STATISTICS IN MEDICINE
Statist. Med. **17**, 1691–1701 (1998)

PRACTICAL ISSUES IN EQUIVALENCE TRIALS

A. F. EBBUTT* AND L. FRITH

European Clinical Statistics, Glaxo Wellcome, Greenford Road, Greenford, Middlesex, UB6 0HE, U.K.

SUMMARY

Equivalence trials aim to show that two treatments have equivalent therapeutic effects. The approach is to define, in advance, a range of equivalence $-d$ to $+d$ for the treatment difference such that any value in the range is clinically unimportant. If the confidence interval for the difference, calculated after the trial, lies entirely within the interval, then equivalence is claimed. Glaxo Wellcome has carried out a series of trials using this methodology to assess new formulations of inhaled beta-agonists and inhaled steroids in asthma. Eleven of these trials are used to review some practical issues in equivalence trials. For the series of asthma trials, a range for peak expiratory flow rate (PEF) from -15 to $+15$ l/min was chosen to be the range of equivalence. This fitted well with physicians' opinions and with previously demonstrated differences between active and placebo. The choice of the size of the confidence interval should depend on the medical severity of the clinical endpoints under consideration and the level of risk acceptable in assuming equivalence if a difference of potential importance exists. From this point of view, a recommendation in the CPMP Note for Guidance on Biostatistics that 95 per cent confidence intervals should be used is inappropriate. Intent-to-treat (ITT) and per-protocol (PP) analyses were compared for the eleven asthma trials. Confidence intervals were always wider for the PP analysis and this was entirely due to the smaller number of subjects included in the PP analysis. There was no evidence that the ITT analyses were more conservative in their estimates of treatment difference. The need to demonstrate equivalence in both an ITT and a PP analysis in a regulatory trial increases the regulatory burden on drug developers. The relative importance of the two analyses will depend on the definitions used in particular therapeutic areas. Demonstrating equivalence in one population with strong support from the other would be preferred from the Industry viewpoint. In trials with regulatory importance, prior agreement with regulators on the role of ITT and PP populations should be sought. Trial designs will need to take account of the estimated size of the PP population if adequate power is needed for both analyses. Careful design in the series of asthma trials, particularly identifying a population of patients with potential to improve, resulted in notable increases in lung function during the course of the trials for both treatments. This provided reassurance that equivalence was not due to a lack of efficacy for both treatments. In one trial equivalence was demonstrated overall but a treatment by country interaction was noted. However, this interaction could not be attributed to differences in patient characteristics or baseline data between the countries. Study conduct was also similar in the different countries. The conclusion was that the interaction was spurious and that the trial provided good evidence of equivalence. © 1998 John Wiley & Sons, Ltd.

INTRODUCTION

Equivalence trials aim to show that two treatments have equivalent therapeutic effects. They are often performed to compare a new treatment with an existing standard treatment or to compare

* Correspondence to: A. F. Ebbutt, European Clinical Statistics, Glaxo Wellcome, Greenford Road, Greenford, Middlesex, UB6 0HE, U.K. E-mail: AFE4362@ggr.co.uk

CCC 0277–6715/98/151691–11$17.50

a new formulation of a treatment with the old formulation. In many areas it is ethically difficult to include a placebo group in the trial and hence the trial will compare two active treatments.

The problems arising in equivalence trials have been clearly laid out in a recent paper.[1] Determination of equivalence depends on defining in advance a range of equivalence $-d$ to $+d$ for the treatment difference such that any value in the range is clinically unimportant. When the trial is complete, a confidence interval for the difference between treatments is calculated, and if this lies entirely within the interval of equivalence then equivalence is claimed. The rationale is that the confidence interval defines a range for the true difference between the treatments, and if this range lies entirely within the pre-defined range of equivalence, then the treatment difference is unimportant. Sample size for the trial must then be determined using methods appropriate to the confidence interval approach and the relevant formula are described in the paper.

Special attention also needs to be paid to careful trial design to provide assurance that both trial treatments are active and not equivalent but inactive. This will include mirroring methodology from earlier trials comparing one of the active treatments with placebo in terms of: inclusion and exclusion criteria; the dosing schedule; the primary endpoint, and the use of concomitant therapy. Randomization and blinding are also fundamental. Some demonstration of efficacy from the trial itself can also provide reassurance, such as a change from baseline in a continuous variable or a level of success in a success/failure outcome that is similar to previous placebo controlled trials.

A number of issues with equivalence trials remain. The choice of the range of equivalence is still a difficult one. An interval much narrower than one which would be widely accepted as showing equivalence would lead to unnecessarily large trials. Too wide an interval will allow treatments which are substantially different to be regarded as equivalent. The size of the confidence interval to be used is an arbitrary one. There is the option to analyse trial results on an 'intention-to-treat' (ITT) or 'per-protocol' (PP) population. The CPMP Note for Guidance in Biostatistics[2] suggests that equivalence should be demonstrated on both populations leading to a greater regulatory burden. The increased confidence in the equivalence conclusion, which results from evidence of drug activity in the trial itself, requires focus during the design stage. Finally, the importance or otherwise of treatment by covariate interactions (particularly with centre or country) needs to be considered.

Glaxo Wellcome has provided effective drugs for asthma for many years and currently markets two inhaled beta-agonists, salbutamol and salmeterol, and two inhaled steroids, beclomethasone and fluticasone. A major initiative has recently been undertaken to replace CFC propellants in asthma inhalers with non-CFC propellants. In addition Glaxo Wellcome has been involved in developing new inhaler devices. As a consequence a series of equivalence studies has been carried out to essentially the same design. This series provides a unique opportunity to investigate the issues described above.

DESIGN OF ASTHMA TRIALS

All eleven trials from the series reported here used a common design. They were all double-blind randomized parallel group trials with half the patients randomized to the new inhaler and half to the old. Patients attended for a two week run-in period for baseline observation and were assessed for at least four weeks after randomization. Since asthma patients fluctuate from day to day depending on exposure to allergens and other stimuli, the primary endpoint was the peak expiratory flow rate (PEF) measured each morning by the patient at home.

In the inhaled steroid trials, patients were taking inhaled steroids at entry, and had no changes in medication in the preceding month. Patients also had reduced lung function compared to that predicted for their sex, age and height (forced expiratory volume in 1 second (FEV1) between 50 per cent and 80 per cent of predicted) and mean morning PEF was less than 90 per cent of the maximum obtained following inhaled salbutamol. Rescue bronchodilator on 4 occasions in the last 7 days of the run-in was also required.

For the beta-agonist trials FEV1 was between 50 per cent and 90 per cent of predicted, mean morning PEF was less than 85 per cent of the maximum obtained following inhaled salbutamol, and rescue bronchodilator was required on 4 occasions in the last 7 days of the run-in. Patients also had no changes in medication in the preceding month.

The range of equivalence for mean morning PEF defined for the trials was -15 to $+15$ l/min. The analysis involved calculation of a 90 per cent two-sided confidence interval for the treatment difference, based on an analysis of covariance approach using run-in PEF, sex, age and country as covariates. The expected standard deviation for PEF based on previous similar studies was 30–40 l/min. A power of at least 80 per cent was required, and, using the methods described in reference 1, a sample size of at least 250 patients was required.

In the trials the ITT population consisted of all patients randomized. There were two major criteria for exclusion from the PP population: about two-thirds of those excluded failed to meet entry criteria, and one-third consumed medication not permitted by the protocol. For withdrawn patients, data available until the day of withdrawal were included in the analysis and hence all patients contributed to the ITT population.

TRIAL RESULTS

In this paper attention is concentrated on the first 4 weeks of treatment. A few trials continued longer than 4 weeks. However, in this review attention was focused on the first 4 weeks to give a consistent time period for efficacy evaluation. Table I shows the results from the 11 equivalence trials. The trial size varied from 212 to 409 patients. In all the trials the 90 per cent confidence interval for the difference in PEF between the treatments was contained in the range -15 to $+15$ l/min.

Typical PEF responses over 4 weeks are shown in Figures 1 and 2 for trials A (beta-agonist) and F (inhaled steroid). There is little treatment difference throughout the 4 week treatment period.

CRITERIA FOR EQUIVALENCE

A number of key issues must be resolved when choosing a criteria for equivalence. Should the range of equivalence be symmetrical around zero? For the series of asthma trials this was considered to be appropriate because the new inhalers being tested would eventually be substituted for existing inhalers. The requirement was to match the two closely in terms of their effects on lung function. Differences in either direction outside an equivalence range symmetrical about zero would imply that drug delivery was different for the new and old inhalers and this would be unsatisfactory. In other situations it may be appropriate to choose an equivalence range which is not symmetric about zero. This would apply where, for example, assurance was needed that the new formulation was not worse than the old but there was less concern about the new formulation being better. The extreme case is a one-sided approach,[1] where attention is focussed

Table I. Results from eleven equivalence trials

Trial	Drug therapy	Number of patients		Estimated treatment difference (l/min)		CI (l/min)		CI width (l/min)	
		ITT	PP	ITT	PP	ITT	PP	ITT	PP
A	Beta-agonist	346	282	1	− 1	− 6 7	− 8 6	13	14
B	Beta-agonist	409	262	− 6	− 7	− 13 0	− 14 1	13	15
C	Beta-agonist	392	302	4	4	− 3 11	− 4 12	14	16
D	Inhaled steroid	399	288	− 1	− 2	− 7 4	− 9 4	11	13
E	Inhaled steroid	366	281	5	3	0 10	− 3 8	10	11
F	Inhaled steroid	421	290	− 2	− 2	− 7 2	− 8 3	9	11
G	Inhaled steroid	412	311	1	2	− 5 7	− 5 8	12	13
H	Inhaled steroid	370	325	− 2	− 1	− 8 3	− 7 5	11	12
I	Inhaled steroid	379	298	− 1	3	− 8 3	− 1 10	11	13
J	Inhaled steroid	346	234	− 2	− 2	− 7 2	− 8 3	9	11
K	Inhaled steroid	212	106	2	3	− 3 6	− 3 9	9	12
	Average	368	270	0·3	0	–	–	11·1	12·8

only on ruling out differences larger than an equivalence limit in one direction only. This can be useful when a new formulation has an improved side-effect profile and the trial tries to ensure that the efficacy is not worse than the standard.

The choice of the range of equivalence is also important. For respiratory trials there are three sources of relevant information. Earlier trials with inhaled beta-agonists and steroids provide information on differences from placebo. In three registration trials using a standard dose of the beta-agonist salmeterol, the average difference from placebo in PEF was 37 l/min. For inhaled steroids, the dose used depends on the severity of the patient's asthma and placebo controlled trials in more serious asthmatics are not easy to perform. Results indicate that over a 4 week period, the steroid effect is less than that of salmeterol and is of the order of 25 l/min. A second key feature is that the immediate effect of a large dose of a short acting beta-agonist gives an upper bound to the PEF-achievable for an individual asthma patient. Typically, a mean increase of up to 70 l/min can be seen following a short acting beta-agonist. Physician opinion of the size of difference which is clinically irrelevant constitutes the third key source of information. Discussion with practitioners suggests that differences less than 15 l/min are considered to be of minor importance – although as with all questions of this type there is variation amongst the experts. In

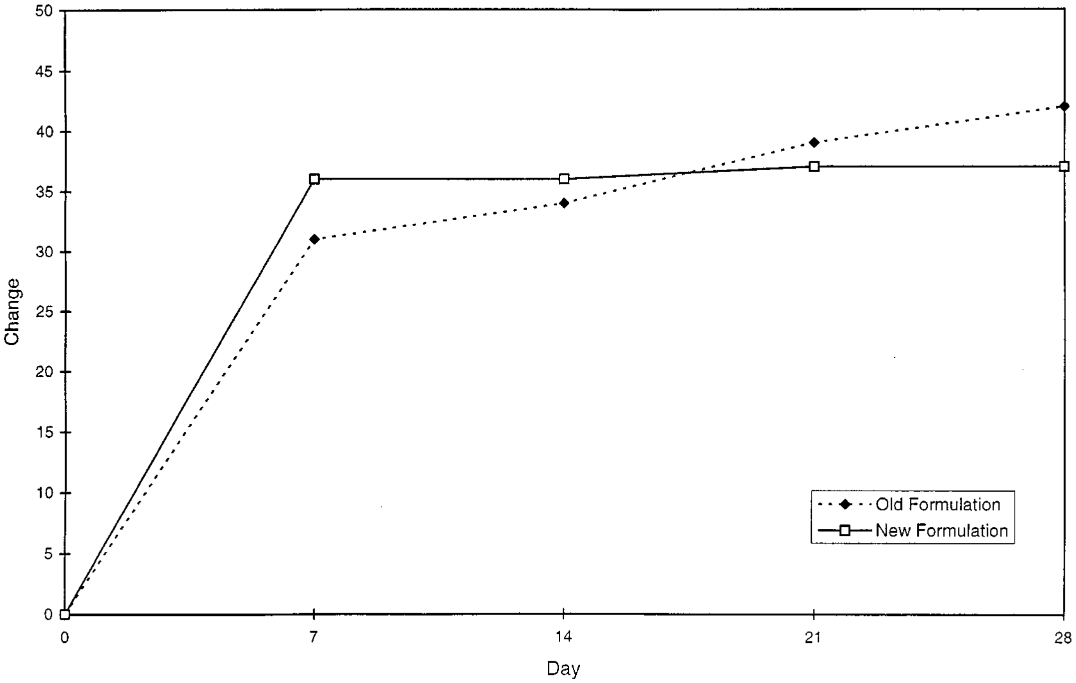

Figure 1. Trial A: beta-agonist change in PEF (l/min) from run-in, weekly mean

planning these asthma trials, the discussion with experts was informal. However, methodologies have been proposed[3] for obtaining probability distributions representing physicians' beliefs about treatment differences which could provide a more analytical approach to evaluating expert opinion.

Based on these considerations, Glaxo Wellcome chose the range of equivalence to be − 15 to + 15 l/min. This is about one-half of the difference expected between active and placebo and one-quarter of the maximum achievable result for a patient. It corresponds to physician opinion and ensures that when formulations are considered equivalent, the difference between them is unlikely to be greater than the difference between active and placebo. Since many equivalence trials may be targeted for use in regulatory submissions, the choice of the range of equivalence is a key regulatory issue. Although the range of − 15 to + 15 l/min appears acceptable in Europe for respiratory trials, there was no opportunity to discuss this in advance. In future, it would be advisable to ensure that the criteria for equivalence are agreed in advance by key regulatory authorities. The trial comparing streptokinase and reteplase[4] is an example where the critical nature of the trial endpoints and the size of the trial made essential widespread discussion of the equivalence criteria prior to starting the trial.

The analysis of equivalence trials is based on the calculation of a confidence interval. Guidelines for pharmacokinetic equivalence (bioequivalence)[5,6] have traditionally used 90 per cent confidence intervals, and mirroring this, the series of asthma trials reported here also used 90 per cent confidence intervals. The CPMP Note for Guidance on Biostatistics[2] suggests the use of 95 per cent confidence intervals. In regulatory situations, the choice is based on the level of risk

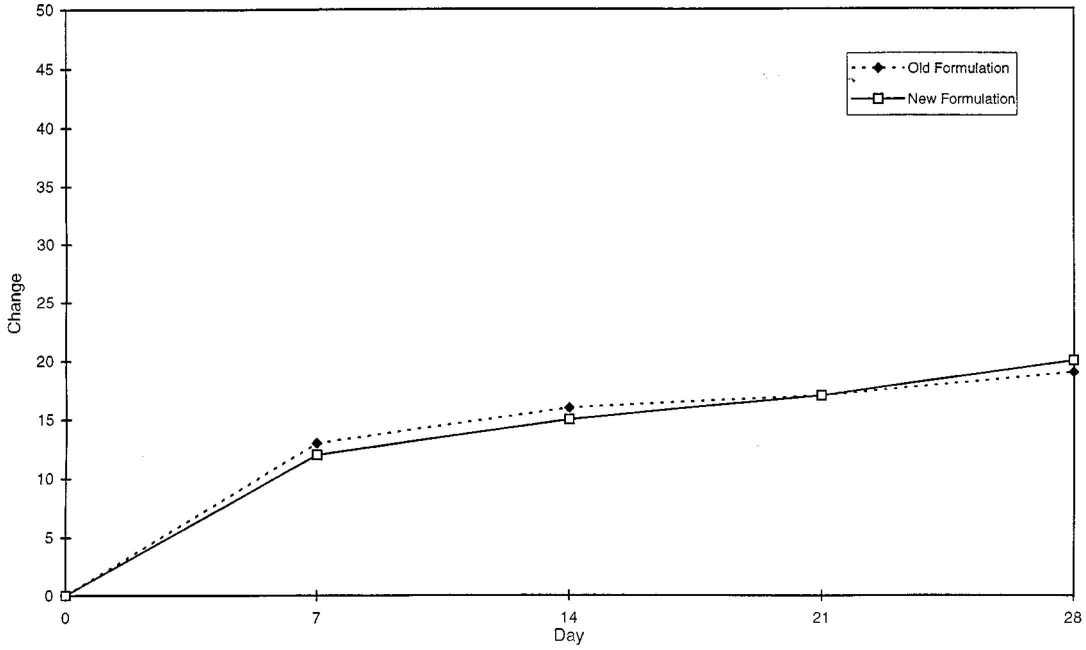

Figure 2. Trial B: inhaled steroid change in PEF (l/min) from run-in weekly mean

regulators are prepared to accept that treatments with a difference of potential importance will be accepted as equivalent. It is likely that this risk will depend on the clinical situation. In some situations where serious clinical events may occur, risk may need to be very carefully controlled and 95 per cent or 99 per cent confidence intervals may be appropriate. In others where minor symptoms are involved a lower level of risk may be appropriate and 90 per cent confidence intervals will be acceptable. As with the choice of a significance level for trials designed to show a difference, there is the need for flexibility rather than a fixed level imposed by a guideline.

A further problem with the interpretation of equivalence trials is that it is seen as a black or white situation. Looking only at the confidence interval implies that a rigid rule will be applied to the results. In an early paper recommending the confidence interval approach,[7] Schuirmann shows that it is equivalent to performing two one-sided tests. With a symmetric range of equivalence, the p-value from the two one-sided tests will summarize the weight of evidence in favour of equivalence and this will be useful additional information. For the ITT analysis of trial A, for example, it is easy to show that a test of the hypothesis that the treatment difference is less than -15 or greater than $+15$ l/min is highly significant ($p < 0{\cdot}01$). This would show that there was strong evidence that the formulations were equivalent.

INTENTION-TO-TREAT AND PER-PROTOCOL ANALYSES

Table I shows the sample sizes for the ITT and PP populations for the 11 equivalence trials, the estimated treatment difference, the 90 per cent confidence intervals and the total confidence interval width. The average sample sizes were 368 in the ITT population and 270 in the ITT population. A sample size of 350 patients in the ITT population provides greater than 90 per cent power to demonstrate equivalence based on a range of equivalence of -15 to $+15$ l/min and an expected standard deviation of 40 l/min. A sample size of 270 in the PP population provides more than 80 per cent power.

Over all the trials the average treatment difference is close to zero for both ITT and PP populations. There is no consistent pattern when estimates are compared for individual trials. These trials provide no evidence that there is a consistent bias in either direction when comparing treatment estimates for the ITT and PP populations.

The total width of the 90 per cent confidence intervals is also shown in Table I. For all trials the width of the confidence interval is greater for the PP population. Inspection of the residual standard deviations for both analyses shows that there is little difference between them. Hence the width of the confidence interval is dominated by the sample sizes in the ITT and PP populations. Clearly on this basis, sample size should be calculated based on the expected patient numbers in the PP population, as this will be the more difficult population in which to demonstrate equivalence.

EVIDENCE OF DRUG ACTIVITY DURING THE TRIAL

A major concern in equivalence trials is the possibility that the treatments may appear equivalent, but may actually be ineffective. Some evidence of clinical effectiveness can be obtained by careful attention to design. A key area for attention is the selection of patients for the trial. In this series of asthma trials, patients were only eligible for entry if they had the potential to improve during the trial. This was determined by: controlling their FEV1 level to be substantially below that predicted for their age, sex and height; ensuring that the morning PEF was less than 85 per cent of the best they could achieve immediately following inhaled beta-agonists; and that patients were still needing rescue medication on 4 of the last 7 days of the run-in-period.

This careful attention to the inclusion criteria allowed the trials to demonstrate improved lung function during the course of the trial. Figure 1 shows the daily PEF for trial A using a beta-agonist. An improvement over baseline of about 40 l/min is shown for both trial treatments and there is little difference between them. Figure 2 shows the results for trial D using an inhaled steroid. Here an improvement of about 20 l/min is shown and again there is little treatment difference.

Clearly, such changes during the trial do not provide conclusive evidence that the treatments are effective. The improvement could be a placebo effect due to the inclusion of patients in the trial process or it could be due to ‘regression to the mean’ since patients are selected with limited lung function at baseline. However, the size of the change, which is similar to the differences seen in previous trials between active and placebo, adds considerable weight to the equivalence claim.

TREATMENT INTERACTIONS

In equivalence trials, as with trials designed to show a treatment difference, there is interest in the consistency of treatment effects across the levels of key baseline covariates. FDA guidelines,[8] for example, ask for effects in different genders, different age groups and different racial groups to be characterized. In addition, consistency across centres or groups of centres is reassuring and a common feature of analyses of multi-centre trials.

Table II. Trial G: baseline data

		Overall		U.K.		Germany		France	
		New	Old	New	Old	New	Old	New	Old
Number of patients		205	214	92	97	57	56	56	61
Sex (%)	M	47	40	44	30	53	48	48	48
	F	53	60	56	70	47	52	52	52
Age (Yrs)	Mean	44	43	39	39	51	51	48	43
	Standard deviation	16	15	14	15	17	15	14	14
Current smoker (%)		17	14	18	14	22	18	7	10
Baseline PEF	Mean	370	391	389	400	349	360	362	406
(l/min)	Standard deviation	101	91	97	83	102	87	102	101

Table III. Trial G: treatment effects by country (over 12 weeks)

Country	Estimated treatment difference (l/min)	CI (l/min)	Significance *P*
U.K.	0	$(-8, 8)$	0·948
Germany	16	$(3, 29)$	0·040
France	-12	$(-25, 1)$	0·139
Overall	1	$(-5, 7)$	0·783

In trial G in the asthma series, treatment was extended to 12 weeks. Equivalence was demonstrated clearly in the overall analysis of the trial both during the first 4 weeks and over all 12 weeks. The trial included centres in the U.K., France and Germany as did many others in the series. A routine exploratory analysis showed a significant treatment by country interaction in the analyses of the 12 week trial period (ITT $p = 0{\cdot}020$; PP $p = 0{\cdot}052$). Baseline data by country and treatment are shown in Table II and treatment effects for the ITT population by country are shown in Table III.

Further analysis showed little in terms of possible explanations for the interaction. The presence of data outliers is a common cause of interactions but this was not the cause in trial G. There were no major differences between the countries in terms of basic demography and history. Overall, the baseline PEF was 21 l/min lower in the new formulation group compared to the old formulation group. This was exaggerated in the French patients. However, an analysis of treatment by baseline interaction showed no evidence that the treatment effect was different for different baseline lung function levels. Similarly, there was no evidence of interaction between treatment and any other baseline factor. No major differences in concomitant medication or medication changes during the trial were seen. The inhalers used in the different countries were from the same manufacturing batches. The recruitment pattern in the three countries was somewhat different. In the U.K., 189 patients were recruited from six centres with one centre contributing two-thirds of the patients, in Germany 12 investigators recruited 113 patients and in France 25 centres recruited 117 patients. The use of small centres in France may have led to a poorer estimate of the treatment difference.

In the series of asthma trials carried out, it is perhaps no surprise that one significant interaction was observed. A spurious interaction has a high probability of occurring on a long series of trials. The interaction observed was reviewed in some detail. If the treatment difference were to depend on country, there could be several explanations. The characteristics of the patients recruited could differ – however there was little evidence of this when demography and baseline characteristics were reviewed and in many previous trials patients in the U.K., France, Germany and other European countries have responded similarly to treatment. In addition there was no evidence of an interaction between baseline factors and treatment, making it unlikely that country was a simply a marker for baseline differences between the populations. Study conduct could differ but in this trial there were similar numbers of and similar types of protocol violations in the different countries and few changes of medication during the trial and similar results were seen for both ITT and PP analyses.

The importance of the observed interaction is difficult to assess, but the general impression is that the treatment by country interaction is spurious and that the claim of overall equivalence is still reasonable.

DISCUSSION

The review of recent Glaxo Wellcome equivalence trials in asthma has thrown light on a number of issues. For asthma trials, there is good background data that helps to focus the choice of the range of equivalence. Physician opinion of the level of difference which is important clinically is still an important feature. However, knowledge of the magnitude of differences from placebo (which may influence physicians' opinions anyway) is particularly useful, as is knowledge of the maximum achievable effect. For most pharmaceutical products knowledge of differences between active and placebo should be available from earlier trials and this will be relevant provided the main design characteristics of the equivalence trials replicate the earlier placebo controlled trials. A useful practical concept is that the range of equivalence should exclude values which would allow the new formulation to be closer to placebo than to the old formulation.

The choice of the size of confidence interval used needs to be flexible depending on the situation. Conventional choices such as 90, 95 or 99 per cent can be employed depending on the clinical situation rather than operating to a fixed rule. Calculation of p-values looking at the analysis from a two one-sided tests approach would provide a guide to the weight of evidence in favour of equivalence.

The principle behind ITT analyses is that all randomized patients are analysed irrespective of adherence to the protocol or completion of the trial. The advantage of this is that it retains the original randomization; it avoids bias which could be introduced if protocol violation or non-completion is related to efficacy and it mirrors what will happen if a treatment is used in practice. For equivalence trials, however, there is concern that an ITT analyses will move the estimated treatment difference towards zero since it will include patients who should not have been in the trial who will get no benefit, or patients who did not get the true treatment benefit because of protocol violation or failure to complete. PP analyses include only those who follow the protocol adequately. This would be expected to detect a clearer effect of treatment since uninformative 'noise' would be removed. This suggests it would be a more suitable population for the evaluation of equivalence, except for the fact that the PP analysis my be biased as it does not contain all the patients originally randomized.

In practice the precise definitions used for ITT and PP populations vary considerably between therapeutic areas. In the asthma series reported here the ITT analyses involved all patients randomized and all data recorded for them. The PP analyses omitted those who failed entry criteria and those who took medication not permitted by the protocol. With these definitions the PP analyses reflected the ITT analyses closely but with wider CIs due solely to the difference in the number of patients included. Essentially the PP analyses are redundant (although documentation of the numbers and reasons for patients violating the protocol remains relevant) and it is reasonable to base decisions on the ITT analyses provided the PP analyses are supportive. In the trial comparing reteplase and streptokinase[4] the primary endpoint was mortality at 35 days. The ITT analyses included all patients randomized (whether or not treatment was started) and the PP analyses included those randomized who started treatment. The trial results quote mortality in both populations to ensure that equivalence is not a consequence of untreated patients who would respond similarly in both treatment groups. It seems clear that the relative importance of the ITT and PP analyses will depend on the disease, the definitions adopted for the populations and the likelihood of bias. For trials with regulatory importance, this needs to be discussed in advance with regulatory authorities.

The data presented here show how careful choice of the inclusion and exclusion criteria can define a trial population which will demonstrate change during the trial. This provides additional assurance that the drugs are showing positive effects during the trial. This is not dissimilar to the choice of criteria that would be made to compare active with placebo. A population unable to change would not be sensible for trials of this kind either and this re-emphasizes the importance of designing equivalence trials to be as similar as possible to earlier placebo controlled trials. When variables such as PEF can be used to assess drug activity, it is easier to show positive effects in the trial. Where endpoints are ordered categorical and can be assessed at baseline and at the end of treatment, similar arguments apply. For binary endpoints such as healed/unhealed or survival endpoints, internal evidence of improvement cannot be obtained. In this case, reference to similar success rates or survival rates in previous placebo controlled trials will help, provided once again that key design features are similar. Evidence of drug activity in secondary endpoints which have discriminated between active and placebo in earlier trials will also help.

The relevance of treatment interactions in equivalence trials and particularly treatment by centre or country interactions is an interesting one. Evidence of an interaction with a major baseline characteristic such as sex, age or baseline severity would, perhaps, be a greater cause for concern. Where both baseline characteristics and study conduct seem similar in the different countries or centres, then there is a higher chance that the interaction is spurious and that equivalence can reasonably claimed from the trial.

REFERENCES

1. Jones, B., Jarvis, P., Lewis, J. A. and Ebbutt, A. F. 'Trials to assess equivalence: the importance of rigorous methods', *British Medical Journal*, **313**, 36–39 (1996).
2. CPMP Working Party. 'Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products', CPMP working party on efficacy of medicinal products Note for Guidance III/3630/92-EN, *Statistics in Medicine*, **14**, 1658–1682 (1995).
3. Freedman, L. S. and Spiegelhalter, D. J. 'The assessment of subjective opinion and its use in relation to stopping rules for clinical trials', *Statistician*, **33**, 153–160 (1983).
4. International Joint Efficacy Comparison of Thrombolytics. 'Randomised, double blind comparison of reteplase double-bolus administration with streptokinase in acute myocardial infarction (INJECT): trial to investigate equivalence', *Lancet*, **346**, 329–336 (1995).

5. 'Guidance: statistical procedures for bioequivalence studies using a standard two-treatment crossover design', Office of Generic Drugs, Food and Drug Administration, Public Health Service, US Department of Health and Human Services, July, 1992.
6. 'CPMP guideline: investigation of bioavailability and bioequivalence', Commission of the European Communities 111/54/89-EN, 1992.
7. Schuirmann, D. J. 'A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability', *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680 (1987).
8. 'FDA Proposed Rule to require sponsors to submit effectiveness and safety data by gender, age and racial subgroup', Federal Register 8/9/95, **60**, No 174, 46794-7.