ORIGINAL ARTICLE

The pros and cons of noninferiority trials

Stuart J. Pocock

Medical Statistics Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT UK

Keywords
control groups,
equivalence,
ethics,
noninferiority,
randomized clinical trials,
trial size

Received 12 January 2001;
revised 1 March 2001;
accepted 13 January 2003

Correspondence and reprints:
stuart.pocock@lshtm.ac.uk

ABSTRACT

Noninferiority trials comparing new treatment with an active standard control are becoming increasingly common. This article discusses relevant issues regarding their need, design, analysis and interpretation: the appropriate choice of control group, types of noninferiority trial, ethical considerations, sample size determination and potential pitfalls to consider.

INTRODUCTION

The term 'noninferiority trial' is commonly used to refer to a randomized clinical trial in which a new test treatment is compared with a standard active treatment rather than a placebo or untreated control group. A prior judgement is made, that for the new treatment to be of merit it only needs to be as good as the active control regarding appropriate outcome measure(s) of response. While the superiority of the new treatment over active control would be an added (perhaps unrealistic) advantage, the clear demonstration of noninferiority in one or more specific criteria of patient response is the desirable goal which motivates such a trial. The term 'equivalence trial' is sometimes used in this context but does not reflect so well the (usually) one-sided nature of this noninferiority question, and is implicitly dismissive of the desirable option that the new treatment could actually be superior to the active control treatment. So the more appropriate term 'noninferiority' is used hereon.

The aim of this article is to present a balanced view of the role of noninferiority trials in the development of safe and effective new treatments. This involves a mix of ethical, scientific, statistical and practical considerations. This article elucidates some of the pros and cons of noninferiority trials and offers some pointers on how to enhance their public health value. The desirability (or not) of a noninferiority trial strategy will depend on the particular circumstances. While general guidance can be given, the relative merits of noninferiority active control trials or placebo-controlled trials aimed at demonstrating superiority for evaluating any specific new treatment rests on a complex of issues requiring wise judgements and continued open debate.

CHOICE OF CONTROL GROUP: ACTIVE VS. PLACEBO

One starting principle is that no patient is denied a known effective treatment by entering a clinical trial. An equally important principle is that the degree of scientific rigour adopted in the evaluation of a new treatment is sufficient to prevent any ineffective, unsafe or inferior treatments obtaining regulatory approval or gaining widespread use. Both principles highlight the ethical responsibility of a society and its medical researchers to facilitate the best possible health care at present and also in the future.

The first principle is more easily grasped because it relates immediately to the individual rights of the next patient. Of importance here is the distinction between a treatment known to be effective, and one thought to be effective, hoped to be more effective, believed to be effective or in widespread use without evidence of

effectiveness. Arguments against the use of placebo controls are put forward because treatment practice involves other active treatments. The question is: how convincing is the evidence that such active treatments are better than placebo in aspects that genuinely benefit patient welfare? While one needs to consider the understandable wish to do something positive for every patient, one needs to draw a clear distinction between desire for benefit in a supposedly active potential control treatment and hard evidence of benefit from previous clinical trials.

So, what is the extent of evidence? Is it 'proof beyond reasonable doubt' of patient benefit derived from several large studies generalizable to the relevant patient population or is it just one or two statistically significant results, perhaps on short-term studies of limited size studying surrogate end points rather than overall patient benefit? $P < 0.05$ for a treatment difference in a clinical trial, does not equate with proof of effect.

Even the relevance of $P < 0.0001$ for a treatment difference can be questioned on several grounds: might the trial have been biased in some aspect of its design or analysis; were the patients, the delivery of treatment, the outcome measure used and the length of follow-up sufficiently relevant to normal clinical practice and patient benefit; was the absolute magnitude of benefit sufficiently large taking account of any adverse side-effects of a treatment; how many trials were performed and on how many patients. That is, in sizing up the evidence that a pre-existing active treatment is superior to placebo one needs to exercise one's constructive critical faculties when appraising its clinical trial evidence for internal validity, external validity, overall patient benefit and extent of research.

Should the overall evidence for patient benefit on a pre-existing active treatment be less than totally convincing, then the dangers of exclusively adopting that treatment as an active control group (instead of a placebo control group) for the evaluation of other new treatments are substantial. In certain areas, this problem is tackled by having both an active control group and a placebo control group, so that noninferiority compared with the former and superiority over the latter can be evaluated within the same trial, or set of trials.

It is important to recognize that approval of a drug by regulatory authorities as being safe and effective for a specific condition does not in itself imply that the use of that drug as an active control without a placebo group would provide a reliable basis for a noninferiority trial of a new drug, (see Temple and Ellenberg 2000).

This ongoing dilemma for clinical trials research can be summarized by the wish to avoid two types of error: a type I error would be the acceptance of a useless treatment into widespread use, and one needs to consider the increased risk of this error occurring by not using placebo controls and instead pursuing a noninferiority (equivalence) trial design with an active control group. The consequences of such an error will depend on the nature of the treatment and disease. If the ineffective treatment has substantial side-effects then great harm could ensue, if it is expensive then it detracts from more fruitful use of health care costs, if it is a safe, useless, cheap 'placebo' for a minor condition then perhaps there is not much harm.

Even if there is an effective active control treatment, there can still be problems in the design, conduct, analysis and interpretation of a noninferiority trial that could lead to such a type I error. Such problems are outlined in the rest of this article.

A type II error is the failure to use an effective active control treatment by adopting a placebo control group instead. As expressed above, the degree of certainty with which such an error occurs depends on the extent of prior knowledge that the active control is truly effective. In addition, the severity of this error will depend on particular circumstances. At one extreme it would be absolutely intolerable to deny a known effective agent that reduces mortality in a rapidly progressing life-threatening condition. However, in a more minor ailment in which recovery often happens on placebo or no treatment, the denial of a known active agent in a short-term placebo-controlled trial, after which all patients can go on to receive active treatment, has much less serious consequences for patient welfare.

In many trials, having a placebo control group does not that mean such patients receive no active intervention. Often, all randomized patients undergo normal accepted care, including other active drugs as appropriate, but the addition of a new treatment is compared with addition of a placebo. It is often debated whether such ancillary care and supplementary drugs should be according to a fixed protocol or pragmatically left to individual clinical judgement as in routine clinical practice. It is also relevant to ask, might the patient have got the active control treatment if they were not included in a clinical trial. In some circumstances the answer is 'no they would not', in which case there appears a perverse twist in the ethical argument whereby trialists are required to adopt more stringent ethical standards than regular treating physicians.

Thus, every time one chooses between active controls or placebo controls in planning a randomized trial, one has to consider the risks of type I and type II error, taking account of the likelihood of them occurring and the severity of the consequences. The consequent decisions are not easily taken: neither passionate one-sided ethical arguments against placebo controls in general nor scientific pleas for mandatory up front demonstration of new treatment superiority over placebo should dominate this thinking and planning. Rather one aims for an ethical balance of the genuine needs of the next patient in a trial to receive good care and the longer term public health need to only allow marketing approval and widespread use of treatments that actually work.

TYPES OF NONINFERIORITY TRIAL

There are many different circumstances that may lead to undertaking a noninferiority trial design. The simplest case is where one wishes to demonstrate (if true) that the efficacy of a new drug is the same as an existing active drug, and one is not anticipating any other differences. This will be particularly plausible if the drugs are of the same class, in which case such a 'me too' drug development could lead to an additional marketable product but not a substantial improvement in therapeutic care. The merits of such a narrow diversity of products within a specific drug class may appear rather small, except as regards company profits. However, even if the average benefits and safety profiles of two drugs in a class appear identical, it is possible that individual patients may benefit more from one drug than the other.

Either before or after marketing approval, there is sometimes a need for a large randomized controlled safety study to evaluate a concern that may have arisen from observational adverse event reporting. For instance, the European Post-Operative NSAID Study Group evaluated in 11 302 patients undergoing surgery the safety of one nonsteroidal anti-inflammatory drug (NSAID) pain relief drug ketorolac compared with two others, diclofenac and ketoprofen. This noninferiority safety study was motivated by concerns in the European regulatory authority, the Committee on Proprietary Medicinal Products.

A more complex and interesting scenario arises when the aim is to demonstrate equivalence (noninferiority) of a new treatment to an active control, while knowing or suspecting that the two treatments will differ in some other important respects. Possibilities here are:

(1) *The new treatment has less side-effects*. For instance, low dose aspirin vs. anticoagulation following thrombolysis after a myocardial infarction may be equivalent as regards recurrence of infarct or cardiac death, but the former produces less bleeding complications. Aspirin is hardly a new drug! However, in the context where anticoagulation had become the norm it was a new alternative strategy.

(2) *The new treatment is less invasive*. Carotid endarterectomy is a surgical procedure for patients at high risk of a stroke. If carotid stenting could be demonstrated as equally efficacious, for many patients it might be the treatment of choice, being less invasive. This is the motivation behind a proposed National Institute of Health-funded trial comparing these two intervention strategies.

(3) *The new treatment is cheaper*. An analogous situation concerns the relative merits of bypass surgery and coronary angioplasty for patients with angina. Trials have shown that the prognosis death and/or myocardial infarction appears similar for both intervention strategies, the former provides better symptomatic relief initially but is more invasive. However, any health care strategy must take costs and cost-effectiveness into account, and much interest in these trials has focussed on the reduced initial costs of an angioplasty and whether that gain is maintained over several years follow-up.

The complexity of these situations arises from the fact that one is looking at a trade-off between efficacy and other issues regarding side-effects, patient acceptability and costs. Although such trials are often presented as noninferiority trials, it is possible that some modest reduction in efficacy may be acceptable alongside the other benefits of a new treatment. One such example of this may be the acellular pertussis vaccines, which are used in preference to whole cell pertussis vaccines because of fewer adverse events, although their efficacy may be somewhat lower.

Another issue is whether any particular trial comparing a new treatment with an active standard treatment should be formulated as a noninferiority trial or not. Given the relative state of ignorance with which one starts any new study, one is often unsure whether to optimistically pursue the prospect of demonstrating a new treatment's superiority or whether to settle for demonstrating noninferiority on the basis that, that it will be still good enough to make the new treatment of some value. Although the statistical power calculations differ somewhat for these two scenarios (i.e. the latter

reverses the roles of null and alternative hypotheses) – the underlying statistical and scientific intent is unaltered. What one wants is a sufficiently large and unbiased study that the true magnitude of treatment difference is estimated precisely. That is, when the trial is completed the point estimate and confidence interval for the appropriate measure(s) of treatment difference contain all the relevant evidence on which to hang claims of noninferiority or superiority.

Rigorous adherence to a single prespecified criterion of noninferiority, except for the convenience of planning the size of a trial, may not necessarily be the most sensible way of interpreting a trial's results. Nevertheless, formulating one's realistic goals, and hence the required number of patients, is an important feature of any noninferiority trial's planning and the next section deals with this topic.

APPROPRIATE GOALS AND SAMPLE SIZES FOR NONINFERIORITY TRIALS

First it is essential to realize that failure to demonstrate a statistically significant difference between two treatments does not allow one to assert that the two treatments are equivalent, or even similar, in their efficacy. Obviously, the fewer patients there are in a trial the less power to detect any meaningful difference so that nonsignificance in a conventional test of a null hypothesis is a hopeless criterion for inferring noninferiority. It would actually encourage the pursuit of smaller trials!

Instead, the most widely accepted approach to determine the required size of a noninferiority trial is to first define the smallest true magnitude of inferiority that would be regarded as unacceptable, assuming that one has already chosen a single primary outcome measure of response for this purpose. Anything truly bad or even worse needs to be detected reliably so that any claim of noninferiority for the new treatment can then be ruled out. However, one is prepared to accept more minor differences from true equivalence as being 'good enough'.

The logical basis here is that even if one carries out an extremely large clinical trial, one never fully proves that two treatments are truly identical in their efficacy. The confidence interval for the treatment difference gets smaller and smaller as the sample size increases, but proof of equivalence would require a confidence interval centred on zero and with zero width. An impossible task!

In a spirit of achievable compromise one sets out to arbitrarily choose this minimum clinically relevant difference, commonly called delta, which if true would deny any claim of a new treatment's noninferiority. We then choose a sample size sufficiently large such that if there is true equivalence of new and control treatments there is a high probability that the confidence interval for treatment difference will be wholly to one side (the good side!) of this 'delta'.

The simplest case to quantify is for a binary response (success or failure). Let π be the anticipated percentage of success on each treatment if true equivalence exists. Let δ be the 'minimum clinically relevant difference'. Suppose that results will be expressed as an estimated percentage of treatment difference with a 95% confidence interval around it. Also, suppose one wants to be 90% sure that if treatments are truly identical then the confidence interval will exclude δ, in a more favourable direction of course.

A simple commonly used formula in this instance is that the required sample size is

$$2n = \frac{4 \times 10.5\pi(100 - \pi)}{\delta^2}$$

More complex refinements exist and alternative but similar formulae exist for other types of outcome data, such as comparison of risk ratios or means of a quantitative measure, but this formula will adequately illustrate the problems of choosing the size of a noninferiority trial.

The difficulty lies in choosing appropriate values for π and δ, especially the latter. For example, consider a noninferiority trial comparing a new drug with omeprazole for treatment of *Helicobacter pylori* infection. The binary response is eradication of infection (yes or no). From past experience with omeprazole, $\pi = 85\%$ was the anticipated eradication rate. For trial planning δ was set at 15%. This means that the new drug would be regarded as noninferior provided that the possibility of its eradication rate being 15% worse than omeprazole could be ruled out (in the sense that the 95% confidence interval for the treatment difference in eradication rates would not include a 15% inferiority relative to omeprazole).

Hence the trial required $2n = (4 \times 10.5 \times 85 \times 15)/15^2 = 238$ randomized patients.

Leaping ahead to the actual results of this trial, the observed eradication rates on new drug and omeprazole were 109 of 126 (86.5%) and 110 of 129 (85.3%), respectively.

The 95% confidence interval for the treatment difference was −7.5% to +9.8%. A difference of −15% is clearly ruled out of consideration, and on that basis the

trial data support the new drug's noninferiority relative to omeprazole. Of course, one still cannot claim with certainty that the new drug is identical in efficacy to omeprazole. After all, the confidence interval for treatment difference does go beyond 5% in both favourable and unfavourable direction. But according to the predefined goal, adequate evidence of noninferiority is deemed to have been achieved.

Now just suppose that the new drug had had an eradication rate of 99 of 126 (78.6%) instead of the above 109 of 126. In that case, the 95% confidence interval for the treatment difference would have been from −16.1% to +2.7%. This would have been a most unhelpful result as one could neither claim noninferiority (as the confidence interval would include −15%) nor could one rule out equivalence of the two treatments (because the confidence interval includes no difference). Fortunately, this did not happen in reality, but it illustrates the inconclusiveness that can easily arise if noninferiority trials are conducted with fairly modest sample sizes.

It seems quite fashionable to choose $\delta = 15\%$ in noninferiority trials. Indeed the regulatory authority did approve such a choice for the above trial, and there are more general regulatory guidelines for choice of delta in such anti-infective trials. But it is hard to come up with an objective reasoning behind this apparently arbitrary often used choice. Why not $\delta = 10\%$ instead? That would require $2n = 535$ patients, more than twice as many. $\delta = 5\%$ might seem a plausibly tight safety margin, on the basis that only the slightest possible inferiority of the new drug should be allowable, but that requires nine times as many patients, a staggering $2n = 2142$ patients.

These calculations are all based on 95% confidence and being 90% sure or ruling out noninferiority. More generally if one requires $100(1 - \alpha)\%$ confidence and wants $100(1 - \beta)\%$ surity, then one requires $4 \times \left(Z_{\alpha/2} + Z_{\beta}\right)^2 \times \pi(100 - \pi)/\delta^2$ patients, where $Z_{\alpha/2}$ and Z_{β} are standardized normal deviates associated with one-tail probabilities $\alpha/2$ and β, respectively.

For instance, with a 90% confidence interval and only 80% surity, each of the above sample sizes is reduced by 40%. However, with the tougher demands set by a 99% confidence interval and being 95% sure of rejecting a difference δ and claiming noninferiority when equivalence truly exists, each of the above sample sizes increases by 70%.

For those more used to power calculations for trials aimed at detecting differences, it is worth noting that the above formulae have reversed the usual concepts of null and alternative hypothesis, and type I and II errors α, β. For a noninferiority trial, $\alpha/2$ is the probability that the $100(1 - \alpha)\%$ confidence interval excludes δ when the null hypothesis, treatment difference $= \delta$, is in fact true. β is the probability that the confidence interval includes δ (or worse) when the alternative hypothesis of no treatment difference is in fact true.

In practice, there seems to be a pragmatic acceptance by trialists and regulatory authorities that fairly generous choices of δ, α and β are allowed in order not to demand inordinately large numbers of patients in noninferiority trials. Should we be concerned therefore that the adoption of noninferiority designs with generously large choices of δ be permitting treatments with more modest but important extents of inferiority to be falsely accepted as noninferior? This is an inherent weakness of noninferiority trials as currently performed. We do take a sizeable risk that some truly inferior treatments will slip through the net.

Perhaps one could draw a distinction between (1) trials comparing two drugs in the same class where there may exist a high prior belief that treatments truly should be equally efficacious and (2) trials comparing quite contrasting treatments, e.g. drugs of differing types, or radically differing intervention strategies where there is no firm grounds on which to anticipate noninferiority. A generous δ leading to a smaller required sample size seems more permissible in the first instance (as was the case from the above *H. pylori* example). The more contrasting the treatments the harder it often is to recruit patients, but that is just the instance when large sample sizes are needed in order to be confident of true noninferiority.

The appropriate choice of δ is particularly important when a noninferiority trial vs. active control is taking place because it is considered unethical to proceed with a placebo control group. The worst that could happen is that the noninferiority margin is set so wide that a new treatment not much (if at all) better than placebo gets accepted as 'noninferior' on such an unduly lose criterion. Hence, it is useful to infer, preferably from past placebo-controlled trials, the magnitude of superiority of the active control over placebo. The choice of δ both for planning trial size and interpreting results of the new trial needs to be substantially smaller than this estimate for several reasons:

(1) The magnitude of superiority of active control over placebo may well be an overestimate. The placebo-controlled evidence may be limited, past trials may have

some biases present and to carry forward one (perhaps lucky) possibly exaggerated effect of active treatments into future planning, reflects a lack of scientific caution.
(2) Without a direct comparison with a placebo-control group, one is using an indirect argument via a comparison with active control to infer that a new treatment is worthwhile. For a whole variety of reasons discussed in the next section (e.g. patient selection, noncompliance), the circumstances of the noninferiority trial may be sufficiently different from the placebo-controlled trials to cast doubt on the appropriateness of the active treatment's apparent magnitude of superiority over placebo. A safety margin of less ambitious efficacy may be in order.
(3) Any new treatment needs to have a certain minimum magnitude of efficacy compared with placebo in order to be of worth, especially if other considerations (side-effects, costs, inconvenience) come into play. Thus, even supposing issues (1) and (2) above did not apply (they usually do though!), one would still want δ to be much smaller than the active control's superiority margin over placebo.

These issues are linked to the earlier arguments concerning the choice between placebo and active controls. The ideal circumstance for a noninferiority trial is when the superiority of active control compared with placebo is irrefutable and well-documented, the noninferiority trial can be conducted in very similar conditions and the choice of delta is small enough to convince one that any new treatment passing such a noninferiority test is truly of therapeutic value.

Fundamentally, many noninferiority trials are not be large enough to satisfy these requirements, meaning that the risk of a type I error (as discussed in section 2 above), false acceptance of a useless treatment, is often greater than it should be.

THE POTENTIAL INFERIORITY OF NONINFERIORITY TRIALS

For conventional clinical trials aimed at exploring the potential superiority of a new treatment over standard treatment (whether placebo or active), many of the pitfalls that can arise operate in the direction of making it harder to detect a genuine treatment difference. Such conservatism leads to the observed treatment difference being a dilution of the true effects under ideal conditions. This is often seen as an appropriate pragmatism, whereby the attempted unbiased comparison of new and standard treatment policies in a practical setting where strict adherence to protocol inevitably does not always happen, means that the real-life benefit of a new treatment is seen for what it really is. That is, a new treatment has to fight its way through the hiccups, failings, frailties and unpredictability of human beings, (both trialists and patients), in order to demonstrate its superiority over standard care.

The great difficulty with noninferiority trials is that their very motivation is to demonstrate the similarity of new and standard treatments, so that all these same problems work towards achieving this goal even if it is not true. The anti-conservatism of a poorly designed and poorly conducted noninferiority trial can greatly enhance the risk of a type I error, the adoption of a useless treatment whose inadequacies could not be detected.

One could argue that the unscrupulous investigator has every intention to undertake a sloppy noninferiority trial. For instance, with selection of inappropriate patients, poor compliance with intended treatments, use of nondiscriminatory outcome measures, inconsistencies between observers, too short a follow-up and a substantial amount of missing data, it would not be surprising if the results showed closely comparable results even if the real treatments properly given to the right patients were substantially different in real patient benefit.

So in noninferiority trials it is especially important to adhere to a well-defined relevant study protocol, and also to document that such adherence is successfully achieved. Some of the principal difficulties to bear in mind are as follows:
(1) *Selection of patients.* It is important to select the type of patient for whom the efficacy of the active control treatment has been clearly established. For instance, were one to deviate, even in part, from the patient population in whom superiority over placebo had previously been demonstrated, then any claim regarding a new treatment's merits could not well distinguish between genuine noninferiority or inappropriate selection of patients. Informative generalizability depends on a representative patient sample of the same kind as had previously demonstrated efficacy for the active control.
(2) *Treatment compliance.* The first requirement is that one chooses a genuinely efficacious active control treatment, and that it be given in the same form, dose and quality as was previously used to demonstrate that efficacy. One then requires that for both new and active treatment groups a satisfactorily high level of patient compliance is achieved, and that appropriate measures

of such compliance are recorded. Any reasons for alteration or discontinuation of treatments need documenting. Also, use of concommitant nonrandomized treatments needs documenting (and possibly standardizing) as any differential use of other efficacious treatments could conceivably mask the inferiority of a new treatment.

(3) *Outcome measures.* One needs to choose outcome measures that reflect genuine patient benefit (i.e. surrogate markers may well not suffice), and which were previously used to demonstrate the efficacy of the active control treatment. Each such measure (or end point) needs consistent well-defined criteria, with appropriate steps to reduce observer variation or bias. In addition, rigorous, objective reporting of adverse events is an important issue, as any noninferiority needs to concern safety and efficacy.

(4) *Duration of treatment and evaluations.* In any noninferiority trial, the randomized treatments need to be given for long enough and the patient response evaluated over a long enough period so that any potential treatment differences have a realistic opportunity to reveal themselves. Due attention needs to be given to the durations of treatment and follow-up in previous trials demonstrating efficacy of the active control treatment, and also the intended duration of treatment in future clinical practice.

(5) *Statistical analysis issues.* Any noninferiority trial requires a well-documented statistical analysis plan. There will often be a single primary outcome measure with a predefined noninferiority criterion and method of analysis, but this should not preclude appropriate secondary analyses of other outcome measures, which could become important if they exhibit any signs of the new treatment's inferiority or if interpretation of the primary outcome findings is not clear-cut.

In major phase III trials aimed at detecting treatment differences, analysis by intention to treat is routinely highlighted. That is, one analyses the complete follow-up results for all randomized patients regardless of their compliance with intended treatment in a spirit of comparing the treatment policies as actually given. Although this may dilute any idealized treatment differences under (unrealistic) circumstances of 100% compliance, that is generally considered wise pragmatism compared with any potential exaggerations of efficacy that could arise from focussing on treatment compliers only. The dilemma for noninferiority trials is that faced with non-negligible noncompliance analysis by intention to treat could artificially enhance the claim of noninferiority by diluting some real treatment difference. More weight may instead be attached to per protocol analyses (which focus on patient outcome amongst compliers only or up until compliance ceases in each patient) in the hope that they may reveal undesirable treatment differences. But this in turn has problems as compliers are a select group of patients who may give a favourably biased view (e.g. if the treatment is not helping, you drop out). These difficulties become particularly problematic if compliance differs between treatment groups.

Thus, for noninferiority trials there is no single ideal analysis strategy in the face of substantial noncompliance or missing data, and both analysis by intention to treat and well-defined per protocol analyses would seem warranted. Such noncompliance will inevitably lead to a degree of concern over any affirmative claims of noninferiority, which feeds back to the need to minimize any such protocol deviations.

In general, statistical considerations in the design, monitoring and analysis of noninferiority are less well established than for superiority trials, making it a fruitful area for further methodological research.

CONCLUDING REMARKS

The most important advantage of a noninferiority trial is that faced with clear evidence of efficacy for an existing standard treatment, it would be ethically unacceptable to proceed with a placebo or inactive control group in the evaluation of a new treatment for the same condition. In any particular circumstance an important reservation before jumping to that conclusion is that such evidence of efficacy really is strong enough to warrant exclusion of placebo controls. Too lax an acceptance of noninferiority trials, with a less than convincing active control treatment, could potentially lead to the adoption of more and more ineffective treatments, and this would be a misguided over-reaction to the ethical concerns in conducting randomized-controlled trials.

So when one has made the right judgement to undertake a noninferiority trial with an active control treatment, all necessary steps need to be taken to ensure that any failings in the trial design, conduct or analysis could not artificially dilute out any real treatment differences. That is, false claims of noninferiority need to be avoided.

Inevitably one can never prove that two treatments are identical, and hence some degree of compromise

is required so that realistically achievable but adequately large numbers of patients are randomized in a noninferiority trial. Thus, any clinically important treatment difference can be demonstrated not to exist with reasonable confidence. One suspects that in too many instances sample size determination for noninferiority trials is based on too generous a criterion of what constitutes a minimum clinically important treatment difference, and this increases the risk that some inferior treatments may gain regulatory approval and widespread use.

So placebo controls may rightly need to be ruled out in certain areas of clinical research, but it would be wrong to rush too enthusiastically into more widespread use of noninferiority (equivalence) trials without full consideration of their inherent problems.

BIBLIOGRAPHY FOR FURTHER READING

1 Blackwelder W.C. ''Proving the null hypothesis'' in clinical trials. Con. Clin. Trials (1982) **3** 345–353.
2 Farrington C.P., Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat. Med. (1990) **9** 1447–1454.
3 Garbe E., Röhmel J, Gundert-Remy U. Clinical and statistical issues in therapeutic equivalence trials. Eur. J. Clin. Pharmacol. (1993) **45** 1–7.
4 International Conference on Harmonisation. Choice of Control Group in Clinical Trials. Federal Register (1999) **64** 51767–51777.
5 Jones B., Jarvis P., Lewis J.A., Ebbutt A.F. Trials to assess equivalence: the importance of rigorous methods. BMJ (1996) **313** 36–39.
6 Roebruck P., Kühn A. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. Stat. Med. (1995) **14** 1583–1594.
7 Senn S. Inherent difficulties with active control equivalence studies. Stat. Med. (1993) **12** 2367–2375.
8 Senn S. Statistical issues in drug development. Wiley, Chichester, 1997.
9 Snapinn S.M. Noninferiority trials. Curr. Cont. Trials Cardiovasc. Med. (2000) **1** 19–21.
10 Temple R., Ellenberg S.S. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Ann. Intern. Med. (2000) **133** 455–470.