BIOMETRICS 37, 589–594
September 1981

READER REACTION

Bioequivalence Testing—A Need to Rethink

Thomas B. L. Kirkwood[1]
Statistics Section, National Institute for Biological Standards and Control, Holly Hill, London NW3 6RB, England

The symmetric confidence interval method for bioequivalence testing, proposed by Westlake (1976, 1979), is founded on a premise which is highly questionable, and in his description of the method, Westlake confuses two distinct statistical issues. The purpose of this note is to clarify the apparent flaws in Westlake's method and to point out a formal similarity between testing drugs for bioequivalence and checking that their potencies conform to specified levels. It is suggested that the adoption of a common statistical approach to the two problems may be advantageous.

A pair of drugs or, more commonly, two alternative formulations of the same drug are said to be 'bioequivalent' when equal amounts of them produce equal therapeutic effects. In place of the extensive clinical trials that would be needed to investigate equality of therapeutic effect directly, decisions on bioequivalence are usually made by comparing univariate biological responses (e.g. area under drug blood-level curve) after administration of supposedly equivalent single doses of the drugs (for a fuller account see Westlake, 1979). Such a test is known as a 'comparative bioavailability trial', and discussion here is confined to this simple case.

Because of experimental error and intrinsic biological variability, true bioequivalence can never be demonstrated exactly. Nor is it meaningful simply to conduct a conventional test of the null hypothesis that the drugs are bioequivalent. As pointed out in this context by Westlake (1972, 1979), a difference which is statistically significant may, nevertheless, be trivially small, while lack of significance may merely be the result of poor reproducibility. A more useful approach is to require that the confidence interval for the mean difference δ between the responses to the drugs is completely contained within some defined range of tolerance about zero (see Westlake, 1972; Metzler, 1974). In the usual statistical approach a confidence interval for δ would be centred on the sample mean difference, $\hat{\delta}$. However, Westlake (1976) proposed a modification to this method, which involves calculating a 'confidence interval' that is constrained to be symmetrical about zero. He claimed that the modification would have the dual advantages of (i) decreasing the 'effective' length of the confidence interval, and (ii) increasing the confidence coefficient. However, Westlake's main argument for the adoption of this approach seems to be based on a misconception. Furthermore, in a later part of his paper, Westlake switched his symmetric-interval method from the context in which it was first developed, namely

[1] Present address: Computing Laboratory, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, England.

inference about the mean difference between two normally distributed random variables, to inference about the ratio of lognormal variables. In doing this, he ignored the fact that the imposition of symmetry on the latter variables must necessarily introduce a form of bias. (For an earlier debate on Westlake's method, see correspondence from N. Mantel and W. J. Westlake in *Biometrics* **33**, 759–760, December 1977.)

The importance of symmetry was justified by Westlake on the grounds that 'most clinicians tend to make their equivalence statements in a symmetrical manner'. He noted that the conventional confidence interval is not, in general, symmetric about zero, and, in view of the preceding statement, he regarded this as a disadvantage. In particular, since the decision on bioequivalence is based on the distance of the further confidence limit from zero, he declared the 'effective' length of the confidence interval to be twice this distance. As an alternative, he suggested that the conventional 95% confidence interval should be replaced by one which is symmetrical about zero and which covers a range accounting for 95% of the area under the likelihood curve. The 'effective' length of this symmetrical interval is obviously less than that of the conventional interval, and Westlake further showed that the probability of it containing δ is always greater than .95 (for $\delta = 0$ this probability is 1, and it tends to .95 only as $\delta \to \infty$).

So, what is wrong with this? The problem is perhaps seen most clearly by considering a hypothetical example. Suppose the limits of the range for accepting bioequivalence are set as ± 10.0, and that a sample of subjects has yielded a mean difference of 7.0 with a conventional confidence interval for δ of (3.5, 10.5). In the conventional approach, bioequivalence is not accepted, and it is concluded that there is probably a genuine though small difference between the drugs (the observed difference is small enough, however, that a larger sample might show the drugs to be acceptably equivalent). With Westlake's method, the confidence interval (3.5, 10.5), which he would claim has an 'effective' length $(-10.5, 10.5)$, is replaced by a symmetrical one, say $(-9.8, 9.8)$. Bioequivalence is accepted, and *information on the difference between the drugs is ignored.*

This example was, of course, chosen to highlight the difference between the methods, but the points it makes are important and general ones. Firstly, while Westlake's symmetrical confidence interval has a shorter 'effective' length, it is actually longer than the conventional interval. (That the confidence coefficient for Westlake's interval varies with δ also underlines its difference from the usual concept of a confidence interval.) Secondly, the probability of accepting bioequivalence with Westlake's approach is *always* higher. In fact, it is easily seen that as $|\delta|$ increases or σ (the standard deviation of responses) decreases, Westlake's method progressively changes, in favour of accepting bioequivalence, from a two-sided to a one-sided approach. When a conventional 95% confidence interval is used, bioequivalence is accepted if $\hat{\delta}$ differs significantly from each of the upper and lower limits at a fixed significance level, the one-sided $2\frac{1}{2}\%$ level. With Westlake's method, bioequivalence is accepted if $\hat{\delta}$ differs from the nearer limit at a one-sided significance level which varies between $2\frac{1}{2}\%$ and 5%. As $|\delta|/\sigma$ increases, or in other words as the evidence for nonequivalence becomes stronger, the criterion for rejecting bioequivalence actually becomes more lax.

The kernel of this dubious strategy is Westlake's attempt to carry over the symmetry of clinicians' bioequivalence statements, which relate only to the setting of tolerance limits, to inference from the data. To impose symmetry thus is inappropriate. The best estimate of the true difference between the drugs is the sample mean difference, not zero, and the confidence interval ought properly to be centred on this. The fact that the symmetrical interval has the higher probability of containing the true difference (assuming that the variables are indeed normal) gives some reassurance that the methods will only seldom

reach different conclusions. However, there are no practical advantages, and some real disadvantages, to Westlake's departure from the conventional approach.

To add to this confusion, Westlake considered another type of asymmetry where the variables are lognormal, and where bioequivalence is defined as occurring when the ratio of their true geometric means is equal to unity. Use of a log transformation makes this situation directly equivalent to the one already considered, but there is the added problem that limits which are symmetrical about zero in log units are asymmetrical when transformed back to limits about unity (e.g. 0.80–1.25). Westlake suggested a different modification here so as still to conform with clinicians' supposed requirements of symmetrical confidence limits. However, this automatically biases the assessment of bioequivalence in favour of accepting ratios less than unity. Since limits for lognormal variables which are *genuinely* symmetrical in terms of the underlying variation do, unfortunately, have the superficial appearance of being asymmetrical, the answer surely lies in better educating our clinical colleagues, and not in distorting the methods of analysis so as merely to hide this problem.

A well-established precedent for this type of limit may be found in the statistically similar context of controlling the potency of biological drugs (see, for example, *European Pharmacopoeia*, 1969, 1971). In bioassay, potency estimates tend to be lognormally distributed, and confidence limits are calculated to be symmetric on a log scale (asymmetric on the scale of potency units). Pharmacopoeial requirements for acceptance of potencies stipulate, for example, that the estimated potency of a drug and its 95% confidence interval should fall within ranges 90%–111% and 80%–125% of the labelled value, respectively. In this case, constraints are placed on both point and interval estimates. This type of logarithmically symmetric tolerance interval is widely accepted by both manufacturer and regulatory authority, and it may be of considerable practical benefit to bring bioequivalence testing closer to the practice of bioassay. The formal similarity of the statistical problems in the two areas suggests that much may be gained by adopting a common methodological approach.

ACKNOWLEDGEMENT

I thank the Editor for his helpful comments on an earlier version of this note.

REFERENCES

European Pharmacopoeia (1969, 1971). Vols I and II. Paris: Maisonneuve S. A.

Metzler, C. M. (1974). Bioavailability—a problem in equivalence. *Biometrics* **30,** 309–317.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences* **61,** 1340–1341.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **32,** 741–744.

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics* **35,** 273–280.

RESPONSE

Kirkwood's comments continue an apparently developing trend, that the criticism of papers on bioequivalence is a suitable pursuit for those who are not familiar with the concept of bioequivalence or with the particular problems that it poses to both manufacturer and drug regulatory agency. The lack of understanding of the meaning of bioequivalence is apparent as early as the second paragraph: bioequivalence is a concept that

applies not to different drugs but only to different formulations of the same drug entity. Kirkwood's note contains a number of quite gratuitous comments and it is tempting to reply in kind. It will probably be more helpful to readers of this journal, however, if I attempt to elucidate the considerations that led me to propose a confidence interval symmetrical about zero (or unity, in the case of a ratio). As I do this, I think it will become clear that most of Kirkwood's objections are irrelevant and that the criticism, 'confusion' of statistical issues, is more appropriate to his comments since they appear to miss the essential point, namely that the confidence interval is proposed as a decision-making device.

Firstly, I should emphasize that the following discussion is presented against the background of the practice in the U.S.A., where a pharmaceutical company, seeking approval of its formulation of a drug, conducts a bioequivalence trial against the standard formulation (usually the originator's) and submits the results to the Food and Drug Administration (FDA) for approval. For a number of years it has been the practice to analyse such trials using an ANOVA in which the key element is a test of the simple hypothesis that for the two formulations the means of several characteristics of the blood-level profile are identical. That this practice is still standard can be verified by perusal of the numerous journals dealing with clinical pharmacology and pharmaceutical sciences.

I have argued for a number of years that the testing of this null hypothesis is irrelevant to the decision that must be made. As a meaningful alternative I have suggested construction of a $1-\alpha$ confidence interval ($\alpha = .05$, traditionally) on the difference or the ratio of the means. The decision procedure is: if the limits of the confidence interval fall within the acceptable limits recommended by the regulatory agency, accept the new formulation; if not, reject it. My next step was to observe that, since the acceptable limits were given in symmetrical form, the use of a confidence interval symmetrical about zero for differences, or about unity for ratios, would increase the manufacturer's chances of success (approval) while still assuring the regulatory agency of a confidence coefficient of at least $1-\alpha$. In my 1976 paper (referenced by Kirkwood) I stated that 'most clinicians tend to make their equivalence statements in a symmetrical manner'. In my experience this is true; but I could have put the case much more strongly by noting that in the U.S.A. the regulatory agency proposes the symmetrical form in its regulations. An examination of the various bioavailability regulations appearing in the *Federal Register*, for example, reveals numerous statements to the effect that the reference and test products should not differ by more than 20% or 30%. Kirkwood's gratuitous comment concerning 'clinicians' supposed requirements of symmetrical confidence limits' can then clearly be seen for what it is. One final point on the symmetrical confidence interval should be made. It is a point that is equivalent, I believe, to one that Kirkwood himself makes. Whatever the true value of the difference or ratio of the means, the probability of accepting the test formulation is always higher with my proposed symmetrical $1-\alpha$ confidence interval than with the conventional $1-\alpha$ confidence interval. Similarly, the probability of acceptance with the symmetrical $1-\alpha$ confidence interval is always *less* than with a conventional $1-2\alpha$ interval. This point is important in the following discussion.

What protection should the regulatory agency seek against approving a new formulation that is not bioequivalent to the standard? Current practice in approving new drugs for efficacy presents a helpful analogy. In this case, one is usually attempting to demonstrate the efficacy of a new drug by testing against a placebo, and it is customary in the U.S.A. to insist that in the test of the null hypothesis (identity of drug and placebo) a statistically significant result at the α level (traditionally .05) be obtained as proof of efficacy. My

interpretation of this is that the regulatory agency is attempting to ensure that if the drug is really the same as placebo there is only a low probability, .05, that it will be approved. Note, however, that since the drug would never be approved for being less efficacious than placebo, the test and its associated critical region should be one-sided. It seems to me that a similar policy should prevail in the decision criterion for bioequivalence. A suitable rule might be: if the difference in means of the two formulations is actually Δ, where $\pm\Delta$ is the allowable range for bioequivalence, then the probability that the $1-\alpha$ confidence interval falls within $\pm\Delta$ should be acceptably small (say .05). That is, in the borderline case the probability of accepting the new formulation as bioequivalent to the standard should be small. I had not formalized this approach when I wrote the 1976 paper, but in the 1979 survey paper (also referenced by Kirkwood) it is mentioned briefly under a discussion of sample-size determination.

If this approach is used it will be seen that the use of a conventional $1-\alpha$ confidence interval with $\alpha = .05$ is unduly conservative since the probability that the interval falls within the $\pm\Delta$ limits when the difference in means is Δ can be shown to be $<\frac{1}{2}\alpha$, or .025. In order to obtain a true analogy with efficacy-testing practice one should use a 90% confidence interval, then the probability of accepting the borderline case is $<.05$. Thus, despite Kirkwood's concerns, it is apparent that the use of my proposed symmetrical 95% confidence interval leads to a decision process which is more stringent than that based on the use of a conventional 90% confidence interval. The use of the latter has much to recommend it: in particular, the fact that it parallels efficacy-testing practice. However, my concern has been that the values .05 for critical regions and 95% for confidence coefficients are so ingrained in traditional practice that it might be hard to obtain universal acceptance of its use. To a regulatory agency, for example, use of a 90% rather than a 95% confidence coefficient might appear to represent a relaxation of its standards whereas, as I have pointed out above, it is completely consonant with the practice of using a one-sided α-level of .05 in efficacy trials. With this background, it should be clear that my recommendation of 95% symmetrical confidence intervals can be viewed as an attempt to bridge the gap from a traditional 95% confidence interval to the 90% confidence interval that is really more appropriate.

I hope that the foregoing remarks shed some light on the decision process involved in bioequivalence testing. In particular, I think it should be clear that the use of the suggested 95% symmetrical confidence interval, far from being a 'dubious strategy' based on a premise which is 'highly questionable', is, in fact, a rather conservative procedure. The other comments of Kirkwood, concerning biased estimation and so on, are not relevant to the decision-making problem faced in bioequivalence trials.

W. J. Westlake
Smith, Kline & French Laboratories,
1500 Spring Garden St,
P.O. Box 7929,
Philadelphia, Pennsylvania 19101, U.S.A.

Note by Editor

The exchange of views by T. B. L. Kirkwood and W. J. Westlake will help to clarify some of the statistical issues involved in bioequivalence testing. Readers may wish to refer to two recent papers in which a Bayesian viewpoint is adopted: Selwyn, Dempster and Hall

(*Biometrics* **37,** 11–21, March 1981), and Mandallaz and Mau (*Biometrics* **37,** 213–222, June 1981). The Bayesian formulation provides a convenient way to review the difference between the symmetrical and conventional confidence interval procedures.

Suppose that the regulatory authority is prepared to regard two formulations as bioequivalent if δ, the true difference in means (say, on a log scale), is within the range $\pm\Delta$. With appropriate assumptions about 'vague priors', the posterior distribution of Δ in any bioequivalence test may be approximated by the usual t distribution centred around the estimate $\hat{\delta}$. Then Westlake's procedure, which involves acceptance if a $100(1-\alpha)\%$ confidence range centred around zero falls within $\pm\Delta$, is equivalent to acceptance when the posterior probability that $-\Delta<\delta<\Delta$ exceeds $1-\alpha$. The conventional procedure favoured by Kirkwood, which is based on the $100(1-\alpha)\%$ confidence range centred around $\hat{\delta}$, is equivalent to acceptance when (i) the probability that $\delta<-\Delta$ is less that $\frac{1}{2}\alpha$, *and also* (ii) the probability that $\delta>\Delta$ is less than $\frac{1}{2}\alpha$. Either of these approaches seems intuitively reasonable, but they are different: hence the possible confusion.