The American Journal of Bioethics, 9(9): 34–48, 2009

ISSN: 1526-5161 print / 1536-0075 online
DOI: 10.1080/15265160903090041

Target Article

Questioning the Methodologic Superiority of 'Placebo' Over 'Active' Controlled Trials

Jeremy Howick, University of Oxford

A resilient issue in research ethics is whether and when a placebo-controlled trial (PCT) is justified if it deprives research subjects of a recognized treatment. The clinicians' moral duty to provide the best available care seems to require the use of 'active' controlled trials (ACTs) that use an established treatment as a control whenever such a therapy is available. In another regard, ACTs are supposedly methodologically inferior to PCTs. Hence, the moral duty of the clinical researcher to use the best methods will favor PCTs. In this target article, I analyze the three reasons for believing that ACTs are inferior to PCTs namely: 1) ACTs lack 'assay sensitivity'; 2) ACTs do not measure absolute effect size; and 3) ACTs require more participants; and I contend that none are acceptable. Consequently the tension between clinical and research ethics dissolves: the moral duty of the clinician to avoid PCTs is unopposed by methodological considerations.

Keywords: placebo, clinical trial, 'active' controlled trial, assay sensitivity, absolute effect size, research ethics, clinical ethics, methodology, randomized trial, RCT, non-inferiority, equivalence

> Let us examine the placebo somewhat more critically, however, since it and 'double blind' have reached the status of fetishes in our thinking and literature. The Automatic Aura of Respectability, Infallibility, and Scientific Savoir-faire which [sic] they possess for many can be easily shown to be undeserved in certain circumstances,
>
> — *Hippocratic Oath—Modern version* (Lasagna, 1964, 360).

EPISTEMOLOGICAL FOUNDATIONS OF THE ETHICAL DEBATE OVER THE USE 'PLACEBO'-CONTROLLED TRIALS

A resilient problem in research ethics is whether, and when, a 'placebo'-controlled trial (PCT) is justified if it deprives some research subjects of recognized therapy. In one regard, standards for ethical clinical practice from the Hippocratic Oath to more modern guidelines (World Medical Association [WMA] 1949; Lasagna 1964; General Medical Council [GMC] 2006) require the clinician to provide the best available treatment. This moral duty would seem to require that the clinician avoid PCTs where there is an established therapy. Instead of PCTs, the ethical clinician should advocate 'active'-controlled trials (ACTs) that compare the new treatment with the best-established treatment.

In another regard, the moral duties of the clinical researcher require her to consider PCTs even when an established treatment is available. It is alleged that ACTs suffer from methodological limitations that make them incapable of detecting effectiveness (Ellenberg and Temple 2000; International Conference on Harmonization [ICH] 2000; Temple and Ellenberg 2000; WMA 2001; Miller and Brody 2002, 7). Poor quality research cannot provide the desired knowledge hence wastes scarce resources and exposes participants to unnecessary risks and burdens (Altman 1980; CIOMS 1993; Emanuel et al. 2000; Halpern et al. 2002). Hence, the moral and professional duties of the clinical researcher will tend to oppose the moral duties of the clinician.

Here, as elsewhere (Ashcroft and ter Meulen 2004; Worrall 2007), ethics and epistemology are inseparable. In this case the alleged methodological differences between PCTs and ACTs imply different ethical duties for clinicians and researchers. However, the alleged methodological advantages of PCTs have been asserted more often than argued for. For instance, a revised Declaration of Helsinki (WMA 2008) states:

> The use of placebo, or no treatment, is acceptable in studies where no current proven intervention exists; or
>
> Where for compelling and scientifically sound methodological reasons the use of placebo is necessary to determine the efficacy or safety of an intervention and the patients who receive placebo or no treatment will not be subject to any risk of serious or irreversible harm.

Acknowledgments: Stephen Senn, John Worrall, Nancy Cartwright, Richard Stevens, and Iain Chalmers offered useful advice on earlier drafts of this paper. Two anonymous referees also provided especially constructive criticism. This paper was revised whilst I was a recipient of a National Institute of Health Research Post-Doctoral bursary from the Department of Primary Health Care (Oxford, 2007–8) and a current (2008–10) MRC/ESRC Interdisciplinary Postdoctoral Fellowship (G0800055) hosted at the Centre for Evidence-Based Medicine, at the Department of Public Health and Primary Care, Oxford.
Address correspondence to Jeremy Howick, University of Oxford, Centre for Evidence-Based Medicine, Old Road Campus, Rosemary Rue Building, Oxford, OX3 7LF, United Kingdom. E-mail: jeremy.howick@dphpc.ox.ac.uk

Yet the authors are silent when it comes to specifying what the "compelling reasons" for requiring PCTs might be, let alone providing any arguments for why we should accept them. Likewise, Miller and Brody (2002) devote one paragraph and cite only two sources (Ellenberg and Temple 2000; Temple and Ellenberg 2000) to justify their claim that ACTs suffer from methodological flaws. The International Conference on Harmonization E10 (ICH 2000) Document, produced and endorsed by the regulatory bodies of the United States, the European Union, and Japan enumerates the alleged methodological flaws with ACTs:

1) ACTs do not always possess 'assay sensitivity', whereas PCTs do (ICH 2000, section 1.5),
2) ACTs do not provide a direct measure of absolute effect size whereas PCTs do (ICH 2000, section 2.1), and
3) ACTs require a larger sample size than PCTs (section 2.4).

But the document fails to defend the claims with sustained arguments. In this paper I aim to address this oversight and examine the arguments supporting the methodological superiority of PCTs. To anticipate, I will contend that none are acceptable. If correct, the tension between clinical ethics and research ethics over the use of PCTs dissolves. The ethical duty of the clinician to avoid PCTs where established therapy is available is unchallenged by methodological considerations. Moreover, ACTs are preferable from a practical point of view. What the average patient, clinician, and policy maker needs to know is not whether a new treatment is better than a placebo, but whether the new therapy is better than what we already have.

PROBLEMS WITH THE ASSAY SENSITIVITY ARGUMENTS AGAINST 'ACTIVE'-CONTROLLED TRIALS

Assay sensitivity is defined as the ability of a trial to distinguish differences between experimental and control therapies. There are, however, two distinct versions of the definition, each leading to different arguments against ACTs. Temple and Ellenberg (2000) define assay sensitivity as: "The ability of a study to distinguish between active [non-placebo] and inactive [placebo] treatments" (457). Others define assay sensitivity as "the ability to distinguish a more effective treatment from a less effective [placebo or not] treatment" (Hwang and Morikawa 1999, 1208; ICH 2000, 7). The first assay sensitivity argument is that PCTs but not ACTs can distinguish between placebos and non-placebos, while the second is that PCTs but not ACTs can distinguish between more effective and less effective treatments. I will examine each in turn.

The motivation for the first assay sensitivity argument is clear: most medical treatments used until at least the mid-19th century were either no better than placebo, or worse (Shapiro and Shapiro 1997; Wootton 2006). Even recently, careful investigation has uncovered that several widely used treatments were useless or harmful (Echt et al. 1991; Hayes et al. 1994; Dwyer and Ponsonby 1996; Herbert et al. 1999; Takala et al. 1999; ALLHAT 2000; Rossouw et al. 2002; Ebell et al. 2004; Evans, Thornton et al. 2007, 7–27). In brief, history teaches us that we cannot always assume that our existing treatments are effective (more effective than placebo).[1] If an experimental treatment demonstrates superiority to an established treatment that itself is less effective than placebo, we cannot conclude that the new agent is effective. Hence, ACTs that employ harmful treatments as controls will not possess assay sensitivity (of the first kind). Put differently, to claim an ACT possesses assay sensitivity, we must assume that the control treatment was effective.

Placebo controlled trials, on the other hand, purportedly do not suffer from these problems. A 'positive' result of a PCT, where the experimental treatment demonstrates superiority to placebo, appears to justify the inference to 'effectiveness' without any external assumptions.

> A well-designed study that shows superiority of a treatment to a control . . . provides strong evidence of the [non-placebo] effectiveness of the new treatment, limited only by the statistical uncertainty of the result. No information external to the trial is needed to support the conclusion of effectiveness (Temple and Ellenberg 2000, 456).

Or so it seems.

I will contend that the first assay sensitivity argument against ACTs is problematic in two ways. Firstly, PCTs can also lack assay sensitivity because actual 'placebo' controls used in clinical trials can be either more or less effective than 'real' placebos and secondly, the argument is severely limited in scope. Most of our current therapies are effective.

Why 'Placebo'-Controlled Trials Suffer from Assay Sensitivity Problems: Actual 'Placebo' Controls can Be More, or Less Effective than 'Real' Placebos

Actual 'placebo' controls used in trials are often 'illegitimate' in the sense that they do not accurately measure the 'placebo' effect. To see why, note that even treatments not considered 'complex' have several components. Therapy for depression involving Prozac (Eli Lilly, Indianapolis, USA) includes fluoxetine hydrochloride, the pill casing, the liquid with which the pill is swallowed, the beliefs and expectations with which the pill is delivered, and perhaps other features. Fluoxetine hydrochloride is the characteristic (non-placebo) feature of treatment for depression; the other features are non-characteristic. To be sure, the division between characteristic and non-characteristic features is sometimes controversial (Howick et al. 2009, unpublished data), but the arguments here are unaffected by any controversy.

The terms *active* (Hróbjartsson 2002) and *specific* (Shapiro and Morris, 1978) are also used in the literature to describe the characteristic features of a treatment. However,

1. We do not necessarily require a 'placebo' controlled trial to demonstrate effectiveness (superiority to 'placebo'). To use a much cited example, we do not need 'placebo' controls to determine that parachute use is more effective than 'placebo', however 'placebo' is defined (Smith and Pell 2003).

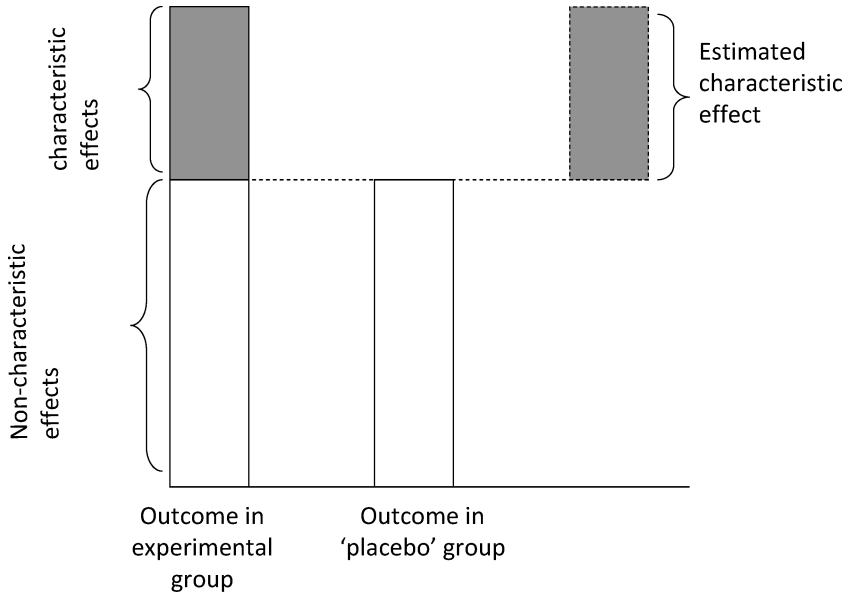

Figure 1. How the absolute measure of characteristic effects are allegedly provided by 'placebo'-controlled trials [PCT].

in at least some cases placebos are active, and moreover their effects can be quite specific (Grünbaum 1986; Benedetti and Amanzio 1997). Indeed recent functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) scans have revealed (specific!) mechanisms of action for both 'placebo' and 'nocebo' (negative 'placebo' side effects) analgesia (Petrovic et al. 2002; Wager et al. 2004; Bingel et al. 2006; Craggs et al. 2007; Kong et al. 2007, 2008; Oken 2008). Indeed one hopes that advances in neuroimaging techniques will shed further light on the mechanism of action, and eventually the conceptualization, of the placebo. In light of these considerations, the term 'characteristic' appears to be less misleading than the other two.

Superiority of the experimental intervention (characteristic + non-characteristic features) over the 'placebo' control (non-characteristic features) is taken as evidence that the characteristic features are positively effective (Figure 1). But this 'equation' will only hold if the 'placebo' control is legitimate, which is to say that it differs from the experimental treatment only in that it does not contain the characteristic features.

Unfortunately, illegitimate 'placebo' controls that either contain characteristic features of their own, or are missing some non-characteristic features of the experimental treatment, are more common than we believe. Although legitimate placebos are notoriously difficult to design for non-drug treatments (Boutron et al. 2005) such as neuroreflexology (Berguer et al. 2008) and exercise (McCann and Holmes 1984; Dunn et al. 2005; Trivedi et al. 2006), legitimate drug 'placebos' are also problematic (Howick 2009; Howick, Golomb et al. 2009. Unpublished data). In one example olive oil was used in 'placebo' controls for cholesterol lowering drugs (Golomb 1995). It was subsequently learned that olive oil had cholesterol-lowering properties of its own, leading to a possible underestimation of the characteristic effects of the experimental drug.

In other cases, 'placebo' control treatments could be illegitimate because they fail to contain the same side effects as the experimental treatment. In an interesting investigation, Moncrieff and colleagues found that what are referred to as *active placebos* (that mimic the side-effects of the experimental treatment) reduce the apparent characteristic benefit of antidepressant drugs (Moncrieff 2003; Moncrieff et al. 2004). The most plausible explanation for this phenomenon is that placebos that do not imitate any side effects are more easily identified as placebos. If a participant believes she is taking a 'mere' placebo rather than the experimental treatment, she will have lower expectations regarding recovery as those in the experimental group (Tuteur 1958; Moncrieff and Wessely 1998; Moncrieff 2003; Moncrieff et al. 2004; Edward et al. 2005; Moncrieff and Kirsch 2005). As a result, any observed difference between the effects of the experimental treatment and the control treatment could be confounded by different expectations. In a related study, Kemp and colleagues (2008) found a correlation between the apparent characteristic effects of schizophrenia drugs and the "strength" of the side effects (4). We would not generally want to assert that a PCT that was confounded by different expectations possessed assay sensitivity. This problem will, of course, be more pronounced where the outcomes are subjective.

Moreover, if a participant is aware that she is taking the 'placebo,' she might drop out of the trial, or conversely, seek 'real' treatment outside the trial. All of these factors could confound the study. Double blinding is far more difficult to achieve in reality than is often assumed. Indeed, two studies

suggest that participants usually guess whether they are taking the experimental or control treatment (Fergusson et al. 2004; Hróbjartsson et al. 2007).

The use of 'active' placebos also highlights the connection between ethics and methodology. Because PCTs employing 'active' placebos are methodologically superior (they help the trial to remain successfully double blind), the ethical clinical researcher will be compelled to recommend them. However, employing 'active placebo' controls raises moral problems. If the 'placebo' control is made 'active' at the expense of including undesirable side effects, then participants will be exposed to harm. Even proponents of PCTs where standard treatment exists such as Miller and Brody (2002), recognize that participants should not be exposed to excessive risks or burdens for the sake of a scientific inquiry. Of course, the side effects induced by the 'active placebo' might not be sufficiently harmful to accuse the study of exposing participants to excessive risks or burdens. In other cases, however, the use of 'active placebos' could well render PCTs unethical even on Miller and Brody's view.

Besides the problem with illegitimate 'placebo' controls, there is another reason why PCTs do not possess assay sensitivity (of the first kind): the effects of 'placebo' controls vary widely. Moerman (1983) found that while the effect of cimetidine for ulcers remained relatively constant across trials, the effectiveness of the 'placebos' in the same trials ranged from 10% to 90% of the drug effect. An update of the review yielded more dramatic results, with the 'placebo' effect ranging from between 0% and 100% of the (again, roughly constant) drug effect (Moerman 2000). Another neuro- gastroenterological study of treatments for irritable bowel syndrome (IBS) found that the 'placebo' response ranged from 16.0% to71.4% of the (roughly constant) experimental treatment (Patel et al. 2005). Although the reason for the substantial variation in 'placebo' effects could be that the placebos in question were illegitimate, placebos could also have inherently variable effectiveness.

The variable 'placebo' effects in trials with roughly constant experimental treatment effect present an assay sensitivity problem for PCTs. Are cimetidine and the IBS treatments effective? The answer depends on which response to 'placebo' we assume to be 'correct'. But this places PCTs on the same footing as ACTs as far as assay sensitivity (of the first kind) is concerned. In order to claim that ACTs possess assay sensitivity we must assume that the established treatment control was effective; in order to claim that PCTs to possess assay sensitivity we must assume that a particular 'placebo' effect is 'correct' (and legitimate).

We might object that it is unreasonable to generalize about the variability of 'placebo' response rate from a handful of studies. This objection can only be answered by further empirical research. Given the vital role of 'placebo' controls in determining whether a treatment is deemed effective, it would be useful for methodologists to investigate the extent to which the effectiveness of 'placebo' controls varies. Furthermore, it suffices for some PCTs to lack assay sensitivity in order for them to be as assay insensitive as ACTs. As I shall now argue, most ACTs possess assay sensitivity.

The Limited Scope of the First Assay Sensitivity Argument

ACTs will lack assay sensitivity when we have good reason to doubt the effectiveness of the established treatment control. In the early 1990s Smith (1991) suggested that 80% to 90% of our treatments lacked a sound evidence base. However, these estimations may have equated 'evidence' with 'randomized evidence', which is a mistake. Many treatments ranging from the Heimlich maneuver to tracheostomies, are undoubtedly effective yet they have not been tested in 'placebo'-controlled randomized trials. More recent research indicates that between 76% (pediatric surgery) and 96% (anesthesia) of our current practice is based on compelling (randomized or non-randomized) evidence (Ellis et al. 1995; Gill et al. 1996; Imrie and Ramey 2000). Since most of our current treatments are effective, most ACTs possess assay sensitivity.

Although Temple and Ellenberg (2000) acknowledge that the assay sensitivity argument is limited in scope, they might object that the problem is more severe than I have indicated. They argue (rather oddly) that a treatment can be undoubtedly effective yet fail to demonstrate this effectiveness reliably: "the effectiveness of [some] drugs that sometimes (or even often) fail to be proven superior to 'placebo' is not in doubt" (458). The alleged problem is neither with trial size (458) nor trial quality (459). Rather, the problem is supposedly that the trial failed to detect the effects of the experimental treatment for some unknown reason:

> In each case . . . the problem [of assay sensitivity] is not identifiable a priori by examining the study; it is recognized only by the observed failure of the trial to distinguish the drug and placebo treatments (459).

The reason they think we should accept that there are treatments with what they call "assay sensitivity problems" (458) is that it would be unlikely for an ineffective treatment to demonstrate positive effects in even 50% of trials:

> even if a drug is statistically significantly superior to placebo in only 50% of well-designed and well-conducted studies, that proportion will still be vastly greater than the small fraction that would be expected to occur by chance if the drugs were ineffective (458).

If Temple and Ellenberg (2000) are correct then the first assay sensitivity argument applies somewhat more widely than I have indicated. However, the scope of the argument would not be extended very much. Temple and Ellenberg mention 11 classes of treatments that have "assay sensitivity problems" (358), provide some evidence for only four of these classes, and only discuss selective serotonin reuptake inhibitors (SSRIs) in any detail. We could grant that there are treatments with "assay sensitivity problems" and still assert that the first assay sensitivity argument was limited in scope.

A deeper problem, however, is that the idea of treatments with "assay sensitivity problems" is hard to swallow. The way we determine that a treatment has an effect is by testing it in well-controlled trials. If trials fail to detect effects, then there is good reason to doubt whether the effect exists. To blame a trial for not detecting the effects begs the question.

The probabilistic argument that an ineffective treatment is unlikely to demonstrate effects in 50% of trials is unacceptable for two reasons. First, it ignores the possibility of systematic bias. Publication bias (Dickersin 1990), funding source bias (Davidson 1986), and bias introduced by under-diagnosed methodological problems such as failure to keep a trial successfully masked (Fergusson et al. 2004; Hróbjartsson et al. 2007) will tend to exaggerate the size of the apparent experimental treatment effect. In cases where the absolute effect size of the treatment is small (as is the case for SSRIs), these hidden biases could well reduce the number of 'positive' studies to well within what we would expect by chance alone. Furthermore, a treatment that is effective in 50% of trials could be could be harmful in the remaining 50%. Systematic reviews (which Temple and Ellenberg [2000] fail to discuss) would be a good way to confirm claims that treatments with "assay sensitivity problems" have overall positive effects. In fact systematic reviews of SSRIs are ambiguous. While the Cochrane Review concludes that there are small differences between active 'placebo' and SSRIs (Moncrieff et al. 2004) other systematic reviews failed to detect statistically significant benefits of SSRIs over 'placebo' (Kirsch and Sapirstein 1998; Kirsch and Moore 2002; Healy 2004, 2006; Kirsch et al. 2008). I am not pronouncing on the debate over the antidepressant effects of SSRIs, but rather disputing the claim that SSRIs have undoubted effects.

The real problem with drugs with "assay sensitivity problems" could, of course, be that the effects of these drugs are so small that many studies fail to detect their effects. In this case another problem arises. Although small effects can sometimes be clinically relevant (if, say, they reduce mortality), this is not always the case. If "assay sensitivity problems" turn out to be problems with the size of the effect, we might have to admit that these effects are not clinically relevant.

Most importantly, the problem with control treatments that might have "assay sensitivity problems" presents a worry for PCTs as much as ACTs. Imagine a new SSRI was developed that, like other SSRIs, had "assay sensitivity problems." If the first few trials in which new SSRI were tested did not possess 'assay sensitivity' (and failed to demonstrate superiority to 'placebo'), it would be dropped before its supposedly undoubted effectiveness was detected. Thus, the potential existence of drugs with "assay sensitivity problems" does not provide us with any reason to prefer PCTs.

The scope if the first assay sensitivity argument is further limited if we require that new agents demonstrate superiority to the control treatments. Even if an established treatment is be ineffective (but is not worse than placebo) or has "assay sensitivity problems," if it demonstrates superiority to its predecessor, we can conclude that the new agent is more effective than placebo. In practice an increasing number of ACTs are conducted as non-inferiority trials.

Briefly (more will follow), a 'non-inferiority' trial is designed to detect whether the experimental intervention is at least of equal (to within some 'margin of equivalence') effectiveness to the control treatment (Figure 2). 'Superiority' trials, on the other hand, are designed to detect whether the experimental treatment is more effective than the control. Provided that the control treatment is not worse than 'placebo,' if a new treatment demonstrates superiority to the control, we can conclude that the new treatment is effective. Hence, superiority ACTs can possess assay sensitivity even if the established treatment is not effective. Of course superiority ACTs could suffer from assay sensitivity problems if, say, the experimental treatment was only slightly better than a control treatment which itself was much less effective than a placebo. However it is unlikely that many of our existing therapies are positively harmful. Hence, if we require that ACTs be conducted as superiority trials, then precious few will be at risk from assay insensitivity. Requiring the new treatment to be superior to a possibly ineffective treatment has an additional important advantage: it tells the patient, doctor, and policy maker what they need to know before deciding to take the new treatment.

I will now argue that ACTs should generally be conducted as superiority rather than non-inferiority trials.

Questioning the Rationale for Non-Inferiority Trials

Non-inferiority trials are justified in some cases. For example, the Extracranial/Intracranial Anastomosis (EC/IC) Bypass Study Group contrasted no treatment with surgery (superficial temporal artery-middle cerebral artery anastomosis) for patients with a high risk of stroke (1985). A non-inferiority test revealed that doing nothing was not worse (to within 3%) than surgery. In this case, and a few others (Sackett et al. 1974), non-inferiority tests have led to the rejection of relatively risky, invasive, and expensive procedures. More recently, however, non-inferiority trials have justified the adoption of many treatments that are no better than our existing treatments, and usually far more expensive (Morgan et al. 2005).

The justification for non-inferiority trials is that treatments can allegedly represent a real advance without offering superiority on the primary outcome. More specifically (Senn 2005; Piaggio et al. 2006):

1) The new treatment might have fewer side effects.
2) The new treatment could be cheaper or less invasive.
3) the new treatment may be necessary in case people develop resistance to existing therapies.

Although the rationale for non-inferiority trials appears sensible, I will argue that non-inferiority trials rarely help us discover whether a new treatment has one of these advantages.

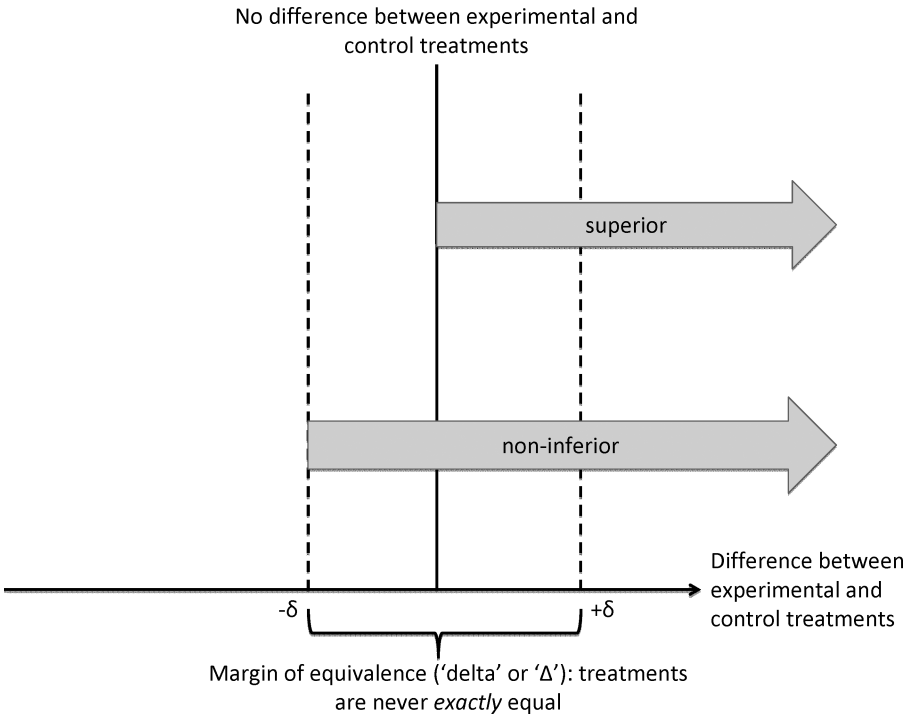

Figure 2. The difference between superiority and non-inferiority.

Comparisons of side effects are often made carelessly. Existing treatments have usually been around for longer, so there will be more extensive data about their side effects. Certainly rare and long-term side effects of the new treatment will be relatively under-studied. Thus comparisons between the side effects of newer and older treatments are often unbalanced. In addition, if the new treatment has a better side effect profile, then we should conduct a superiority test of the relevant side effects. It is, of course, possible to run a superiority test for the side effects of interest and a non-inferiority test for the main outcome simultaneously.

Then, if the new treatment is supposed to be more tolerable because it is less invasive or more convenient—say it involves one daily dose instead of two—then the benefits of the new regimen should result in a superior outcome (Garattini and Bertele 2007). For instance, we would expect participants taking one dose per day to adhere better to the regime. The superior adherence should translate to better outcomes. If not, then it is unclear whether the apparent improved convenience is of any value. At least in principle, apparently less convenient or more invasive regimes could improve the primary outcome, perhaps by enhancing the 'placebo' response.

Next, even if we allow some non-inferior treatments in case people develop resistance to our existing therapies or an unexpected side effect is discovered, it does not follow that we need dozens of similar therapies. Yet, dozens of roughly equivalent treatments is just what indiscriminate use of non-inferiority trials encourages. For instance, there are currently more than six SSRI antidepressants, and numerous other pharmaceutical antidepressants (tricyclic agents, monoamine oxidase inhibitors (MAOIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), noradrenergic and specific serotonergic antidepressants (NASSAs), norepinephrine (noradrenaline) reuptake inhibitors (NRIs), and norepinephrine-dopamine reuptake inhibitors). In addition there are many non-pharmaceutical treatments used to treat depression, including St. John's wort, cognitive behavioral therapy (CBT), exercise, and self-help. None of these treatments have demonstrated consistent superiority to others in trials, although the administration of some (e.g. exercise) is admittedly very different from others. Even if it were useful to have a few of these treatments available in case one of them suddenly turned out to be harmful or because patients somehow developed resistance, it is difficult to justify so many.

Finally, non-inferiority trials present an ethical problem for the clinician. If the experimental treatment is at best roughly equal, but could be worse, then the best available therapy is the existing one. It is unclear whether the ethical clinician should allow her patient to risk receiving an inferior treatment.

In brief, non-inferiority trials cannot be deemed worthwhile without special justification. Accordingly, institutional review boards (IRBs) should investigate requests to approve non-inferiority trials more carefully. Market constraints might make it difficult in practice to officially restrict the number of non-inferior treatments, but it does not follow that non-inferiority trials are morally justified.

To recap what has been argued thus far, PCTs suffer from assay sensitivity problems much like ACTs because 'placebo' controls can be illegitimate, and their effectiveness varies widely. In addition, the scope of the first assay sensitivity argument is limited to non-inferiority ACTs where we have reason to doubt the effects of the control treatment.

Before considering the second assay sensitivity argument, I will say a few words about three-armed trials that include experimental, existing treatment, and 'placebo' groups. One might think that three-armed trials can be used to compare the experimental treatment with an existing therapy and test whether the trial possessed assay sensitivity. However, the three-armed solution remains problematic for the clinician who seems morally compelled to use the best available treatment. Besides, the three-armed solution is only an improvement over regular ACTs if we believe that adding the 'placebo' group will make the trial assay sensitive, which the above discussion suggests is not the case.

The Second 'Assay Sensitivity' Argument

Recall that the term *assay sensitivity'* is also defined as the ability of a trial to distinguish between more effective and less effective (placebo or not) treatments. The second assay sensitivity argument is then that PCTs but not ACTs are able to detect differences between experimental and control treatments. Since the structure of superiority ACTs and superiority PCTs are identical as far as detecting differences are concerned, this argument applies exclusively to non-inferiority ACTs. Hence, the second assay sensitivity argument is thus also limited in scope because it only applies to (often unjustified) non-inferiority ACTs.

I will contend that although the purpose of non-inferiority trials is not to detect differences, they are as capable as superiority trials to detect differences. Although some discussion of classical statistics is necessary to understand why I believe this argument fails, I will keep the current discussion intuitive and as non-technical as possible. The reader is referred to the Appendix and other sources for a more detailed treatment (Dunnett and Gent 1977; Blackwelder 1982; Hwang and Morikawa 1999; Temple and Ellenberg 2000; Armitage et al. 2002 636–639; Gomberg-Maitland et al. 2003; Piaggio et al. 2006; Senn 2007).

Both superiority and non-inferiority can be tested using confidence intervals. A confidence interval represents a range within which the mean difference between experimental and control treatments is likely to lie (Figure 3). If the entire confidence interval lies to the right side of the line representing no difference (the solid vertical line in Figure 3), then we can conclude (in a probabilistic sense) that the experimental treatment is superior. If the entire confidence interval lies to the right of the lower bound of the equivalence margin (-∂ in Figure 3) then we can conclude (again in a probabilistic sense) non-inferiority. It is true that a positive result of a non-inferiority trial does not provide evidence of a difference—a treatment could be non-inferior and also

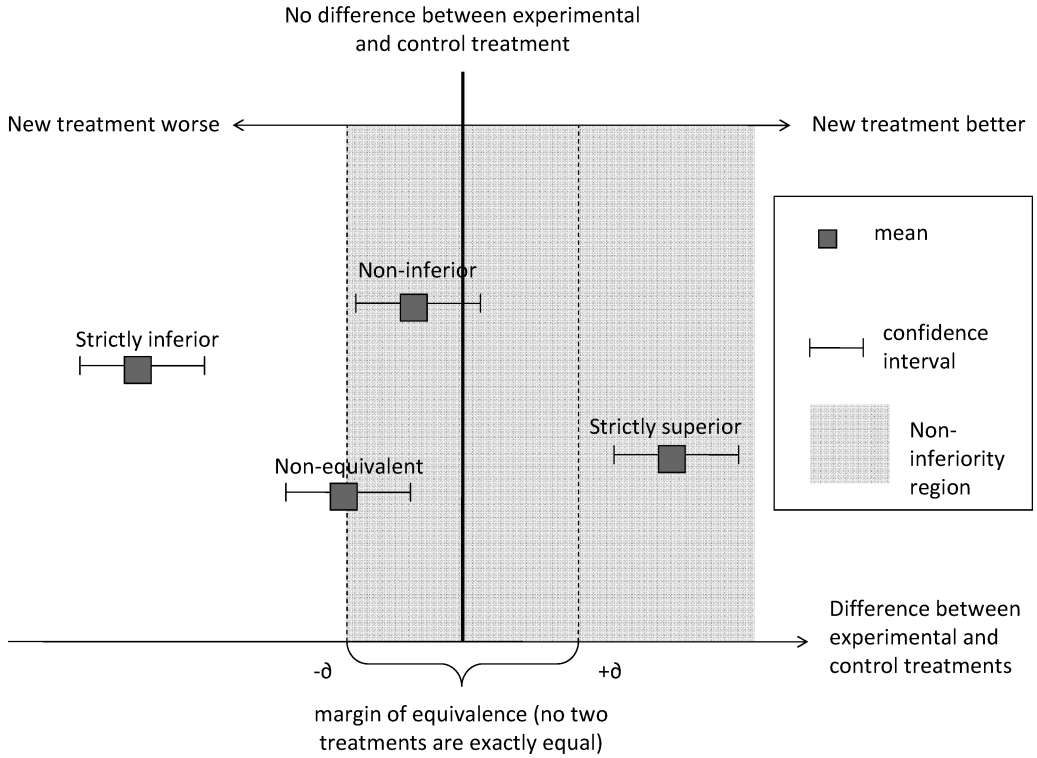

Figure 3. The difference between superiority and non-inferiority trials using confidence intervals.

roughly equal (the confidence interval could lie within the 'equivalence margin' bound by $-\partial$ and $+\partial$ in Figure 3). It is also true that the purpose of a non-inferiority trial is not to detect a difference—we would be happy if the experimental treatment were roughly equal. It does not follow, however, that a non-inferiority trial cannot detect a difference. Even if the purpose of the trial is to detect non-inferiority, if we find that the confidence interval lies entirely to the either side of the line representing no difference, then the trial will have provided evidence that a treatment difference exists.

One might object that the second assay sensitivity argument against ACTs is only apparent if we consider the null and alternative hypotheses used in superiority and non-inferiority trials. Since I have used confidence intervals without referring to the null and alternative hypotheses, the objection proceeds, I have concealed the problem. Because a full response to this objection is somewhat more technical I treat it in the appendix. However, given that the confidence interval and hypothesis test methods are equivalent (Armitage et al. 2002, 636–639), it would certainly be surprising if it turned out that consideration of the null and alternative hypotheses revealed that non-inferiority trials are incapable of detecting differences. Thus, the second assay sensitivity argument, like the first, fails as a justification for the claim that PCTs are methodologically superior to ACTs.

I will now examine the argument that PCTs (but not ACTs) provide a measure of absolute effect size.

CHALLENGING THE VIEW THAT 'PLACEBO'-CONTROLLED TRIALS PROVIDE A MEASURE OF ABSOLUTE EFFECT SIZE

It is often alleged that PCTs are superior to ACTs on the grounds that only the former provide a measure of absolute effect size:

> The placebo-controlled trial measures the total pharmacologically mediated effect of treatment. In contrast an active controlled trial . . . measures the effect relative to another treatment . . . The absolute effect size information is valuable (ICH 2000, 18, emphasis added).

I will argue that two unwarranted assumptions must be made in order to assert that PCTs provide a measure of absolute effect size. The first, 'additivity', is that the 'placebo' and nonplacebo components of a treatment add (like vectors) rather than interact (like compounds in a chemical reaction). The second assumption is that 'placebo' controls are legitimate.

Questioning the Assumption of Additivity

'Additivity' is the assumption that the various treatment factors combine like vectors rather than chemicals in a reaction. For example, we can dissect a force propelling a billiard ball in the northeasterly direction into its northward and eastward components. Since the component forces act independently we can deduce the resultant force if we know the magnitude and direction of its components (Mill 1843 [1973], I.v.1). If additivity held in PCTs, then if we knew the combined effect of the characteristic and non-characteristic features (measured in the experimental group), and we also knew the effect of the non-characteristic features (measured in the 'placebo' group), then we could deduce the 'absolute' effect of the characteristic features (Figure 1).

But it is unclear why we should assume that the characteristic and non-characteristic treatment features add rather than interact. Certainly additivity does not usually apply to the combination of chemical, biological, or even non-mechanical physical causes. For example, the combination of hydrogen and oxygen produces water, which does not retain the properties of either hydrogen or oxygen (Mill [1843] 1973 I.v.1). In fact there is some evidence to support the view that 'placebo' and non-placebo components of a treatment process interact rather than add.

Evidence that Additivity Does Not Generally Hold

Evidence for interactions between characteristic and non-characteristic features is sparse. Yet the paucity of evidence must not be taken as evidence that interactions are rare. Aside from a few interesting articles (Kleijnen et al. 1994; Kirsch 2000; Kaptchuk 2001), the assumption of additivity has been largely ignored. It would be helpful for methodologists to investigate the issue further. Until the assumption has been examined more carefully, it is difficult to draw conclusions about the prevalence of interactions. With that in mind, the modest intent of this section is to show that additivity cannot be taken for granted.

In the studies cited earlier about the variable effects of placebos for cimetidine and IBS-therapy placebos, the characteristic features did not add to the non-characteristic features. Rather, they interacted in such a way that the characteristic benefit tapered off as the strength of the 'placebo' (non-characteristic) features increased. If the characteristic and non-characteristic components were additive, changing the non-characteristic features would not affect the characteristic features (changing the magnitude of the eastward force on a billiard ball will not affect its northward motion).

In other cases, the magnitude of the non-characteristic features has been manipulated experimentally and the resulting characteristic effectiveness changed. For instance, Hughes and colleagues (1989) investigated the effects of nicotine gum for smoking cessation. The trial involved 77 participants who provided their informed consent to have a 50/50 chance of receiving nicotine or 'placebo' gum, and that they might or might not be told the contents of their gum. They were not told that they could be deceived. (It is unlikely that IRBs would approve such a trial today.) The participants were then assigned one of six groups (Table 1). Two groups (groups 1 and 2) were told they would receive 'placebo' (and thus had low expectations regarding recovery), two groups (groups 3 and 4) were told they would receive real nicotine gum (and thus had higher expectations regarding recovery), while the remaining two

Table 1. The extended balanced 'placebo' design used in the Hughes et al. (1989) study of the effects of nicotine gum

	Received	
	Treatment	No treatment
TOLD		
Treatment	1	2
No treatment	3	4
Neither (treatment delivered in double blind conditions)	5	6

groups (groups 5 and 6) were delivered either 'placebo' or nicotine gum under 'double blind' conditions (and thus had 'medium' expectations regarding recovery). Only one group from each pair (groups 1, 3, 5) was actually given nicotine gum; the other group was given placebo.

All participants, especially those in the two groups were told they were given a 'placebo' were encouraged to use the gum whenever the urge to smoke occurred.

The outcome measures were proportion of participants who smoked no cigarettes during the week, proportion who smoked on fewer than 2 days per week, number of days smoked per week, and number of cigarettes smoked per week. These were calculated based on assessments measured 1 and 2 weeks after they attempted to quit, and were measured in three ways. First, the participants self-reported how many cigarettes they smoked. Second, a designated observer (usually a spouse) reported the participant's smoking habits during the week. Third, a breath sample of carbon monoxide was taken to verify claims of complete abstinence.

The characteristic benefit of 'real' nicotine gum was not a constant that added to the benefit of the varying non-characteristic features. Rather, the characteristic effects tapered off as the strength of expectations increased (Figure 4). The statistical evidence for interactions between instructions and the overall outcome was significant ($P = 0.01$). Several other studies employing similar designs also indicate that additivity does not hold. Analgesic response (Levine and Gordon 1984), amphetamine effects (Mitchell et al. 1996), and subjective feelings of intoxication (Ross et al. 1962) have all been shown to decrease as the strength of expectations increases. One study even indicates that the increased 'strength' of non-characteristic features of treatment with naloxone can change the apparent benefit of the characteristic features from positive to negative (Levine and Gordon 1984).

It is also possible for non-characteristic and characteristic features to combine synergistically. For example, Freud argued that charging a hefty fee might act as a catalyst for what are commonly thought of as the characteristic features of Freudian psychoanalysis (Grünbaum 1986, 24; Freud et al. [1913] 2001, volume 12).

We need not resort to mysterious explanations to account for interactions. Many ailments can only be relieved by so much. If the expectations alone produce this maximum effect (or something close to it), then there will be little room for drug-induced improvement. Once a headache is completely relieved, taking another pill will not relieve it further. The mechanism that accounts for a maximum drug response is often understood (Aronson 2007). To oversimplify, the maximum response is related to the

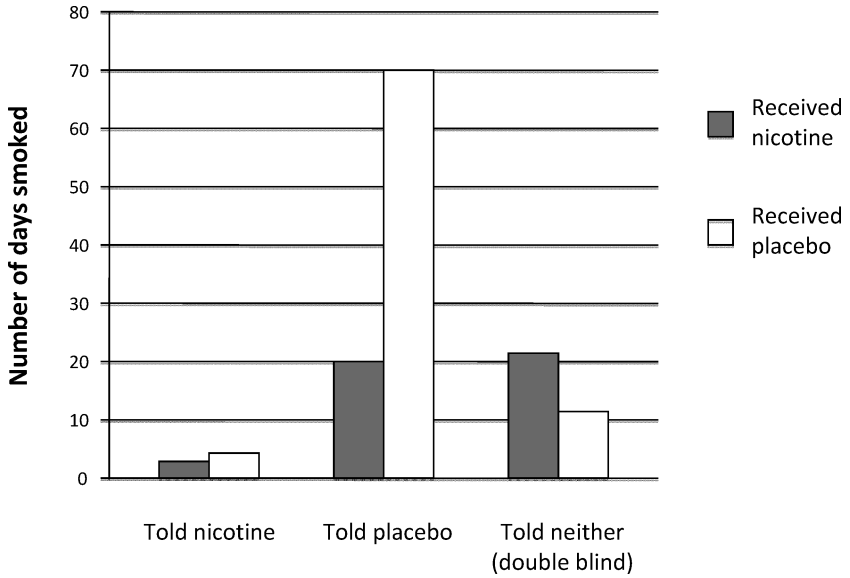

Figure 4. Smoking behavior by instruction and drug group. The results are cumulative across the two weeks where assessments were made. Based on Hughes and colleagues (1989).

maximum number of receptors cells have for the drug to attach-to. Once the receptors are all occupied, there is no further room for improvement. If expectations and beliefs spur the body to produce an agent that occupies these receptors then the drug will not have its otherwise significant effect. Synergistic interactions can also be explained. There is evidence that placebos for pain increase the levels of endogenous opioids (Benedetti and Amanzio 1997; ter Riet et al. 1998). The increased opioid level could stimulate interaction with the characteristic features to increase the effects by interacting synergistically with the active treatment.

These examples suffice to conclude that additivity cannot be taken for granted. Even if additivity held all the time, however, another assumption must be made to support the claim that PCTs provide a measure of absolute effect size: the performance of the 'placebo' controls must be legitimate and perform consistently. But we saw above that 'placebo' controls are not always legitimate. Likewise, the Moerman (2000) and Patel (2005) studies indicate that the wide variability of 'placebo' control treatments (legitimate or not) can determine how effective an experimental treatment appears effective. It would be a strange definition of 'absolute effect size' indeed that was compatible with the effect size changing drastically from study to study.

In brief, the assumptions (a) that the characteristic and non-characteristic treatment features add rather than interact, and (b) that the 'placebo' controls are legitimate can rarely, if ever, be jointly made. As a result, the claim that PCTs provide an absolute measure of effect size cannot be maintained.

QUESTIONING THE CLAIM THAT 'PLACEBO'-CONTROLLED TRIALS REQUIRE SMALLER SAMPLE SIZES

Some claim that PCTs are advantageous because they require a smaller sample size than ACTs (ICH 2000, section 2.4). It is an ethical requirement to use the smallest possible sample size since smaller trials are cheaper and expose fewer participants to risk. However, it is unclear that the sample size issue in particular and practical considerations in general weigh in on the side of PCTs.

Firstly, only non-inferiority and not superiority ACTs allegedly require a larger sample size than PCTs. Then, the supposed reason why PCTs require a smaller sample size is that the equivalence margin (Figure 2) is often much smaller than the treatment difference that a PCT is designed to detect. Yet if a PCT is designed to detect a difference that is the same size as the equivalence margin, then it will require an equally large sample size.

Besides, there are several practical considerations that reduce the force of the claim that PCTs are preferable because they require a smaller sample size even if it were true. For one, potential participants may be more likely to consent to a trial where they are certain to receive an 'active' treatment than they are if they might get a 'placebo'. Similarly, PCTs might face a more acute threat from the fact that participants seem to be quite good at detecting which group they are in despite efforts to keep the trial blind (Fergusson et al. 2004; Hróbjartsson et al. 2007). A participant in a PCT who guesses she is taking the 'placebo' might well drop out or covertly seek treatment outside the trial. In another regard, a participant in an ACT who guesses they are in the control group are already (supposedly) taking the best available treatment, and will have less incentive to drop out or seek outside treatment.

Next, a further study is required in order to apply the results of a PCT. In order to make an informed choice about whether to use the new treatment, the patient, practitioner, or policy maker must know how the new treatment compares with the best existing treatments not merely how it compares with placebo. This information would have to be obtained from an additional ACT, or an indirect comparative study. The human and financial burden of the additional study would have to be added to the cost of the PCT before asserting that PCTs are preferable because they require fewer participants than ACTs. In reality, of course, the further studies are rarely done. If not, then we must take into account the risk of doing harm, or allocating scarce resources to an inferior treatment when assessing the relative practical benefits of PCTs.

Lastly, what the average patient, practitioner, and policy maker needs to know in order to decide to use a new treatment is how it compares with the best existing treatment, not how it compares with 'placebo.' In short, the alleged practical advantages of PCTs have been exaggerated, while the practical disadvantages of PCTs have been overlooked.

CONCLUSION: A RE-ASSESSMENT OF THE RELATIVE METHODOLOGICAL QUALITY OF 'PLACEBO'-CONTROLLED TRIALS

Even if taken on their own terms, both assay sensitivity arguments are strictly limited in scope to non-inferiority trials where the control treatment is ineffective or has "assay sensitivity problems." Moreover, neither assay sensitivity argument is acceptable. The first fails because PCTs also sometime lack assay sensitivity. The second is based on a conflation of the difference between the purpose and properties of non-inferiority trials. The argument that PCTs provide a measure of the absolute effect of the characteristic features of the experimental treatment relies on rarely warranted assumptions that 'placebo' controls are legitimate and that characteristic and non-characteristic treatment features add. Lastly, practical considerations, including sample size, often support ACTs rather than PCTs. Judged as arguments that ACTs are methodologically to PCTs, all three arguments must be judged as failures. Claims (Ellenberg and Temple 2000; ICH 2000; Temple and Ellenberg 2000; WMA 2001; Miller and Brody 2002) about the methodological superiority of PCTs over ACTs are therefore unjustified.

Consequently, the apparent tension between clinical and research ethics as far as the use of 'placebo' controls

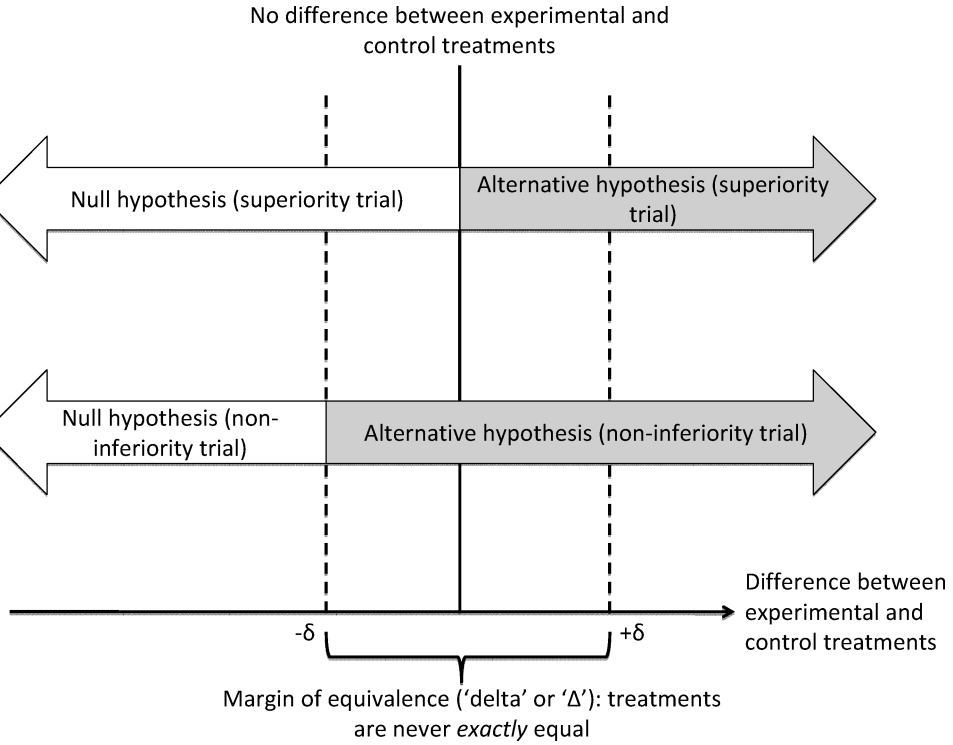

Figure 5. Illustration of null and alternative hypotheses in one-tailed superiority and one-tailed non-inferiority trials.

dissolves: methodological considerations do not support the use of PCTs where there is an established therapy. The ethical duty of the clinician to provide the best care (and avoid PCTs) where there is an established treatment available stands unchallenged by the moral duties of the clinical researcher to use the best method. The Declaration of Helsinki should retract the statement that PCTs can be justified on methodological grounds (where there is an available established therapy) and IRBs should dismiss claims that PCTs are methodologically superior to ACTs as grounds to approve PCTs where standard therapy is available.

APPENDIX: MORE DETAILED EXPLANATION OF WHY THE SECOND ASSAY SENSITIVITY ARGUMENT FAILS

In classical hypothesis testing, hypotheses are not confirmed, but they can be 'rejected' (in a probabilistic sense). We therefore attempt to reject the 'null hypothesis' that represents the opposite of what we would like to establish. If we succeed, this is generally taken to support the alternative hypothesis. In a one-tailed superiority PCT the null hypothesis is that there is no difference between experimental treatment and placebo, or that the experimental treatment is less effective than placebo. The alternative hypothesis will be that the experimental treatment is superior to the 'placebo' (Figure 5).

A non-inferiority trial is designed to determine whether the experimental treatment is at least (roughly) equal in effectiveness to the control treatment. Therefore, we seek to rule out the null hypothesis that the experimental treatment was less effective by at least some minimum amount than the control treatment. The alternative hypothesis in a non-inferiority trial is that the experimental treatment was of equal or greater effectiveness (Figure 5).

Whether a trial is good at detecting differences depends on the degree of risk of Type I and Type II errors. A Type I error, or 'false positive', is the error of rejecting a true null hypothesis (and accepting a false alternative hypothesis). A Type II error, or 'false negative', is the mistake of failing to reject a false null hypothesis (and not accepting a true alternative hypothesis). In a superiority trial, a Type I error is the mistake of accepting a positive difference when there is none, while a Type II error is the mistake of accepting no difference or inferiority when the experimental treatment is superior. A Type I error in a non-inferiority trial is the mistake of accepting rough equality or superiority when the experimental treatment is strictly inferior, while a Type II error is the mistake of accepting strict inferiority when the experimental treatment is roughly equal or superior.

Both Type I and Type II errors can be controlled for and specified in advance of a trial. To sensitize a superiority trial to differences, we reduce the Type I error rate. To sensitize a non-inferiority trial to differences, we must reduce the Type II error rate (see Appendix).

With this in mind, it is straightforward to show that non-inferiority trials are as good (or bad) at detecting differences

as superiority trials. The following discussion follows Anderson (2006, 75). Anderson postulates four conditions required to assume assay sensitivity. Using 'D' to denote "difference between intervention and control group", and T to denote "trial", the conditions for asserting assay sensitivity (of the second kind) are,

1) D
2) T indicates D
3) D → (T indicates D)
4) not-D → not (T indicates D)

The first condition tells us that, ontologically speaking, there is a difference between the interventions being compared. The second tells us that the trial indicated a difference. The third tells us that the trial would indicate a difference if there were one. The fourth tells us that if there is no difference, then the trial will not indicate a difference. In the real world, of course, the modality of the conditionals in (3) and (4) is not necessity—actual trials deal in probability.

The four conditions for assay sensitivity will be satisfied in a superiority trial when:

1) There is a difference (the experimental treatment is superior to placebo).
2) The trial indicates a difference (there is a 'positive' result).
3) The Type II error rate is sufficiently low (the trial did not wrongly suggest no difference).
4) The Type I error rate is sufficiently low (the trial did not wrongly indicate a difference).

In a non-inferiority trial, the four conditions for affirming assay sensitivity will be satisfied when:

1) There is a difference (the experimental treatment is superior to placebo).
2) The trial indicates a difference (there is a 'positive' result).
3) The Type I error rate is sufficiently low.
4) The Type II error rate is sufficiently low.

As along as we are able to reduce both the Type I and Type II error rates sufficiently, both superiority and non-inferiority trials can be made equally assay sensitive. Anderson concludes, and he is surely correct, that, "Contrary to the assay sensitivity argument, there is not an absolute difference between PCTs and ACTs with respect to . . . the assay sensitivity assumption" (Anderson 2006, 78). ■

REFERENCES

Altman, D. G. 1980. Statistics and ethics in medical research. Misuse of statistics is unethical. *British Medical Journal* 281(6249): 1182–1184.

Anderson, J. A. 2006. The ethics and science of placebo–controlled trials: Assay sensitivity and the Duhem–Quine thesis. *Journal of Medicine and Philosophy* 31: 65–81.

Antihypertensive and Lipid–Lowering Heart Attack Trial (ALLHAT) Collaborative Research Group. 2000. Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: The antihypertensive and lipid–lowering treatment to prevent heart attack trial (ALLHAT). *Journal of American Medical Association* 283(15): 1967–1975.

Armitage, P., Berry, G., and Matthews, J. N. S. 2002. *Statistical Methods in Medical Research*. Oxford, UK: Blackwell Science.

Aronson, J. K. 2007. Concentration–effect and dose–response relations in clinical pharmacology. *British Journal of Clinical Pharmacology* 63(3): 255–257.

Ashcroft, R., and ter Meulen, R. 2004. Ethics, philosophy, and evidence based medicine. *Journal of Medical Ethics* 30(2): 119.

Benedetti, F., and Amanzio, M. 1997. The neurobiology of placebo analgesia: From endogenous opioids to cholecystokinin. *Progress in Neurobiology* 52(2): 109–125.

Berguer, A., Kovacs, F., Abraira, V., Mufraggi, N., Royuela, A., Muriel, A., et al. 2008. Neuro–reflexotherapy for the management of myofascial temporomandibular joint pain: A double–blind, placebo–controlled, randomized clinical trial. *Journal of Oral Maxillofacial Surgery* 66(8): 1664–1677.

Bingel, U., Lorenz, J., Schoell, E., Weiller, C., and Buchel, C. 2006. Mechanisms of placebo analgesia: RACC recruitment of a subcortical antinociceptive network. *Pain* 120(1–2): 8–15.

Blackwelder, W. C. 1982. Proving the null hypothesis in clinical trials. *Controlled Clinical Trials* 3: 345–353.

Boutron, I., Moher, D., Tugwell, P., Giraudeau, B., Poiraudeau, S., Nizard, R., et al. 2005. A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *Journal Clinical Epidemiology* 58(12): 1233–1240.

CIOMS. 1993. *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Geneva: World Health Organization.

Craggs, J. G., Price, D. D., Verne, G. N., Perlstein, W. M., and Robinson, M. M. 2007. Functional brain interactions that serve cognitive–affective processing during pain and placebo analgesia. *Neuroimage* 38(4): 720–729.

Davidson, R. A. 1986. Source of funding and outcome of clinical trials. *Journal General Internal Medicine* 1(3): 155–158.

Dickersin, K. 1990. The existence of publication bias and risk factors for its occurrence. *Journal of American Medical Association* 263(10): 1385–1389.

Dunn, A. L., Trivedi, M. H., Kampert, J. B., Clark, C. G., and Chambliss, H. O. 2005. Exercise treatment for depression: Efficacy and dose response. *American Journal of Preventive Medicine* 28(1): 1–8.

Dunnett, C. W., and Gent, M. 1977. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics* 33(4): 593–602.

Dwyer, T., and Ponsonby, A. L. 1996. Sudden infant death syndrome: After the back to sleep campaign. *BioMedical Journal* 313(7051): 180–181.

Ebell, M. H., Siwek, J., Weiss, B. D., Woolf, S. H., Susman, J., Ewigman, B., et al. 2004. Strength of recommendation taxonomy (SORT): A patient–centered approach to grading evidence in the medical literature. *American Family Physician* 69(3): 548–556.

Echt, D. S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias–Manno, D., Barker, A. H., et al. 1991. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. *New England Journal of Medicine* 324(12): 781–788.

Edward, S. J., Stevens, A. J., Braunholtz, D. A., Lilford, R. J., and Swift, T. 2005. The ethics of placebo–controlled trials: A comparison of inert and active placebo controls. *World Journal of Surgery* 29(5): 610–614.

Ellenberg, S. S., and Temple, R. 2000. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 2: Practical issues and specific cases. *Annals of Internal Medicine* 133(6): 464–470.

Ellis, J., Mulligan, I., Rowe, J., and Sackett, D. L. 1995. Inpatient general medicine is evidence based. A–Team, Nuffield Department of Clinical Medicine. *Lancet* 346(8972): 407–410.

Emanuel, E. J., Wendler, D., and Grady, C. 2000. What makes clinical research ethical? *Journal of American Medical Association* 283(20): 2701–2711.

Evans, I., Thornton, H., and Chalmers, I. 2007. *Testing Treatments: Better Research for Better Healthcare.* London, UK: British Library.

Extracranial–Intracranial Bypass Study Group. 1985. Failure of extracranial–intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an international randomized trial. *New England Journal of Medicine* 313(19): 1191–1200.

Fergusson, D., Glass, K. C., Waring, D., and Shapiro, S. 2004. Turning a blind eye: The success of blinding reported in a random sample of randomized, placebo controlled trials. *BioMedical Journal* 328(7437): 432.

Freud, S., Strachey, J., Freud, A., and Richards, A. [1913] 2001. *The Standard Edition of the Complete Psychological Works of Sigmund Freud,* vol. 24, indexes and bibliographies. London, UK: Vintage.

Garattini, S. and Bertele, V. 2007. Non–inferiority trials are unethical because they disregard patients' interest. *Lancet* 370(9602): 1875–1877.

Gill, P., Dowell, A. C., Neal, R. D., Smith, N., Heywood, P., Wilson, A. E. 1996. Evidence based general practice: A retrospective study of interventions in one training practice. *BioMedical Journal* 312(7034): 819–821.

General Medical Council (GMC). 2006. Good medical practice. *The Duties of a Doctor Registered with the General Medical Council*. London, UK: GMC.

Golomb, B. A. 1995. Paradox of placebo effect. *Nature* 375(6532): 530.

Gomberg–Maitland, M., Frison, L., and Halperin, J. L. 2003. Active–control clinical trials to establish equivalence or noninferiority: Methodological and statistical concepts linked to quality. *American Heart Journal* 146(3): 398–403.

Grünbaum, A. 1986. The placebo concept in medicine and psychiatry. *Psychological Medicine* 16(1): 19–38.

Halpern, S. D., Karlawish, J. H., and Berlin, J. A. 2002. The continuing unethical conduct of underpowered clinical trials. *Journal of American Medical Association* 288(3): 358–362.

Hayes, M. A., Timmins, A. C., Yau, E. H., Palazzo, M., Hinds, C. J., and Watson, D. 1994. Elevation of systemic oxygen delivery in the treatment of critically ill patients. *New England Journal of Medicine* 330(24): 1717–1722.

Healy, D. 2004. *Let Them Eat Prozac.* New York, NY: New York University Press.

Healy, D. 2006. Did regulators fail over selective serotonin reuptake inhibitors? *BioMedical Journal* 333(7558): 92–95.

Herbert, P. C., Wells, G., and Blajchman, M. A. 1999. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion requirements in critical care investigators, Canadian Critical Care Trials Group. *New England Journal of Medicine* 340(6): 409–417.

Howick, J., Golomb, B., Dieppe, P., and Enkin, M. 2009. Unpublished data.

Hróbjartsson, A. 2002. What are the main methodological problems in the estimation of placebo effects? *Journal of Clinical Epidemiology* 55(5): 430–435.

Hróbjartsson, A., Forfang, E., Haahr, M. T., Als–Nielsen, B., and Brorson, S. 2007. Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology* 36(3): 654–663.

Hughes, J. R., Gulliver, S. B., Amori, G., Mireault, G. C., and Fenwick, J. F. 1989. Effect of instructions and nicotine on smoking cessation, withdrawal symptoms and self–administration of nicotine gum. *Psychopharmacology Berlin* 99(4): 486–491.

Hwang, I. K. and Morikawa, T. 1999. Design issues in noninferiority/equivalence trials. *Drug Information Journal* 33: 1205–1218.

International Conference on Harmonization (ICH). 2000. Harmonized tripartite guideline. Choice of control group and related issues in clinical trials. *International Conference on Harmonization*, E 10. San Diego, CA: Centre for Biologics Evaluation and Research.

Imrie, R., and Ramey, D. W. 2000. The evidence for evidence–based medicine. *Complement Therapeutic Medicine* 8(2): 123–126.

Kaptchuk, T. J. 2001. The double–blind, randomized, placebo–controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology* 54(6): 541–549.

Kemp, A. S., Schooler, N. R., Kalali, A. H., Alphs, L., Anand, R., Awad, G., et al. 2008. What is causing the reduced drug–placebo difference in recent schizophrenia clinical trials and what can be done about it? *Schizophrenia Bulletin*, e-pub a head of print; 1-6.

Kirsch, I. 2000. Are drug and placebo effects in depression additive? *Biological Psychiatry* 47(8): 733–735.

Kirsch, I., Deacon, B. J., Huedo–Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. 2008. Initial severity and antidepressant benefits: A meta–analysis of data submitted to the Food and Drug Administration. *PLoS Medicine* 5(2): e45.

Kirsch, I., and Moore, T. 2002. The Emperor's new drugs: An analysis of antidepressant medication data submitted to the U. S. Food and Drug Administration. *Prevention & Treatment*, 5.

Kirsch, I., and Sapirstein, G. 1998. Listening to Prozac but hearing placebo: A meta–analysis of antidepressant medication. *Prevention & Treatment,* I.

Kleijnen, J., de Craen, A. J., van Everdingen, J., and Krol, L. 1994. Placebo effect in double–blind clinical trials: A review of interactions with medications. *Lancet* 344(8933): 1347–1349.

Kong, J., Gollub, R. L., Polich, G., Kirsch, I., Laviolette, P., Vangel, M., et al. 2008. A functional magnetic resonance imaging study on the neural mechanisms of hyperalgesic nocebo effect. *Journal of Neuroscience* 28(49): 13354–13362.

Kong, J., Kaptchuk, T. J., Polich, G., Kirsch, I., and Gollub, R. L. 2007. Placebo analgesia: Findings from brain imaging studies and emerging hypotheses. *Reviews of Neuroscience* 18(3–4): 173–190.

Lasagna, L. 1955. The controlled clinical trial: Theory and practice. *Journal of Chronic Disease* 1(4): 353–367.

Lasagna, L. 1964. Hippocratic oath—Modern version. Available at: http://www.pbs.org/wgbh/nova/doctors/oath_modern.html (accessed February 9, 2009).

Levine, J. D. and Gordon, N. C. 1984. Influence of the method of drug administration on analgesic response. *Nature* 312(5996): 755–756.

McCann, I. L., and Holmes, D. S. 1984. Influence of aerobic exercise on depression. *Journal of Personality and Social Psychology* 46(5): 1142–1147.

Mill, J. S. [1843] 1973. *A System of Logic, Ratiocinative, and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation.* Toronto, Canada: University of Toronto Press.

Miller, F. G., and Brody, H. 2002. What makes placebo–controlled trials unethical? *American Journal of Bioethics* 2(2): 3–9.

Mitchell, S. H., Laurent, C. L., de Wit, H. 1996. Interaction of expectancy and the pharmacological effects of d–amphetamine: Subjective effects and self–administration. *Psychopharmacology Berlin* 125(4): 371–378.

Moerman, D. E. 1983. General medical effectiveness and human biology: Placebo effects in the treatment of ulcer disease. *Medical Anthropology Quarterly* 14(4): 3, 13–16.

Moerman, D. E. 2000. Cultural variations in the placebo effect: Ulcers, anxiety, and blood pressure. *Medical Anthropology Quarterly* 14(1): 51–72.

Moncrieff, J. 2003. A comparison of antidepressant trials using active and inert placebos. *International Journal of Methods in Psychiatric Research* 12(3): 117–127.

Moncrieff, J., and Kirsch, I. 2005. Efficacy of antidepressants in adults. *BioMedical Journal* 331(7509): 155–157.

Moncrieff, J. and Wessely, S. 1998. Active placebos in antidepressant trials. *British Journal of Psychiatry* 173: 88.

Moncrieff, J., Wessely, S., and Hardy, R. 2004. Active placebos versus antidepressants for depression. *Cochrane Database Systems Review* (1): CD003012.

Morgan, S. G., Bassett, K. L., Wright, J. M., Evans, R. G., Barer, M. L., Caetano, P. A., et al. 2005. Breakthrough drugs and growth in expenditure on prescription drugs in Canada. *BioMedical Journal* 331(7520): 815–816.

Oken, B. S. 2008. Placebo effects: Clinical aspects and neurobiology. *Brain* 131(Pt 11): 2812–2823.

Patel, S. M., Stason, W. B., Legedza, A., Ock, S. M., Kaptchuk, T. J., Conboy, L., et al. 2005. The placebo effect in irritable bowel syndrome trials: A meta–analysis. *Neurogastroenterol Motility* 17(3): 332–340.

Petrovic, P., Kalso, E., Petersson, K. M., and Ingvar, M. 2002. Placebo and opioid analgesia–imaging a shared neuronal network. *Science* 295(5560): 1737–1740.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., and Evans, S. J. 2006. Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *Journal of American Medical Association* 295(10): 1152–1160.

Ross, D. F., Krugman, A. D., Lyerly, S. B., and Clyde, D. J. 1962. Drugs and placebos: A model design. *Psychological Reports* 10: 383–392.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., et al. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *Journal of American Medical Association* 288(3): 321–333.

Sackett, D. L., Spitzer, W. O., Gent, M., and Roberts, R. S. 1974. The Burlington randomized trial of the nurse practitioner: Health outcomes of patients. *Annals of Internal Medicine* 80(2): 137–142.

Senn, S. J. 2005. Active control equivalence studies. In *Encyclopaedic Companion to Medical Statistics,* eds. B. S. Everett, C. R. Palmer, and H. Arnold, 19–22.

Senn, S. J. 2007. *Statistical Issues in Drug Development.* Hoboken, NJ: Wiley.

Shapiro, A., and Shapiro, E. 1997. The placebo. Is it much ado about nothing? In *The Placebo Effect: An Interdisciplinary Exploration*, ed. A. Harrington.

Shapiro, A. and Morris, L. A. 1978. *The Placebo Effect in Medical and Psychological Therapies. Handbook of Psychotherapy and Behavioural Change: An Empirical Analysis.* S. L. Garfield and A. E. Bergin, Eds. New York, John Wiley & Sons, 369–410.

Smith, G. C., and Pell, J. P. 2003. Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomized controlled trials. *British Medical Journal* 327(7429): 1459–1461.

Smith, R. 1991. Where is the wisdom. . . ? *British Medical Journal* 303(6806): 798–799.

Takala, J., Roukonen, E., Webster, N. R. Nielsen, M. S., Zandstra, D. F., Vundelinckx, G., et al. 1999. Increased mortality associated with growth hormone treatment in critically ill adults. *New England Journal of Medicine* 341(11): 785–792.

Temple, R., and Ellenberg, S. 2000. Placebo-controlled trials and active-control trials in the evaluation of new treatments: Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 133(6): 455–463.

ter Riet, G., de Craen, A. J., de Boer, A., and Kessels, A. G. 1998. Is placebo analgesia mediated by endogenous opioids? A systematic review. *Pain* 76(3): 273–275.

Trivedi, M. H., Greer, T. L., Grannemann, B. D., Church, T. S., Galper, D. I., Sunderajan, P., et al. 2006. TREAD: Treatment with Exercise Augmentation for Depression: Study rationale and design. *Clinical Trials* 3(3): 291–305.

Tuteur, W. 1958. The double blind method: Its pitfalls and fallacies. *American Journal Psychiatry* 114(10): 921–922.

Wager, T. D., Rilling, J. K., Smith, E. E., Sokolik, A., Casey, K. L., Davidson, R. J., et al. 2004. Placebo–induced changes in FMRI in the anticipation and experience of pain. *Science* 303(5661): 1162–1167.

World Medical Association (WMA). 1949. *World Medical Association International Code of Medical Ethics. Policy, WMA*. Ferney-Voltaire: WMA.

World Medical Association (WMA). 2008, October. The Declaration of Helsinki. Geneva, Switzerland: World Medical Association. Available at www.wma.net/e/policy/63.htm (accessed July 7, 2009).

Wootton, D. 2006. *Bad Medicine: Doctors Doing Harm Since Hippocrates.* Oxford, UK: Oxford University Press.

Worrall, J. 2007. Evidence in medicine. *Compass,* 2(6): 981–1022.

Appendix. The difference between Type I and Type II errors for superiority (PCT) and non-inferiority (ACT) trials

	Evidence for difference	
	Superiority (PCT)	Non-inferiority (ACT)
Low Type I error rate (false positive)	Yes*	No
Low Type II error (false negative)	No	Yes

*"Yes" means that there is evidence of a difference; "No" means that there is no evidence for difference. For example, a superiority trial with a low Type I error rate provides good evidence for a difference.