The Statistician (1988) 37, pp. 51–58

General approaches to the problem of bioequivalence

JOHN O'QUIGLEY[1]* & CLAUDE BAUDOIN[2]

[1]*Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington Seattle, Washington 98104 U.S.A. and* [2]*Unité 21 de l'Institut National de la Santé et de la Recherche Médicale (INSERM)—16, avenue Paul Vaillant Couturier 94807 Villejuif Cedex France*

Abstract. Apart from the Bayesian approach to the problem of demonstrating bioequivalence, all other existing approaches can be grouped under four headings (Rocke, 1984). We reconsider the problem from the standpoint of fiducial probability and illustrate how this leads to the immediate deduction of Rocke's findings. Furthermore, by taking this standpoint, much light is thrown on many of the other issues and controversies of the subject.

1 Introduction

Alternative formulations of the same drug or a pair of similar drugs are said to be 'bioequivalent' if they produce, in some sense, equivalent therapeutic effects. In practice we will want to assess to what extent such drugs, or different formulations, are, in normal conditions of use, equivalent.

Authors (Westlake, 1972, 1976, 1979, 1981; Dunnett & Gent, 1977; Selwyn *et al.*, 1981; Kirkwood, 1981; Blackwelder, 1982; Rocke, 1984; Hauck & Anderson, 1983, 1984) have pointed out that a test of the usual null hypothesis is inappropriate since small and clinically insignificant differences may be detected with large sample sizes. Furthermore, as is always carefully underlined in introductory statistics courses, failure to reject the null hypothesis can in no way lead to its affirmation. Considerable controversy (Mantel, 1977; Westlake, 1977, 1981; Kirkwood, 1981) has arisen over the appropriateness of the different approaches. Comparative studies (Mandallaz & Mau, 1981) using the two-period crossover design (Grizzle, 1965) and leaning on simulations have led to somewhat ambiguous conclusions. In fact leaving aside the Bayesian method developed by Selwyn *et al.* (1981) and specific applications such as the equivalence problem with binomial outcome (Dunnett & Gent, 1977) or with ordered categorical data (Mehta, Patel & Tsiatis, 1984) the various methods can be seen to be closely related and to come under four broad headings.

Before looking at these we recall the main ideas. Let δ (≥ 0) measured on some scale represent the true difference between the two population treatment means. Should any true difference be negative, it only remains to reverse the direction of this difference to ensure that $\delta \geq 0$. In practice we will estimate δ by $\hat{\delta}$ and try to make inferences regarding δ. As pointed out we will be unable to infer $\delta = 0$ and even were we able to infer $\delta \neq 0$ this is of little practical assistance. Thus we introduce $\Delta > 0$ as being the maximum value for δ of negligible practical interest. The value of Δ chosen would in practice necessitate considerable discussion and this in itself is probably no bad thing.

The four broad methods are briefly summarised in Section 2 and the relationship between them in Section 3. In Section 4 we briefly recall the main ideas of the fiducial argument, giving these a pictorial representation in Section 5. This representation

* On leave from Unité 292 INSERM, France.

enables the immediate deduction of Rocke's findings (1984) and gives a context in which many of the results and controversies of the bioequivalence question can be more clearly understood.

2 Methods

2.1 Westlake's symmetric confidence interval method

Westlake (1972, 1976) put the problem in the following way. Denote the new formulation of the drug by N, the standard by S, μ_N and μ_S being the respective treatment means. In earlier work Westlake referred to these means as the mean amount of drug absorbed, although other authors, and subsequently Westlake himself, considered these population parameters to denote some mean response however measured. Thus, and this point does not seem to have given rise to much discussion in the bioequivalence literature, treatments may be bioequivalent in some respects and not in others.

Some thought is needed to understand the notation of Westlake since at first sight there appears to be a lack of distinction between population parameters and their estimates based on the data. For instance he suggests the data from a trial be used to construct a confidence interval with specified confidence coefficient of the form

$$\mu_S+C_2\leq\mu_N\leq\mu_S+C_1$$

and to reject the hypothesis of bioequivalence if the interval is, in some accepted sense, too wide. In a numerical example this expression becomes

$$0{\cdot}727\mu_S\leq\mu_N\leq1{\cdot}15\mu_S$$

and from a classical viewpoint we ought to feel unhappy with such an expression, since there is nothing obviously random in it. This is the clue to a point, first understood by Westlake, and that is that the intervals we are dealing with are not confidence intervals in the classical sense. Their motivation stems from the observation that we are in a decision making context and that the classical methods of statistical decision making do not, in an appropriate way, address the problems being posed.

Westlake further observed that in practice the applied worker, be they clinician or pharmacologist, tend to make equivalent statements in a symmetrical manner so that if the absolute value of $\delta=\mu_N-\mu_S$, or possibly log δ where $\delta=\mu_N/\mu_S$, is less than some given value then we have bioequivalence. He, therefore, introduced the idea of the 'effective' length of the confidence interval which is not C_1-C_2 but $2K$ where $K=\min\{|C_1|,|C_2|\}$. This motivated the introduction of symmetric intervals.

These are obtained as follows. Let $\bar{x}_N$ and $\bar{x}_S$ be the sample means and s^2, the estimated residual variance based on $n-1$ degrees of freedom. Let

$$t(u;n-1)=\{(n-1)^{1/2}\,\beta(1/2,(n-1)/2)\}^{-1}\{1+u^2/(n-1)\}^{-n/2}$$

where

$$\beta(z,w)=2\int_o^{\pi/2}\sin^{2z-1}(\theta)\cos^{2w-1}(\theta)\,d\theta$$

i.e. the density for a t-variate on $n-1$ degrees of freedom. Further suppose that

$$\int_{K_2}^{K_1}t[\sqrt{n}\{(\bar{x}_N-\bar{x}_S)-\delta\}/s\sqrt{2};n-1]du=0{\cdot}95$$

for some K_1 and K_2. Then Westlake (1976) shows by simple algebra that if we can find K_1 and K_2 such that

$$(K_1+K_2)\sqrt{2s^2/n}=2(\bar{x}_S-\bar{x}_N)$$

then the above interval, viewed as a confidence interval for μ_N, will be symmetrical about μ_S. Westlake suggested we determine K_1 and K_2 by trial and error although algorithms and tables have since been provided (Spriet & Beiler, 1978). He further demonstrates that the confidence coefficient is in fact always greater than $1-\alpha$, if, for instance, we construct a $100(1-\alpha)\%$ confidence interval.

In practice then we will calculate a $(1-\alpha)100\%$ confidence interval for δ, on the basis of $\hat{\delta}$, symmetric about zero. Reducing α will increase the interval size. The limits $(-\Delta,\Delta)$ will be encroached on simultaneously, at which point α can be interpreted as the degree of significance against the null hypothesis—absence of bioequivalence.

2.2 Kirkwood's method

Kirkwood's approach (1981) stemmed from his view that testing for bioequivalence and checking that drug potencies conform to specified levels should share a common statistical approach. He disagreed with the symmetrical intervals of Westlake for a number of reasons and constructed an example whereby we would conclude bioequivalence using symmetrical intervals and yet the data give strong indication for a difference between the treatments, albeit small.

Kirkwood underlined an apparent paradox in the symmetrical interval method, whereby, as Δ/s increases and the evidence for non-equivalence becomes stronger, the criterion for accepting bioequivalence actually becomes more lax (here Kirkwood talks about rejecting rather than accepting bioequivalence and this is presumably an error). The reason for this, as noted by Kirkwood, is that, under symmetry and assuming $\hat{\delta}\neq 0$, a two-sided interval progressively becomes one-sided.

His proposed method would be to calculate a $(1-\alpha)100\%$ confidence interval for δ on the basis of $\hat{\delta}$. If the interval is entirely contained within $(-\Delta,\Delta)$ then bioequivalence is concluded. In order to obtain a significance level it suffices to vary α noting, once again, that this varies inversely with interval size. As α decreases, at some point, one of the limits $(-\Delta,\Delta)$ will be encroached upon and we can take this corresponding α to be the degree of significance.

2.3. Westlake's one-sided method

Westlake (1981) takes issue with Kirkwood, emphasising that in the context of bioequivalence, confidence intervals are being used as an aid to decision making and should not necessarily be given a rigorous classical interpretation. Thus no philosophical problem is raised should we simultaneously, on the basis of a single data set, conclude that treatments differ significantly and that they are also bioequivalent. Kirkwood's second point prompted Westlake to suggest that use of a $(1-\alpha)100\%$ confidence interval is, in fact, unduly conservative. Since we can suppose, without loss of generality, that $\hat{\delta}>0$ we need only concern ourselves with Δ (and not $-\Delta$). Westlake suggests we ought then work with a $(1-2\alpha)100\%$ confidence interval, the main reasons for not doing so being more to do with conservatism and traditional practice than statistical. Using this approach the degree of significance will be obtained as before by reducing α up until encroachment of the, now one-sided interval, at the point Δ.

2.4 Rocke's method

Rocke felt that much of the controversy could be cleared up by formulating the problem squarely in terms of a statistical test. The traditional null and alternative

hypotheses change roles and we test H_Δ:$|\delta|>\Delta$ against H_0:$|\delta|<\Delta$. Under the null hypothesis H_Δ we will consider the distribution of $\hat{\delta}$ to be centered around Δ and under this distribution a critical region will be defined between $-\delta_0$ and δ_0. The value of δ_0 will depend not only on the variance of $\hat{\delta}$ but also on α. We then note that if the real difference is even greater than Δ then the effect of the distribution being centered around Δ can only be conservative. Secondly, the problem being expressed in terms of absolute values, the same reasoning follows through if we deal with negative quantities. Detailed discussion is given by Rocke (1984), Anderson & Hauck (1983) and by Hauck & Anderson (1984). Similar ideas are exploited in the case of proportions (Dunnett & Gent, 1977; Blackwelder, 1982). From a classical viewpoint this method is in many ways the only one (of those considered here) with a solid foundation in statistical theory, at least as conceived by Neyman & Pearson. It has however been pointed out that the other methods do give rise to formal statistical tests, and this no less so because they find their expression in the language of confidence intervals. Even so the problem stated in the terms expressed by Rocke does clear up some confusion and assists the practitioner who, for example, finds the null hypothesis of Mandallaz & Mau (1981) the alternative of Hauck & Anderson (1984) and vice versa!

3 Relationship Between the Methods

Denote the methods as method i ($i=1,\ldots,4$) where the relevant method is described in 2.i of the previous section. Following Rocke (1984) let

$$T(x)=\int_x^\infty t(u;n-1)du.$$

Suppose $\hat{\delta}\leq\Delta$ (assuming without loss of generality that $\hat{\delta}$ is positive), then under method 1 bioequivalence will be concluded whenever

$$T\{(\Delta-\hat{\delta})/s\}+T\{(\Delta+\hat{\delta})/s\}\leq\alpha. \tag{3.1}$$

For method 2 bioequivalence will be concluded whenever

$$2T\{(\Delta-\hat{\delta})/s\}\leq\alpha. \tag{3.2}$$

For method 3 bioequivalence will be concluded whenever

$$T\{(\Delta-\hat{\delta})/s\}\leq\alpha. \tag{3.3}$$

For method 4 bioequivalence will be concluded whenever

$$T\{(\Delta-\hat{\delta})/s\}-T\{(\Delta+\hat{\delta})/s\}\leq\alpha \tag{3.4}$$

Similar expressions are given by Rocke in the case $\hat{\delta}>\Delta$. In either event, defining as p_i the significance level obtained for the ith method, some algebra (Rocke, 1984) shows that

$$p_2>p_1>p_3>p_4.$$

This result is made much more transparent by appealing to the idea of fiducial probability. This is considered in the next section where the associated idea of a pictorial representation of a confidence interval for translation families is anticipated, and used in the remaining section. It is our view that this device leads to much clarification of many of the issues raised by the apparently different approaches to bioequivalence.

4 The Fiducial Argument

Let T be a sufficient statistic for the parameter θ and let

$$pr(T<t)=F(t,\theta)$$

$$f(t,\theta)=\partial F(t,\theta)/\partial t.$$

Denote the range of possible values for θ by R_θ and that for $f(t,\theta)$ continuous, t fixed,

$$\partial f(t,\theta)/\partial\theta \leq 0 \qquad \theta\in R_\theta\ ,\ \theta\geq\theta^*$$

$$\partial f(t,\theta)/\partial\theta \geq 0 \qquad \theta\in R_\theta\ ,\ \theta\leq\theta^*$$

$$\partial^2 f(t,\theta)/\partial\theta^2<0 \qquad \theta\in R_\theta\ ,\ \theta=\theta^*.$$

Once again considering t fixed, define θ^L and θ^U such that

$$f(t,\theta^L)=f(t,\theta^U) \text{ and}$$

$$F(t,\theta^U)-F(t,\theta^L)=1-\alpha.$$

The $(1-\alpha)100\%$ fiducial interval for θ is then defined as (θ^L,θ^U).

The area of fiducial inference is certainly no less controversial than that of bioequivalence. However, controversy in the first case arises when θ is of dimension greater than one or when θ is other than a translation parameter, invariant under the group of linear transformations (see Fraser, 1961). This second condition is satisfied if we can factorise the density $g(x|\theta)$ of the data x in the form (Fisher, 1934)

$$g(x|\theta)=h(\theta-\theta^*)\ \phi(A)$$

where A is ancillary. In the cases considered here it has been assumed that $\sigma_N^2=\sigma_S^2$, in which case the ancillary statistic A is just $s^2(n_N^{-1}+n_S^{-1})$ where n_N and n_S are the respective sample sizes for the formulations N and S. Were we not to suppose $\sigma_N^2=\sigma_S^2$ then the various techniques, advanced as solutions to the Behrens-Fisher problem, could be applied to the area of bioequivalence, leading to results other than those yet obtained.

Our aim here though is not to propose new solutions but simply to show how, from the standpoint of fiducial inference, additional light can be thrown upon these solutions currently in use. For the methods described in Section 2 the above factorisation is usually appropriate, possibly after having transformed the original data.

For our purposes we will assume that the constraint $f(t,\theta^L)=f(t,\theta^U)$, generally deemed necessary for generating a fiducial interval, be relaxed. In the next section then, we see that the methods put forward in Section 2 can be formulated in the above terms and that a pictorial representation of this formulation leads to an immediate deduction of Rocke's findings.

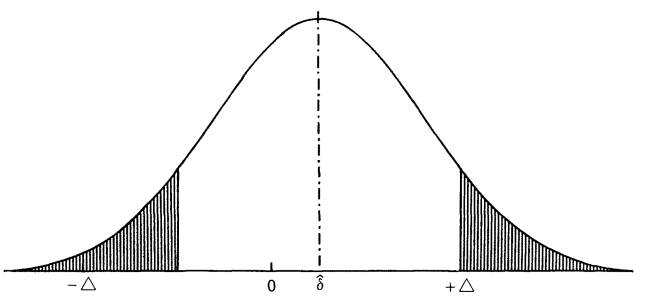

Fig. 1. *P*-value for method 2 using 'confidence intervals' symmetric about $\hat{\delta}$.

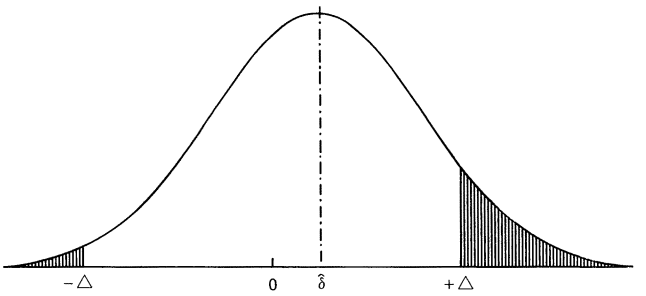

Fig. 2. *P*-value for method 1 using 'confidence intervals' symmetric about 0.

5 The Calculation of *p*-values

In a strictly decision making context we would fix α in advance and for method 4 for instance carry out a test at level α. For the other methods confidence intervals for δ of size $1-\alpha$ would be calculated and if these are wholly contained within the interval $(-\Delta,\Delta)$ then the conclusion of bioequivalence is made. In practice we will more often than not calculate *p*-values. For method 2 (assuming without loss of generality that $\hat{\delta}$ is positive) we will calculate that $(1-\alpha)100\%$ confidence interval whose right-hand limit just touches Δ. The *p*-value will then be equal to twice the probability associated with values greater than Δ and this is illustrated graphically in Fig. 1. For method 1 the $(1-\alpha)100\%$ confidence interval is centered around zero and once again we vary α until the right-hand limit of the confidence interval touches Δ, by symmetry the left-hand limit will also touch $-\Delta$. This is shown in Fig. 2 and note that we use the same curve centered about $\hat{\delta}$ and not about zero in accordance with the fiducial argument. For method 3 illustrated in Fig. 3 it is clear we end up with half the *p*-value associated with method 2. Conceptually method 4 operates in quite a different way in calculating a rejection region under the alternative hypothesis (i.e. $|\delta|>\Delta$) which in a classical sense is now viewed as the null hypothesis. As with the confidence interval approaches we proceed less formally by calculating the probability (*p*-value) associated with the interval $(-\hat{\delta},\hat{\delta})$. The symmetry in Fig. 4 means that the shaded area corresponding to the *p*-value between $-\hat{\delta}$ and $\hat{\delta}$ under the broken curve is the same as that between Δ and $\Delta+2\hat{\delta}$ under the unbroken curve. If we denote by p_i the *p*-value corresponding to the *i*th method we only need look at Figs. 1–4 to see straightaway that

$$p_2>p_1>p_3>p_4$$

the central result obtained from Rocke (1984).

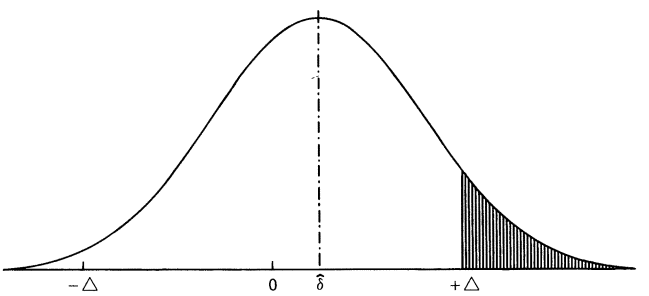

Fig. 3. *P*-value for method 3 using one sided 'confidence interval'.

Figs 1–4 are helpful in appreciating other results and controversies arising in the area of bioequivalence. The simulations of Mandallaz & Mau (1981) led them to conclude $p_2 > p_1$ and as already pointed out this is immediate upon comparing Figs 1 and 2, where we see that the critical region (unshaded area) for method 1 contains that for method 2. A little thought is needed here since the critical regions are unshaded whilst the associated p-values correspond to the shaded areas. In terms of these figures it is also very much easier to understand the Westlake method for calculating symmetrical intervals. In effect we start with Fig. 2, progressively transferring portions of shaded area from left to right until the resulting figure looks like Fig. 1, although with a different Δ since here the shaded areas are not equal.

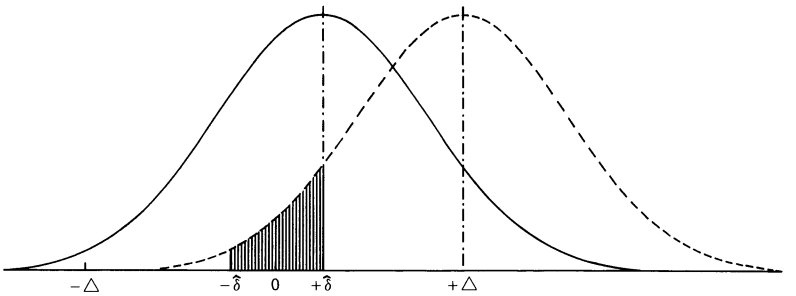

Fig. 4. p-value for method 4 based upon conventional test.

Selwyn & Hall (1985) have strongly criticised method 4 on two grounds. Firstly they say there is no intuitive rationale for defining a p-value to be the difference of two tail areas and secondly, an anomaly noted by Rocke (1984) whereby for fixed $\hat{\delta}$ and Δ $p_4 \to 0$ as $s^2 \to \infty$, should in their view disqualify the method. The first criticism does not seem to rest on any statistical or logical foundation although the second was considered sufficiently serious by Rocke (1985), one of the method's main proponents, to suggest its abandonment in favour of method 3. This contrasts with the conclusion of his 1984 paper where, despite this anomaly, he considered the overall properties of method 4 indicated its preference to method 3. Our view is that this problem needs further thought and that it would be hasty to abandon a method with many attractive aspects. The anomaly is readily appreciated by considering Fig. 4 where, if we keep $\hat{\delta}$ and Δ fixed and increase s^2 the curves become flatter and flatter. We see that $p_1, p_2 \to 1$, $p_3 \to 1/2$ and $p_4 \to 0$. Even so, if the standard error of $\hat{\delta}$ is as great as Δ, let alone several times greater, it is hard to imagine the serious practitioner carrying out a formal test of bioequivalence. In other words the problem is probably of academic rather than practical interest.

References

ANDERSON, S. & HAUCK, W.W. (1983) A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics*, A12, pp. 2663–2692.

ANDERSON, S. & HAUCK, W.W. (1985) Letter to the Editor, *Biometrics* 41, pp. 561–563.

BLACKWELDER, W.C. (1982) proving the null hypothesis in clinical trials, *Controlled Clinical Trials*, 3, pp. 345–353.

DUNNETT, C.W. & GENT, M. (1977) Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables, *Biometrics*, 33, pp. 593–602.

FISHER, R.A. (1934) Two new properties of mathematical likelihood, *Proceedings of the Royal Society Series A* 144, pp. 285–307.

FRASER, D.A.S. (1961) The fiducial method and invariance, *Biometrika*, 48, pp. 261–280.
GRIZZLE, J.E. (1965) The two period change over design and its use in clinical trials, *Biometrics,* 21, pp. 467–480.
HAUCK, W.W. & ANDERSON. S. (1984) A new statistical procedure for testing equivalence in two-group comparative bioavailability trials, *Journal of Pharmacokinetics and Biopharmaceutics*, 12, pp. 83–91.
KIRKWOOD, T.B.L. (1981) Bioequivalence testing: a need to rethink (reader reaction), *Biometrics*, 37, pp. 589–591.
MANDALLAZ, D. & MAU, J. (1981) Comparison of different methods for decision making in bioequivalence assessment, *Biometrics*, 37, pp. 213–222.
MANTEL, N. (1977) Do we want confidence intervals symmetrical about the null value? (Letter to the editor), *Biometrics*, 33, p. 759.
MEHTA, C.R., PATEL, N.R. & TSIATIS, A.A. (1984) Exact significance testing to establish treatment equivalence with ordered categorical data, *Biometrics*, 40, pp. 819–825.
ROCKE, D.M. (1984) On testing for bioequivalence, *Biometrics*, 40, pp. 225–230.
ROCKE, D.M. (1985) Reply to correspondence, *Biometrics*, 41, p. 563.
SELWYN, M.R., DEMPSTER, A.P. & HALL, N.R. (1981) A Bayesian approach to bioequivalence for the 2×2 changeover design, *Biometrics*, 37, pp. 11–21.
SELWYN, M.R., HALL, N.R. & DEMPSTER, A.P. (1985) Letter to the editor, *Biometrics,* 41, p. 561.
SPRIET, A. & BEILER, D. (1978) Tables to facilitate determination of symmetrical confidence intervals in bioavailability trials with Westlake's method, *European Journal of Drug Metabolism and Pharmacology*, 2, pp. 129–132.
WESTLAKE, W.J. (1972) Use of confidence intervals in analysis of comparative bioavailability trials, *Journal of Pharmaceutical Sciences*, 61, pp. 1340–1341.
WESTLAKE, W.J. (1976) Symmetrical confidence intervals for bioequivalence trials, *Biometrics*, 32, pp. 741–744.
WESTLAKE, W.J. (1977) Reply to letter to the Editor from N. Mantel, *Biometrics*, 33, p. 760.
WESTLAKE, W.J. (1979) Statistical aspects of comparative bioavailability trials, *Biometrics*, 35, pp. 273–280.
WESTLAKE, W.J. (1981) Response to bioequivalence testing: a need to rethink (Reader Reaction Response), *Biometrics*, 37, pp. 591–593.