STATISTICS IN MEDICINE
*Statist. Med.* 2003; **22**:169–186 (DOI: 10.1002/sim.1425)

# Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics

Ralph B. D'Agostino Sr.[1,2,*,†], Joseph M. Massaro[1,2,3] and Lisa M. Sullivan[1,3]

[1] *Boston University Statistics and Consulting Unit, 111 Cummington Street, Boston, MA 02215, U.S.A.*
[2] *Harvard Clinical Research Institute, 930 Commonwealth Avenue, Boston, MA 02215, USA*
[3] *Boston University, Biostatistics, 715 Albany Street, Boston MA 02118, USA*

SUMMARY

Placebo-controlled trials are the ideal for evaluating medical treatment efficacy. They allow for control of the placebo effect and are most efficient, requiring the smallest numbers of patients to detect a treatment effect. A placebo control is ethically justified if no standard treatment exists, if the standard treatment has not been proven efficacious, there are no risks associated with delaying treatment or escape clauses are included in the protocol. Where possible and justified, they should be the first choice for medical treatment evaluation. Given the large number of proven effective treatments, placebo-controlled trials are often unethical. In these situations active-controlled trials are generally appropriate. The non-inferiority trial is appropriate for evaluation of the efficacy of an experimental treatment versus an active control when it is hypothesized that the experimental treatment may not be superior to a proven effective treatment, but is clinically and statistically not inferior in effectiveness. These trials are not easy to design. An active control must be selected. Good historical placebo-controlled trials documenting the efficacy of the active control must exist. From these historical trials statistical analysis must be performed and clinical judgement applied in order to determine the non-inferiority margin $M$ and to assess assay sensitivity. The latter refers to establishing that the active drug would be superior to the placebo in the setting of the present non-inferiority trial (that is, the constancy assumption). Further, a putative placebo analysis of the new treatment versus the placebo using data from the non-inferiority trial and the historical active versus placebo-controlled trials is needed. Useable placebo-controlled historical trials for the active control are often not available, and determination of assay sensitivity and an appropriate $M$ is difficult and debatable. Serious consideration to expansions of and alternatives to non-inferiority trials are needed. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: control group; clinical trial; placebo control; active control; equivalence; non-inferiority; assay sensitivity

## 1. INTRODUCTION

The randomized clinical trial (RCT) is one of the most important advances in the twentieth century [1–3]. Its importance grew as evidence-based medicine became the norm for establishing efficacy of drugs, biologics and medical devices. In the early 1900s the efficacy of

* Correspondence to: Ralph B. D'Agostino, Boston University Statistics and Consulting Unit, 111 Cummington Street, Boston, MA 02215,U.S.A.
† E-mail: Ralph@bu.edu

*Received October 2002*
*Accepted October 2002*

medical treatments was based on anecdotal evidence, often gathered on one or several patients (medical reports and case series). Some treatments had profound effects such that evidence based on few patients was convincing (for example, penicillin). In general this was not the case. Later, more rigorous studies followed in which several patients were given the same treatment and evaluated. Many of these studies, however, were uncontrolled. Bradford Hill pointed out the problems of these and set the stage for RCTs in the medical arena [4]. Others illustrated the importance of RCTs and the potential deception of uncontrolled clinical trials by contrasting the 'positive results' reported in uncontrolled trials versus RCTs [5–7]. Spilker gave a review in four major clinical areas: psychiatry; depression; respiratory distress, and rheumatoid arthritis [5]. In each area, a substantially higher proportion of positive findings were reported in uncontrolled trials as compared to RCTs. For example, in psychiatric therapy trials, 83 per cent of uncontrolled trials reported positive findings, as compared to only 25 per cent of RCTs [6]. In rheumatoid arthritis trials, 62 per cent of uncontrolled trials reported positive findings, as compared to only 25 per cent of RCTs [7]. The RCT can distinguish the effects of a medical treatment from other effects, such as spontaneous changes in the course of the disease, the body's natural healing, improvement due to participating in a study (that is, the placebo effect), and biases in observation and measurement. Few now doubt the virtues of RCTs for assessing medical treatment efficacy.

The United States' Food and Drug Administration (FDA) emphasizes the need for RCTs for medical treatment (drugs, biologics and devices) approval. For example, the Code of Federal Regulations (CFR) Title 21, Part 314, outlines the procedures for applications to the FDA for approval to market new drugs and Section 126 outlines the criteria of 'adequate and well-controlled' studies [8]. Focus is on the RCT. The same emphasis holds in the international setting. The International Conference on Harmonisation (ICH) is attempting to consolidate procedures for the registration of pharmaceuticals in the European Union, Japan and the United States. The ICH E9 guidance document discusses statistical principles for clinical trials [9]. The ICH E10 guidance document discusses the selection of appropriate controls in clinical trials [10, 11]. The latter document describes five types of controls (placebo, no treatment, dose–response, active and historical), and outlines the advantages and disadvantages of each. The first four controls are concurrent controls. These controls in randomized clinical trials are preferable to historical controls as patients for both the test and control treatments are drawn from the same population and studied under similar conditions, thereby minimizing bias in the comparison. Of all the possible RCTs, to many the ideal is the placebo-controlled RCT.

In the absence of effective treatments, placebo-controlled RCTs are uncontroversial. When, however, a proven effective treatment exists, the ethics of the placebo-controlled trials are questionable. In this setting, the attacks against placebo-controlled trials are many and substantial [12–15]. Of most importance is the Declaration of Helsinki [16]. Article II.3 of this states 'In any medical study, every patient – including those of a control group, if any – should be assured of the best proven diagnostic and therapeutic method. This does not exclude the use of inert placebo studies where no proven diagnostic or therapeutic methods exists'. Many interpret this to mean that when an effective treatment exists the use of a placebo is unethical and should not be included in a RCT. Others, including prestigious groups such as the American Medical Association and the World Health Organization, leave room for the possible use of placebo-controlled RCTs under certain circumstances (see Section 2) [17–21].

The active-controlled trial has been one response to the attack on placebo-controlled trials. Here the new experimental treatment is compared to a proven active control treatment. The

 

new treatment may not be superior to the active treatment in terms of efficacy, but it may be equivalent. Borrowing ideas from the field of bioequivalency, medical researchers including clinicians and statisticians developed equivalency trials with their design issues and the necessary statistical testing procedures [22–27]. Upon further clarification of the issues, it became clear that what was desired were non-inferiority trials (or more precisely, non-inferiority active-controlled RCTs), even if the term 'equivalency trials' is often used. The objective of a non-inferiority clinical trial is to establish that the effect of the new treatment, when compared to the active control, is not below some pre-stated non-inferiority margin.

The designing, implementation and analysis of non-inferiority trials have presented substantial challenges and issues for the pharmaceutical, biologics and medical device industries. The FDA and its scientists are well aware of these [11, 28, 29]. In our roles as academic consultants, industry sponsors are constantly seeking advice to decide when a non-inferiority trial is warranted, to clarify for them the unique design concepts and the issues involved, to help design, implement and perform the trial and ultimately to aid in the analysis and interpretation of the study. In this paper we focus on the *design concepts and issues* involved. We illustrate these with real world examples, many that we have encountered.

In Section 2 we review the usefulness of the placebo-controlled trial and the situations where they may be justified, even when proven active treatments exist. Section 3 discusses two major issues in active-controlled non-inferiority trials: (i) the statistical hypotheses and tests involved in a non-inferiority trial and (ii) the selection of the non-inferiority margin. The latter includes discussion of clinical meaningfulness, assay sensitivity (which relates to establishing that the active treatment and in turn the experimental treatment would have been superior to placebo had a placebo been used in the trial), and the fear of what is called 'biocreep'. Section 4 concerns the putative placebo analysis as a means of establishing that the new treatment is superior to placebo. Section 5 deals with selecting the appropriate sample to use for the statistical analysis. In Section 6 we discuss the role of interim analysis. Then in Section 7 we expand the non-inferiority trial to consider safety issues and also review some alternatives to non-inferiority trials. Finally, in Section 8 we give a brief closing discussion and some recommendations.

## 2. PLACEBO-CONTROLLED TRIALS

An appropriate control group is always essential and, when feasible, a placebo control is optimal. Figures 1 and 2 demonstrate the problem when a study does not contain a placebo control. The comparison of the active control $C$ with the test treatment $T$ in Figures 1 and 2 indicates that the two treatments are similar. However, if a placebo group is not included in the study, then one can never be sure if the new treatment is better than the placebo, as Figure 1 indicates, or not different from the placebo, as Figure 2 indicates. Figure 1 corresponds to both $C$ and $T$ being effective, Figure 2 to neither being effective.

Historically, a placebo control group was the usual optimal control group for establishing efficacy of an experimental treatment. It has been the basis for many FDA approvals. Superiority of the experimental treatment over placebo in two well controlled and performed RCTs justified approval. At times it was essential to establish that the trial had sensitivity (or sometimes called assay sensitivity) and an active control was added as, for example, in analgesic studies [30, 31]. Here the comparison of the active control to the placebo was an

 

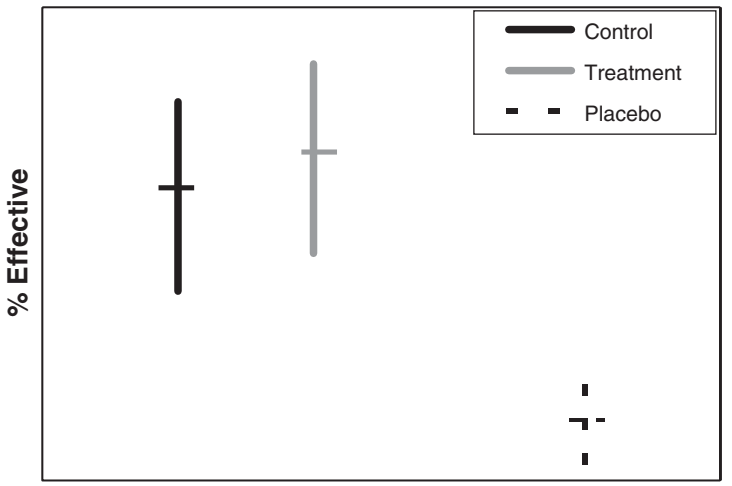


Figure 1. Comparison of test treatment ($T$) with active control ($C$) and unobserved placebo ($P$) ($T$ and $C$ superior to $P$).

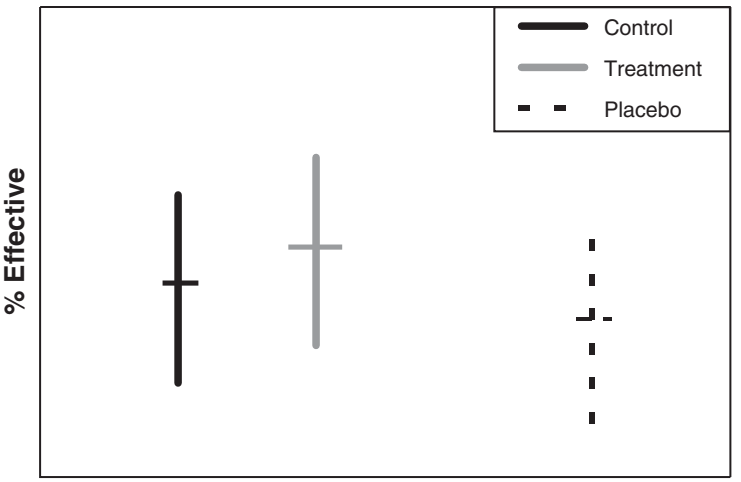


Figure 2. Comparison of test treatment ($T$) with active control ($C$) and unobserved placebo ($P$) ($T$ and $C$ not superior to $P$).

essential component of the analysis. The comparison of the active control to the experimental treatment was not required. The ideal was a study with a placebo, an active control and an experimental treatment.

Now with the large array of proven effective treatments, ethical considerations cast doubts on the appropriateness of using a placebo control. Dose response trials are possible alternatives, but they also raise ethical problems since the low dose may not be any different than a placebo. *So when is a placebo control justified in the presence of proven active treatments?* We agree with Ellenberg and Temple [21]. 'that placebo controls are ethical when delaying

 

or omitting available treatment has no permanent adverse consequences for the patient and as long as patients are fully informed about the alternatives'. We also believe escape clauses should be included in the protocol.

An active control arm may be included in the RCT, but the active control is there for reasons such as assay sensitivity. It is not necessary for comparison with the experimental treatment. Thus for many over-the-counter drug situations such as pain, headaches, upset stomach and the treatment of the common cold, placebo-controlled trials are ethical. Ellenberg and Temple [20, 21] discuss numerous prescription drug situations involving, for example, antidepressants and short term trials (such as some anti-hypertensive trials), and *settings where the available* '*effective treatment*' *may not be uniformly accepted as standard treatment* and so placebo-controlled trials are justified.

## 3. ACTIVE-CONTROLLED TRIALS/NON-INFERIORITY TRIALS

Now let us move to the situation where the placebo control is considered unethical or for some other reason is deemed inappropriate. This leads us to active-controlled trials in which the experimental treatment is compared directly to a proven effective active control. If the sponsor believes the experimental treatment is superior to the active control, then a standard superiority trial with the objective of showing that the experimental treatment is statistically and clinically superior to the active control is appropriate.

What, however, if anticipated superiority is not the case? Then a non-inferiority trial (that is, a trial with the objective of showing that the experimental treatment is statistically and clinically not inferior to the active control) may be appropriate. A sponsor of an experimental treatment may logically decide to conduct a non-inferiority trial even when he believes the active control's efficacy cannot be surpassed. Why? The new product may offer safety advantages. For example, a new anti-infective product may produce no resistant bacteria, a new respiratory distress product for premature infants may be synthetic as opposed to animal derived and pose less risk, a new asthma treatment inhaler may have no chlorofluorocarbons in contrast to the standard product [23]. In the case of HIV treatments, new products may have simpler regimens promoting adherence and potentially reducing resistance. It is even possible that costs, marketing and potential profits are the underlying reasons. For example, the costs of the new product may be less expensive or the sponsor may have better access to the markets.

### *3.1. Statistical algorithm for assessing non-inferiority*

The statistical algorithms for assessing non-inferiority (and equivalency) are in Blackwelder's paper [22]. We give a brief summary here and in Table I. Let $T$ and 'Test' represent the value of the efficacy variable for the new (experimental) treatment. Similarly let $C$ and 'Control' and $P$ and 'Placebo' represent the values of the efficacy variable for the active control and placebo, respectively. Further, say we have a trial where higher values of this efficacy variable are desirable. The standard null and alternative hypotheses for proving non-inferiority are

$$H_0\colon\ C - T \geqslant M\ (C \text{ is superior to } T)$$
$$H_1\colon\ C - T < M\ (T \text{ is not inferior to } C)$$

 

Table I. Hypotheses for a non-inferiority trial.

| |
|---|
| $H_0$: $C - T \geqslant M$ ($C$ superior to $T$) |
| $H_1$: $C - T < M$ ($T$ not inferior to $C$) |

Here $T$ is the new treatment, $C$ is the active control and $M$ is the non-inferiority margin.

Here, $M$ is the non-inferiority margin, that is, how much $C$ can exceed $T$ with $T$ still being considered non-inferior to $C$ ($M > 0$). The null hypothesis states that the active control $C$ exceeds the experimental treatment $T$ by at least $M$; if this cannot be rejected, then the active control is considered superior to the experimental treatment with respect to efficacy. The alternative hypothesis states that the active control may indeed have better efficacy than the experimental treatment, but by no more than $M$. In such a case, we say the investigational product is *not inferior* to the active control. Rejection of the null hypothesis is needed to conclude non-inferiority.

One of the major issues today in non-inferiority clinical trials is the choice of $M$. We discuss this in Section 3.2. We should note here that the above displays the statistical hypotheses as differences between the treatments. The hypotheses could be in terms of means or proportions of successes. Also, depending on the application the hypotheses could be stated in terms of ratios ($C/T \geqslant M$), logs ($\log C - \log(T) \geqslant M$), etc.

In order to assess if non-inferiority is met (that is, whether the null hypothesis is rejected) we can perform a one-sided hypothesis test at $\alpha$ level of significance. Equivalently, we can compute a $100(1-2\alpha)$ per cent two-sided confidence interval for the difference $(C-T)$. If the confidence interval's upper bound is less than $M$, then with $100(1-2\alpha)$ per cent confidence, we say the active control is more efficacious than the investigational product by no more than $M$, hence allowing us to claim non-inferiority of the experimental product as compared to the active control at an $\alpha$ level of significance.

### *3.2. Choosing the non-inferiority margin M*

Prior to mounting the active-controlled non-inferiority trial (or at least before the blinding of the trial is broken) we need to state the non-inferiority margin $M$, that is, how close the new treatment $T$ must be to the active control treatment $C$ on the efficacy variable in order for the new treatment to be considered non-inferior to the active control. The ICH documents offer two guidelines [10]:

1. The determination of the margin in a non-inferiority trial is based on both *statistical reasoning and clinical judgement*, and should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.
2. This non-inferiority margin cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of a placebo-controlled trial.

*While the first guideline mentions 'clinical judgement' we have never seen a case where this has actually been employed.* There is often talk that $C$ and $T$ should be within some percentage of one another (for example, the sponsor says 20 per cent while the FDA says 10

 

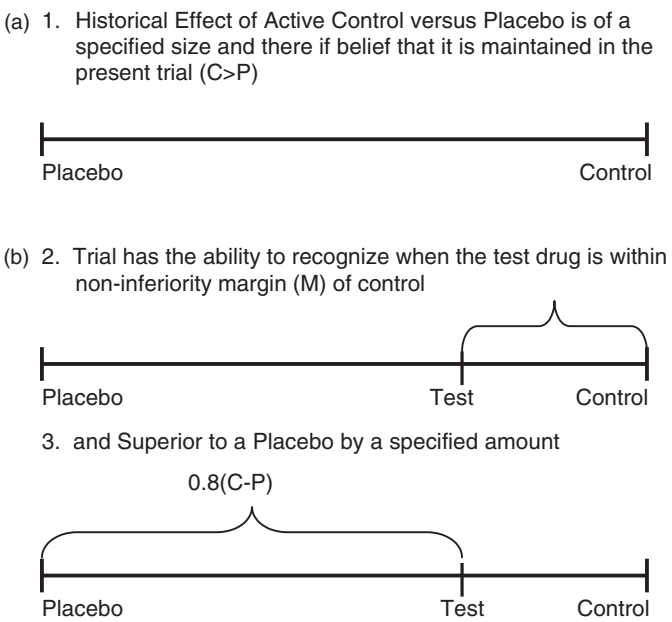


Figure 3. Considerations in the determination of non-inferiority margin $M$. ($a$) Assay sensitivity in a non-inferiority trial. The ability of a specific trial to detect a difference between treatments if one exists (that is, assay is working and can detect a difference). ($b$) Assessment of non-inferiority and putative placebo comparison.

per cent), clinical judgement does not seem to be the deciding factor. Rather, the determination becomes a statistical discussion usually focusing on trying to extract information from historical data. To the dismay of some, the statisticians seem to have taken control of this issue.

Attempts have been made to take a statistical approach; specifically to combine data from historical placebo-controlled trials of the active drug $C$ and determine $M$ so that it reflects the uncertainty in the historical data and is not greater than the smallest reliable effect size of the active treatment versus a placebo [32, 33].

*3.2.1. Our summary of the determination of the non-inferiority margin M.* In our review of the field, the determination of $M$ must address three steps or issues. We present them here and display them in Figure 3.

First, in the non-inferiority trial we must have assurance that the active control would have been superior to a placebo if a placebo were employed. This is the need to demonstrate or establish *assay sensitivity*. The use of past placebo-controlled trials often accomplishes this. We must have available historical data in which it has been established that the active control $C$ is superior to the placebo $P$. Further, we must evoke a very strong assumption, the *constancy assumption*, namely, that the historical difference between the active control and placebo is assumed to hold in the setting of the new trial if a placebo control had been used. This is step 1 in Figure 3.

 

Second, the non-inferiority active-controlled trial should demonstrate that the new treatment $T$ is within the non-inferiority margin $M$ of the active control $C$ (step 2 in Figure 3). This margin should have clinical relevance.

Third, it is then necessary to use the $C$ versus $T$ data (step 2 of Figure 3) in conjunction with the $C$ versus $P$ historical placebo-controlled trial data (step 1 of Figure 3) to demonstrate that $T$ is superior to $P$. This step is the *putative placebo comparison*. In conjunction with this step it is often necessary to establish that not only is the new treatment superior to the placebo, but that it also *retains at least a certain amount of the superiority of the active control over placebo* (say, 80 per cent or 50 per cent). Figure 3, step 3, illustrates this last step. If we think of $(C-P)$ as representing the difference between the active control and the placebo and $(T-P)$ as the difference between the new treatment and the placebo, then the amount retained by the new treatment is $(T-P)/(C-P)$. Jones *et al.* favour 50 per cent [32]. This seems to be where the clinical community is leaning.

One way of viewing $M$ is that it should be no larger than $X(C-P)$ where $C$ and $P$ are based on historical placebo-controlled trials of the active control $C$ versus the placebo $P$ and $X$ is 1 minus the amount of the difference $(C-P)$ we desire to retain with the experimental treatment (for example, $X=1-0.8=0.2$ or $1-0.5=0.5$).

To employ the above, the historical difference $(C-P)$ in Figure 3 must be estimated and this estimate must incorporate the variability in the historical data. Ideally, good historical placebo-controlled data from more than one study are available. In such an ideal situation $(C-P)$ could be estimated as follows. Estimate $C-P$ for each study and its corresponding two-sided 95 per cent confidence interval. Of all the confidence intervals, use the 'smallest' lower bound (that is, the lower bound that yields the smallest value of $C-P$). This is the most conservative estimate of $(C-P)$. Another approach would be to perform a meta-analysis of the historical studies and use the average estimate of $(C-P)$ or the lower confidence limit. Hauck and Anderson [24] discuss more formal approaches for estimating $(C-P)$ and $M$ from previous active versus placebo trials, accounting for both within-trial and across-trial variability. At the present time there is no universally accepted way of doing this.

*3.2.2. Some caveats.* These are caveats:

1. *Assay sensitivity*. As we mentioned above, in some areas, such as the analgesic field, there is a need to include both a placebo control and an active control in the same trial in order to ensure assay sensitivity [30]. No matter how much historical data exists there is no assurance that the next trial will have assay sensitivity. One can argue, for those fields, the use of historical data does tell us about the historical difference between the active and placebo controls, but not necessarily anything about assay sensitivity for the non-inferiority trial.
2. *Constancy assumption*. With the rapid changes in medical practice and standard of care we may not be correct in saying that the historical difference between the active control and placebo is valid for the present day. In our experience this constancy assumption is often a major issue, at times putting an end to a discussion for a formal determination on $M$.
3. *Variability of* $(C-P)$. Suppose the estimate of $C-P$ differs markedly across previous active versus placebo clinical trials. Which is the most appropriate estimate to use for

 

determining $M$ for the non-inferiority study? To be conservative, the smallest estimate of $(C-P)$ should be used, but is that too conservative? What if the smallest estimate of $(C-P)$ is not statistically significant? What if the smallest difference is a case where assay sensitivity was not established?

4. *Small number of available historical placebo-controlled studies.* Historical placebo-controlled trials are often not plentiful; it is the experience of the authors that for many indications, only one historical placebo-controlled trial exists. The estimate of $(C-P)$ from only one study often is called into question by regulatory agencies since there is not an adequate estimate of the variability of estimate of $C-P$.
5. *No available placebo-controlled studies.* In our experience there are cases where there are no placebo-controlled studies. In such situations, one may try to work with previous dose response studies of the active control where the marketed dose of the active control was compared with a low dose. Here the low dose effect may or *may not* be an adequate substitute for a placebo effect.

*3.2.3. Biocreep.* Biocreep is the phenomenon that can occur when a slightly inferior treatment becomes the active control for the next generation of non-inferiority trials and so on until the active controls become no better than a placebo. This is a real possibility, except it is easy to address. The active control comparator should always be the 'best' comparator.

*3.2.4. Two examples. Example 1*: *No available placebo-controlled trials*. Studies in vancomycin-resistant-enterococcal (VRE) infection (where the outcome is success defined as cure of the infection) often use the marketed product linezolid as the active control comparator. Unfortunately, there are no published placebo-controlled studies of linezolid. The results of a study have been published comparing high dose (that is, the marketed dose) linezolid versus low dose. The results showed a difference in success rates of 14 per cent. While this approach is conservative and may underestimate the true $C-P$, it is better than a simple guess at $C-P$. The best value to use for $M$, however, is still not clear. For example, is one-half of 14 per cent too conservative? At the very least, 7 per cent will lead to very large sample sizes, which is problematic due to the very small number of patients with VRE. In this particular example, because there is only one study comparing the marketed dose with a low dose, the reliability of the estimate is also questionable.

*Example 2*: *M and the history of anti-infective trials*. The choice of a margin is quite difficult and somewhat controversial in anti-infective trials. To underscore this fact, consider a non-inferiority anti-infective trial comparing an experimental product to an active control (the non-inferiority study design is quite common for anti-infectives, given the large number of generic and non-generic anti-infectives already marketed). Suppose the outcome is cure or improvement of infection (dichotomous 'success') at the 'test-of-cure visit' (which occurs at a predetermined time interval after the last application of study treatment). Although there are no official guidelines for the choice of $M$, a common recommendation from regulatory agencies is to use $M=10$ per cent, regardless of the specific type or severity of infection. Until recently, however, the FDA considered a 'step function' for $M$. Here $M=0.10$ (or 10 per cent) when it was thought that the cure rate of the active control and investigational drugs were $\geqslant 90$ per cent, an $M$ of 15 per cent when the cure rate was thought to be between 80 and 90 per cent, and an $M$ of 20 per cent when the cure rate was 80 per cent or below. The

 

FDA no longer suggests this step-down function for non-inferiority trials and has disclaimed it on its web site.

The removal of the step-down function, and the unofficial FDA guideline of an $M$ of 10 per cent, caused a major concern in the anti-infective industry [34]. The FDA is now being more conservative with $M$ because of its concern over biocreep. This concern is understandable. However, the concern over biocreep can be counteracted by the FDA regulation of the choice of a comparator in such trials (for example, always use the 'best' comparator). Overall the anti-infective industry is very concerned with using $M = 0.10$, especially in rare, serious infections, since the sample size, cost and time implications can be enormous. For example, if the success rate of both treatments is assumed to be 70 per cent and a non-inferiority margin of $M = 15$ per cent is used in the trial, then the number of *evaluable* subjects required is approximately 400 (this assumes a one-sided significance level of 0.025 and power of 0.90). The sample size increases to approximately 900 evaluable subjects when the non-inferiority margin is reduced $M = 10$ per cent. Enrolling such numbers of patients can be practically impossible for rare, serious infections.

## 4. PUTATIVE PLACEBO ANALYSIS

Assay sensitivity of the active control is determined from the historical active- versus placebo-controlled trials. In the above, the putative placebo comparison of the new experimental treatment to the placebo was satisfied by requiring that the new experimental treatment retains a portion of the active control's superiority to the placebo. A second approach due to Lloyd Fisher [35] has been published by Hasselblad and Kong [36]. This method involves estimating the effect of the new experimental treatment compared to the placebo by a set of ratios as follows:

$$T \text{ versus } P = T/P = T/C \times C/P$$

$T/C$ and $C/P$ can be, for example, the relative risks comparing treatments. Note $T/C$ is from the non-inferiority trial and $C/P$ is from a meta-analysis of the historical placebo-controlled trials, so the $C$s are from different data sets. The approach is very clever for from the above we can in fact obtain an estimate of the variance of the effect of the new treatment to placebo. We obtain this simply by taking logs

$$\ln(T/P) = \ln(T/C) + \ln(C/P)$$

and

$$\text{var}(\ln(T/P)) = \text{var}(\ln(T/C)) + \text{var}(\ln(C/P))$$

Here var denotes variance. Note that all the quantities on the right side of the equations are obtainable from existing data. Odds ratios can be dealt with using ratios directly. Others have suggested similar methods [24, 37] and even a Bayesian approach has been developed [38].

For Hasselblad and Kong the active control versus the placebo control comparison (assay sensitivity) is obtained by a meta-analysis and the new treatment versus the placebo comparison (putative placebo comparison) is obtained using the method just described.

 

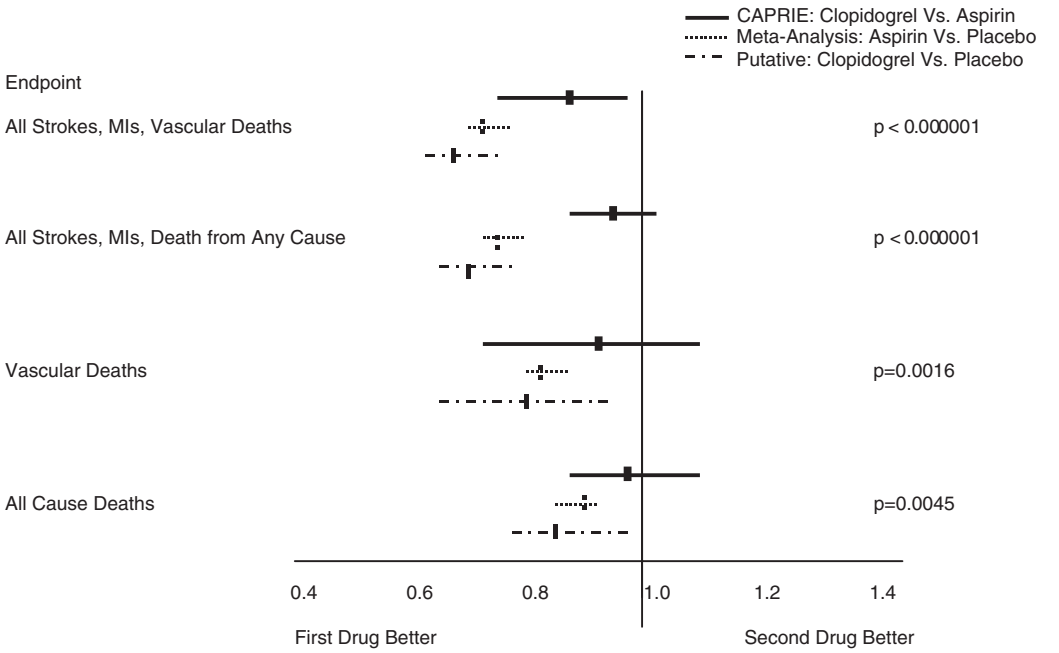


Figure 4. Clopidogrel versus synthetic placebo control: odds ratios and 95 per cent confidence intervals.

### *4.1. CAPRIE trial*

Hassleblad and Kong use as one of their major examples the remarkable clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE) study [39]. This study was not designed as a non-inferiority study, rather it appears to be a superiority study with clopidogrel hypothesized to be superior to aspirin. The subjects of the study were patients with myocardial infarction (MI), ischaemic stroke (IS) or peripheral arterial disease (PAD). The primary endpoint of the study was a composite endpoint of the incidence of MI, IS or vascular death over 1 to 3 years of follow-up (mean follow-up of 1.91 years; total sample size = 19185). There was no placebo control group in this trial. An intent-to-treat analysis showed that clopidogrel patients had an annual 5.32 per cent risk of IS, MI or vascular death versus an annual risk of 5.83 per cent in aspirin patients (risk reduction of 8.7 per cent with a 95 per cent confidence interval of 0.3–16.5 per cent; $p = 0.043$).

To estimate assay sensitivity and perform the putative placebo analysis of clopidogrel as compared to placebo on the composite endpoint, a meta-analysis of all published and unpublished studies through 1990 was conducted. Placebo-controlled aspirin studies from the Antiplatelet Trialists' Collaboration (APTC) [40] in patients with prior MI, prior stroke or transient ischaemic attack, or intermittent claudication were included. The total number of studies included in the meta-analysis was 41, and each had clear definitions of endpoints and well defined statistical methodology. There was no significant heterogeneity across these 41 studies with respect to the aspirin–placebo treatment difference, justifying the meta-analysis [35]. Figure 4 displays odds ratios and their 95 per cent confidence intervals for various

 

cardiovascular endpoints for (a) clopidogrel versus aspirin as estimated from the CAPRIE trial, (b) aspirin versus placebo as estimated from the meta-analysis (assay sensitivity), and (c) clopidogrel versus placebo estimated using the meta-analytic methods of Hasselblad and Kong (putative placebo). The *p*-values given in the table *are reported to* assess the significance of the clopidogrel versus placebo odds ratio.

### *4.2. Concerns with CAPRIE study application of putative placebo analysis*

While the meta-analysis approach by Fisher/Hasselblad and Kong is a major advancement, its application, especially to the CAPRIE study, raises many problems that call into question this particular application and the general application of the method. In fact the problems and concerns are universal to the practice of using historical controlled studies.

1. *Assay sensitivity*. Many of the APTC studies were performed before the 1990s when the CAPRIE study was undertaken. Today, the baseline (placebo) rates for the cardiovascular events would most likely be lower than what the meta-analysis produced. A true measure of the assay sensitivity of aspirin versus placebo may not be obtainable from the data.
2. *Constancy assumption.* Again, because of the age of the APTC studies and the changes in the health care systems, new medical practices, changes in the recognition and diagnosis procedures, and even changes in the disease process, it is hard to argue that the constancy assumption holds.
3. *Consistency of diagnosis and retrospective attainment of primary endpoint*. The primary outcome of the CAPRIE trial was the composite endpoint of IS, MI or vascular death. Many of the APTC trials did not have IS, MI or vascular death as an outcome and hence did not collect its incidence. The original investigators were asked to generate *retrospectively* this data for the subjects in their study, leading to potential immeasurable biases in the results. It is hard to imagine another setting in the drug approval process where retrospective data of this nature would be acceptable.
4. *Heterogeneity across study populations*: *effect of peripheral arterial disease* (*PAD*). Another issue is the potential for heterogeneity across study populations. In the APTC trials there was no significant difference between aspirin and placebo for the PAD subjects. Yet, in a subset analysis of the CAPRIE trial, the *only entry group* that attained statistically significant differences for clopidogrel versus aspirin was the PAD group. Figure 5 shows the relative risk reduction for the CAPRIE trial by stratified by entry criteria. The significance of the PAD group is not shared by the other entry groups (MI or IS) [39]. Further, there was a significant interaction between treatment and entry groups. The overall significance of the CAPRIE trial may relate to a subset for which the significant difference between aspirin and placebo has not been demonstrated.
5. *Interpretation of the p-levels from the putative placebo analysis*. The reported level of significance for the putative placebo comparison is given in Figure 4 as $p < 0.000001$. This *p*-level is obtained from using the meta-analysis data that contains all the problems enumerated above. This *p*-level does not have the interpretation that comes from a well-controlled RCT that would normally be presented in a drug approval application and should not be interpreted as such.

 

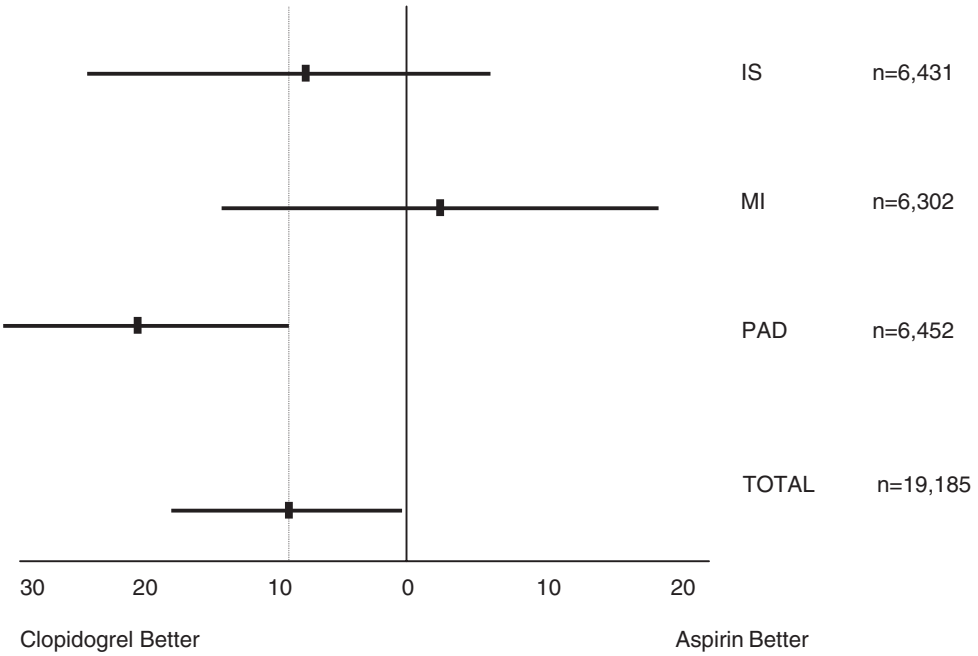


Figure 5. Relative risk reduction by qualifying condition.

### *4.3. Concerns with historical active versus placebo control trials*

In general, when considering historical placebo-controlled trials to estimate the effect of an experimental treatment versus placebo, the following questions need to be considered:

1. Is the disease/condition being studied now the same as in the historical studies? Have there been changes in the diagnosis of the disease? Course of the disease?
2. Have there been changes in what is the standard of care or treatment of the disease/condition?
3. Is the same population being studied? Are the new study subjects from similar settings, same age, same gender etc.?
4. Is the dose and route of administration of the active control in the historical trials the same as in the current study?
5. Are the outcomes and modes of data collection consistent across studies? Will we allow retrospective collection of outcomes in the historical controlled studies? If so, what is the effect on our estimate of assay sensitivity?
6. Which historically placebo-controlled studies do we use? If not all, how do we select the ones of interest? If some studies showed no significant effect of active control versus placebo (due to lack of efficacy of the active drug or due to a high placebo effect), then do we include these studies? What are the implications of their inclusion or exclusion?

In general, inherited biases in using historical placebo-controlled studies cannot be fixed by statistical adjustments and must be carefully thought out.

 

## 5. APPROPRIATE ANALYSIS SAMPLES: INTENT-TO-TREAT VERSUS PER-PROTOCOL

In an intent-to-treat analysis (ITT), patients are analysed according to the treatment to which they were assigned, regardless of whether they received the assigned treatment [41]. Some definitions of ITT exclude patients who never received treatment. The per-protocol (PP) analysis includes all patients who completed the full course of assigned treatment and who had no major protocol violations. In non-inferiority trials the ITT analysis tends to be 'liberal'. That is, by inclusion of those who do not complete the full course of the treatments, the ITT tends to bias towards making the two treatments (new treatment and active control) look similar. The PP removes these patients and is more likely to reflect differences between the two treatments.

The above suggests that the PP analysis may be preferable in the non-inferiority trial setting. Such logic is in direct opposition to current thinking that the ITT should be the preferable analysis. The current thinking of regulatory agencies is that the study objective should be achieved in both the ITT and PP populations, especially in a non-inferiority trial. The Committee on Proprietary Medical Products Points-to-Consider (CPMP, 2000)[42] specifically states: '…similar conclusions from both the Intent-to-Treat and Per-Protocol are required in a non-inferiority trial'. This approach makes sense, as the ITT tends to give an (albeit, perhaps not ideal) estimate of the overall effect that the experimental treatment will have on the population, since not all people taking the experimental product in the population will take it for the full course as prescribed. The PP results estimate the overall effect of the full course of experimental treatment. Both sets of results are important and should be considered when assessing if the study objective is met. Thus, the sponsor of an experimental product should ensure sufficient sample size exists in both the ITT and PP samples. Because there are fewer patients in a PP analysis, sample size computations should be performed to ensure sufficient numbers of subjects in the PP population and then increased for the ITT population. Jones *et al.* [32] provide formulae for determining required samples sizes for equivalence and non-inferiority studies for both normally distributed and binary outcomes.

## 6. INTERIM ANALYSES

Interim analyses are often performed in superiority trials in order to stop the trial in the case where the experimental treatment is causing more harm than good, or in the case where the experimental treatment is far superior to the control. In the former case, the study is stopped for safety concerns; in the latter case, the study is stopped in order to approve the experimental treatment and get it to market as quickly as possible. Also, it is considered unethical to continue giving patients the control when the experimental treatment is far superior. For interim efficacy analyses, the alpha is split across the analyses so the overall alpha spent is no more than 0.05, often using the Lan–DeMets spending function [43].

Interim analyses are also important in a non-inferiority trial for safety reasons, either to ensure the experimental treatment is not doing more harm than good, or that it is superior with regard to specific adverse events. For efficacy reasons one can argue that there is no real necessity for interim analyses in a non-inferiority trial. This is because there is no real ethical issue with seeing the trial to completion from an efficacy perspective, since we will not

 

find that the experimental treatment is superior to the active control (since the active control usually has a high success rate to begin with) but rather that the experimental treatment is not inferior to the active control. In other words, there is no real ethical need to 'get the experimental treatment to the market' as quickly as possible from an efficacy perspective. Nevertheless, if an interim efficacy analysis is desired, it can be done where the alpha is split across the interim analyses. For example, suppose a study is conducted with one planned interim analysis and one final analysis. It is recommended that the alpha (say, 0.05) be split in such a way that a two-sided 99.95 per cent confidence interval is calculated in the first interim analysis, and a 95.05 per cent confidence interval is calculated in the final analysis. As with a superiority trial, it is desired to keep as much alpha as possible for the final analysis; a Bonferroni-type even split of the alpha could lead to not claiming non-inferiority when in fact there really is non-inferiority. The issue with this approach is that, in reality, the experimental treatment would need to be vastly superior to the active control in order to declare non-inferiority in the first interim analysis, given the high degree of confidence (and hence a very wide confidence interval that in all likelihood will cover the non-inferiority margin $M$). Thus, if a sponsor truly believes the experimental treatment is not superior to the active control, then most likely the cost and time to carry out an interim analysis for efficacy will not be justified.

## 7. EXPANSION OF AND ALTERNATIVES TO THE NON-INFERIORITY TRIAL

A new medical treatment may have benefit even if it does not have efficacy superiority over the active control. These benefits can include cost, safety or some indication-specific reason. Here we discuss briefly strategies for establishing efficacy of a new experimental treatment that does not depend solely on non-inferiority trial methods discussed above.

### *7.1. Safety superiority*

Often, in our experience with non-inferiority trials, a sponsor will have a composite objective of showing that the experimental treatment, while being non-inferior to the active control, has a significantly lower incidence of specific adverse events or adverse events in a specific body system (for example, a pharmacological stress-inducing agent for magnetic resonance imaging of the heart may have fewer cardiovascular events than the active control). The question arises as to whether or not the alpha level (for example, 0.05) needs to be split evenly across the two objectives (non-inferiority and safety) using a Bonferroni approach. Our recommendation to the sponsor is to perform the assessment of efficacy non-inferiority and safety superiority each at an alpha of 0.05. With such an approach, the overall alpha across the study is controlled at 0.05.

In such a non-inferiority trial with an added objective of safety superiority, the sample size must be determined to ensure adequate power in both the efficacy and safety objectives. This can be achieved by calculating the sample size separately for each objective, and then choosing the maximum of the two sample sizes to carry out the study.

Using safety as an outcome in what may be considered an efficacy trial setting raises issues and problems. We do not suggest the above will automatically work. We suggest it as a point and possibility for discussion.

 

### *7.2. Different setting*

Another strategy is to locate the study in a region or setting where the active control is not the usual standard of a care and perform a superiority trial of the experimental treatment against the best standard of care for the region. This may well require moving the study to a non-U.S. or non-European setting. Sponsors often point out for such trials that the control treatment (standard of care) may be a much better treatment than is usually given in the area. This type of study is full of ethical concerns and can raise serious problems for the sponsor; just ask the sponsors of such trials who have had the trials reported in the *New York Times.*

There is a twist to this latter strategy that can work reasonably well. Specifically, the control treatment is an approved efficacious drug that is not the best in the field. Such a superiority trial is ethically sound.

### *7.3. Historical control without a placebo control*

There are situations where the standard of care is untested and the populations are extremely small. This is the situation with, for example, Fabrye disease. There are only 3500 cases in the entire United States. There are no placebo-controlled trials for the major endpoints of renal disease or cardiovascular disease. There are, however, studies where experimental treatments have demonstrated superiority on surrogate endpoints. Under the accelerated approval process, a phase IV study is needed to demonstrate clinical benefit. A non-inferiority trial with the new treatment versus the 'unproven' standard of care drug or a superiority placebo-controlled trial is suggested. Interpretation of the former is problematic and attempting to perform the latter may be unethical and not feasible given the drugs have already produced acceptable phase III studies. Because the disease is well studied in special locations (specialized clinics, hospitals), the possibility of obtaining good historical (contemporary) control data appears feasible. Comparison of these with matched series of cases treated with the experimental treatment may be a valid means of treatment evaluation.

## 8. DISCUSSION AND RECOMMENDATIONS

Placebo-controlled trials are the ideal for evaluating medical treatment efficacy. They allow for control of the placebo effect and are most efficient requiring the smallest numbers of patients to detect a treatment effect. A placebo control is ethically justified if no standard treatment exists, if the standard treatment has not been proven efficacious, there are no risks associated with delaying treatment and/or escape clauses are included in the protocol. Where possible and justified to use, they should be the first choice for medical treatment evaluation.

There are instances when a placebo is not justified or is unethical, and in these situations an active control is used. Non-inferiority trials offer a possibility for evaluating the efficacy of new treatments versus active controls. Large samples are usually needed for these studies. In addition serious considerations are needed for deciding upon the active control, determining the non-inferiority margin $M$, assessing assay sensitivity, evaluating the constancy assumption, and carrying out the study non-inferiority hypothesis testing and putative placebo analysis.

In assessing assay sensitivity and in determining the non-inferiority margin $M$, the presence of multiple, well carried out, historical trials comparing the active control to placebo that are

 

consistent with the present active control trial under study will alleviate many problems. The presence of such historical trials will allow for a reliable estimate of the efficacy difference between the active control and placebo, thereby allowing for: (a) estimation of the difference between the experimental treatment and placebo using methods such as the Fisher/Hasselblad and Kong method; and (b) assignment of $M$ as a fraction (for example, one-half) of the difference between the active control and placebo. However, in our experience, the presence of such historical trials is rare for a number of conditions, thereby making determination of assay sensitivity and an appropriate $M$ difficult and debatable. Serious consideration to expansions of and alternatives to non-inferiority trials are needed.

## REFERENCES


1. Fisher LD. Advances in clinical trials in the twentieth century. *Annual Review of Public Health* 1999; **20**: 109–124.
2. Harrington DP. The randomized clinical trial. *Journal of the American Statistical Association* 2000; **95**: 312–315.
3. Smith R. Fifty years of randomized controlled trials. *British Medical Journal* 1998; **317**:1166.
4. Hill AB. *Statistical Methods in Clinical and Preventive Medicine*. Oxford University Press: New York, 1962.
5. Spilker B. *Guide to Clinical Trials*. Wiley: New York, 1991.
6. Foulds GA. Clinical research in psychiatry. *Journal of Mental Science* 1958; **104**:259–265.
7. O'Brien WM. Indomethacin: a survey of clinical trials. *Clinical Pharmacological Therapy* 1968; **9**:94–107.
8. Adequate and well-controlled studies. Code of Federal Regulations, 21, Part 314.126. Revised as of 1 April 2000. Washington, DC, U.S. Government Printing Office, 2000.
9. International Conference on Harmonisation. Statistical principles for clinical trials (ICH E 9). Food and Drug Administration, DHHS, 1998.
10. International Conference on Harmonisation. Guidance on choice of control group and related design and conduct issues in clinical trials (ICH E 10). Food and Drug Administration, DSSH, July 2000.
11. Department of Health and Human Services, Food and Drug Administration [Docket No. 99D-3082], International Conference on Harmonisation. Choice of Control Group in Clinical Trials (E10). Federal Register vol. 64, No. 185, 51767–51780.
12. Rothman KL, Michels KB. The continued unethical use of placebo controls. *New England Journal of Medicine* 1994; **331**:393–398.
13. Freedman B, Weijer C, Glass KC. Placebo orthodoxy in clinical research I: Empirical and methodological myths. *Journal of Law and Medical Ethics* 1996; **24**:243–251.
14. Freedman B, Weijer C, Glass KC. Placebo orthodoxy in clinical research II: Ethical, legal, and regulatory myths. *Journal of Law and Medical Ethics* 1996; **24**:252–259.
15. Angell M. The ethics of clinical research in the Third World (editorial). *New England Journal of Medicine* 1997; **337**:847–849.
16. World Medical Association Declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subject. *Journal of the American Medical Association* 1997; **277**:925–926.
17. American Medical Association Council on Ethical and Judical Affairs. The use of placebo controls in clinical trials. Report 2-A-1996. American Medical Association: Chicago, 1996.
18. Grof P, Lapierie YD, Akhter MI, Campbell M, Gottfries CG, Khan I, Lembeger L, Müller-Derlingshausen B, Woggon B. Clinical evaluation of psychotropic drugs for psychiatric disorders: principles and proposed guidelines. In *WHO Expert Series on Biological Psychiatry*, vol 2. Hogrefe & Hubert: Seattle, 1993; 28–29.
19. Council for international Organizations of Medical Sciences. International Ethical Guidelines for Biomedical Research Involving Human Subjects. Council for International Organizations of Medical Sciences: Geneva, 1993.
20. Temple R, Ellenberg SS. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine* 2000; **133**:455–463.
21. Ellenberg SS, Temple R. Placebo-controlled trials and active controlled trials in the evaluation of new treatment. Part 2: practical issues and specific cases. *Annals of Internal Medicine* 2000; **133**:464–470.
22. Blackwelder WC. Proving the null hypothesis. *Controlled Clinical Trials* 1982; **3**:345–353.
23. Ebutt AF, Frith L. Practical issues in equivalency trials. *Statistics in Medicine* 1998; **17**:1691–1701.
24. Hauck WW, Anderson S. Some issues in design and analysis of equivalence trials. *Journal of the Drug Information Association* 1998; **33**:109–118.
25. Hwang IK, Toshiko M. Design issues in noninferiority/equivalence trials. *Journal of the Drug Information Association* 1999; **33**:1205–1218.


 

26. Rohmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**:1703–1714.
27. Siegel JP. Equivalence and noninferiority trials. *American Heart Journal* 2000; **139**:S166–S170.
28. Guidelines for the format and content of the clinical and statistical sections of new drug applications. U.S. Department of Health and Human Services, Public Health Service, Food and Drug Administration, Rockville, MD, 1998.
29. Temple R. Problems in interpreting active control equivalence trials. *Accountability in Research* 1996; **4**: 267–275.
30. D'Agostino RB, Heeren TC. Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls. *Statistics in Medicine* 1991; **10**:1–6.
31. Lasagna L. Placebos and controlled trials under attack (editorial). *European Journal of Clinical Pharmacology* 1979; **15**:373–374.
32. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36–39.
33. Holmgren EB. Establishing equivalence by showing that a prespecified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 1999; **9**:651–659.
34. Shales DM, Mollering RC Jr. The United States Food and Drug Administration and the End of Antibiotics. *Clinical Infectious Diseases* 2002; **34**:420–422.
35. Fisher LD. Active control trials: what about a placebo? A method illustrated with clopidogrel, aspirin and placebo. *Journal of the American College of Cardiology* 1998; **31**:49A.
36. Hasselblad V, Kong DF. Statistical methods for comparison of placebo in active-control trials. *Drug Information Journal* 2001; **35**:435–449.
37. Fleming TR. Evaluation of active control trials in AIDS. *Journal of Acquired Immune Deficiency Syndromes* 1990; **3(Supp 2)**:S82–S87.
38. Simon R. Bayesian design and analysis of active control clinical trials. *Biometrics* 1999; **55**:18–21.
39. CAPRIE Steering Committee. A randomized, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* 1996; **348(16)**:1329–1339.
40. Antiplatelet Trialists' Collaboration. Collaborative overview of randomized trials of antiplatelet therapy-I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *British Medical Journal* 1994; **308**:81–106.
41. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*, 2nd edn. PSG Publishing Co: Littleton, MA, 1985.
42. Committee on Proprietary Medical Products Point-to-Consider. Points to consider on switching between superiority and non-inferiority. CPMP, 2000.
43. DeMets DL, Lan KKG. Interim analysis: the alpha spending function approach. *Statistics in Medicine* 1994; **13**:1341–1352.

 