, and different combinations of parameters from the model (7). Under this 'moment-based' approach to bioequivalence assessment, differences in means and variances are combined into one 'aggregate' statistic for the assessment of population and individual bioequivalence. If the upper ninety-five percent bound on the aggregate statistic falls below a preset equivalence margin, bioequivalence is demonstrated. Such a procedure also allows for widening (or narrowing) of the equivalence margin based upon variation observed in the study.

Bootstrap (Schall, 1995)[67] or Bayesian (Sheiner, 1992)[65] based assessment of the quantiles of the composite endpoint were initially proposed; however, estimation procedures for such an aggregate statistic using approximation procedures involving the Cornish-Fisher Expansion (Bickel and Doksum, 1977)[69] and methods for the linear combination of independently χ^2-distributed variables (Huitson, 1955; Fleiss, 1971; Howe, 1974; Harville, 1976; Burdick and Sielken, 1978; Graybill and Wang, 1980; Lu *et al.*, 1988; Ting *et al.*, 1990; Wang, 1990; Burdick and Graybill, 1992)[70-79] were developed in Holder and Hsuan (1993a and 1993b)[80,81].

Practical strategies for population and individual bioequivalence assessment under this approach were developed in Schall and Williams (1996)[19]. Application to the moment-based criterion of most interest was developed in greater detail by Hyslop *et al.* (2000)[82], and an alternative parametric procedure was described by Kimanini and Potvin (1997)[83].

Consideration of these ideas led the FDA Biopharmaceutical Science Division to form a bioequivalence working group in the mid-1990's. This body (composed of FDA representatives from clinical, scientific, and statistical disciplines) was tasked with determining whether a public health risk under the average bioequivalence approach could exist and if so to determine a method or methods to evaluate bioequivalence in a manner to protect the public health. A description of the ideas under discussion may be found in Anderson (1993)[84], Hauck and Anderson (1994)[85], Anderson (1995)[86], Anderson and Hauck (1996)[87], Hauck (1996)[88], Hauck *et al.* (1996)[89], Patnaik *et al.* (1997)[90], Gould (1997)[91], Chen (1997)[92], and Anderson and Hauck (1996)[87] but will not be discussed further in this paper.

It should be noted that many other approaches were considered during the debate on bioequivalence. Testing procedures for assessing differences in means and variances simultaneously (though not as a composite

endpoint) were developed in Bauer and Bauer (1994)[56], Bauer and Keiser (1996)[93], and Ghosh *et al.* (1996)[94]. Stepwise procedures (testing for equivalence in means between formulations followed by testing for equivalence in variances) were described in Endrenyi and Schulz (1993)[95], Endrenyi (1994)[96], Vuorinen and Turunen (1996)[97], Vuorinen (1997)[98], Guilbaud (1999)[51], and Gould (2000)[99]. Unbiased, optimal tests for bioequivalence assessment were described in Munk (1993)[100], Hsu *et al.* (1994)[101], Brown *et al.* (1997)[102], and Wang (1999)[103], and multivariate, optimal assessment of bioequivalence (e.g. for AUC and Cmax simultaneously) were described in Berger (1992)[104], Berger and Hsu (1996)[105], Chinchilli and Elswick (1997)[106], and Munk and Pfuger (1999)[107]. Testing directly for differences in profiles were described in Mauger and Chinchilli (2000)[108].

Though statistically valid, under the approach to inference described by Hauck *et al.* (1995)[109], multivariate procedures were not of direct interest to the bioequivalence debate. The other approaches seem to have little additional benefit in practical bioequivalence assessment relative to those the FDA were considering (Senn, 2000)[110] and thus seem to not have impacted upon the debate.

We now turn to the conclusion of the bioequivalence debate beginning with the draft FDA guidance on population and individual bioequivalence released for public comment in 1997.

MOST RECENT DEVELOPMENTS: IMPLEMENTING POPULATION AND INDIVIDUAL BIOEQUIVALENCE

The US Food and Drug Administration's decision following the debate on whether population and individual bioequivalence was needed to protect public health and the approach chosen for assessment were announced in draft guidance released in 1997 (FDA Guidance, 1997)[111] based on the principles discussed in Schall and Williams (1996)[19]. Previously discussed approaches to moment-based assessment of population and individual bioequivalence were established as follows for studies conducted prior to approval and following approval of new chemical entities.

Average bioequivalence was deemed insufficient to protect the public health as it assessed only the difference in formulation means, did not adjust for the variance of narrow therapeutic drug products and highly variable drug products, and did not account for assessment of subject-by-formulation interaction. No clear evidence of therapeutic failure had been established over the five years in which the 1992 FDA guidance had been in effect (Barrett *et al.*, 2000)[9]. Conventional two-period, randomised, well-controlled, crossover designs were established as the design to be performed in the assessment of population bioequivalence for approval of bioequivalence in formulation changes prior to approval of the new drug product (FDA Guidance, 1997)[111]. Two-sequence (RTRT, TRTR), randomised, well controlled, replicate designs were chosen as the design to be performed in the assessment of individual bioequivalence for approval of new formulations following approval of the new drug product for both generic manufacturers and those manufacturers wishing to make formulation changes following approval. Replicate designs were required for the assessment of individual bioequivalence so that within-subject estimates of variance were estimable along with the subject-by-formulation interaction (FDA Guidance, 1997)[111]. Requirements for adequate washout between study periods was again required to ensure that carryover effects were negligible, and outliers were again deemed to be indicative of either product failure or sub-population-by-formulation interactions. Rate and extent of bioavailability were again measured by Cmax and AUC, respectively.

Overall, the FDA draft Guidance (1997)[111] involved little change in study design for sponsors conducting trials to establish bioequivalence of a new commercial formulation relative to that used in clinical trials under the population bioequivalence approach to inference. The new draft guidance however required replicate designs for changes following approval - a more complex design for the majority of drug products. Also, under this approach to inference, log_e-transformed AUC and Cmax were to be analysed separately using a two stage (mixed effect, restricted maximum likelihood) linear model including terms for sequence, period, and formulation in the model in accordance with model (7) for a replicate design.

Population bioequivalence is assessed using the following aggregate statistic (FDA Guidance, 1997)[111].

$$\frac{(\mu_T-\mu_R)^2 + \sigma^2_T - \sigma^2_R}{max(0.04, \sigma^2_R)} \qquad (8)$$

where $\sigma^2_T = \sigma^2_{WT} + \sigma^2_{BT}$ and $\sigma^2_R = \sigma^2_{WR} + \sigma^2_{BR}$. Note that this aggregate statistic can be constructed using a mixed model from a two period crossover design (with appropriate

modification to model (7)) and does not require the use of a replicate design.

Individual bioequivalence is assessed using the following aggregate statistic (FDA Guidance, 1997)[111].

$$\frac{(\mu_T-\mu_R)^2+\sigma^2_D+\sigma^2_T-\sigma^2_R}{max(0.04,\ \sigma^2_R)} \qquad (9)$$

Because the within-subject variance of each formulation cannot be separately estimated from between-subject variance estimates in most two-period crossover designs of the form { TR, RT }, a replicate design is required.

At least one thousand five hundred (two thousand samples were recommended in the FDA Guidance, 1997)[111] bootstrap samples (Efron and Tibshirani, Chapter 25, 1993)[112] preserving the number of subjects in each sequence are derived, and the above mixed model is fit to each bootstrap sample. The appropriate aggregate statistic, either (8) or (9), is derived based on the model estimates for each bootstrap sample; note that the denominator for each bootstrap's aggregate statistic is chosen based on the point estimate from the model estimates of the original data set. The nonparametric percentile method (Efron and Tibshirani, 1993)[112] is then used to calculate a upper ninety five percent bound for the quantity of interest. It was required that the upper ninety-fifth percent bound for the metric of interest fall below predetermined regulatory bounds (1.7443 and 2.4943 for population and individual bioequivalence, respectively) for both AUC and Cmax for bioequivalence to have been demonstrated.

Responses to release of the US FDA's draft guidance (1997)[111] were plentiful from academia. Scientific flaws of the new procedure for individual bioequivalence were noted (Endrenyi *et al.*, 1998)[113] as being:

1. The numerical tradeoff of distance between within-subject variances and the means was strongly asymmetric. Developed in more detail in Endrenyi and Hao (1998)[114], it was found that a small change in within-subject variances, could allow for a change in means, which would still permit a conclusion of bioequivalence but which would expose a large number of individual patents to risk of therapeutic failure or overexposure to drug.
2. The scaling of the criterion to within-subject variance potentially declares the equivalence of formulations liberally. Again, scaling to variance could allow a proportion of individual patients to exceed safe or therapeutic levels of drug product when switched to a new medication.
3. Computational uncertainty of estimation, both in the models used to assess population and individual bioequivalence [see model (7), and the nonparametric-percentile bootstrap method (Efron and Tibshirani, Chapter 25, 1993)[112]] used to assess inference were noted as being of potential concern when near the predetermined acceptance bound.

Subsequent work describing the properties of the subject-by-formulation interaction in Endrenyi and Tothfalusi (1999)[60] determined that this subject-by-formulation interaction statistic was directly confounded under a restricted-maximum-likelihood based estimation approach, with within-subject variation. This would be expected under such a constrained likelihood based procedure and is clinically meaningful in that between-subject variation is known to be confounded with extent of bioavailability (Hellriegel *et al.*, 1996)[26]. Method-of-moment based estimates for $\hat{\sigma}^2_D$ (Endrenyi *et al.*, 2000)[115] are unbiased, but the variance of $\hat{\sigma}^2_D$ is still related to σ^4_W where W denotes within-subject variation under the Huyhn-Feldt condition.

On practical grounds, Endrenyi *et al.* (1998)[113] and Endrenyi and Midha (1998)[116] determined that:

1. Average bioequivalence had not been observed to fail to protect the public health as no objective, adequately demonstrated reports to this effect had been published.
2. It was noted that the available data made public by the FDA from replicate designs for estimates of subject-by-formulation interaction ($\hat{\sigma}^2_D$) were not sufficient to demonstrate a clinical need for the assessment of individual bioequivalence. Furthermore, bioequivalence studies were not conducted in the patient population of concern, and so clinical safety/therapeutic failure could not reasonably be assessed in a non-patient population. As such, comparison of between- and within-subject variances had not been demonstrated to be clinically relevant surrogate markers for therapeutic inefficacy and or unacceptable safety profile.

Other academic responses to the draft FDA guidance (1997)[111] by Senn (1998)[6] noted that it was inappropriate

for generic drug products to be held to a stricter standard (i.e. be subject to assessment of differences in within-subject variance) when innovator drug products were not (i.e. only held to the assessment of differences in total-subject variance between formulations) The new procedures were also noted as being illogical in terms of risk assessment (Senn, 2000)[110]. Patients are more at risk when they start a new treatment than when they switch to a new formulation following ongoing treatment implying that standards for population bioequivalence should be more stringent than individual bioequivalence.

Lindley (1998)[117], following on from ideas originally discussed in Westlake (1986)[118] and Hwang (1996)[119], argued that bioequivalence determination involved making a decision and proposed the use of Bayesian decision theory (Lindley, 1971)[120] in bioequivalence assessment. Lindley discussed two potential decisions: δ_1 to declare bioequivalence, and δ_0 to deny bioequivalence based upon a measure of equivalence, θ. Under this approach, a loss function $L_\theta = u(\delta_0, \theta) - u(\delta_1, \theta)$, predetermined based on agreement between the sponsoring company and regulatory authorities (Lindley and Singpurwalla, 1991)[121], is assessed where $u(\delta, \theta)$ is a utility function measuring the worth of δ when the uncertain value is θ. Expected loss can be derived using a prior distribution for θ and Bayes' rule using available software packages and applied to bioequivalence assessment, and structure of the problem can easily be extended to multiple bioequivalence measures (i.e. for AUC and Cmax) Lindley (1998)[117] discussed a straightforward method for choice of prior distributions and describes previous work impacting choice of sample size (Lindley, 1997)[122] under such an approach.

Industry responses on the scientific merits of the FDA draft guidance (1997)[111] were similarly negative and were primarily based on retrospective analysis of existing replicate design data sets. Key findings are summarized below. These results were presented at the American Association of Pharmaceutical Scientists held a joint workshop with the FDA on the topic of bioequivalence from 16-18 March 1998 in Washington D.C. at which a few industry representatives were able to speak to the scientific issues behind the proposal. Of particular note, preliminary analyses of SmithKline Beecham's existing database of previously performed replicate design studies (Zariffa *et al.*, 1998)[123] revealed that:

1. Large differences in means (not permitted under the average bioequivalence criteria) were permitted under the new population and individual approaches to bioequivalence when offset by decreased test variance or scaling to variance of the reference formulation. This was particularly the case for highly variable drug products.
2. Also, substantial subject-by-formulation interaction variation could be masked in the aggregate individual bioequivalence criteria by decreased within-subject test formulation variation relative to within-subject reference formulation variation.
3. Behavior of the individual bioequivalence statistic when variation for the reference product nears the cutoff (0.04, see (8) and (9)) was inconsistent with logical inference concerning bioequivalence.
4. Results for Cmax were far less consistent between average, population, and individual criteria than AUC suggesting that uniform criteria for rate and extent of bioavailability might not be appropriate.

Additionally, a practical benefit of the new criteria (Zariffa *et al.*, 1998)[123] for sponsors of bioequivalence studies was noted. For highly variable drug products, a substantial decrease in sample size was possible under the new criteria as scaling to the reference formulation's variation. Thus for highly variable products, while a replicate design was required for assessment of bioequivalence, decreased resources would be necessary for sponsors to determine if a formulation was bioequivalent (assuming that subject-by-formulation interaction was negligible).

Subsequent analyses (Patterson *et al.*, 1998)[124] revealed that the effect of scaling to reference product variation was not substantial until coefficients of variation on the order of thirty to forty percent were observed for population and individual bioequivalence. While precision of the estimated variance components was remarkably poor in the existing data sets, it was observed that the magnitude of subject-by-formulation interaction was greater for Cmax than AUC, and the magnitude of subject-by-formulation interaction appeared to increase with increasing magnitude of within-subject variation. The choice of the restricted maximum likelihood estimation procedure in the FDA Guidance (1997)[111] was potentially related to these findings.

Other published industry responses to the FDA draft guidance (1997)[111] were similarly negative. Schumaker

and Metzler (1998)[125] conducted an analysis of a replicate design study using two formulations of phenytoin. Notable conclusions of this trial were that, as means between formulations were equivalent and within-subject variation across formulations was homogeneous, no evidence existed for individual bio-inequivalence. This result implied that individual bioequivalence could be assessed using procedures based upon the usual two period crossover study design, and that the imposition of additional rules for bioequivalence was not necessary. Additionally, as a previously known problematic drug substance was involved, this implied that the existing FDA Guidance (1992)[1] was sufficient to protect public health.

Responses from European authorities on bioequivalence were similarly negative on the scientific merits of the proposed FDA Guidance (1997)[111]. Steinijans and Diletti (1997)[126] noted that comparison of within-subject variance between formulations was possible but that the use of these methods had not properly been considered and that use of the new proposed FDA population and individual bioequivalence criteria had not been justified on clinical grounds. On statistical issues, Steinijans and Diletti (1997)[126] encouraged the consideration of alternative inferential procedures to the use of the bootstrap. Lastly, Steinijans and Diletti (1997)[126] encouraged the FDA to expand its working group and to gain consensus among a wider audience. Later published reports, produced in this period, may be found in Hauschke and Steinijans (2000)[127] and Kimanani *et al.* (2000)[128].

In summary, according to the FDA (private communication, 1998), a total of twenty-four individuals provided a total of two hundred forty separate comments to the FDA draft Guidance (1997)[111] broken down into the following categories:

General:

1. Individual bioequivalence is not justified because the current practice of average bioequivalence has worked well.
2. Subject-by-formulation interactions are unimportant.
3. Individual bioequivalence should not be required for all drugs.
4. Patients should be used in bioequivalence studies rather than healthy volunteers.
5. An individual bioequivalence criterion will not assure inter-changeability between two generic products.

Resources:

1. More time to complete a replicate design study.
2. Increased cost of bioequivalence studies.
3. Increase in blood volumes and drug exposure, with possible reduction in availability of subjects.
4. More technical and procedural problems.
5. More subjects have to be recruited because of the high dropout rate.
6. The proposed statistical methods are complicated and would need sophisticated computer software.

Process:

1. The development of the new approaches should be coordinated through the International Conference on Harmonization.
2. An experimental period is proposed where average bioequivalence is the primary assessment method and the proposed population/individual criteria will be alternatives, which may be left to the sponsor's choice.
3. Future studies on approved drugs should use the average bioequivalence approach.
4. The SUPAC-IR document is not referenced.
5. Lack of harmony between this guidance and the existing food effect guidance.
6. Consensus should be obtained between FDA, Industry, and Academia before broader implementation of the new criteria.

Application:

1. The proposed approach may be suitable only for highly variable drugs, narrow therapeutic index drugs, drugs with long half-life, or special cases where safety and efficacy profiles are greatly affected by the absorption rate of the drug.
2. It is not clear under which situations certain study designs should be used, e.g. single versus multiple dose.
3. It is not clear that an individual bioequivalence applies during the IND (i.e. pre-approval) phase of drug development.
4. Application of the proposed approach to other clinical pharmacology studies (i.e. drug-drug interactions).
5. Same criterion for metabolites.
6. It would not be necessary to use the individual bioequivalence approach for all drugs, i.e. if the

residual variance estimated in a two-period crossover is adequately small, concerns need not arise about either within-subject variances or subject-by-formulation interaction.

Criterion :

1. A disaggregate criterion might be a better alternative.
2. The mean/variance tradeoff might allow products in the market with substantial mean differences.
3. The proposed individual bioequivalence criterion is asymmetric.
4. Justification should be provided for equal weighting of means and variances, as well as for grouping squared differences and differences of squares in the same equation.
5. The interpretation of the aggregate criterion based on trans-formed values is not straightforward.
6. Why is the criterion not expressed in a more readily interpretable manner?

Methodology:

1. Bootstrap introduces randomness.
2. Bootstrap may be biased if only one is reported.
3. 1500 bootstraps may not be enough.
4. There are many ways to produce random numbers.

Numerous miscellaneous public comments were also received by FDA and will not be discussed in this paper.

In 1998, FDA subsequently formed what was termed a 'Blue Ribbon Panel' of academic and industry representatives to advise the FDA Working Group on implementation of population and individual bioequivalence in practice. The Pharmaceutical Research and Manufacturers Association (PhRMA) formed a parallel expert panel to assess the issues involved and prepare a joint industry statement on the merits of the proposal. A summary of FDA rejoinders to the concerns of industry, academia, and international regulators may be found in Chen *et al.* (2000)[129] and in Williams *et al.* (2000a and 2000b)[130,131].

After considering the public comments on the draft (1997)[111] guidance and after once consulting the Blue Ribbon Panel (October 1998), FDA re-issued two draft guidances on the topic of bioequivalence in August 1999 (replacing the draft guidance issued in 1997). These two guidances described when to perform a relative bioavailability, population, or individual bioequivalence study (FDA Guidance, 1999a)[132] for drug products in solution, suspensions, aerosols and for topical administration and for the more usual immediate-release and modified-release drug products. General guidance for study design (discussed earlier in this article) was provided. A novel aspect of the guidance was the suggestion that a two-year data collection period for all drugs would be mandated when the guidance was finalised. During this period, all sponsors would be required to perform a replicate design study in order to gain market access, and the sponsoring company would have the option of what criteria to choose to assess bioequivalence (Average or Population bioequivalence for sponsors applying for approval of a new product; Average or Individual bioequivalence for sponsors applying for approval of a new formulation of a product already approved for the market).

FDA acknowledged in the new draft guidance (FDA Guidance, 1999a)[132] that narrow therapeutic index drugs should be held to a stricter equivalence criteria than the usual twenty-percent range required in the existing FDA Guidance (1992)[1]. For these drug products, a ten-percent acceptance range on the log_e-scale (corresponding to an equivalence range of 0.90-1.11) was required.

The second draft guidance from FDA (1999b)[133] described in more detail the study designs, models, and approaches to statistical inference for average, population, and individual bioequivalence relative to the 1997 draft guidance, but departed from the original approach only in minor respects. Requirements for power and sample size was described in more detail in this draft guidance relative to the original 1997 draft guidance; however, the main departure was in the method for assessment of statistical inference.

This draft guidance (1999b)[133] required the use of the Cornish-Fisher expansion (FDA Guidance, 1999b; Hyslop *et al.*, 2000)[82,133] for the assessment of population and individual bioequivalence based upon estimates derived using a method-of-moments based estimation approach. In contrast, the use of restricted maximum likelihood estimation was required for assessment of average bioequivalence in studies employing a replicate design but was to be used in the assessment of population and individual bioequivalence only in the case of data sets with 'substantial' missing data. Bootstrap based inference was also relegated to the status of a 'back-up' procedure,

to be used only in instances where the Cornish-Fisher expansion and method-of-moments based estimation could provide misleading results.

Academic responses to the FDA draft (1999a and 1999b)[132,133] guidances were not as plentiful. Longford (1999)[134] discussed the concept that Phase 3 pivotal safety/efficacy trials should establish whether treatment effects were of limited variance in the population of interest. If variance was small and only a small proportion of patients could be placed at risk when a novel formulation was introduced, this was not as concerning as those situations where higher levels of variance suggested that a significant proportion of patients would be placed at risk of therapeutic failure. Longford (2000)[135] also introduced an alternative procedure for the assessment of individual bioequivalence based upon a linear combination of independent χ^2 variates where inference could be assessed in small samples using a bootstrap based procedure or in large samples based upon a normal approximation.

Other academic sources (Senn, 2000)[110] held that average bioequivalence should suffice based upon grounds of 'practicality, plausibility, historical adequacy, and purpose' and 'because we have better things to do'. Additionally, Senn (2000)[110] notes that statisticians have 'a bad track record in bioequivalence', that 'the literature is full of ludicrous recommendations from statisticians', that 'Regulatory recommendations (of dubious validity) have been hastily implemented', and that 'Practical realities have been ignored'.

Lastly, other academic authorities called for publication of data pertaining to the validity and applicability of the new methods (Colburn and Keefe, 2000)[136].

Innovator and generic industry responses to the newest (1999) draft guidance were however more plentiful. PhRMA's expert panel published its work (Barrett *et al.*, 2000)[9] and concluded that:

1. The clinical relevance of σ^2_D and its use as a surrogate marker for switchability could be studied by a targeted clinical pharmacology trial constructed to provide the best evidence of σ^2_D.
2. Trade-offs between parameters, scaling, and the maximum allowable difference (Hauck *et al.*, 1996)[89] could be addressed by the use of an ordered testing procedure.
3. Generic-to-generic switching could be addressed through the use of simulation studies.
4. To maintain the spirit of global harmonization, it is reasonable to expect that FDA and PhRMA will continue to engage in dialogue with other regulatory agencies and solicit their involvement. PhRMA's expert panel further recommended that simulation studies be used to assess the use of alternative statistical procedures (Dragalin and Fedorov, 1999a and 1999b; Lin, 1989, 1992, and 2000; Gould, 2000)[99,137-141] relative to the FDA draft guidance (1999a and 1999b)[132,133].

Additional industry responses (Patterson and Zariffa, 1999; Zariffa and Patterson, 1999; Patterson and Zariffa, 2000; Zariffa *et al.*, 2000; Zariffa and Patterson, 2000)[57,58,142-144] described the practical application of population and individual bioequivalence and the behaviour of the proposed criteria based upon actual data and simulation studies. These works concluded that:

Some of the expected features of current proposed population and individual bioequivalence criteria, such as the mean-variance tradeoffs, have been observed in the current database. However, collection of more data in an unsystematic manner will not result in clear answers to the questions of interest. Simulation studies should be utilized to enhance understanding of factors impacting the assessment of bioequivalence and to consider alternative criteria for assessment. Such additional simulation assessment should be undertaken prior to the implementation of a mandatory data collection period. Market access should not be permitted using any new criteria until it is clearly demonstrated that the new criteria offer substantive benefit and no added risk to public health.

There are currently sufficient doubts on both sides of the debate as to the validity of issues raised by the opposing viewpoints. As such, some manner of further study, conducted with scientific rigor, is called for. Given the complex interplay between the many factors at work, it is necessary to clearly outline the goals of proposed further studies to avoid misleading results. Of particular interest is the overall question of added-value, 'Do the proposed criteria reliably address substantial limitations of average bioequivalence and if so, can we be assured they do not in turn introduced additional limitations which could potentially be more serious?' A combination of simulation studies and data collection may be relevant to laying some of the issues to rest.

International regulatory responses deemed the concepts of population and individual bioequivalence to be un-necessary as average bioequivalence had protected the public health. Representative of this view were the presentations by Ormsby (1999)[145] and Pound (1999)[146] at the FDA/AAPS 1999 Workshop on 'Individual Bioequivalence: Realities and Implementation' co-sponsored by the International Pharmaceutical Federation, Canadian Society for Pharmaceutical Sciences, and the Therapeutic Products Program, Health Canada. Ormsby (1999)[145] noted that until subject-by-formulation interaction had been proven to be indicative of therapeutic failure and the causes identified, average bioequivalence (which had served to protect the Canadian public since its introduction with over 2500 generic products introduced to the market) would continue to be the standard. Pound (1999)[146] described alteration of the average bioequivalence decision rules with changes in type 1 error rate and or acceptance range indicated for narrow therapeutic drug products or those thought to be 'dangerous' in clinical practice.

FDA responses to the questions of interest were plentiful at the FDA/AAPS 1999 Workshop on 'Individual Bioequivalence: Realities and Implementation' but have been listed previously in this section and will not be re-iterated in this paper. The chief outcome of the conference was the realisation that little evidence existed to warrant the use of the new bioequivalence methods based on sufficient and adequate safety of patients in the marketplace under average bioequivalence and that subject-by-formulation had not been established as a surrogate marker for therapeutic failure (i.e. there was no 'smoking gun') in an extensive review of replicate design data sets. Population and individual bioequivalence were referred to as a 'theoretical' solution to a 'theoretical' problem.

The FDA Blue Ribbon Panel present at the meeting voted to consider a mandatory data collection period using replicate designs for bioequivalence assessment only for drugs most likely to have a subject-by-formulation interaction, modified release drug products and highly variable drugs, and to allow market access only using the established decision rules of average bioequivalence. This view was subsequently endorsed by the FDA's Advisory Committee of Pharmaceutical Science in September 1999.

Subsequent reports described the rationale behind assessment of subject-by- formulation interaction in the assessment of individual bioequivalence (Hauck *et al.*, 2000)[147]. The concept of a large interaction, a $\hat{\sigma}^2_D$ greater than 0.0225, was held to be a conservative measure and potentially indicative of significant subgroup-by-formulation interactions. Another report (Singh *et al.*, 1999)[148] established population-modeling based procedures for assessing bioequivalence in those drug products where pharmacokinetic measures such as AUC and Cmax cannot be used as surrogate markers for safety and efficacy.

A report was subsequently produced (Meyer *et al.*, 2000)[149] to provide an example of a data set from a replicate design using two marketed immediate release formulations of methylphenidate (indicated for the treatment of sleep disorders). The innovator version had been admitted to the market following a full clinical development programme; however, the generic version of methylphenidate was admitted following only *in vitro* dissolution testing (under an exception to the normal average bioequivalence requirements) and had not been held to the average bioequivalence standard. Following reports of therapeutic failure in patients switched to the generic product, a replicate design bioequivalence study was conducted in twenty volunteers. AUC and Cmax of the formulations were bioequivalent under the average bioequivalence approach; however, Cmax showed slightly higher within-subject variance for the test formulation relative to reference was claimed to exhibit a nominally high level of $\hat{\sigma}^2_D$ which passing under average bioequivalence failed to demonstrate individual bioequivalence under the method proposed by FDA. Other prospectively performed studies for the assessment of individual bioequivalence were published in Bekersky *et al.* (1999)[150], Cerutti *et al.* (1999)[151], Canafax *et al.* (1999)[152], and Yacobi *et al.* (2000)[153], but these studies did not identify a difference in formulations.

FDA guidance (2000b)[3] finalized in October 2000 indicated that the agency would adopt the recommendations of the Pharmaceutical Sciences Advisory Committee (1999). This guidance recommended the use of replicate designs for highly variable and modified release drug products; however, market access is granted if and only if the study demonstrates average bioequivalence. Sponsors conducting the study may use population or individual bioequivalence approaches to inference if justification is sufficient to meet FDA review.

Procedures for review of the data generated by these replicate designs to assess the need and appropriateness for population and individual bioequivalence are under consideration.

CONCLUSION

Bioequivalence studies evolved in the 1960's and 1970's to meet the practical needs of consumers in having access to inexpensive efficacious products and to meet the needs of producers in supplying markets with such products without the extensive costs associated with a full clinical development plan and the delay associated with long clinical studies. On a practical level therefore, their genesis was practical, economic, and driven by legislation to allow market access under strictly regulated conditions. In parallel, scientific advances in drug manufacturing and the science of clinical pharmacology and pharmacokinetics and statistics made it possible to assess differences in mean response between formulations based on small, well-controlled, crossover studies in normal healthy volunteers.

Therapeutic failures in the 1970's prompted extensive research into the science of bioequivalence. This continued in the 1980's and culminated in the establishment of the techniques for judging formulations bioequivalent based on similarity of mean rate and extent of bioavailability between different formulations of the same drug product. This average bioequivalence approach has served to protect the public health since its adoption by the US Food and Drug Administration in 1992 and has quickly spread to all parts of the globe.

However, average bioequivalence compares only the mean rate and extent of bioavailability between formulations and does not compare between- or within-subject variances between formulations. Nor does average bioequivalence assess individual similarity of responses or establish whether different subgroups among the general population will react differently to different formulations. Theoretical solutions to these theoretical problems with the average bioequivalence approach prompted extensive research on the topics of population and individual bioequivalence in the 1990s. The FDA issued draft guidances for public comment in 1997 and 1999 prompting even more extensive international debate among regulators, academia, and industry.

As of the year 2000, no consensus among regulators, academia, and industry has been established as to the appropriateness and applicability of population and individual bioequivalence, though such approaches are intuitively appealing. Average bioequivalence continues to provide a worldwide standard for demonstrating that two products are sufficiently similar to be interchangeable in the marketplace.

ACKNOWLEDGEMENTS

The author gratefully acknowledges support and advice from Byron Jones, N'evine Zariffa, Stephen Senn, Walter Hauck, and Darryl Downing and thanks them for many helpful discussions and support for research on the topic of bioequivalence.

REFERENCES

1. FDA Guidance. Statistical procedures for bioequivalence studies using a standard two treatment crossover design, 1992, 1.
2. FDA Guidance. Bioavailability and Bioequivalence Studies for Orally Administered Drug Products: General Considerations, 2000b, 1.
3. FDA Guidance. Waiver of *In Vivo* Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms, 2000a, 1.
4. Schuirmann, D.J., **J. Pharmacok. Biopharm.**, 1987, 15, 657.
5. Jones, B. and Kenward, M.G., In; Design and Analysis of Crossover Trials. Chapman and Hall, London, 1989,1.
6. Senn, S., **Lancet,** 1998, 352, 85.
7. Hauck, W.W. and Anderson, S., **Stat. Med.**, 1986, 5, 203.
8. Steinijans, V.W., Hartmann, M., Huber, R. and Radtke, H.W., **Int. J. Clin. Pharm. Ther. Toxicol.**, 1991, 29, 323.
9. Barrett, J.S., Batra, V., Chow, A., Cook, J., Gould, A.L., Heller, A., Lo, M.W., Patterson, S.D., Smith, B.P., Stritar, J.A., Vega, J.M. and Zariffa, N., **J. Clin. Pharmacol.**, 2000, 40, 571..
10. Benet, L.Z. and Goyan, J.E., **Pharmacotherapy,** 1995, 15, 433.
11. Colaizzi, J.L. and Lowenthal, D.T., **Clin. Ther.**, 1986, 8, 370.
12. Phillips, K.F. **J. Pharmacok. Biopharm.**, 1990, 18, 137.
13. Barrett, J.S., Batra, V., Chow, A., Cook, J., Gould, A.L., Heller, A., Lo, M.W., Patterson, S.D., Smith, B.P., Stritar, J.A., Vega, J.M. and Zariffa, N. **J. Clin. Pharmacol.**, 2000, 40, 561.
14. Ansbacher, R., **Contraception**, 1990, 43, 139.
15. Blume, H.H. and Midha, K.K., Eds., In; Bio-International: Bioavailability, Bioequivalence, and Pharmacokinetics, Medpharm., Stuttgart, 1993,13.
16. Midha, K.K., Rawson, M.J. and Hubbard, J.W., **J. Pharm. Sci.**, 1997, 86, 1193.

17. Midha, K.K., Rawson, M.J. and Hubbard, J.W., **J. Pharm. Sci.**, 1997, 6, 87.
18. Boddy, A.W., Snikeris, F.C., Kringle, R.O., Wi, G.C-G., Oppermann, J.A. and Midha, K.K., **Pharm. Res.**, 1995,12, 1865.
19. Schall, R. and Williams, R.L., **J. Pharmacok. Biopharm.**, 1996, 24, 133.
20. Patterson, H.D. and Thompson, R., **Biometrika**, 1971, 58, 545–554.
21. Laird, N.M. and Ware, J.H., **Biometrics**, 1982, 38, 963.
22. Gaffney, M., **J. Pharm. Sci.**, 1992, 81, 315.
23. Chinchilli, V.M. and Esinhart, J.D., **Stat. Med.**, 1996,15, 1619.
24. Vonesh, E.F. and Chinchilli, V.M., In; Linear and Nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York, 1997, 1.
25. Hinkelmann, K. and Kempthorne, O., In; Design and Analysis of Experiments, Vol. I, John Wiley and Sons, New York, 1994, 1.
26. Hellriegel, E.T., Bjornsson, T.D. and Hauck, W.W., **Clin. Pharmacol. Ther.**, 1996, 60, 601.
27. Hwang, S., Huber, P.B., Hesney, M. and Kwan, K.C., **J. Pharm. Sci.**, 1978, 67, IV.
28. Peace, K.E., In; Proceedings of the American Statistical Association, Biopharmaceutical Chapter, 1986, 63.
29. Anderson, S. and Hauck W.W., **J. Pharmacok. Biopharm.**, 1990, 18, 259.
30. Hauck, W.W. and Anderson, S., **Int. J. Clin. Pharmacol. Ther. Toxicol.**, 1992, 30, 181.
31. Welleck, S., **Biometrical J.**, 1993, 35, 47.
32. Hwang, J.T.G. and Wang, W., **Biometrika**, 1997, 84, 893.
33. Schuirmann, D.J., **Drug Inform. J.**, 1990, 24, 315.
34. Welleck, S., **J. Biopharm. Stat.**, 1997, 7, 17.
35. Esinhart, J.D. and Chinchilli, V.M., J. Biopharm. Stat., 1994, 4, 39.
36. Chinchilli, V.M., **J. Biopharm. Stat.**, 1996, 6, 1.
37. Brown, E.B., Iyer, H.K. and Wang, C.M., **Stat. Med.**, 1997, 16, 803.
38. Esinhart, J.D. and Chinchilli, V.M., **Int. J. Clin. Pharmacol. Ther.**, 1994, 32, 26.
39. Hauck, W.W, Bois, F.Y., Hyslop, T., Gee, L. and Anderson, S., **Stat. Med.**, 1997, 16, 441.
40. Grahnen, A., Hammarlund, M. and Lundqvist, T., **Eur. J. Clin. Pharmacol.**, 1984, 27, 595.
41. Pitman, E.C.G., **Biometrika**, 1939, 31, 9.
42. Morgan, W.A., **Biometrika**, 1939, 31, 13.
43. Brown, M.B. and Forsythe, A.B., **J. Amer. Stat. Assn.**, 1974, 69, 364.
44. Balakrishnan, N. and Ma, C.W., **J. Stat. Comput. Simul.**, 1990, 35, 41.
45. Bristol, D.R., **J. Biopharm. Stat.**, 1991, 1, 185.
46. Mallet, A., **Biometrika**, 1986, 73, 645.
47. Cornell, R.G., **Commun. Stat. Theor. Meth.**, 1991, 20, 1099.
48. Chow, S.C. and Liu, J., In; Design and Analysis of Bioavailability and Bioequivalence Studies, 2nd Edn., Marcel Dekker, New York, 2000, 1.
49. Wang, W., **J. Amer. Stat. Assn.**, 1997, 88, 939.
50. Guilbaud, O., **J. Amer. Stat. Assn.**, 1993, 88, 939.
51. Guilbaud, O., **Drug Inform. J.**, 1999, 33, 455.
52. Kepner, J.L. and Randles, R.H., **J. Amer. Stat. Assn.**, 1982, 77, 475.
53. Bhoj, D.S., **Biometrika**, 1979, 66, 681.
54. Ekbohm, G., **Biomed. J**, 1981, 23, 261.
55. McCulloch, C.E., **Commun. Stat. Theor. Meth.**, 1987, 16, 1377.
56. Bauer, P. and Bauer, M.M., **Biomed. J**, 1994, 6, 643.
57. Zariffa, N.M–D., Patterson, S.D., Boyle, D. and Hyneck, H., **Stat. Med.**, 2000, 19, 2811.
58. Zariffa, N.M–D. and Patterson, S.D., In; AAPS/FDA Workshop on Biopharmaceutics in the New Millenium: Regulatory Approaches to Bioavailability and Bioequivalence. 2000, 1.
59. Bristol, D.R., **J. Biopharm. Stat.**, 1991, 1, 237.
60. Endrenyi, L. and Tothfalusi, L., **Pharm. Res.**, 1999, 16, 186.
61. Davidian, M. and Giltinan, D.M., In; Nonlinear Models for Repeated Measurement Data. Chapman and Hall, London. 1995, 1.
62. Ekbohm, G. and Melander, H., **Biometrics**, 1989, 45, 1249.
63. Patterson, H.D., **Biometrika**, 1950, 39, 32.
64. Harville, D.A., **J. Amer. Stat. Assn.**, 1977, 72, 320. .
65. Sheiner, L.B., **Stat. Med.**, 1992, 11, 1777.
66. Schall, R. and Luus, H.G., **Stat. Med.**, 1993, 2, 1109.
67. Schall, R., **Biometrics**, 1995, 51, 615.
68. Sheiner L.B., Beal, S.L. and Sambol, N.C., **Clin. Pharm. Ther.**, 1989, 46, 63.
69. Bickel, P.J. and Doksum, K.A., In; Mathematical Statistics. Holden Day, San Fransisco, 1977, 1.
70. Huitson, A., **Biometrika**, 1955, 42, 471.
71. Fleiss, J.L., **J. Amer. Stat. Assn.**, 1971, 66, 142.
72. Howe, W.G., **J. Amer. Stat. Assn.**, 1974, 69, 789.
73. Harville, D.A., **Biometrics**, 1976, 32, 403.
74. Burdick, R.K. and Sielken, R.L., **J. Amer. Stat. Assn.**, 1978, 73, 632.
75. Graybill, F.A. and Wang, C.H., **J. Amer. Stat. Assn.**, 1980, 75, 869.
76. Lu, T.C., Graybill, F.A. and Burdick, R.K., **J. Stat. Plan. Inf.**, 1988, 18, 35.
77. Ting, N., Burdick, R.K., Graybill, F.A., Jeyaratnam, S. and Lu, T-F.C., **J. Stat. Comput. Simul.**, 1990, 35, 135.
78. Wang, C.M., **Biometrics**, 1990, 46, 187.
79. Burdick, R.K. and Graybill, F.A., In; Confidence intervals on variance components. Marcel Dekker, New York, 1992, 1.
80. Holder, D. and Hsuan, F., **Biometrika**, 1993, 80, 835.
81. Holder, D. and Hsuan, F., **Drug Inform. J.**, 1993, 29, 965.
82. Hyslop, T., Hsuan, F. and Holder, D.J., **Stat. Med.**, 2000, 19, 2885.
83. Kimanani, E.K., Lavigne, J. and Potvin, D., **Stat. Med.**, 2000, 19, 2775.

84. Anderson, S., **Biopharm. Rep.**, 1993, 2, 1..
85. Hauck, W.W. and Anderson S., **J. Pharmacok. Biopharm**, 1994, 22, 551.
86. Anderson, S., **Drug Inform. J.**, 1995, 29, 961.
87. Anderson, S. and Hauck W.W., **Int. J. Clin. Pharmacol. Ther.**, 1996, 34, 369.
88. Hauck, W.W., **Int. J. Clin. Pharmacol.Ther.**, 1996, 34, 535.
89. Hauck, W.W, Chen, ML., Hyslop, T., Patnaik, P., Shuirmann, D. and Williams, R. for the FDA Individual Bioequivalence Working Group, **Int. J. Clin. Pharmacol. Ther.Toxicol.**, 1996, 34, 535.
90. Patnaik, R.N., Lesko, L.J., Chen, ML., Williams, R.J. and the FDA Individual Bioequivalence Working Group. **Clin Pharmacok.**, 1997, 33, 1.
91. Gould, A.L., **J. Biopharm. Stat.**, 1997, 7, 23.
92. Chen, M.L., **J. Biopharm. Stat.**, 1997, 7, 5.
93. Bauer, P. and Kieser, M., **Biometrika**, 1996, 83, 934.
94. Ghosh, K., Hesney, M., Sarkar, S. and Thiyagarajan, B., In; Proceedings of the American Statistical Association, Biopharmaceutics Chapter, 1996, 77.
95. Endrenyi, L. and Schulz, M., **Drug Inform. J.**, 1993, 27, 195.
96. Endrenyi, L., **Int. J. Clin. Pharmacol. Ther.**, 1994, 32, 497.
97. Vuorinen, J. and Turunen, J., **Stat. Med.**, 1996, 15, 2635.
98. Vuorinen, J., **Stat. Med.**, 1997, 16, 2229.
99. Gould, A.L., **Stat. Med.**, 2000, 19, 2721.
100. Munk, A., **Biometrics**, 1993, 49, 1225.
101. Hsu, J.C., Hwang, J-T.G., Liu, H-K. and Ruberg, S.J., **Biometrika**, 1994, 81, 103.
102. Brown, L.D., Hwang, J.T.G. and Munk, A., **Annals Stat.**, 1997, 25, 2345.
103. Wang, W., **J. Amer. Stat. Assn.**, 1999, 94, 880.
104. Berger, R.L., **Technometrics**, 1992, 24, 295.
105. Berger, R.L. and Hsu, J.C., **Stat. Sci.**, 1996, 11, 283.
106. Chinchilli, V.M. and Elswick, R.K., **J. Biopharm. Stat.**, 1997, 7, 113.
107. Munk, A. and Pfuger, R., **J. Amer. Stat. Assn.**, 1999, 94, 1311.
108. Mauger, D.T. and Chinchilli, V.M., **Stat. Med.**, 2000, 19, 2855.
109. Hauck, W.W., Hyslop, T., Anderson, S., Bois, F.Y. and Tozer, T.N., **Clin. Res. Reg. Affairs**, 1995, 12, 249.
110. Senn, S. In; Conference Proceedings: Challenging Statistical Issues in Clinical Trials. Decisions and Bioequivalence, 2000, 1.
111. FDA Draft Guidance. *In vivo* bioequivalence studies based on population and individual bioequivalence approaches. 1997
112. Efron, B. and Tibshirani, R.J., In; An Introduction to the Bootstrap, Chapman and Hall, New York, 1993, 1.
113. Endrenyi, L., Amidon, G.L., Midha, K.K. and Skelly, J.P., **Pharm. Res.**, 1998, 15, 1321.
114. Endrenyi, L. and Hao, Y., **Int. J Clin. Pharmacol. Ther. Toxicol.**, 1998, 36, 1.
115. Endrenyi, L., Taback, N. and Tothfalusi, L., **Stat. Med.**, 2000, 19, 2867.
116. Endrenyi, L. and Midha, K.K., **Eur. J. Pharm. Sci.**, 1998, 6, 271.
117. Lindley, D.V., **Stat. Sci.**, 1998, 13, 136.
118. Westlake, W.J., In; Peace, K., Eds., Biopharmaceutical Statistics for Drug Development, Marcel Dekker, New York, 1986, 329.
119. Hwang, J.S., **Statistica Sinica**, 1996, 6, 663.
120. Lindley, D.V., In; Making Decisions, 2nd Edn., John Wiley and Sons, New York, 1971, 1.
121. Lindley, D.V. and Singpurwalla, N.D., **J. Amer. Stat. Assn.**, 1991, 86, 933.
122. Lindley, D.V., **The Statistician**, 1997, 46, 129.
123. Zariffa, N.M-D., Patterson, S.D., Boyle, D. and Hyneck, H., In; FDA/AAPS Workshop on Bioavailability and Bioequivalence, 1998.
124. Patterson, S.D., Zariffa, N.M-D., Boyle, D. and Hyneck, M., In; Annual Midwest Biopharmaceutics Workshop, 1998.
125. Schumaker, R.C. and Metzler, C.M., **Drug Inform. J.**, 32, 1063-1072.
126. Steinjans, V.W. and Diletti, E., **J. Biopharm. Stat.**, 1998, 7, 31.
127. Hauschke, D. and Steinijans, V.W., **Stat. Med.**, 2000, 19, 2769.
128. Kimanani, E.K. and Potvin, D., **J. Pharmacok. Biopharm.**, 1997, 25, 595.
129. Chen, M-L., Patnaik, R., Hauck, W.H., Schuirmann, D.J., Hyslop, T. and Williams, R., **Stat. Med.**, 2000, 19, 2821.
130. Williams, R.L., Adams, W., Chen, M-L., Hare, D., Hussain, A., Lesko, L., Patnaik, R., Shah, V., and FDA Biopharmaceutics Coordinating Committee., **Eur. J. Drug Metab. Pharma.**, 2000, 25, 7.
131. Williams, R.L., Patnaik, R.N. and Chen, M-L., **Eur. J. Drug Metab. Pharm.**, 2000, 25, 13.
132. FDA Draft Guidance., BA and BE Studies for Orally Administered Drug Products: General Considerations, 1999.
133. FDA Draft Guidance., Average, Population, and Individual Approaches to Establishing Bioequivalence. Based on a Biopharmaceutics Classification System, 1999.
134. Longford, N.T., **Stat. Med.**, 1999, 18, 1467.
135. Longford, N.T., **Statistica Neerlandica**, 2000, 14, 14.
136. Colburn, W.A. and Keefe, D.L., **J. Clin. Pharmacol.**, 2000, 40, 559.
137. Dragalin, V. and Fedorov, V.. In; SB BDS Technical Report 1999-03.
138. Dragalin, V. and Fedorov, V., In; SB BDS Technical Report 1999-04.
139. Lin, L., **Biometrics**, 1989, 45, 255.
140. Lin, L., **Biometrics**, 1992, 48, 599.
141. Lin, L., **Stat. Med.**, 2000, 19, 255.
142. Patterson, S.D. and Zariffa, N.M-D., In; AAPS International workshop on Individual Bioequivalence: Realities and Implementation,1999.

143. Zariffa, N.M-D. and Patterson, S.D., In; AAPS International workshop on Individual Bioequivalence: Realities and Implementation, 1999.
144. Patterson, S.D. and Zariffa, N.M-D., In; Annual Meeting of the American Society of Clinical Pharmacology and Therapeutics, 2000.
145. Ormsby, E.D., In; AAPS International workshop on Individual Bioequivalence: Realities and Implementation, 1999.
146. Pound, N.J., In; AAPS International workshop on Individual Bioequivalence: Realities and Implementation. 1999.
147. Hauck, W.W., Hyslop, T., Chen, M.L., Patnaik, R., Williams, R.L., and the FDA Population/Individual Bioequivalence Working Group., **Pharm. Res.**, 2000, 17, 375.
148. Singh, G.J-P., Adams, W.P., Lesko, L.J., Shah, V.P., Molzon, J.A., Williams, R.L. and Pershing, L.K., **Clin. Pharmacol. Ther.,** 1999, 66, 346.
149. Meyer, M.C., Straughn, A.B., Jarvi, E.J., Patrick, K.S., Pelsor, F.R., Williams, R.L., Patnaik, R., Chen, M.L. and Shah, V.P., **Pharm. Res.,** 2000, 17, 381.
150. Bekersky, I., Dressler, D., Colburn, W. and Mekki, Q., **J. Clin. Pharmacol.**,1999, 39, 1032.
151. Cerutti, R., Rivolta, G., Cavalieri, L., Di Giulio, C., Grossi, E., Vago, T., Baldi, G., Righini, V. and Marzo, A., **Pharm. Res.**, 1999, 39, 193.
152. Canafax, D.M., Irish, W.D., Moran, H.B., Squiers, E., Levy, R., Pouletty, P., First, M.R. and Christians, U., **Pharmacol.**, 1999, 59, 78.
153. Yacobi, A., Masson, E., Moros, D., Ganes, D., Lapointe, C., Abol-fathi, Z., LeBel, M., Golander, Y., Doepner, D., Blumberg, T., Co-hen, Y. and Levitt, B., **J. Clin. Pharmacol.**, 2000, 40, 826.