STATISTICS IN MEDICINE, VOL. 12, 2367–2375 (1993)

INHERENT DIFFICULTIES WITH ACTIVE CONTROL EQUIVALENCE STUDIES*

STEPHEN SENN

Medical Department, Pharmaceutical Division, CIBA-GEIGY AG, 4002 Basle, Switzerland

SUMMARY

A simple model is used to investigate the relevance of 'competence' to active control equivalence studies (ACES). It is shown that to the extent that such trials are successful the results of such trials must raise doubts regarding their competence. ACES are thus more problematic than classical clinical trials and the problems with such studies cannot be solved simply by exchanging the usual roles of null and alternative hypotheses.

INTRODUCTION

For conditions for which there is a treatment which is at least partially effective it may be unethical to run clinical trials in which patients are given a placebo. On the other hand it may be unreasonable to expect that every addition to the therapeutic armoury should prove itself superior (on average) to an existing treatment. A new treatment which, were it the only one available, would be no better than the existing standard, may nevertheless be useful because it may be the case that some patients who are resistant to treatment with the standard treatment, or who cannot tolerate it, will benefit from taking the new treatment. Of course conversely some patients may benefit from the standard therapy who do not from the new treatment, for if this is not the case the new treatment will indeed be superior to the old. Under such circumstances patient choice is increased by the availability of the new therapy.

Makuch and Johnson[1] have proposed the designation active control equivalence study (ACES) for a type of study which they have described as follows: 'Rather than being oriented towards detecting a significant difference between two treatments, these new trials are directed toward showing that an experimental treatment is 'equivalent' in efficacy to a standard therapy' (p. 503). The justification for running such trials is perhaps in terms of the practical considerations given above. Nevertheless, it is now generally recognized that such trials pose considerable problems in interpretation and that these problems are not merely statistical.[2,3] For example, such trials can never be truly blind to the same degree as trials designed to prove superiority, since it is not necessary to know the treatment code in order to bias data towards equivalence.[4] This is not an argument against blinding such trials, since, even in a study designed to prove equivalence, a difference may nonetheless be demonstrated and under such circumstances it would be useful to be reassured that this could not be due to bias.

* Presented at the Society for Clinical Trials/International Society for Clinical Biostatistics Joint Meeting, Brussels, Belgium, July 1991.

0277–6715/93/242367–09$09.50

In this note, which is a companion paper to one previously published in *Statistics in Medicine*,[3] a simple probabilistic model will be developed to examine the problem of ACES in terms of the notions of fairness and competence. It is not claimed, however, that the treatment of the problem here is in anyway definitive and one of the main objectives of the paper is to encourage others to take up the challenge of discussing the problem of equivalence in more formal terms.

FAIRNESS AND COMPETENCE

An experiment comparing two treatments will be designated 'fair' if it accords the treatments equal status and deals with them even-handedly. To be fair it should be as nearly as possible symmetrical in all aspects except one: the treatments themselves. If the experiment itself cannot be perfectly fair with regard to allocation of subjects to treatments it should use a method of analysis, for example analysis of covariance, which makes such imbalance irrelevant. Willingness to randomize can be regarded as a declaration of fair intention on the part of the investigator. Fairness is determined on external grounds: it has nothing to do with the results. We cannot decide that the trial was blind or randomized by looking at the outcomes. We might, of course, for a trial in which deception had been practised clumsily, obtain evidence that the trial had not been randomized or blinded. On the whole, however, fairness is an issue of scientific trustworthiness.

Competence, on the other hand, can only be partially determined at best on external grounds: its most convincing demonstration is internal. A competent trial is a trial which can detect a difference between treatments where it exists but we can never be certain as to what it would take to determine such a difference. The investigator may study previous trials in the same or similar indications with similar treatments in order to determine what features he ought to incorporate into his study, but however diligently he does this he may not succeed in identifying all features which are necessary to his study. It may require, for example, that the subjects be possessed of some hidden or unknown quality in whose presence the treatments, although otherwise similar, behave differently. For example the standard treatment may be effective for many patients but scarcely at all for those enrolled in the trial because they are genetically unsuitable. Since clinical trial protocols scarcely specify the rule by which patients *will* be chosen for entry to the trial but simply, via inclusion criteria, the means by which they *may* be chosen, it will usually be difficult to assess the probability that patients treated do not differ from those in whom the drug was developed in some essential but unmeasured respect. Or it may be that the new treatment is highly toxic if used with a given concomitant treatment but that the trial protocol excludes such treatment.

There is, however, one circumstance under which we can assume a trial is competent to find a difference and that is if it finds one. This is thus the paradox of the ACES: if the stated object of the trial protocol of proving equivalence is achieved we have no proof that it was competent. If we have proof of its competence then the trial will not have demonstrated equivalence.

This problem will be investigated in terms of a simple (semi-Bayesian) model below.

A SIMPLE MODEL

Consider an ACES with two treatments. Let C stand for the condition that the trial is competent to find a difference between treatments given that they are not equivalent, and C' for the condition that it is not; and let E stand for the condition that two treatments are equal, and E' for the condition that they are not. Now suppose that only two outcomes are possible: D and D'. In terms of standard statistical conventions, D and D' might be observable values of a statistic

(although a rather unusual one) whereas E and E' and C and C' correspond to values of unobservable parameters.

The conjunction of two conditions is represented as a product. Thus $(E'\ C)$ is the combination corresponding to a competent trial in two treatments which are not equivalent.

The various likelihoods are as follows:

$$\begin{aligned} P(D|E'C) &= \pi & P(D'|E'\ C) &= 1-\pi \\ P(D|E'\ C') &= \theta & P(D'|E'\ C') &= 1-\theta \\ P(D|EC) &= \phi & P(D'|EC) &= 1-\phi \\ P(D|EC') &= \phi & P(D'|EC') &= 1-\phi, \end{aligned} \tag{1}$$

$$\text{where } 1 \geqslant \pi > \theta \geqslant \phi \geqslant 0.$$

Note that under this formulation of the likelihoods it is irrelevant as to whether the trial is competent or not given that the treatments are equivalent, because given that the treatments are equivalent the likelihoods are identical whether or not the trial is competent. Alternatively, we could regard the combination EC as impossible. Whether we do so or not is partly a matter of convention and depends on whether or not we regard the non-equivalence of the treatments as being part of the essence of competence itself. We shall take up this point again below.

Note also that we require $\pi > \theta$ because otherwise an incompetent experiment is at least as competent as a competent experiment, but this requirement is merely a linguistic one and might easily be relaxed to allow $\theta > \pi$ since all that happens in that case is that C and C' change roles. However, to have a problem worth investigating we require that $\pi \neq \theta$.

Now, suppose that the trialist assigns himself a prior probability that the treatments are equivalent, $P(E) = \beta$, and that the trial is competent given that the treatments are not equivalent, $P(C|E') = \alpha$. Given the formulation of the likelihoods above, the value of $P(C|E)$ need not be considered. We might logically regard this as being equal to zero but any value at all will give the same result.

The model is, of course, rather crude, being expressed in terms of dichotomies. For example competence might be a matter of degree rather than kind and θ might be allowed to vary between π (extreme competence) and ϕ (extreme incompetence) according to some prior distribution. We shall consider some of the difficulties associated with the crudeness of the model in the discussion at the end, but for the moment consider the consequences of the model as it stands.

Given the likelihoods and prior probabilities, we may calculate:

$$\begin{aligned} P(DE') &= P(DE'C) + P(DE'\ C') = P(E')\,(C|E')\,P(D|E'C) + P(E')\,P(C'|E')\,P(D|E'\ C') \\ &= (1-\beta)\,[\alpha\pi + (1-\alpha)\theta] \\ P(DE) &= P(DEC) + P(DEC') = \beta\phi \\ P(D'E) &= P(D'EC) + P(D'EC') = \beta(1-\phi) \\ P(D'E') &= P(D'E'C) + P(D'E'\ C') = \alpha(1-\beta)(1-\pi) + (1-\alpha)(1-\beta)(1-\theta). \end{aligned}$$

Use of Bayes' theorem then yields a posterior probability of the non-equivalence of the treatments given that D is observed:

$$P(E'|D) = \frac{(1-\beta)[\alpha\pi + (1-\alpha)\theta]}{\beta\phi + (1-\beta)[\alpha\pi + (1-\alpha)\theta]}. \tag{2}$$

On the other hand, given that D' is observed the posterior probability of equivalence is given by

$$P(E|D') = \frac{\beta(1-\phi)}{\alpha(1-\beta)(1-\pi) + (1-\alpha)(1-\beta)(1-\theta) + \beta(1-\phi)}. \tag{3}$$

Finally, we might also be interested in the posterior probability of the trial being competent given that the treatments are not equivalent but given that we have observed D'. For this we may note that $P[(C|E')|D'] = P(CE'|D')/P(E'|D')$ and that $P(E'|D') = 1 - P(E|D')$.

Now since

$$P[(CE')|D'] = \frac{\alpha(1-\beta)(1-\pi)}{\alpha(1-\beta)(1-\pi) + (1-\alpha)(1-\beta)(1-\theta) + \beta(1-\phi)}, \tag{4}$$

we have

$$P[(C|E')|D'] = \frac{\alpha(1-\pi)}{\alpha(1-\pi) + (1-\alpha)(1-\theta)}. \tag{5}$$

It may be queried as to why we are interested in the condition $C|E'$ rather than simply C. There are at least two reasons. First, by concentrating on $C|E'$ we avoid a difficulty, namely that of having to consider what we mean by a trial which is competent to find non-equivalence where in fact equivalence obtains. Second, it may be argued that the competence of the trial only affects the observed result if the treatments are in fact non-equivalent, a feature which is reflected in the formulation of the likelihoods.

CONSIDERATION OF THESE RESULTS

Suppose that, in violation of assumption (1), $\theta = \pi$, which would imply that all experiments were competent. We ought to find, therefore, that the value of α would be irrelevant. This does, in fact, turn out to be the case, for substitution of π for θ in (2) and (3) yields

$$P(E'|D) = \frac{(1-\beta)\pi}{\beta\phi + (1-\beta)\pi} \tag{6}$$

$$P(E|D') = \frac{\beta(1-\phi)}{\beta(1-\phi) + (1-\beta)(1-\pi)}. \tag{7}$$

These are, of course, the familiar expressions for posterior probabilities in cases where the competence of experiments is taken for granted, and they may be obtained equivalently by substituting $\alpha = 1$ in (2) and (3).

I shall assume, however, that θ is not in general equal to π, and in particular that although π may be arbitrarily increased and ϕ decreased by designing better and better experiments, the joint effect of α and θ represents a hidden element in any experiment which is beyond the experimenter's ability to control. Consider, therefore, what happens to (2) and (3) when $\pi \to 1$ and $\phi \to 0$. Under such circumstances we find that

$$P(E'|D) \to 1 \tag{8}$$

but that

$$P(E|D') \to \frac{\beta}{(1-\alpha)(1-\beta)(1-\theta) + \beta}. \tag{9}$$

We thus see that there is an asymmetry in refuting the hypothesis of equality of treatments and failing to do so.

Matters may be taken a little further, however, since it is not clear where knowledge of α might come from. Suppose that $\theta = 0$ and $\alpha = 0$: then substitution into (9) yields β as the limit of $P(E|D')$. Now we might argue that this is to be unduly pessimistic about α. Surely our experience can allow us to give a higher value than 0 to α? Consider, however, what experience in this experiment tells us about α.

Suppose we define the support given for the competence of the experiment given that the treatments are not equal but that no difference is observed, $S[(C|E')|D']$, as the difference between the posterior and prior conditional probabilities of competence. Thus

$$S[(C|E')|D'] = P[(C|E')|D'] - P[(C|E')],$$

from which

$$S[(C|E')|D'] = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\alpha)(1-\theta)} - \alpha = \frac{-\alpha(1-\alpha)(\pi-\theta)}{\alpha(1-\pi)+(1-\alpha)(1-\theta)}. \tag{10}$$

Now, it will be seen that, except where $\alpha = 1$ (which implies that all experiments are competent given non-equivalence of the treatments) or where $\theta \geqslant \pi$, relation (10) is negative; and in fact, as $\pi \to 1$, (10) $\to -\alpha$ and hence (5) $\to 0$. There thus can be no support from this experiment for the value of α unless we succeed in proving that the treatments are not equivalent. The more we appear to have proved the equivalence of the treatments, the more we ought to doubt the competence of our experiment.

Graphical illustrations of some examples of these results are given in Figures 1 to 3. Figure 1 shows the posterior probability of non-equivalence given an observed difference for two values of α: $\alpha = 1$ corresponding to a trial for which, *a priori*, competence may be assumed with absolute confidence; and $\alpha = 0{\cdot}5$ for a trial in which the competence is in some doubt. The value of ϕ has been set to 0·05, corresponding to the common value used in tests of significance (but see discussion below) and θ has also been set to 0·05. Given the value of ϕ, this corresponds to extreme incompetence. The prior probability of equivalence, β, has been set to 0·5. The posterior probability of E' is plotted as a function of π. As π increases, this posterior probability approaches 0·952 for the case of $\alpha = 1$ and 0·913 for the case of $\alpha = 0{\cdot}5$. Clearly, for this example, the prior competence or otherwise of the experiment has little effect on the interpretation. To increase the posterior probability of non-equivalence beyond these limits we need to have trials in which ϕ is decreased.

Figure 2 shows a plot of the posterior probability of equivalence given an observed lack of difference for $\phi = \theta = 0{\cdot}05$, $\beta = 0{\cdot}5$ and $\alpha = 1$ against π. As π increases the posterior probability approaches 1. Figure 3 on the other hand considers an identical setting of parameters to Figure 2 except that $\alpha = 0{\cdot}5$. Here, as π increases the posterior probability of equivalence approaches 2/3. In this case decreasing ϕ will not improve matters if the consequence of this is that θ decreases also. Also illustrated in Figure 3 is the posterior probability of competence given an observed lack of difference and given genuine difference between treatments as a function of π. It will be seen that as π approaches 1 this declines to 0.

DISCUSSION

Some caution is indicated in interpreting these results. The results show an asymmetry between equivalence and non-equivalence in clinical trials. To a certain extent, however, this asymmetry is

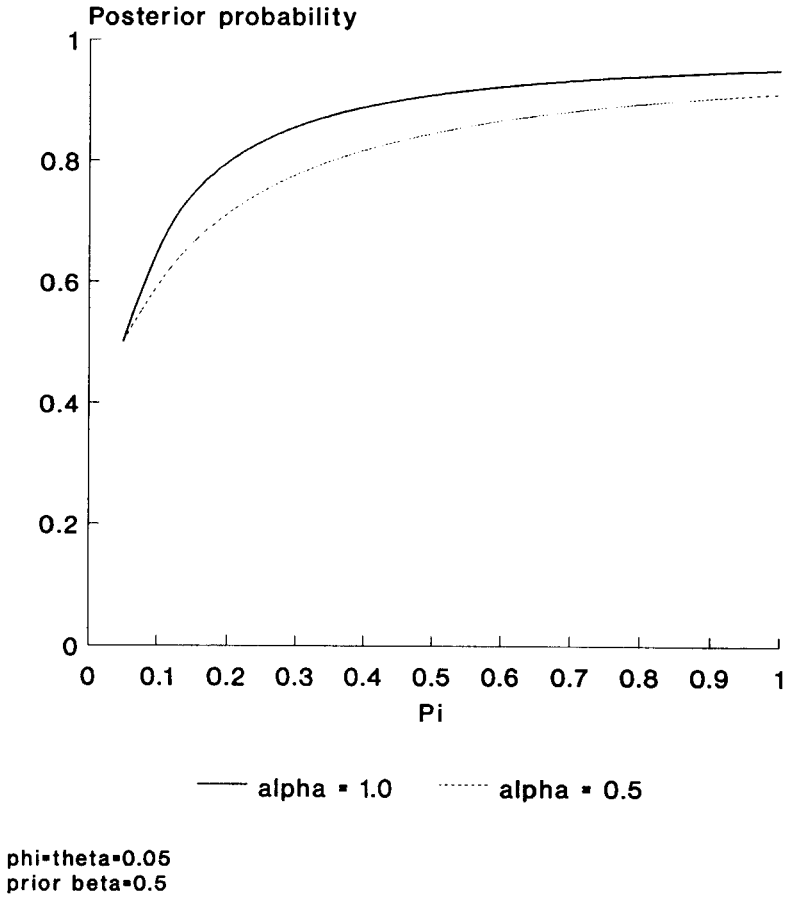

Figure 1. Posterior probabilities of non-equivalence given an observed difference

built into the formulation which takes it for granted that the purpose of a clinical trial is to show that at least some patients benefit from treatment – not that all will. If it were to be regarded as being the object of a clinical trial to prove that all patients, or even all patients of a given type, benefited from a given treatment (even if benefit were only measured in terms of a probability of being cured), then the issue of competence would be one which would also affect trials in which demonstrated superiority of one treatment to another were the outcome. One might, for example, then ask whether the patients in the trial were essentially similar to future patients on whom the successful treatment would be used. This is then a 'competence' issue of sorts.

I have argued elsewhere that clinical trials require a 'falsificationist' view:[3] that they are means by which statements of the sort 'the treatments are always equal' may be shown to be false. To demonstrate that such a statement is false it is sufficient to show that it is not true for a given group of patients. To go further than this and suggest that the treatments are unequal for other patients requires a means of establishing a general pattern from particular instances, a process which is known in philosophy as 'induction'.

It should also be noted that although a 'falsificationist' view of clinical trials would also superficially appear to be more in tune with frequentist methods than with Bayesian ones, the model shows some problems with the frequentist formulation where an alternative hypothesis

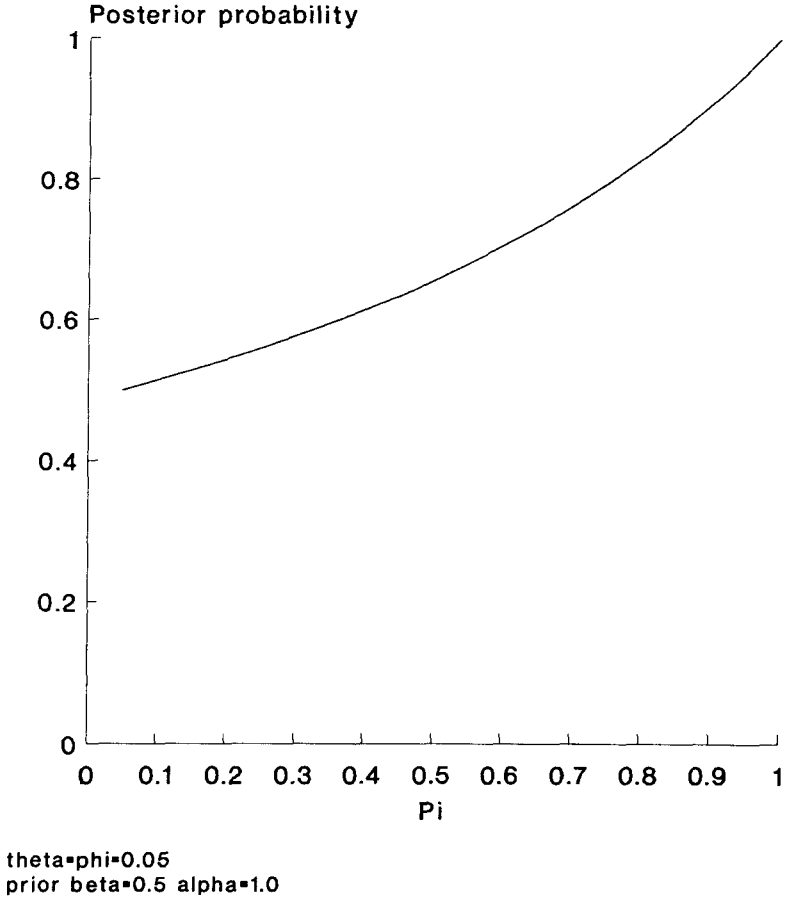

Figure 2. Posterior probability of equivalence given a lack of observed difference

exists. In the discussion of Figure 1, for example, it was noted that the posterior probability of non-equivalence given an observed difference could only be increased beyond a certain limit by reducing ϕ. Now if the event D were 'significant difference' and the event D' were 'non-significant difference' then ϕ and π might correspond to size and power in a Neyman–Pearson formulation. This would show that, given a large number of patients, we should eventually be better off reducing the nominal size of our test rather than continuing to increase power, a result which is related to the Lindley paradox.[5] It must also be pointed out that we have avoided some of the difficulties which the famous paradox poses for frequentist methods by only allowing ourselves to observe a simple dichotomy D or D'. Where we carry out a test of significance we replace an observed statistic capable of taking on many values by a label 'significant' or 'not-significant', and this is not a legitimate summary of the evidence where a null hypothesis is being tested against a fixed alternative or even a well defined class of alternatives.

It must also be conceded that specification of such alternatives is an essential part of the Neyman–Pearson approach, and also, in a sense, of Bayesian approaches. The Fisher significance test does not involve an alternative hypothesis and, indeed, Fisher regarded test statistics as having logical priority to alternative hypotheses and not vice versa as is the case in the Neyman–Pearson theory (Reference 6, p. 246). Lindley showed, however, that even very weak

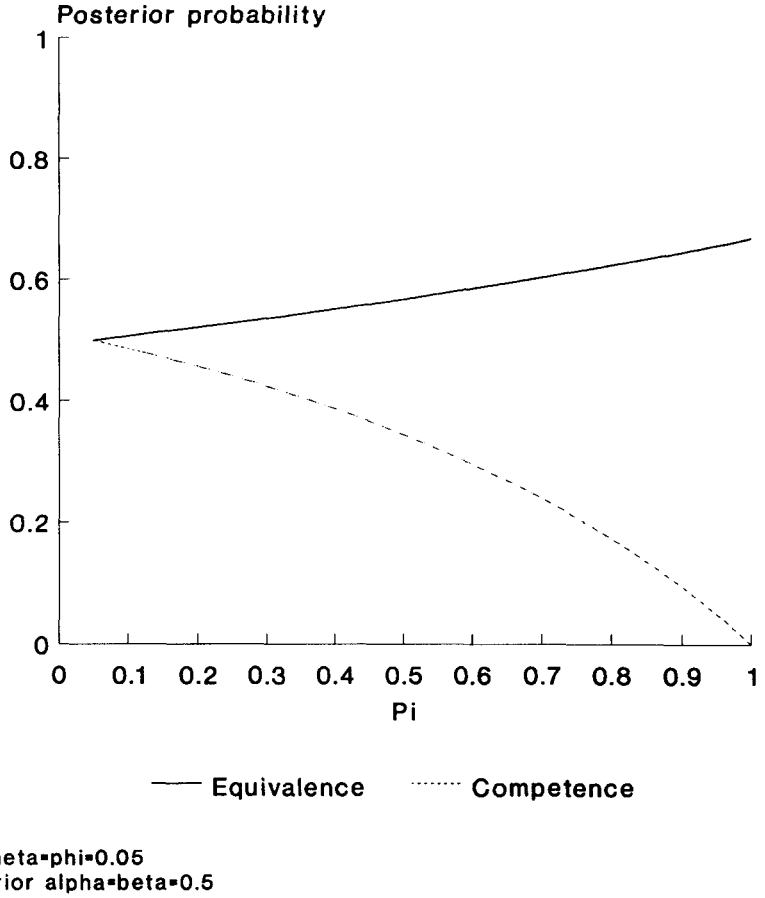

Figure 3. Posterior probabilities given a lack of observed difference

theories about alternatives could cause problems for the significance test,[5] although his own formulation of the alternative hypothesis rather disguises the fact that different alternative theories will produce very different results. Johnstone[7] provides an interesting account of the standard Bayesian criticisms and Barnard[8] provides a defence of Fisher.

Two aspects of the model presented are, perhaps, worthy of attention. First, Fisher (Reference 9, p. 45) considered that 'A test of significance contains no criterion for "accepting" a hypothesis' (by hypothesis he meant a true null hypothesis, which in this context would be the hypothesis of equality) and this at least implies that within his system of statistics, a demonstration of equivalence cannot be handled in the same way as a 'proof' of a difference. Some have seen this as a weakness. Most modern Bayesian formulations do not recognize such a distinction; nor – although it is a point which many commentators, including at least one famous Bayesian (Reference 10, p. 181), seem to have missed – does the Neyman–Pearson formulation. Provided that the 'null' hypothesis is declared to be the hypothesis that the treatments differ by at least some amount, then a Neyman–Pearson decision rule may (in certain circumstances) be constructed which allows, given a suitable value of the test statistic, the acceptance of the alternative hypothesis of equivalence. The model above suggests, however, that there may be some merit in

recognizing a fundamental difference between equivalence and not-equivalence and it may be that the fact that the Fisherian approach does so is a strength not a weakness.

The second point of interest is that the model may illustrate a phenomenon explicitly drawn attention to by Popper and Miller.[11] They have investigated the nature of probabilistic support and come to the following conclusion: 'Although evidence may raise the probability of a hypothesis above the value it achieves on background knowledge alone, every such increase in probability has to be attributed entirely to the *deductive connections* that exist between the hypothesis and the evidence' (p. 569, original italics). In their view any inductive 'support' is countersupport.

The results above show at least a superficial parallel to these findings; for, given an assumption concerning α and the other assumptions implicit in the formulation above, deductive support may be produced for the theory of equivalence. To the extent, however, that the equivalence of treatments under this assumption is supported by the evidence, the assumption itself regarding α can only be countersupported. The support offered by d' to the hypothesis of equivalence is genuine, deductive but entirely conditional, whereas the support offered by d to the hypothesis of non-equivalence is genuine, deductive and unconditional. These two forms of support, therefore, take place on entirely different levels.

If this interpretation is accepted, then the result which follows is what is offered as the major conclusion of this paper: there is a fundamental logical difference between concluding, as a result of running a clinical trial, that the effects of treatments are different, and concluding that they are the same. My own view is that no amount of mathematical juggling can remove this distinction.

ACKNOWLEDGEMENT

I thank two anonymous referees for helpful comments on an earlier draft of this paper.

REFERENCES

1. Makuch, R. and Johnson, M. 'Issues in planning and interpreting active control equivalence studies', *Journal of Clinical Epidemiology*, **42**, 503–511 (1989).
2. Temple, R. 'Government viewpoint of clinical trials', *Drug Information Journal*, **16**, 10–17 (1982).
3. Senn, S. J. 'Falsificationism and clinical trials', *Statistics in Medicine*, **10**, 1679–1692 (1992).
4. Senn, S. J. 'Clinical trials and epidemiology', *Journal of Clinical Epidemiology*, **43**, 628–632 (1990).
5. Lindley, D. V. 'A statistical paradox', *Biometrika*, **44**, 187–192 (1957).
6. Fisher, R. A. Letter to C. I. Bliss in Bennet, J. H. (ed.), *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*, Oxford University Press, Oxford, 1990.
7. Johnstone, D. J. 'Tests of significance in theory and practice', *The Statistician*, **35**, 491–498 (1986).
8. Barnard, G. A. 'Tests of significance in theory and practice: discussion', *The Statistician*, **35**, 499–502 (1986).
9. Fisher, R. A. *Statistical Methods and Scientific Inference* (1956), reprinted in Bennett, J. H. (ed.), *Statistical Methods, Experimental Design and Scientific Inference*, Oxford University Press, Oxford, 1990.
10. Savage, J. 'Bayesian statistics', in Ericson, L. (ed.) *The Writings of Leonard Jimmie Savage – A Memorial Selection*, The American Statistical Association and the Institute of Mathematical Statistics, 1981.
11. Popper, K. and Miller, D. 'Why probabilistic support is not inductive', *Philosophical Transactions of the Royal Society of London, Series A*, **321**, 569–591 (1987).