STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2785–2799 (DOI: 10.1002/sim.743)

Statistical issues in bioequivalance

Stephen Senn*,†

Department of Epidemiology and Public Health, Department of Statistical Science, University College London, Gower Street, London, U.K.

SUMMARY

There has been much work recently on individual bioequivalence and the topic has attracted considerable controversy. Some previous controversies regarding average bioequivalence are examined. It is argued that a contributory factor in these controversies may have been confusion over the purpose of bioequivalence trials. It is concluded that this purpose needs further clarification before guidelines for individual bioequivalence can be established and indeed that such guidelines may turn out to be unnecessary. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Those unfamiliar with the pharmaceutical industry are sometimes surprised that bioequivalence is an issue at all. It seems obvious to some that generic and brand name formulations of the same molecule must be the same product, but for some irrelevant packaging and a considerable difference in price. It is, after all, a central tenet of chemistry that if two molecules have identical constituent atoms in identical arrangements they are identical. In fact, however, the history of pharmacology has been one of surprises in this respect. With benefit of hindsight, and our more modern stereoscopic view of chemistry, we now understand that such identity of arrangement requires identity in three dimensions and not just two. We no longer find it amazing that left hand and right hand forms of the same molecule (optical isomers) should show different effects in man. Yet, to the chemists of the 19th century this must have seemed a strange and puzzling phenomenon. In fact we also now appreciate that different formulations of the same products can differ greatly in term of their *bioavailability*, 'the rate and extent to which a dose of a test drug reaches the systemic circulation' (reference [1], p. 23). These differences in bioavailability are, of course, translated into differences in potency so that if two formulations do not have the same bioavailability and are not, therefore, bioequivalent, they are likely to differ in potency with attendant risks of lack of

*Correspondence to: Stephen Senn, Department of Epidemiology and Public Health, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.
†E-mail: stephens@public-health.ucl.ac.uk

effectiveness or lack of tolerability if the evidence of suitability of the one is mistakenly assumed to apply to the other.

These concerns are not just theoretical but can turn out to be justified in practice. For example, a trial in which two different dry powder formulations of formoterol, both developed by the same sponsor, were compared showed a four-fold difference in potency, to the surprise of those who had designed the trial [2]. As a consequence, where the evidence of efficacy and tolerability of an already registered formulation of a drug is given as a justification for the safety and efficacy of a new formulation, regulators generally require sponsors to show that the new formulation is effectively equivalent to the standard. By this means the sponsor for the new formulation, who might be a generic manufacturer or the original innovator seeking registration of an alternative formulation, hope to obviate the necessity of repeating an expensive and time consuming full development programme.

The means by which such equivalence has been proved until now has been using so-called bioequivalence studies. These often but not invariably use healthy volunteers, and again usually, but not always, a cross-over trial to compare the concentration time profile in the blood of the active ingredient from the two formulations. Typically, 12 to 30 subjects are studied, blood samples being taken frequently from each subject, and the concentration–time profiles are then compared using certain summary measures. Particularly important amongst these summary measures is the area under the concentration–time curve, the AUC, although frequently the observed concentration maximum, $C_{\max}$ and sometimes the time to reach this maximum $T_{\max}$ are also studied.

In the rest of this paper two statistical controversies regarding the analysis of such cross-over trials will be discussed in some detail. The first is an old one concerning the appropriate approach to comparing so-called average bioequivalence and the second a more recent controversy regarding individual bioequivalence. Both of these concepts will be clarified in due course. Some aspects of these two issues have received some previous discussion in *Statistical Issues in Drug Development* [3]. The reader who is interested in knowing more about the background to bioequivalence is referred to that text and the useful articles by Steinijans and Hauschke [4, 5]. Although this is not essential to the discussion, it will implicitly be assumed in what follows that two formulations are being compared in a cross-over study as regards the measure AUC. We now consider these two controversies in turn.

2. AVERAGE BIOEQUIVALENCE

2.1. Basic position

Whatever statistical test or procedure then follows, a common approach to comparing the bioavailability of two products involves the following stages:

(i) running an AB/BA cross-over in which so-called test (T) and reference (R) formulations are compared;
(ii) log transforming the AUCs measured in the trial;
(iii) fitting a linear model to the log-AUCs in which subject and period effects are eliminated to produce an estimate, l, of the 'formulation effect', the log-relative bioavailability, λ;
(iv) estimating the standard error of the estimated formulation effect;

(v) comparing the results to pre-established limits of equivalence, δ_1 (a lower limit) and δ_2 (an upper limit). It is customary now to use the limits $\log(0.8)$ and $\log(1.25)$ on the log-AUC scale, so that $\delta_1 = -\delta_2$ where $\delta_2 = \delta = 0.223$.

The precise details of the circumstances under which the comparison of point estimate and limits of equivalence will lead to a declaration of equivalence vary from scheme to scheme. Some approaches will be discussed in due course.

It is sometimes erroneously maintained that what is being compared is the median bioavailability of the two formulations. The justification for this observation is that the location parameter of the log-Normal distribution is, in fact, the median. Hence, if two distributions of Normally distributed log-AUCs are being compared as regards location, it is the medians of the distributions of the two original observations that are being compared. In fact, this argument is false. Fitting the subject effect is a form of conditional analysis and the only level at which any assumption regarding Normality is required is at the level of the disturbance terms for the linear model. Indeed, as is well known, the analysis of a cross-over trial can be reduced to an analysis of so-called basic estimators, these basic estimators being the contrast of interest defined for an individual subject [6]. Thus, analysis can proceed by calculating first for each subject the difference in log-AUC, test minus reference, averaging these differences within sequences and then averaging the sequence means. If, as is usually the case, an answer is wanted in terms of relative AUC, this result can then be anti-logged.

An absolutely equivalent approach would be to calculate ratios T to R for each subject, obtain geometric means within sequences and then geometric means over the two sequences. Whichever route is taken to calculating this relative bioavailability, this calculation in itself carries with it no implication regarding any population of any sort. In particular, the mean, or for that matter the median, bioavailabilities of the test and reference formulations are completely irrelevant. It is the relative bioavailability that is being directly measured on each subject and inference is directly about this. This discussion is deferred for the moment and will be resumed later when considering so-called individual bioequivalence. For the moment, it will simply be assumed that the analysis of the bioequivalence data produces a single statistic, l, and associated standard error and that this statistic is an estimate of the relative bioavailability λ, of the test compared to the reference formulation. The simplest interpretation of λ is then if 'strict additivity' applies and this relative bioavailability is assumed identical for every subject in the trial and indeed if this additivity is assumed to apply beyond the trial to patients as well. (This is one justification for using healthy volunteers rather than patients to study bioequivalence.) It is the analysis of bioequivalence under these circumstances that corresponds to what has sometimes been referred to as the comparison of *average bioequivalence.*

In the following paragraphs, various approaches to the use of the point estimate of relative bioavailability for deciding whether average bioequivalence applies or not will be described. To simplify such discussion it will be assumed that the nuisance parameter, the variance of the relative bioavailability, is known, so that inference may be based upon the Normal distribution rather than the t-distribution. The complication that relaxing this assumption introduces will be considered briefly when discussing 'optimal' Neyman–Pearson approaches.

2.2. Westlake's symmetric confidence intervals approach

Westlake pointed out that conventional hypothesis testing is inappropriate in the context of bioequivalence [7, 8]. This is because if the null hypothesis of strict equality, $H_0 : \delta = 0$, is

tested and the type I error rate of the associated test is controlled, then it is the probability of falsely declaring inequivalence that is being controlled. However, to the extent that the Neyman–Pearson approach is accepted as being relevant, it would seem to be appropriate to control the error of falsely declaring equivalence. Instead, Westlake suggested basing a decision on 95 per cent confidence intervals. However, instead of using conventional confidence intervals symmetric (on the log-scale) about the point estimate, he proposed using confidence intervals centred on the point of exact equivalence.

Let W, be the upper Westlake limit. Then, since the limits are symmetrical about $\log(1)=0$, $-W$ is the lower Westlake limit. W is the root of the equation

$$\Phi\left(\frac{W-l}{\mathrm{SE}(l)}\right)-\Phi\left(\frac{-W-l}{\mathrm{SE}(l)}\right)=0.95 \tag{1}$$

where, $\Phi(\,)$ is the Normal distribution function and $\mathrm{SE}(l)$ is the standard error for the point estimate of (log) relative bioavailability. O'Quigley and Baudoin [9] point to a simple interpretation in terms of a Fisherian fiducial distribution for the relative bioavailability parameter. (This fiducial distribution can, in turn, be thought of as a Bayesian posterior distribution for an uninformative prior.) Consider the case where the upper Westlake limit coincides exactly with the upper limits of equivalence so that $W=\delta$. Since the limits are symmetric about 0, it thus follows that the lower Westlake limit coincides with the lower limit of equivalence. Then in that case, the probability of equivalence, on this interpretation, is exactly 95 per cent. Thus, provided that the Westlake limits lie exactly within the limits of equivalence, the probability of equivalence is at least 95 per cent.

2.3. Kirkwood's conventional confidence limits approach

Kirkwood proposed as an alternative to Westlake's approach that conventional 95 per cent limits centred on the point estimate should be calculated instead [10]. Thus two limits

$$K_{\mathrm{L}}=l+\mathrm{SE}(l)\Phi^{-1}(0.025)\quad\text{and}\quad K_{\mathrm{U}}=l+\mathrm{SE}(l)\Phi^{-1}(0.975) \tag{2}$$

are calculated, where $\Phi^{-1}(\,)$ is the inverse distribution function for the Normal.

The fiducial/Bayesian interpretation of this is that the interval contained by the limits is the most likely (highest posterior density) region such that the true relative bioavailability lies with 95 per cent probability within the region [9, 11]. Thus the difference between Westlake's and Kirkwood's approach, on this interpretation, is that the former accepts equivalence if the probability of equivalence is at least 95 per cent, whereas the latter accepts equivalence if the limits of equivalence are within the 'shortest' 95 per cent region.

Consider the case where the upper Kirkwood limit is just equal to the upper limit of equivalence so that $K_{\mathrm{U}}=\delta$. Then, under the fiducial/Bayes interpretation, the probability of 'superavailability' (that test is more bioavailable than reference) is exactly equal to 0.025. However, unless the point estimate is 0, in which case Westlake and Kirkwood intervals coincide then, either the point estimate is below 0, in which case $K_{\mathrm{L}}<-\delta$ and equivalence will not be declared, or $l>0$, in which case $K_{\mathrm{L}}>-\delta$, from which it follows that the probability of 'subavailability' is less than 0.025. Thus Kirkwood's approach can be seen as being more stringent than Westlake's in that it makes a separate requirement that the probabilities of subavailability and superavailability should each be less than 0.025. Westlake's approach, however, merely requires that their sum should not exceed 0.05 [11].

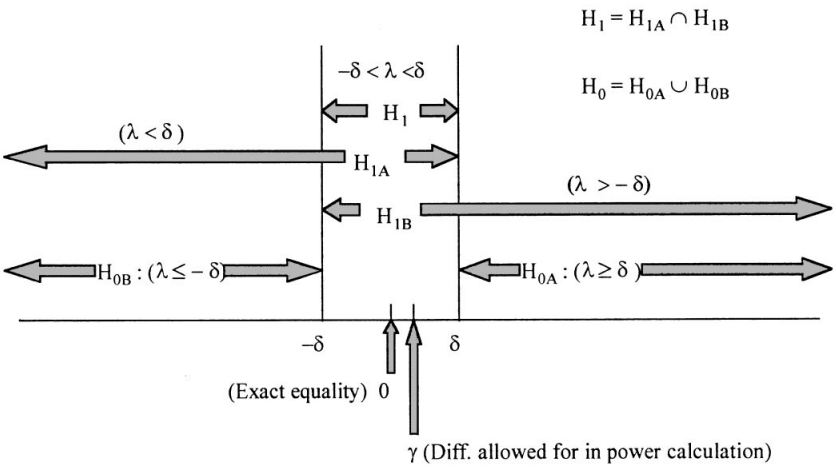

Figure 1. Illustration of null and alternative hypotheses in connection with bioequivalence testing.

2.4. *Schuirmann's two one-sided test approach* (*TOST*)

Schuirmann proposed an approach whereby the problem was formulated in terms of two independently performed one-sided hypothesis tests [12]. The first test is a test of the null hypothesis that the test formulation is superavailable against the alternative hypothesis that it is not. Referring to Figure 1, which is taken from *Statistical Issues in Drug Development* [3], this is a test of $H_{0A}:\lambda \geqslant \delta$ against $H_{1A}:\lambda < \delta$. Independently of the results of this first test, a second test, of the null hypothesis that the test formulation is subavailable against the alternative that it is not, is carried out. This is a test of $H_{0B}:\lambda \leqslant -\delta$ against $H_{1A}:\lambda > -\delta$. If both null hypotheses are rejected, this leads to the assertion of both alternatives and hence their intersection H_1. If both tests are carried out at the 2.5 per cent level then this is equivalent to Kirkwood's confidence interval approach. In fact, Schuirmann proposed carrying out each test at the 5 per cent level and this is then equivalent to calculating conventional 90 per cent confidence intervals and checking that these lie between the limits of equivalence. At the time of writing this is the internationally agreed approach to bioequivalence.

2.5. *Comparison of the above three procedures*

If we consider the two confidence intervals as testing approaches also, then we can construct critical values for the point estimate for all three procedures. In what follows we consider the upper critical values for the point estimate only. The Westlake critical values are the most difficult to establish. However, W is the root of equation (1) for fixed l and $\mathrm{SE}(l)$. Considering W then as a function of l for given $\mathrm{SE}(l)$ so that we may write $W(l;\mathrm{SE}(l))$, the root of a further equation

$$\delta - W(l;\mathrm{SE}(l)) = 0 \tag{3}$$

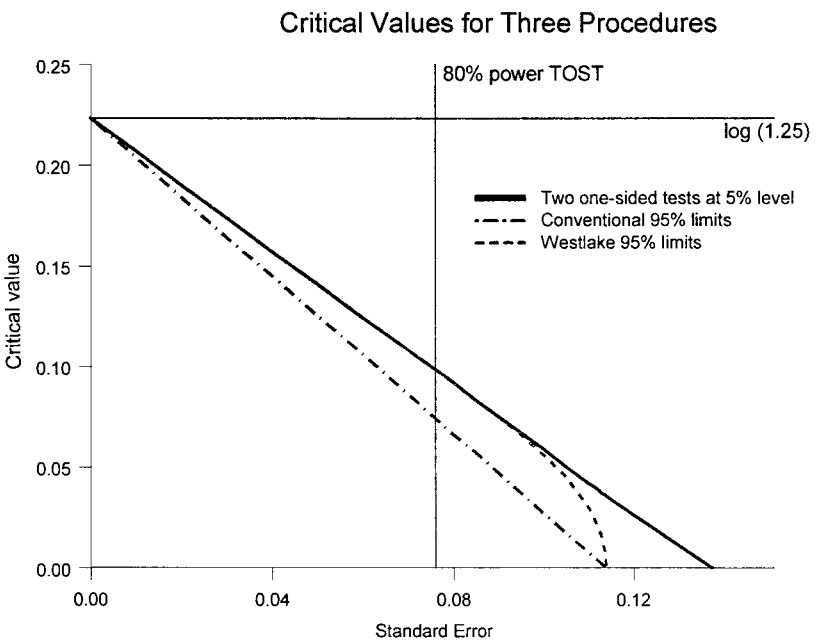

Figure 2. Critical values for Westlake, conventional (Kirkwood) and TOST procedures: critical values of log relative bioavailability as a function of the standard error.

provides the solution for the critical value of l in terms of SE(l). The critical values for the Kirkwood and Schuirmann procedures are given by

$$\delta - \mathrm{SE}(l)\Phi^{-1}(0.975) \quad \text{and} \quad \delta - \mathrm{SE}(l)\Phi^{-1}(0.95) \tag{4}$$

respectively.

Figure 2, shows a plot of critical values for all three procedures as a function of the standard error. It is noticeable that for high precision (low standard error) the critical value for the Westlake procedure is effectively the same as that for Schuirmann's TOST procedure, whereas at moderate precision it departs, eventually joining Kirkwood's procedure for low precision. To understand why this is so, return to the fiducial interpretation of O'Quigley and Baudoin [9]. Westlake's procedure requires that the sum of the probabilities of superavailability and subavailability should be less than 5 per cent. However, if precision is extremely high, the critical value of the point estimate will be close to the upper limit of equivalence. That being so, the probability of subavailability is effectively zero so that the probability of superavailability can be allowed to reach 5 per cent. However, Schuirmann's approach is to carry out each test at the 5 per cent level, hence the similarity of the two approaches. As, however, the standard error increases, the critical value of Westlake's approach is pushed closer and closer to zero. Eventually, at the value of 0 the probabilities of subavailability and superavailability are identical and equal to 2.5 per cent. However, this is equivalent to observing that conventional 95 per cent limits are equal to the limits of equivalence, which is Kirkwood's approach.

2.6. A Neyman–Pearson type test

If Figure 2 is considered, it will be seen that for all three procedures, the upper critical value eventually approaches zero. Since the procedures are symmetric in terms of critical values and the lower critical value is simply the negative of the upper one, at this point the two critical values meet and the procedures thus have a type I error rate or size of zero. This is simply understood in terms of confidence intervals. If the standard error is large enough, the width of the confidence intervals will exceed the width of the region of equivalence. Thus, under these circumstances, whatever the point estimate, equivalence cannot be declared. Thus if we seek a procedure that under all circumstances, whatever the standard error, provides the same type I error rate, whether 5 per cent or 2.5 per cent, then these procedures clearly do not fit the bill.

Consider again Figure 1. One approach is to work directly with the union hypothesis H_0 and the intersection hypothesis H_1 and to try and develop a test in a single step using these. This is an example of the problem of general interval testing which has been extensively developed and discussed in an important but neglected paper by Mehring [13], building on work by Karlin [14], who shows that a general 'monotone test procedure' has the required property for the wide variety of probability distributions defined by the Pólya family.

Consider a pair of symmetric critical values $-c$ and c for l. Now suppose that the test formulation is just superavailable so that $\lambda = \delta$. The power function for a test based on l, is then

$$\Phi\left(\frac{c-\delta}{\mathrm{SE}(l)}\right) - \Phi\left(\frac{-c-\delta}{\mathrm{SE}(l)}\right) \tag{5}$$

If the power function is set equal to 0.05 and the resulting equation is solved for c, it turns out that this defines a test of level 0.05. This procedure will be referred to as the NP test.

Illustration of the properties of the NP test will be deferred until a further procedure is introduced. However, it should be pointed out here that whereas the Westlake, Kirkwood and TOST procedures can be readily adapted for the realistic case where the standard error is not known but has to be estimated by using the t-distribution, attempting to do the same for the NP procedure is not possible. For example, replacing (5) by a power function based on the t-distribution as proposed by Anderson and Hauck [15] does not produce tests of correct size, the theory of monotone tests for the Pólya distributions not applying to the relevant shifted t-distribution [13]. Some rather ingenious and complicated schemes have been proposed to deal with this problem [16, 17]. For reasons that will be explained below, however, these schemes are unlikely to win popular support.

2.7. Lindley's expected loss approach

A recent proposal of Lindley's for determining bioequivalence [18], ostensibly a comment on the NP approach of Berger and Hsu, but in reality criticizing only the TOST procedure that they themselves sought to replace, is to work directly via loss-functions. Lindley supposes that a loss function of the form

$$L(\lambda) = A - (A+B)\exp\{-(1/2)\lambda^2/c^2\} \tag{6}$$

is associated with a declaration of equivalence, where A, B and c are suitable constants. This function has loss $-B$ at exact equivalence, when $\lambda = 0$, and loss A at extreme inequivalence

when λ is infinite. As Lindley points out, the scale of A and B is arbitrary so that their sum can be set equal to 1. Lindley further supposes that the loss at the boundary of the limit of equivalence, when $\lambda = \delta$ should be zero. This implies that

$$c = \sqrt{(-\delta^2/\{2\ln(A)\})} \tag{7}$$

Lindley considers the case where the ratio of B to A is 19 to 1. For $\delta = \log(1.25) = 0.223$, (7) yields a solution for the value of c of 0.697. These combinations of parameters imply that the loss of declaring equivalence when the true relative bioavailability is 1.38 is the same as in failing to declare equivalence when the ratio is 1.

Next Lindley considers the expected loss

$$\int_{-\infty}^{\infty} L(\lambda)\, p(\lambda)\, \mathrm{d}\lambda = A - c(c^2+\sigma^2)^{-1/2} \exp\{-l^2/\{2(c^2+\sigma^2)\}\} \tag{8}$$

where $p(\lambda)$ is the posterior distribution of λ and σ^2 is the posterior variance of this distribution. For an uninformative prior we should have $\sigma = \mathrm{SE}(l)$. Since the critical value for such a procedure is obtained at the point of indifference when the loss function is zero, (8) may be set equal to zero and solved accordingly. This yields the solution

$$l = \sqrt{((\sigma^2+c^2)[\log\{c^2/(c^2+\sigma^2)\} + \delta^2/c^2])} \tag{9}$$

2.8. Comparison of TOST, NP and Lindley's approach

Figure 3 gives a comparison of Lindley's approach to the TOST and NP approaches as a function of the standard error of l. In the case of Lindley's approach this standard error is the standard error of the posterior distribution, although given an uninformative prior this is, in fact, the same as the sample standard error.

What is remarkable is how radically different these procedures are for large values of the standard error. For infinite precision, they all three accept a declaration of equivalence for any value of l inside the region of equivalence. As the standard error increases, however, Lindley's procedure differs markedly, being considerably more liberal than the other two. A bioequivalence declaration is in practice possible provided that the standard error does not exceed 2.29. On the other hand, if TOST is applied, a declaration of equivalence is only possible if $l < \log(1.25)/\Phi^{-1}(0.95) = 0.136$. The TOST and the NP procedure agree fairly closely until the standard error exceeds 0.1, from which point onwards the NP procedure is more liberal. In fact the acceptable critical value for the NP procedure actually rises eventually with increasing standard error and can even eventually exceed δ. In other words a declaration of equivalence can be accepted even though the point estimate is outside the limit of equivalence. It is extremely improbable that a regulator would accept the declaration of equivalence under such circumstances.

Lindley himself suggests that the loss function he considers may not be steep enough. This would appear to be almost certainly the case, since, for example, for a true relative bioavailability of 3, which it would seem to be absolutely catastrophic for any regulator to accept, the loss is only 0.66, or only 13 times that of failing to declare equivalence when equivalence is perfect. In fact, Lindley's procedure represents such a radical departure from current approaches that to implement it would probably require a complete rethink of accepted limits of bioequivalence as well.

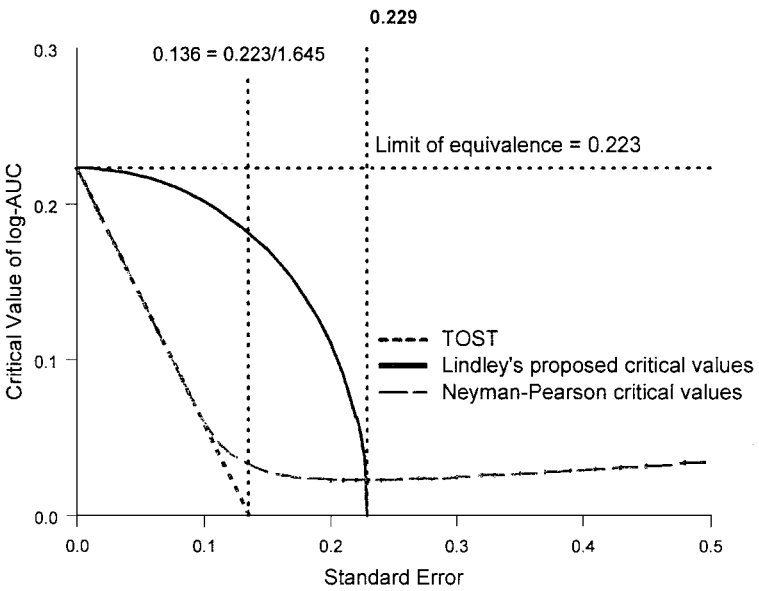

Figure 3. Comparison of TOST, NP and Lindley procedures in terms of critical value for the point estimate for log relative bioavailability as a function of the standard error.

It might also be argued that one should be careful in accepting more liberal procedures than those currently applied. Both the Lindley and NP approaches represent more formal attempts to apply decision rules than those currently employed, Lindley's approach being the most complete in this respect. However, as mentioned by Lindley himself, an implicit assumption is that only two choices can be made: between rejecting the new formulation once and for all and accepting it once and for all. However, although an acceptance is in some sense irreversible, in that it will have the irrevocable consequences of exposing some patients to the new formulation, a rejection is not, since the regulator would usually accept further experimentation on the sponsor's part. Thus, there is really a three-way choice between 'accept now', 'study further' and 'abandon'. This more extensive decision problem warrants further careful examination. However, it seems at least intuitively plausible that the value of further study will be greater when the standard error is large. The fact, therefore, that the greatest differences between the various approaches occurs for large standard errors should warn that caution is appropriate.

However, if the specific implementation of the Bayesian decision-analytic approach proposed by Lindley does not seem particularly attractive, it does not therefore follow that his general approach is without interest. On the contrary, bioequivalence is one of the most structured areas in clinical drug development and as such is a promising candidate for applying Bayesian decision analysis.

2.9. Summary of average bioequivalence

There is a surprising variety of statistical approaches to bioequivalence. These approaches also show a disturbing degree of practical disagreement. One possible explanation is that this

disagreement arises because consensus on the objective of bioequivalence requirements has not been agreed. For example, even remaining within the Bayesian framework, and assuming an uninformative prior, Westlake's approach can be regarded as being appropriate if the objective is that the posterior probability of equivalence should exceed some value, conventional confidence intervals if neither the posterior probability of superavailability nor that of subavailability should be less than some particular value, and Lindley's approach given an explicit consideration of losses [18]. Whatever the reason, however, this disagreement should encourage a degree of caution in extending methods for average bioequivalence to other aspects of equivalence.

3. INDIVIDUAL BIOEQUIVALENCE

3.1. Background

It is theoretically possible that two formulations could be bioequivalent in some average sense but still not produce identical results when applied to a given patient. One reason might be that one formulation might be more variable than another. There are two plausible possible causes of such variability: one is to do with quality of manufacture and the other is to do with route of administration.

If a generic manufacturer were unable to reproduce the same degree of control of the manufacturing process that the innovator company achieved, then between- and within-batch variability might be increased. The generic manufacturer might nevertheless be able to adjust and calibrate the process so that on average a similar bioavailability was produced to the brand-name manufacturer. If that were so it is then conceivable that a product that was passed as being equivalent using any of the techniques would then not be as 'prescribable', as Anderson and Hauck put it [19]. Also, if the reference formulation was a suppository but the test formulation was a tablet, then it is likely that a higher fraction of the former would be absorbed compared to the latter. The unit doses could be adjusted to make the average bioavailability the same. However absorption from the oral formulation will be subject to the influences of a number of body parts (for example, mouth, oesophagus, stomach) that have no influence on the suppository's absorption. Hence variability might be greater.

A further source of variation is conceivable. It might be the case that in a given patient population, two formulations might have similar bioavailability both in terms of location and dispersion but that subgroups of patients could be identified for whom the test was more bioavailable than the reference and groups for whom the reverse was the case. In the terminology of Anderson and Hauck, the formulations would then not be 'switchable' since a patient might experience a difference in effect in being switched from one to the other [19]. Note, however, that if this is the case strict additivity no longer applies. A consequence then is that the observed identity of the two formulations is population specific. Alternative populations could be found in which the formulations did not show the same average bioavailability and a logical consequence of accepting the importance of this phenomenon would seem to be studying bioequivalence in patients rather than volunteers [20].

In 1997 the American Food and Drug Administration (FDA) put forward a draft 'Guidance for Industry', which it has since (1999) updated [21], proposing means of addressing these further concerns. A number of papers from an associated working group have also ap-

peared [22–24]. The guideline is addressed to sponsors of new drug applications (NDAs) and abbreviated new drug applications (ANDAs) and proposes that average bioequivalence should be replaced by notions of *population* and *individual* bioequivalence, the former addressing the issue of the variability of formations as well as their average bioavailability. The guideline states, "based on extensive intramural and extramural discussions, we now recommend that the average BE be supplemented by two new approaches, termed *population* and *individual bioequivalence*' (reference [21], p. 3). The guideline then states, 'the population and individual approaches reflect differences in the objectives of BE testing at various stages of drug development. These differences are embodied in the concepts of *prescribability* and *switchability. Prescribability* refers to the clinical setting in which a practitioner prescribes a drug product to a patient for the first time. ... *Switchability* refers to the setting in which a practitioner transfers a patient from one drug product to another (reference [21], p. 3).

It is not proposed to discuss here the requirement for population bioequivalence. The reader who is interested in this topic is referred to a recent paper of Grieve's [25]. Instead the new requirement for individual bioequivalence will be considered. However, neither the technical recommendations of the FDA guidance documents nor various alternatives that others have put forward will be considered here. Instead, an examination will be made as to the possible practical purpose of requirements for individual bioequivalence. In fact it will be argued that such an examination shows that there is no regulatory purpose in requesting proof of individual bioequivalence [20] and that as such, the relative merits of one scheme compared to another are neither here nor there.

3.2. The purpose of drug regulations and of bioequivalence studies

Hauck and Anderson have introduced an important concept in the notion of prescribability [19]. Unfortunately its implications have been insufficiently appreciated. Establishing that drugs are prescribable is the essential purpose of drug development and regulation. Prescribability implies that an acceptable risk is run by a naive patient taking the drug in view of the benefit that the drug confers. There is no regulatory requirement to sponsors in putting a new drug on the market to prove that it is the best drug for this indication, still less does the regulator automatically act to remove drugs that are on the market if a better treatment has now been developed. It seems that it is thus generally accepted that drugs can be allowed to exist together on the market even if patients might suffer some loss in being switched from one to another.

Bioequivalence is not an end in itself. The purpose of a bioequivalence trial is to circumvent the need for an expensive and unnecessary full development when this cheap alternative scheme of study is feasible. However, suppose that a choice is made between a full development and a bioequivalence study. That choice in itself says nothing about the value of the drug. If it is accepted that a sponsor may choose the full development without being required to address bioequivalence, it then follows that the only consistent standard by which the drug's acceptability should be judged is prescribability. Switchability is not an end in itself. This is not merely an academic argument. Concrete examples can be found. Both Novartis and Astra-Zeneca have developed and registered formulations of formoterol, a drug originally discovered by Yamanouchi, Novartis being the innovator in this respect with Foradil® [26] and Astra-Zeneca following with Oxis® [27]. However, Astra-Zeneca has developed its formulation independently using a free-standing dossier. Was Astra-Zeneca required to prove switchability

of Oxis with Foradil? Presumably not. Now suppose that another company decides to develop a third generic formulation using an equivalence trial. (There is a technical obstacle to doing this in that, since the drug is inhaled, the bioequivalence route is probably not appropriate but this is not really relevant to the argument.) Why should switchability be a requirement when two formulations compete in the same market on grounds of their prescribabilty alone?

It might be argued that this simply points to the inadequacy of existing approaches. If switchability is not addressed in the case of free-standing dossiers, this indicates that a current legitimate regulatory concern is not being addressed and needs to be. The fact that it is not addressed cannot be allowed to stand in the way of 'improved' bioequivalence regulations. Were this argument to be conceded, it would have consequences for every drug development. Patients are switched from one treatment to another all the time. To continue our example, suppose that a sponsor develops a new beta-agonist, 'fivemoterol' and proves its superiority to formoterol, using this as justification for registration. It may have done this using clinical trials in which patients had been previously treated with either salbutamol, salmeterol or terbutaline but had not been treated with either formoterol or fivemoterol but were randomly allocated to one or the other. The issue of switching patients from formoterol to fivemoterol would thus *never* have been investigated. However, once the drug is marketed such switches may take place.

3.3. The losses of naive and experienced patients

It could be argued that the primary purpose of regulating drugs for quality, safety and efficacy is to protect the interests of naive patients (those who have yet to receive the treatment covered by regulation) in the following sense. It is the regulators' responsibility to address the expected risk and benefit of the naive patient in deciding to take the drug and judge its suitability for registration accordingly. There is a red herring here. It is of course the case that side-effects of a drug may reveal themselves after considerable use. This does not, however, affect the argument regarding the interests of naive patients. Consider a patient who has already taken two months of a three-month course of treatment. It might be argued that the regulator's duty to naive patients would not cover the danger faced in the remaining month. This is incorrect. It is the expected risk benefit for a full course of treatment that is relevant, not just for the first millisecond in which a patient is exposed to the drug. The responsibility to naive patients is a responsibility for a full course of treatment.

Unless naive patients can have a reasonable expectation of some efficacy at an acceptable risk then drug regulation has failed them; but all experienced patients were naive at one time or another so that unless appropriate guarantees can be given to naive patients there will, in fact, be no others to whom the regulations can apply. Thus it seems that when we consider whether a drug is *prescribable* or not, it is primarily the needs of naive patients we should have in mind. After all, if the disease is chronic so that long-term palliative therapy is involved, patients who have already received the drug in question will have their own experience of having received it and are less in need of regulatory protection to guide their choice.

Consider a population of patients for whom two formulations of the same drug are considered identically *prescribable* using a given regime: one is a registered drug (the brand-name B) but that the other is not (the generic G). Suppose, for argument's sake, that a once daily regime is given and that the standard prescription is one month. In order to simplify the

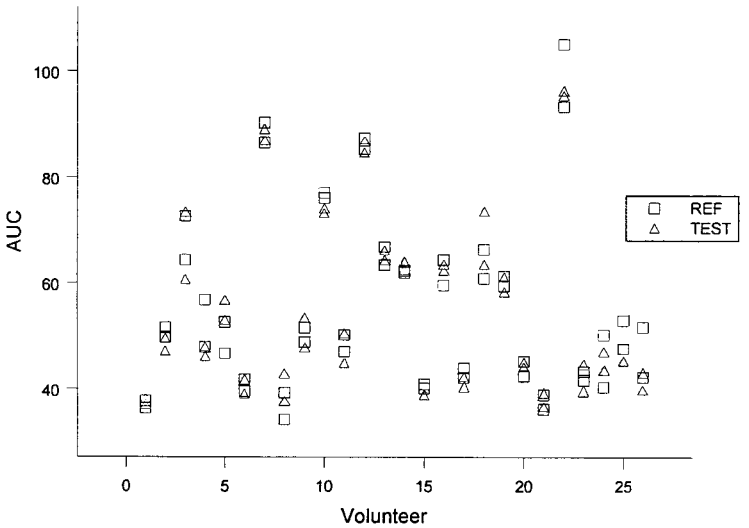

Figure 4. Plot of data in Shumaker and Metzler [28]. AUC for phenytoin for reference (REF) and test (TEST) formulations for 26 healthy volunteers treated in a four period cross-over trial.

argument it will be assumed, which in practice is unlikely to be exactly true, that if a patient has an acceptable steady-state AUC with a given formulation in a given month this will be observed for any exact prescription of exactly the same formulation. Let the probability that a patient chosen at random has an acceptable steady-state AUC be θ. Since the drugs are prescribable, θ must be sufficiently large by accepted regulatory standards. Because the two formulations are equally prescribable it makes no difference to the value of θ whether B or G is given. Now let the joint probability that a patient chosen at random would have an acceptable AUC under B and also under G be α_{BG}, under B but not under G be α_B, under G but not under B be α_G and under neither be α. Note that

$$\begin{aligned} \alpha_{BG} + \alpha_B + \alpha_G + \alpha &= 1 \\ \alpha_{BG} + \alpha_B = \alpha_{BG} + \alpha_G &= \theta \\ \alpha + \alpha_B = \alpha + \alpha_G &= 1 - \theta \\ \alpha_B &= \alpha_G \end{aligned} \tag{10}$$

Complete switchability implies $\alpha_{BG} = \theta$, $\alpha = 1 - \theta$ and $\alpha_B = \alpha_G = 0$, whereas complete independence would imply $\alpha_{BG} = \theta^2$. However, in practice this latter case is extremely unlikely, not least because a plausible reason for lack of acceptability of patient's AUC will be due to personal permanent 'main effect' characteristics of the patient rather than formulation by patient interaction. For example, the patient may have poor absorption or impaired elimination or be extremely small or unusually large. Thus, it is likely in practice that $\alpha_{BG} > \theta^2$.

For example, Figure 4 plots data from a trial comparing two formulations of phenytoin in healthy volunteers reported by Schumacher and Metzler [28]. Each volunteer was given each formulation twice. The switchability of these formulations was demonstrated, however, what is of interest here is the way in which bioavailability varies from subject to subject.

Now consider a subject adequately treated under B. The probability that this subject will be adequately treated under G is α_{BG}/θ. However, since $\alpha_{BG} > \theta^2$, this probability is greater than θ, which must in itself be greater than the regulatory standard. It could be argued that this is irrelevant since it does not guarantee that a patient switched from B to G will not suffer a loss that patient would not otherwise have suffered. This is, of course, true, but to take this line is to take a prescriptive rather than a permissive view of drug regulation: to argue that the regulator decides what must be prescribed, not what may be.

In fact, it can be argued that the existence of a non-switchable G increases the options available to the physician. If this formulation is not registered, then the probability of an acceptable formulation being eventually found for a patient is θ, whereas if it is registered it is $\theta + \alpha_G$.

3.4. The argument form Economics

There is no doubt that the existence of generic formulations drives prices down and that this has a beneficial effect on the budgets of health-care reimbursers. Of course, innovators need an inducement to innovate and this is what patents are supposed to provide. However, the position that the purpose of drug regulations should be, *inter alia*, to increase the protection offered to an innovator from a competitor with a product of equivalent quality, safety and efficacy is hard to sustain. For no disease is the population of sufferers permanent and for many there is a rapid turnover. For example, for antibiotics a single course of treatment may suffice. The population of patients for whom switchability is at all relevant may be tiny. Yet the proposed FDA guidance does not exclude antibiotics. Thus regulations that are designed to decide which drug *may* be prescribed could be used to prevent naive patients receiving cheaper and equally effective formulations because experienced patients might suffer some loss in switching, despite the fact that (a) such patients could be protected by quite other means, for example, by forbidding reimbursers from forcing physicians to switch (any say, only allowing them to dictate what prescriptions naive patients will have reimbursed) and (b) the risk run by such patients is likely to be less than that run by naive patients.

4. CONCLUSIONS

We have seen that the field of bioequivalence has been one that has been marked by controversy. The debate on appropriate approaches to average bioequivalence has continued unabated for over a quarter of a century. The controversy regarding individual bioequivalence is likely to be just as heated. However, the history of average equivalence suggests that before detailed statistical guidance is developed a more fundamental examination of the purpose of bioequivalence studies may be needed.

It needs to be remembered that bioequivalence is not an end in itself but a means to an end. What exactly that end is needs careful examination. That guidelines for individual bioequivalence will drive up the price of pharmaceuticals and statistical advice seems highly plausible; that they will be of any practical benefit to patients less so.

ACKNOWLEDGEMENTS

I am grateful to Professor John Lewis and Professor Dennis Lindley for helpful comments. The views expressed are my sole responsibility.

REFERENCES

1. Winslade J, Hutchinson DR, *Dictionary of Clinical Research*. Brookwood Medical Publications: Brookwood, 1992.
2. Senn SJ, Lillienthal J, Patalano F, Till D. An incomplete blocks cross-over in asthma: a case study in collaboration. In *Cross-over Clinical Trials*, Vollmar J, Hothorn L. (eds). Fischer: Stuttgart, 1997; 3–26.
3. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Chichester, 1997.
4. Steinijans VW, Hauschke D. International harmonization of regulatory bioequivalence requirements. *Clinical Research and Regulatory Affairs* 1993; **10**:203–220.
5. Hauschke D, Steinijans VW. Cross-over trials for bioequivalence assessment. In *Cross-over Clinical Trials*. Vollmar J, Hothorn L (eds). Fischer: Stuttgart, 1997; 27–40.
6. Senn SJ. *Cross-over Trials in Clinical Research*. Wiley: Chichester, 1993.
7. Westlake WJ. The use of confidence intervals in comparative bioavailability trials. *Journal of Pharmaceutical Sciences* 1972; **61**:1340–1341.
8. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. *Biometrics* 1976; **32**:741–744.
9. O'Quigley J, Baudoin C. General approaches to the problem of bioequivalence. *Statistician* 1988; **37**:51–58.
10. Kirkwood TBL. Bioequivalence testing—a need to rethink. *Biometrics* 1981; **37**:589–591.
11. Armitage P. Editorial note. *Biometrics* 1981; **37**:593–594.
12. Schuirmann DJ. A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**:657–680.
13. Mehring GH. On optimal tests for general interval-hypotheses. *Communications in Statistics – Theory and Methods* 1993; **22**:1257–1297.
14. Karlin S. Decision theory for Pólya type distributions. Case of two actions 1. *Third Berkeley Symposium on Probability and Statistics* 1956; **1**:115–129.
15. Anderson S, Hauck WW. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics*, *A* 1983; **12**:2663–2692.
16. Berger R, Hsu J. Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Statistical Science* 1996; **11**:283–319.
17. Brown LD, Hwang JTG, Munk A. An unbiased test for the bioequivalence problem. *Annals of Statistics* 1997; **25**:2345–2367.
18. Lindley DV. Decision analysis and bioequivalence trials. *Statistical Science* 1998; **13**:136–141.
19. Anderson S, Hauck WW. Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 1990; **18**:259–273.
20. Senn SJ. In the blood: proposed new requirements for registering generic drugs. *Lancet* 1998; **352**:85–86.
21. Centre for Drug Evaluation and Research. Guidance for Industry: Average, Population, and Individual Approaches to Establishing Bioequivalence. US. Department of Health and Human Services Food and Drug Administration, 1999.
22. Hauck WW. Chen ML, Hyslop T, Patnaik R, Schuirmann D, Williams R. Mean difference vs. variability reduction: tradeoffs in aggregate measures for individual bioequivalence. FDA Individual bioequivalence group. *International Journal of Clinical Pharmacology and Therapeutics* 1996; **34**:535–541.
23. Patnaik RN, Lesko LJ, Chen ML, Williams RL. Individual bioequivalence. New concepts in the statistical assessment of bioequivalence metrics. FDA Individual bioequivalence working group. *Clinical Pharmacokinetics* 1997; **33**:1–6.
24. Schall R, Williams RL. Towards a practical strategy for assessing individual bioequivalence. Food and Drug Administration individual bioequivalence working group. *Journal of Pharmacokinetics and Biopharmaceutics* 1996; **24**:133–149.
25. Grieve AP. Joint equivalence of means and variances of two populations. *Journal of Biopharmaceutical Statistics* 1998; **8**:377–390.
26. Anderson GP. Formoterol: pharmacology, molecular basis of agonism, and mechanism of long duration of a highly potent and selective beta 2-adrenoceptor agonist bronchodilator. *Life Sciences* 1993; **52**:2145–2160.
27. Selroos O. The pharmacologic and clinical properties of Oxis (formoterol) Turbuhaler. *Allergy* 1998; **53**(42 Suppl):14–19.
28. Shumaker RC, Metzler CM. The phenytoin trial is a case study of 'individual' bioequivalence. *Drug Information Journal* 1998; **32**:1063–1072.

