Trials to assess equivalence: the importance of rigorous methods

B Jones, P Jarvis, J A Lewis, A F Ebbutt

The aim of an equivalence trial is to show the therapeutic equivalence of two treatments, usually a new drug under development and an existing drug for the same disease used as a standard active comparator. Unfortunately the principles that govern the design, conduct, and analysis of equivalence trials are not as well understood as they should be. Consequently such trials often include too few patients or have intrinsic design biases which tend towards the conclusion of no difference. In addition the application of hypothesis testing in analysing and interpreting data from such trials sometimes compounds the drawing of inappropriate conclusions, and the inclusion and exclusion of patients from analysis may be poorly managed.

The design of equivalence trials should mirror that of earlier successful trials of the active comparator as closely as possible. Patient losses and other deviations from the protocol should be minimised; analysis strategies to deal with unavoidable problems should not centre on an "intention to treat" analysis but should seek to show the similarity of results from a range of approaches. Analysis should be based on confidence intervals, and this also carries implications for the estimation of the required numbers of patients at the design stage.

The gold standard in clinical research is the randomised placebo controlled double blind clinical trial. This design is favoured for confirmatory trials carried out as part of the phase III development of new medicines. Because of the number and range of medicines already available, however, new medicines are increasingly being developed for indications in which a placebo control group would be unethical. In such situations one obvious solution is to use as an active comparator an existing drug already licensed and regularly used for the indications in question. Some authors have questioned whether placebo controlled trials are used excessively and unethically,[1-3] and such views would reinforce the trend towards using active comparators. Others have proposed that, once licensed, new drugs should be compared with existing treatments for the same indication in order to examine their relative cost effectiveness and that large randomised trials are the appropriate tool.[4]

When an active comparator is used the expectation may sometimes be that the new treatment will be better than the standard, and the objective is to demonstrate this fact unequivocally. This situation is similar to using a placebo control and poses no special methodological problems. More probably, however, the new treatment is simply expected to match the efficacy of the standard treatment but have advantages in safety, convenience, or cost; in some cases the new treatment may have no immediate advantage but may present an alternative or second line therapy. Under these circumstances the objective of the trial is to show equivalent efficacy—the so called "equivalence" trial. Such trials have been referred to as "active control equivalence studies"[5] or "positive control studies."[6]

This paper describes the methodological issues that surround equivalence trials and explains their implications. We explain why equivalence trials generally need to be larger than their placebo controlled counterparts; why their standard of conduct needs to be especially high; why the handling of withdrawals, losses, and protocol deviations needs more care than usual; and why different approaches to analysis and interpretation are appropriate. A proper appreciation of these issues ensures that when equivalence trials are conducted they reach the scientific standards necessary for reliable conclusions to be drawn.

There are two fundamental methodological features of equivalence trials which underlie the general approach to their design and analysis, and these will be addressed first. These features distinguish equivalence trials from trials whose aim is to detect a difference between two treatments and which are referred to here as "comparative" trials.

Confidence intervals and sample size

The first feature relates to the statistical methods used for analysis and the consequences for determining the required number of patients. In a comparative trial the standard analysis uses statistical significance tests to determine whether the null hypothesis of "no difference" may be rejected, together with confidence limits to place bounds on the possible size of the difference between the treatments. In an equivalence trial the conventional significance test has little relevance: failure to detect a difference does not imply equivalence;[7] a difference which is detected may not have any clinical relevance and may correspond to practical equivalence. The relevance of the confidence interval, however, is easier to see. This defines a range for the possible true difference between the treatments, any point of which is reasonably compatible with the observed data. If every point within this range corresponds to a difference of no clinical importance then the treatments may be considered to be equivalent.

It is important to emphasise that absolute equivalence can never be demonstrated: it is possible only to assert that the true difference is unlikely to be outside a range which depends on the size of the trial, the results of the trial, and the specified probabilities of error. If we have predefined a range of equivalence as an interval from $-\Delta$ to $+\Delta$ we can then simply check whether the confidence interval centred on the observed difference lies entirely between $-\Delta$ and $+\Delta$. If it does, equivalence is demonstrated; if it does not, there is still room for doubt.

Possible results of the comparison of a confidence interval with a predefined range of equivalence are shown in figure 1, and the importance of not basing conclusions on statistical significance can also be seen in this figure. Any confidence interval which does not overlap zero corresponds to a statistically significant difference.

This intuitive procedure of checking whether a confidence interval lies within a range of equivalence does in fact correspond to a significance testing procedure, but one in which the roles of the usual null and alternative hypotheses are reversed. In comparative trials the null hypothesis is that there is no difference between the treatments. The alternative hypothesis is that a difference exists. In equivalence testing the relevant null hypothesis is that a difference of at least Δ exists, and the trial is targeted at disproving this in favour of the alter-

Department of Medical Statistics, School of Computing Sciences, De Montfort University, Leicester LE1 9BH
B Jones, *professor of medical statistics*
P Jarvis, *senior lecturer in medical statistics*
J A Lewis, *visiting professor in medical statistics*

Glaxo Wellcome Ltd, Greenford, Middlesex UB6 0HE
A F Ebbutt, *director of European clinical statistics*

Correspondence to:
Professor J A Lewis, Medicines Control Agency, London SW8 5NQ.

BMJ 1996;313:36–9

native that no difference exists. This formulation is important in validating the intuitive confidence interval procedure, and it also helps in calculating sample sizes. The formulas for calculating sample sizes for normally distributed and binary data are provided in the appendix. Values need to be specified for the range of equivalence (Δ) and the probabilities of type I and II errors (α and β, respectively). An important point to note is that if a $100(1-2\alpha)\%$ confidence interval is used to decide on equivalence then the significance level is α—that is, the probability of the type I error is α. So, for example, if a 95% interval is used then $\alpha = 0.025$. The choice of Δ is difficult and requires extensive debate with knowledgable clinical experts, and the chosen Δ should generally be smaller than in a comparative trial. In comparative trials against placebo, Δ is often set equal to a difference of undisputed clinical importance, and hence may be above the minimum difference of clinical interest by a factor of perhaps two or more; there may be scientific reasons to expect a treatment to have more than a minimal effect. However, when comparing a new agent with a standard comparator it is necessary to show that the new agent is sufficiently similar to the standard to be clinically indistinguishable. This entails using smaller values of Δ than were used to detect the effect of the standard relative to placebo. A factor of two does not seem inappropriate, leading to sample sizes roughly four times as large as those in similar comparative trials.

The selection of α and β follows similar lines as for comparative trials. The use of a 95% confidence interval in an equivalence trial, as recommended by the European Committee for Proprietary Medicinal Products in its note for guidance on biostatistics,[8,9] corresponds to a value for α of 0.025. However, β is treated identically, and is generally set to 0.1 (to give a power of 0.90) or 0.2 (to give a power of 0.8).

The distinction between one sided and two sided tests of statistical significance also carries over into the confidence interval approach. For a one sided test equivalence is declared if the lower one sided confidence limit exceeds $-\Delta$. This approach is indicated when the objective is to ensure that the new agent is not inferior to the standard. Equivalence or superiority are both regarded as positive outcomes.

Internal validity of trials

The second special feature affecting the equivalence trial is the lack of any natural internal check on its validity.[6] In a comparative trial there is a strong incentive to remove any sloppiness in design, conduct, and analysis because such sloppiness is likely to obscure any differences between the treatments. As a consequence, the detection of a treatment difference not only implies that a difference exists but also that the trial was of sufficient quality to detect it. Such an incentive and natural check on quality are lacking in an equivalence trial, where the finding of equivalence may arise either from true equivalence or from a trial with poor discriminatory power—a trial which was too small, for example, or one in which most patients were likely to improve spontaneously without medical intervention.

The finding in a trial that two treatments are equivalent does not require that both treatments were effective; it is equally compatible with the alternative that neither was. In any equivalence trial, therefore, it is vitally important to have some means of confirming that both treatments were indeed effective. We need to be certain that if a third placebo arm had been included both active treatments would have been shown to be superior to placebo.

The degree of certainty can be increased only by paying careful attention to the design of the equivalence trial, by being strict about matters of conduct, and by making additional checks during analysis. The active comparator is usually a licensed medicine which has been evaluated in controlled trials against placebo, perhaps during the phase III studies used to support its marketing application. If the equivalence trial mirrors as closely as possible the methods used in these earlier placebo controlled trials then confidence in its results will be increased, since the methods have been positively validated in a similar context.

Important design features to follow as closely as possible are the inclusion and exclusion criteria (defining the patient population), the dosing schedule of the standard treatment, the use of concomitant medication and other interventions, and the primary response variable and its schedule of measurements. During analysis it is valuable to show similarities between the equivalence trial and the earlier comparative trials in terms of patient compliance, the response during any run in period, and the scale of patient losses and the reasons for them.

The two major features covered so far provide the background for some brief comments on other considerations in the design, conduct, and analysis of equivalence trials.

Design and conduct

The amount of information available to plan an equivalence trial will generally exceed the amount available at the

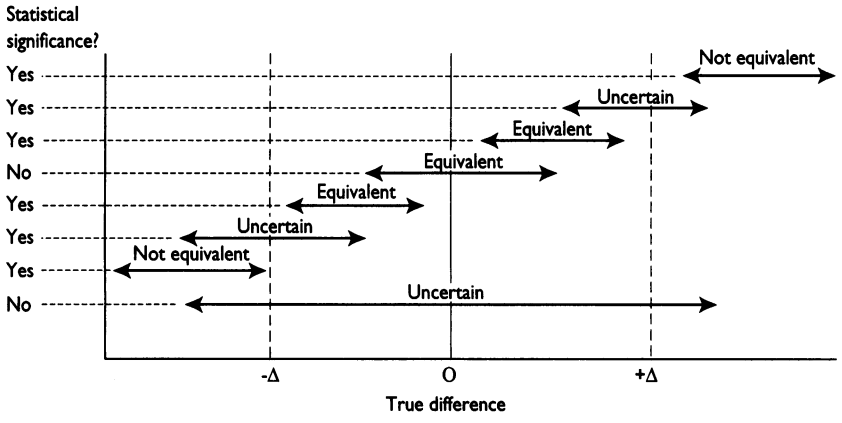

Fig 1—*Examples of possible results of using the confidence interval approach: $-\Delta$ to $+\Delta$ is the prespecified range of equivalence; the horizontal lines correspond to possible trial outcomes expressed as confidence intervals, with the associated significance test result shown on the left; above each line is the decision concerning equivalence*

Example of sample size calculation

Two inhalers used for the relief of asthma attacks are to be assessed for equivalence. They will be considered equivalent if the 95% two sided confidence interval for the treatment difference, measured using morning peak expiratory flow rate (l/min), falls wholly within the interval ±15 l/min —that is, $\Delta = 15$ and $\alpha = (1-0.95)/2$. From a previous trial the prior estimate of δ^2, the between subject variance of morning peak expiratory flow rate, is 1600 $(l/min)^2$. The sample size of each group is to be such that there is a power of 0.80 that the inhalers will be deemed equivalent if they are, in fact, identical. To use the formula for normally distributed data given in the appendix, we note that

$$z_{(1-\alpha)} = z_{(0.975)} = 1.96$$

and $z_{(1-\frac{\beta}{2})} = z_{(0.90)} = 1.28$

from tables of the normal distribution, so

$$n = \frac{2 \times 40^2}{15^2}[1.96 + 1.28]^2 = 149.3 \approx 150.$$

Each group should contain 150 patients.

Example: Assessment of equivalence

Two inhalers, R and T, used for the relief of asthma attacks were compared in an equivalence trial using morning peak expiratory flow rate (l/min) as the primary measurement. The range of equivalence was set at ±15 l/min—that is, $\Delta = 15$. The results of the trial were as follows:

Mean morning peak expiratory flow rate on treatment

R = 420 l/min (150 patients)

T = 417 l/min (150 patients)

Mean difference between R and T, $\bar{d} = 3$

Estimated standard error of the mean difference,

$SE(\bar{d}) = 4$

The 95% confidence interval for the true difference ranges from $-1.96\ SE(\bar{d})$ to $+1.96\ SE(\bar{d})$, where 1.96 is the appropriate value from tables of the normal distribution (that is, $z_{(0.975)}$). This interval is −4.8 to 10.8 and lies entirely within the range of equivalence of −15 l/min to +15 l/min and so equivalence is confirmed.

time of planning earlier trials of the active comparator. There should be little excuse, therefore, for poor design. Double blinding of medication may pose extra difficulties but is no less important than in comparative trials, and randomisation is equally important. Inclusion and exclusion criteria must be carefully chosen on the basis of prior experience of the active comparator to ensure that the trial contains patients likely to respond to the active comparator and hence avoid a conclusion of equivalence through non-response. Care in this choice should be mirrored in the response observed to the trial treatments. The level of success for success/failure outcomes should be similar to that seen in previous trials of the active comparator. For more quantitative endpoints, improvements from baseline in the course of the trial provide some assurance that the trial treatments have both been effective.

The dosing regimen and period of dosing of the active comparator should reflect the standard manner of use known to be effective on the basis of earlier clinical trials; and there should be a sound rationale for the choice of the potentially equivalent dosing regimen of the new medication. If the doses chosen for both agents are too high then patients may reach an upper threshold in response, leading to a conclusion of equivalence which may not carry over to the doses more likely to be used in practice. Unreasonably low doses may lead to similar false conclusions, through lack of response. It is sometimes necessary to check that all patients can tolerate one or both treatments in order to maintain patient numbers and hence power, and this should be done during a run in period before randomisation.

The use in all patients of a standard dose of concomitant medication with known beneficial effects can also result in patients reaching their upper threshold of response and hence lead to the masking of treatment differences. Alternatively, if the use of concomitant medication is flexible, greater use in one arm of the trial may produce a bias towards equivalence. Similar biases towards equivalence can arise from the use of "rescue" medication in patients in whom treatment fails—that is, from patients who withdraw from randomised treatment because of lack of efficacy. These issues are closely connected with the means adopted for dealing with such patients in the analysis.

Analysis

The most difficult issue relating to the analysis of an equivalence trial concerns which patients and which data from these patients to include. The most common approaches to the analysis of randomised trials are "intention to treat" and "per protocol" analyses. A fuller discussion of intention to treat can be found in Lewis and Machin,[10] and a severe criticism in Salsburg.[11]

In an intention to treat analysis patients are analysed according to their randomised treatment, irrespective of whether they actually received the treatment. Patients may fail to take a treatment altogether, may be given the wrong treatment, or may violate the protocol in some other way, but under an intention to treat analysis this does not affect matters. The strength claimed for such an analysis is that it is pragmatic—that is, that it mirrors what will happen when the treatment is used in practice. In a comparative trial, where the aim is to decide if two treatments are different, an intention to treat analysis is generally conservative: the inclusion of protocol violators and withdrawals will usually tend to make the results from the two treatment groups more similar. However, for an equivalence trial this effect is no longer conservative: any blurring of the difference between the treatment groups will increase the chance of declaring equivalence.

A per protocol analysis compares patients according to the treatment actually received and includes only those patients who satisfied the entry criteria and properly followed the protocol. This approach might be expected to enhance any difference between the treatments rather than diminishing it, because of the removal of uninformative "noise." Unfortunately it is possible to envisage circumstances under which the exclusion of patients in a per protocol analysis might bias the results towards a conclusion of no difference—for example, if patients not responding to one of the two treatments dropped out early. For this reason the subgroup of patients excluded from a per protocol analysis should be examined carefully to explore whether any biases of this nature might have occurred. Indeed, if the two treatments produce a different pattern of withdrawal for adverse events or lack of effect then this in itself is evidence that they are not entirely equivalent.

In an equivalence trial it is probably best to carry out both types of analysis and hope to show equivalence in either case. In preparation for this policy it is important to collect complete follow up data on all randomised patients as per protocol, irrespective of whether they are subsequently found to have failed entry criteria, withdraw from trial medication prematurely, or violate the protocol in some other way. Such a rigid approach to data collection allows maximum flexibility during later analysis and hence provides a more robust basis for decisions.

With respect to other aspects of analysis, equivalence trials are similar in nature to comparative trials.

The result of the analysis of the primary endpoint should be one of the following:

- that the confidence interval for the difference between the two treatments lies entirely within the equivalence range so that equivalence may be concluded with only a small probability of error;
- that the confidence interval covers at least some points which lie outside the equivalence range, so that differences of potential clinical importance remain a real possibility and equivalence cannot safely be concluded; and
- that the confidence interval is wholly outside the equivalence range (though this is likely to be rare).

Discussion

The most common failing of reported equivalence studies is that they are planned and analysed as if they were comparative studies, and the lack of a statistically significant difference is then taken as proof of equivalence. The material covered in this paper should make it clear that such an approach is likely to lead to wrong conclusions.

Improvements to the standards of this type of research could be encouraged if journal editors and ref-

erees adopted a more critical attitude. The following is a suggested minimal set of criteria against which to judge reports of clinical trials in which the equivalence of two treatments is claimed.

- The size of the trial should be based on a null hypothesis of non-equivalence and an alternative hypothesis of equivalence.
- Conclusions should be drawn on the basis of an appropriate confidence interval using the prespecified criteria of equivalence used in the sample size calculation.
- The results of both intention to treat and per protocol analyses should be presented.
- There should be adequate evidence on the rigour of the trial and of the similarity of important features of design to those of earlier comparative trials which showed useful clinical effects.
- The trial data should provide some evidence of the efficacy of the treatments; this might be success rates similar to those of previous trials, or clinically important changes from baseline treatments.
- Some of these aspects could most easily be covered by insisting that papers submitted to journals referred to published trials of the standard comparator against placebo with similar methods. Referees should also be familiar with the special difficulties surrounding equivalence trials in the relevant clinical area.
- Improving the standards of equivalence trials has consequences for the resources required. Such trials will become larger and their monitoring will become more labour intensive in order to ensure they are conducted in close accordance with the protocol, so minimising the occurrence of biases towards a conclusion of no difference.

Funding: BJ and PJ thank Glaxo Wellcome Ltd for providing a research grant.
Conflict of interest: None.

Appendix: Sample size and power formulas

NORMALLY DISTRIBUTED DATA (COMPARISON OF MEANS)

We assume that subjects are randomised into two treatment groups of equal size n, the groups being denoted by R (reference treatment) and T (test treatment). Let μ_R and μ_T denote the expected mean values of the normally distributed observations in groups R and T, respectively, and let s^2 be an estimate of σ^2, the variance of the observations, assumed to be the same in the two groups.

In the confidence interval approach equivalence is concluded if the interval falls entirely within two prespecified tolerance limits, $-\Delta$ and $+\Delta$. If $\bar{x}_R$ and $\bar{x}_T$ denote the observed means of the reference and treatment groups respectively, then, provided n is reasonably large, the two sided 100 $(1-2\alpha)$% confidence interval for $\mu_R - \mu_T$ is

$$\bar{x}_R - \bar{x}_T \pm z_{(1-\alpha)}\sqrt{2s^2/n}$$

where $z_{(1-\alpha)}$ is the 100$(1-\alpha)$% point of the normal distribution. That is, if X has the standard normal distribution with mean 0 and variance 1 then,

$$Pr(X \leq z_{(1-\alpha)}) = 1 - \alpha.$$

When the confidence interval (or significance testing) approach is used to assess equivalence, two sorts of mistake can occur: we can decide that the treatments are equivalent when they are not (the type I error with probability α) or we can decide the treatments are not equivalent when they are (the type II error with probability β). These definitions are an exact switch of those applying to conventional significance testing.[12] The values of α and β depend on the size of the true difference between the treatment means $\delta = \mu_R - \mu_T$. The value of α reaches a maximum on the boundary of the range of equivalence (that is, when $|\delta| = \Delta$) and this is the value of α used in calculations. The value of β is usually calculated at the point of equivalence (that is, at $\delta = 0$). The corresponding power of the trial, $1-\beta$, is the probability of correctly declaring equivalence when $\delta = 0$.

The sample size and the power formulas for a 100 $(1-2\alpha)$% two sided interval are as follows.

The null hypothesis is $H_0 : |\mu_R - \mu_T| \geq \Delta$ (inequivalence)

The alternative hypothesis is $H_1 : -\Delta < \mu_R - \mu_T < \Delta$ (equivalence)

$$n = \frac{2s^2}{\Delta^2}\left[z_{(1-\alpha)} + z_{(1-\frac{\beta}{2})}\right]^2$$

$$Power = 2\Phi\left(\frac{\Delta}{\sqrt{s^2\left(\frac{2}{n}\right)}}\right) - z_{(1-\alpha)} - 1$$

where $\Phi(x)$ denotes $Pr(X \leq x)$ and X has the standard normal distribution with mean 0 and variance 1.

For a 100$(1-\alpha)$% one sided interval the corresponding formulas are:

$H_0 : \mu_R - \mu_T \geq \Delta$ (inequivalence)

$H_1 : \mu_R - \mu_T < \Delta$ (equivalence)

$$n = \frac{2s^2}{\Delta^2}\left[z_{(1-\alpha)} + z_{(1-\beta)}\right]^2$$

$$Power = \Phi\left(\frac{\Delta}{\sqrt{s^2\left(\frac{2}{n}\right)}}\right) - z_{(1-\alpha)} - 1$$

BINARY DATA (COMPARISON OF PERCENTAGES)

Using notation found in Pocock,[13] we define p to be the overall percentage of successes to be expected if the treatments are equivalent and use Δ to define the range of equivalence for the difference in percentage success rates. Other notation is unchanged.

The required size of each treatment group and the power can be calculated as follows[14]:

- Two sided case:

$$n = \frac{2p(100-p)}{\Delta^2}\left[z_{(1-\alpha)} + z_{(1-\frac{\beta}{2})}\right]^2$$

$$Power = 2\Phi\left(\frac{\Delta}{\sqrt{p(100-p)\left(\frac{2}{n}\right)}} - z_{(1-\alpha)}\right) - 1$$

- One sided case:

$$n = \frac{2p(100-p)}{\Delta^2}\left[z_{(1-\alpha)} + z_{(1-\beta)}\right]^2$$

$$Power = \Phi\left(\frac{\Delta}{\sqrt{p(100-p)\left(\frac{2}{n}\right)}} - z_{(1-\alpha)}\right) - 1$$

1 Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med* 1994;331:394-8.
2 Taubes G. Use of placebo controls in clinical trials disputed. *Science* 1995;267:25-6.
3 The use of placebo controls. *N Engl J Med* 1995;332:60-2.
4 Henry D, Hill S. Comparing treatments: comparison should be against active treatments rather than placebo. *BMJ* 1995;310:1279.
5 Makuch RW, Pledger G, Hall DB, Johnson MF, Herson J, Hsu J-P. Active control equivalence studies. In: Peace K, ed. *Statistical issues in drug research and development*. New York: Marcel Dekker, 1990:225-62.
6 Temple R. Government viewpoint of clinical trials. *Drug Information Journal* 1982;16:10-7.
7 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
8 Lewis JA, Jones DR, Röhmel J. Biostatistical methodology in clinical trials—a European guideline. *Statistics in Medicine* 1995;14:1655-7.
9 CPMP Working Party on Efficacy of Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. Note for guidance III/3630/92-EN. *Statistics in Medicine* 1995;14: 1658-82.
10 Lewis JA, Machin D. Intention to treat—who should use ITT. *Br J Cancer* 1992;68:647-50.
11 Salsburg D. Intent to treat: the reductio ad absurdum that became gospel. *Pharmacoepidemiology and Drug Safety* 1994;3:329-35.
12 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical and continuous outcomes in two group comparisons. *BMJ* 1995;311:1145-8.
13 Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley, 1983.
14 Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports* 1978;62:1037-40.