# Comparative Effectiveness Clinical Trials in Psychiatry: Superiority, Noninferiority, and the Role of Active Comparators

*Andrew C. Leon, PhD*

**ABSTRACT**

The Agency for Healthcare Research and Quality, part of the US Department of Health and Human Services, has issued several Requests for Applications to conduct comparative effectiveness research (CER). Many of the applications will involve randomized controlled clinical trials that include an active comparator. The inclusion of an active comparator has implications for clinical trial design.

Despite a common misperception, a clinical trial result of *no significant difference* between active treatment groups does not imply equivalence or noninferiority. A noninferiority trial, on the other hand, can directly test whether one active treatment group is noninferior to the other. For example, noninferiority of an inexpensive generic could be tested in comparison with a novel, more costly intervention. Although seldom used in psychiatry, noninferiority clinical trials could play a fundamental role in CER. Features of noninferiority and the nearly ubiquitous superiority designs are contrasted. The noninferiority margin is defined and its application and interpretation are discussed.

Evidence of noninferiority can only come from well-designed and conducted noninferiority CER. Sample sizes needed in noninferiority trials and in superiority trials that include an active comparator are substantially larger than those needed in trials that can utilize a placebo control in their scientific design. As a result, trials with active comparators are more costly, require longer recruitment duration, and expose more participants to the risks of an experiment than do trials in which the only comparator is placebo.

*J Clin Psychiatry 2011;72(10):1344–1349*
*© Copyright 2011 Physicians Postgraduate Press, Inc.*

*See also Commentary on page 1350.*

***Submitted:*** *March 3, 2010; accepted May 10, 2010.*
***Online ahead of print:*** *February 8, 2011 (doi:10.4088/JCP.10m06089whi).*
***Corresponding author:*** *Andrew C. Leon, PhD, Weill Cornell Medical College, Departments of Psychiatry and Public Health, Box 140, 525 East 68th St, New York, NY 10065 (acleon@med.cornell.edu).*

The Agency for Healthcare Research and Quality, part of the US Department of Health and Human Services, issued several recent requests for applications (RFAs) that call for comparative effectiveness research (CER).[1] These RFAs were motivated primarily by the efforts for health care reform in the United States and funded by the American Recovery and Reinvestment Act of 2009. The RFAs adopt the Federal Coordinating Council for Comparative Effectiveness Research definition of CER, "the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in 'real-world' settings … [This] is not meant to exclude randomized trials; however, these trials would need comparator arms other than placebo and be representative of populations seen in 'real-world' practice."[2(pp5,16)] The definition conforms to those of both the Congressional Budget Office report on Research on Comparative Effectiveness of Medical Treatment[3] and the Institute of Medicine report on Initial National Priorities on Comparative Effectiveness Research.[4] Much of the research supported by this initiative will involve randomized controlled trials (RCTs), yet meta-analyses and observational studies will also be used.

The Agency for Healthcare Research and Quality effort is expected to broaden the research agenda, moving away from a predominant focus on efficacy trials to include more effectiveness trials for evaluation of therapeutics. Efficacy trials evaluate a treatment effect in a rarefied sample under ideal circumstances. These trials include the industry-funded trials that are conducted for regulatory review and, until recently, many of the National Institute of Mental Health–funded trials. Effectiveness trials (or pragmatic trials), on the other hand, evaluate the treatment effect in real-world settings among a broader range of participants.[5]

Several recent National Institute of Mental Health–funded effectiveness trials were conducted with *active comparators*. These include Sequenced Treatment Alternatives to Relieve Depression (STAR*D),[6] Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) for schizophrenia,[7] and Lithium Treatment–Moderate dose Use Study (LiTMUS) for bipolar disorder.[8] Each was designed as a superiority CER trial, ie, to test whether one intervention was *better than* another.

Here I consider the implications of including an active comparator for clinical trial design. Aspects of the superiority trial design will be contrasted with those of the noninferiority design. The noninferiority design is seldom used in psychopharmacologic research for regulatory submissions, but it could play a fundamental role in CER. For example, a study might seek to determine if an inexpensive generic intervention were inferior to a more costly, novel intervention or if a brief psychotherapy were inferior to a 12-month psychotherapeutic intervention. The results of the CATIE schizophrenia trial,[7] for example, have been misinterpreted by some as if it were a noninferiority trial; ie, designed to show that one intervention is *not worse than* another.[9] However, a superiority clinical trial result of no statistically significant difference between 2 treatment groups does not imply equivalence. For a trial to make a claim of equivalence (or noninferiority), the protocol and a priori hypothesis must indicate that it is not a superiority trial.[10]

This article will consider the choice of a noninferiority margin, the required sample sizes, and an unintended consequence of including active

© Copyright 2011 Physicians Postgraduate Press, Inc. © Copyright 2011 Physicians Postgraduate Press, Inc.

comparators in either an efficacy or an effectiveness trial. The duration of the trial and the broader inclusion criteria, elements of trial design that distinguish efficacy and effectiveness trials, will not be discussed. Nevertheless, what is described below applies to both long- and short-term trials, those that use either broad or narrow inclusion criteria, and those conducted in either community-based settings or academic medical centers.

## TERMINOLOGY

The following distinction is made in terminology as applied below. An *investigational intervention (I)* that is examined in an efficacy trial designed for regulatory submission, among other audiences, will typically be an intervention that has not received approval for the indication being studied. In contrast, an *I* that is evaluated in a comparative effectiveness trial is likely to be an intervention that rests on solid empirical evidence of efficacy and is perhaps more novel than the *active comparator* (*A*).

A priori research hypotheses specify the directional relations between *I* and *A*. Here, Δ is the endpoint population treatment group difference in pre-post change on a severity rating ($\Delta = A - I$), d* is a threshold of *clinical meaningfulness* and δ is the *noninferiority margin*, which represents a margin of clinical indifference.[11] *A* is clinically superior to *I* if $\Delta > d^*$. *I* is clinically superior to *A* if $\Delta < -d^*$. *I* is *noninferior* to *A* if $A - I < \delta$. *A* is *noninferior* to *I* if $A - I > -\delta$. In some situations, noninferiority will include *equivalence* or *clinical superiority* (Figure 1). An investigator must choose among these hypotheses (ie, *I is clinically superior to A*, *A is clinically superior to I*, *I is noninferior to A*, or *A is noninferior to I*) at the design stage of an RCT, prior to study initiation. It is unlikely that a trial would test noninferiority in both directions; instead, the direction would be specified (eg, *Is I worse than the current standard of care, A?*). The null hypothesis that corresponds to each of these research hypotheses is stated somewhat differently for statistical testing, as described below.

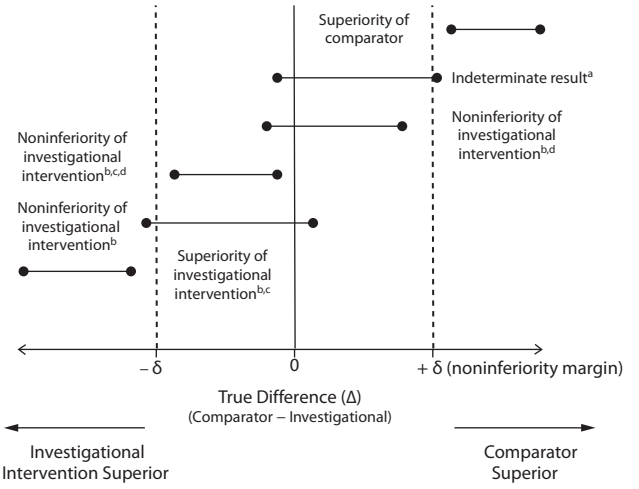
## SUPERIORITY TRIALS

Consider an efficacy trial that includes 3 cells: *I*, placebo (*P*), and *A* (where *A* has previously been shown to be superior to *P*). Efficacy in a particular patient population is examined by contrasting *I* and *P*. The assay sensitivity of the trial, its ability to discriminate between an intervention that is effective and one that is not effective, is examined by contrasting *A* and *P* in a sample from that population. A third contrast examines the relative effectiveness (or safety) of *I* and *A*. The inclusion of *P* in such a trial is justified because there is clinical equipoise in choosing between *I* and *P*. That is, despite the investigators' hopes and expectations, there is no evidence that *I* is superior to *P*.

### Assay Sensitivity

Consider, as an alternative design, a 2-arm superiority study in which the null hypothesis is $H_0$: $I = A$. $H_0$ can be rejected if *I* is superior to *A* or if *I* is inferior to *A*. If $H_0$ is not rejected, the result of *no difference* can be interpreted either as both treatments' being effective or as neither treatment's being effective, but the 2 interpretations cannot necessarily be disentangled. Although positive change would seem to imply positive treatment effects, we would not know what effect *P* would have had in this particular study, given the sample, the assessment procedures, and other study idiosyncrasies. In a 3-arm efficacy trial, however, *P* provides a context in which to test assay sensitivity. That is, *P* will help determine whether the trial was designed and implemented in such a way that differences between effective and ineffective interventions might be detected. Assay insensitivity might stem from, among other things, an inadequate sample size, poor retention, or inappropriate inclusion criteria or dosing.

**Figure 1. Confidence Intervals for Differences Between Investigational Intervention and Active Comparator in Hypothetical Noninferiority Trials**




[a]The indeterminate result stems from the CI that overlaps both 0 and +δ, the noninferiority margin. [b]Noninferiority of the investigational intervention requires that the CI fall entirely to the left of +δ, the noninferiority margin. [c]Superiority of the investigational intervention requires that the CI fall entirely to the left of 0. If such a result is seen in a noninferiority trial, and a 2-stage testing procedure was specified (first a test of noninferiority, then a test of superiority), the evidence would support both noninferiority and superiority (see US Department of Health and Human Services[19]). [d]Assuming that the equivalence margin ranges from –δ to +δ, results will provide evidence not only of noninferiority but also of equivalence.

### Sample Sizes for Superiority Trials

In the design of a 3-arm trial, the sample size required for adequate statistical power must be driven by the contrast with the smallest hypothesized effect. That smallest effect, most likely the comparison of *I* and *A*, would only be detected if the study were designed to test that contrast. To understand the problematic nature of the *I* versus *A* comparison, consider the basis for sample size determination in a clinical trial.

There are 4 components of statistical power analyses in a superiority RCT: the α level (typically .05, although it will

© Copyright 2011 Physicians Postgraduate Press, Inc. © Copyright 2011 Physicians Postgraduate Press, Inc.

be smaller in the case of trials with multiple primary efficacy measures[12]), statistical power (typically 80% or 90%), sample size, and the population effect size (eg, Cohen's *d*, a standardized group difference). For simplicity, assume that the trial is designed with equal cell sizes and that the number of participants is based on that needed to provide statistical power of 80%, with a 2-tailed α level of .05. Further assume that a *t* test will be used for 2-group comparisons on pre-post changes in severity ratings, which are approximately normally distributed. The sample size in an efficacy trial must be estimated such that the trial has appropriate statistical power to detect the smallest *clinically meaningful* effect of *I* versus *P*. To place this in the context of trials for schizophrenia, the effect size metric can be transformed into Positive and Negative Syndrome Scale (PANSS)[13] units by applying a standard deviation of 20 for PANSS change (based on Marder and Meibach[14]). For instance (based on Cohen[15]), 393 participants are needed per group to detect an effect size of $d = 0.20$ (where .20 SD units = 4 PANSS units), 64 participants are needed per group to detect an effect size of $d = 0.50$ (10 PANSS units), and 26 participants are needed per group to detect an effect size of $d = 0.80$ (16 PANSS units). Fewer subjects are needed to detect larger effects.

These effect sizes can be put in perspective by considering the results of a meta-analysis of 38 placebo-controlled RCTs of second-generation antipsychotics for schizophrenia.[16] The analysis involved 7,323 participants with a mean $d = 0.51$, which can serve as a benchmark. Approximately 64 participants per group would be needed for 80% statistical power to detect an effect of this magnitude. However, in an effort to be prudent such that a somewhat smaller effect would be detected, suppose that a trial is designed to detect $d = 0.40$, which would require 100 participants/group. (That sample size can be calculated as: $N/\text{group} = 16/d^2$; hence: $16/.40^2 = 100$.)[17,18] Yet, a study designed to detect $d = 0.40$ for *I* versus *P* will have substantially less power to detect the smaller difference in *I* versus *A*. That is, with 100/group there is 80% statistical power for $d = 0.40$ (8 PANSS units), but only 69% power for $d = 0.35$ (7 PANSS units), 56% for $d = 0.30$ (6 PANSS units), 42% for $d = 0.25$ (5 PANSS units), and 29% for $d = 0.20$ (4 PANSS units).

In contrast, if the goal of the *I* versus *A* comparison is to show *no difference*, a poorly powered trial would seemingly provide an excellent opportunity to achieve the goal. However, a superiority trial, particularly one that is underpowered, does not provide scientific evidence of equivalence or noninferiority. This is because a null hypothesis can be *rejected*; it cannot be *accepted*. Bias in superiority trials typically favors the null hypothesis. A nonsignificant result could stem from a poorly designed or implemented study (eg, inadequate sample size, inappropriate dosing, or excessive attrition). If a 3-arm superiority trial is to examine *I* versus *A*, it must be designed with adequate power for that contrast in particular. However, if the objective is to demonstrate noninferiority, a superiority trial design cannot be used.

## NONINFERIORITY TRIALS

An *A* might be included in an efficacy trial to examine assay sensitivity or in CER to evaluate the relative efficacy or safety of the *I*. Alternatively, in a noninferiority trial, *A* could be included to address the question, "Is *I worse than A*?" To test this question, the protocol must define a noninferiority margin ($\delta$), preferably in a way that is accepted by the clinical community. The noninferiority design might be implemented for 1 of 2 reasons. In a 2-arm efficacy trial, it might be used to show that the magnitude of the difference between *A* (with demonstrated efficacy) and *I* provides support for the superiority of *I* to *P* and indirectly provides evidence of efficacy.[19] This approach could be used to obviate the need to expose participants to *P* when an efficacious treatment is already available. Alternatively, a noninferiority design might be used in CER to show that a less costly medication is not worse than the current standard with regard to safety and/or efficacy. The choice of δ might very well differ in those 2 settings.

Why is a study with this design referred to as a noninferiority trial and not as an equivalence trial? The objective of an equivalence trial is to show that 2 interventions differ by no more than a specified amount, in either direction (ie, $\pm\delta$); one intervention is not better and not worse than the other perhaps stemming from the terminology of bioequivalence studies. In this instance, a 2-sided confidence interval (CI) would be used to examine equivalence. In contrast, the objective of a noninferiority trial is to show that *I* is not worse than *A* by more than a prespecified amount (δ, the margin of indifference). Such a trial involves a 1-sided CI (Figure 1).

The null hypothesis in a noninferiority trial is that *A is superior to I* ($H_0$: $A - I \geq \delta$). The alternative hypothesis is that *I is not inferior to A* ($H_1$: $A - I < \delta$).

Noninferiority is supported (ie, $H_0$ is rejected) whether, based on the data, *I* is superior, equivalent, or noninferior to *A*. The inferential errors in noninferiority tests almost appear to be inverted relative to those of superiority trials. A false positive result (type I error) in a noninferiority trial is seen when one incorrectly concludes noninferiority. A false negative result (type II error) in a noninferiority trial occurs when one fails to conclude noninferiority when, in fact, the treatments are similar.[19] Unlike a superiority trial in which poor study design and implementation favor the null hypothesis, many of the deficits in a noninferiority trial favor conclusion of noninferiority (the alternative hypothesis). For example, inappropriate inclusion criteria, inadequate dosing, small sample sizes, unreliable assessment procedures, noncompliance, and excessive attrition all can contribute to a finding of noninferiority.[20,21] For that reason, it is preferable that each aspect of the noninferiority trial design mirror that of completed superiority trials for the same intervention and indication.

As stated earlier, the noninferiority trial protocol must explicitly state the magnitude of the noninferiority margin, and, of course, this must be done prior to the start of the trial. δ represents the largest *clinically acceptable*

© Copyright 2011 Physicians Postgraduate Press, Inc.
© Copyright 2011 Physicians Postgraduate Press, Inc.

difference between groups. It is a challenge to get consensus on the magnitude of a treatment group difference that characterizes "not worse than." Unquestionably, δ must be smaller than the effect of *A* versus *P* and, it could be argued, much smaller than a clinically meaningful difference (d*) used in the design of a superiority trial. Consider again the meta-analysis of second-generation antipsychotics in which $d = 0.51$ for the PANSS change score.[16] With a PANSS change SD = 20, $d = .51$ represents approximately 10 PANSS units. It would be convenient to simply reduce that effect by 50% to define δ such that δ = .25 = 5 PANSS units. Yet that strategy would only be a reasonable choice if there were a consensus among clinical researchers that a 5-points difference in PANSS change would, in fact, be a clinically acceptable difference; that is, 5 points in PANSS change would not represent a notable difference.

The recent US Food and Drug Administration (FDA) Guidance Document on noninferiority trials described a method for identifying *δ* for a trial in which an *I,* which has no evidence of efficacy, is to be compared with *A*.[19] Initially all relevant RCTs that reported tests of *A* versus *P* are identified. The effect sizes are examined for consistency and aggregated into one summary measure. The investigator then specifies the proportion of that *A* versus *P* effect that must be preserved by *I* (in the *A* versus *I* comparison) to demonstrate that a meaningful effect of *I* over *P* would likely have been seen, had the study included *P*. That proportion is used to estimate δ. However, it is not clear how this strategy could be adapted for CER, in which, presumably, 2 efficacious interventions are compared.

### Sample Sizes for Noninferiority Trials

The required sample size for a noninferiority trial is substantially larger than that for a superiority trial. This disparity is based on the difference in magnitude between d* and δ, the latter being much smaller (at the trial design stage). Some might argue that δ need only be slightly smaller than d*. I, however, believe that there is a gray zone between d* and δ in which the results are equivocal: smaller than *clinically meaningful* but greater than *clinically acceptable*. For example, if d* = .40 were used to define a clinically meaningful difference, I would not accept δ = .39 as a margin of indifference. (I would accept, instead, perhaps δ = .15 or δ = .20). In part, this has to do with the imprecision of outcome measures used in psychiatry.

As a rule of thumb, if δ is half the size of d* from a superiority trial, the noninferiority sample size must be 4-fold higher.[20,21] This general guideline applies to both binary and continuous outcomes. However, despite the convenience of this axiom, 50% of a clinically *meaningful* difference will not necessarily represent a clinically *acceptable* difference. The choice of δ requires input from clinicians, statisticians, and perhaps patients and their families—if they are familiar with the metric (eg, PANSS units). Examples of sample sizes required *per group* for various noninferiority margins for continuous outcomes are 6,280 (δ = .05), 1,570 (δ = .10), and 698 (δ = .15).[22] (These estimates assume an α level of .025 for the 1-sided noninferiority confidence interval and 80% power.)

## AN UNINTENDED CONSEQUENCE OF INCLUDING AN ACTIVE COMPARATOR

It is conceivable that an *A* might be included in an efficacy trial strictly to examine the assay sensitivity of the trial (*A* versus *P*) and that the study would not be powered for another contrast. Nevertheless, the data collected for a test of assay sensitivity would provide the data needed for a comparison of *I* and *A* interventions. This being the case, it is unlikely that reviewers would ignore the opportunity to examine those data. Consider, for example, the Vanda Pharmaceuticals regulatory submission for iloperidone as a brief case study of one problem introduced by including an *A*. Vanda conducted a randomized double-blind RCT to evaluate the efficacy, safety, and tolerability of iloperidone, an atypical antipsychotic for schizophrenia it licensed from Novartis. Participants were randomized to receive 28 days of iloperidone, ziprasidone, or *P*.[23,24] This trial followed an unsuccessful attempt by Novartis to assemble sufficient empirical evidence to gain regulatory approval for iloperidone. An active comparator (haloperidol in 1 trial or risperidone in the other 2 trials) was included to examine assay sensitivity in the negative Novartis trials. The FDA reviewed the totality of the iloperidone evidence, including the Novartis experience, and in 2008 issued a nonapprovable letter to Vanda Pharmaceuticals.[25] The decision was based on the results of at least 2 RCTs in which iloperidone was superior to *P* yet at the same time significantly worse than the active comparator. Although the comparison with the active comparator was included for examining assay sensitivity (ie, *A* versus *P*), regulators used those data to determine if iloperidone were worse (ie, less helpful or more harmful) than each *A*. The FDA concern was based on several considerations; apparently, in part, on safety issues, ie, should a drug be made available to consumers if it has been shown to be significantly worse than a drug that is currently available? Would availability of such a drug pose a public health risk, particularly when dealing with a serious and persistent chronic illness such as schizophrenia? It remains the case, however, that federal regulations do not require evidence of superiority or noninferiority to an *A*; instead, they require evidence of superiority (typically to *P*). After further evaluation, the FDA ultimately approved iloperidone to treat adults with schizophrenia.[26]

## DISCUSSION

Implications of including an active comparator in comparative effectiveness and efficacy trials have been discussed using 2- and 3-arm trials as examples. An efficacy trial typically includes a superiority contrast (*I* versus *P*), and it must be designed to detect a clinically meaningful difference (d*). If an active comparator is included, an efficacy trial might also test assay sensitivity, contrasting *A* versus *P*.

© Copyright 2011 Physicians Postgraduate Press, Inc. © Copyright 2011 Physicians Postgraduate Press, Inc.

A third contrast, which may or may not be a component of an efficacy trial design but is fundamental in CER, is *I* versus *A*. This comparison could examine either superiority or noninferiority, but the choice between the 2 must be stated a priori. Comparative effectiveness research, for example, might use a noninferiority design to compare a costly brand-name medication with a generic. For this aspect of the trial, the design must prespecify a clinically acceptable difference (δ), which operationalizes *not worse than*.

The results of the CATIE trial have been interpreted by some as evidence of noninferiority (or equivalence) of older antipsychotics to the second-generation antipsychotics for patients with schizophrenia,[27] when, in fact, the study was designed as a superiority trial. A nonsignificant superiority contrast, as seen in CATIE, does not demonstrate equivalence or noninferiority. The primary aim of the CATIE schizophrenia trial posited neither a noninferiority hypothesis nor a noninferiority margin.

Evidence of noninferiority can only come from well-designed and well-conducted noninferiority CER. Poor design and implementation in a noninferiority trial could, in fact, favor noninferiority. Pocock[20] described the ideal circumstances for a noninferiority trial. First, the investigators must select an *A* that has well-documented and convincing evidence of superiority over *P*. Second, the noninferiority trial must be conducted under conditions similar to those of superiority trials with regard to subject selection, dosing, trial duration, and primary outcome. Third, the choice of the noninferiority margin must be small enough that clinicians and researchers will be convinced that the *I* has therapeutic value.

Fulfilling a noninferiority trial objective is especially challenging when evaluating interventions for depression and anxiety disorders, for which only approximately 50% of RCTs for *A* (with known efficacy) are positive.[19] It leaves open the very real possibility that a noninferiority finding might be based on comparison with an *A* that would not have actually separated from *P* had it been included in the trial. Trial idiosyncrasies (eg, the sites, sample characteristics, or assessors) can have more influence on results than we would like. A noninferiority finding is more plausible if evidence of assay sensitivity is provided. Such evidence can only come from within the trial, and it requires inclusion of *P* as a third cell.[28] For that reason, the inclusion of *P* in CER warrants serious consideration for interventions such as antidepressants and anxiolytics.

Clinical trials that include *A*'s tend to require more resources, both financial and human, because a larger sample size is necessary to detect the smaller differences expected in trials that do not have a *P*. As a result, CER trials require more time for recruitment and expose more participants to the risks of an experiment than do 2-arm, placebo-controlled efficacy trials. This applies to both superiority and noninferiority trials, albeit more so to the latter, because a noninferiority margin is typically substantially smaller than the clinically meaningful difference guiding a superiority trial. However, there are important benefits of designs that involve active comparators. For instance, a larger sample size provides more safety data by virtue of the additional person-time of exposure to the *I*. Moreover, the inclusion of an active comparator could attract potential participants and, among those who do enroll, it could enhance retention.

In conclusion, active comparators are essential in some studies but not in others. The CER design necessitates an active comparator. Despite this fact, the Psychiatric Drugs Division of the FDA does not currently require that trials have an active comparator, and therefore many efficacy trials do not include one. The European Medicines Agency does have such a requirement for antidepressants and antipsychotics.[29,30] One benefit of the third arm in a placebo-controlled efficacy RCT is that it provides a context in which to examine the assay sensitivity of the trial. That test is particularly valuable with indications for which approved mediations have high failure rates.[19] However, once data from the active comparator cell are available in a 3-arm trial, it leaves the investigational intervention vulnerable to failure in a test of comparative efficacy or noninferiority. Even if such a comparison is not prespecified by the sponsor, it is in the interest of the public health for reviewers to determine if a novel intervention is indeed inferior to the current standard of care. Investigators must very carefully weigh the costs and benefits of including an active comparator in their trials.

***Drug names:*** haloperidol (Haldol and others), iloperidone (Fanapt), lithium (Lithobid and others), risperidone (Risperdal and others), ziprasidone (Geodon).

***Author affiliation:*** Department of Psychiatry, Weill Cornell Medical College, New York, New York.

***Potential conflicts of interest:*** (past 12 months) **Dr Leon** has served on the data and safety monitoring boards of AstraZeneca, Dainippon Sumitomo America, Pfizer, and Vanda; (2006–2007) has served as a consultant/advisor to the US Food and Drug Administration, MedAvante, the National Institute of Mental Health (NIMH), and Roche; and holds equity in MedAvante.

***Funding/support:*** This research was supported, in part, by grants from the NIMH (MH060447 and MH068638).

***Acknowledgment:*** The author gratefully acknowledges Lori L. Davis, MD (University of Alabama School of Medicine, Birmingham), Barbara Milrod, MD (Weill Cornell Medical College, New York, New York), Nina Schooler, PhD (State University of New York Downstate Medical Center, Brooklyn), and an anonymous reviewer for valuable comments on an earlier draft of this manuscript. Dr Schooler has provided consultation or served on advisory boards for AstraZeneca, Bristol-Myers Squibb, Dainippon Sumitomo, Eli Lilly, Hoffman-LaRoche, Lundbeck, Merck/Schering-Plough, Ortho-McNeil Janssen, and Pfizer. She has received grant/research support from AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Lundbeck, Ortho-McNeil Janssen, and Pfizer. Drs Davis and Milrod report no financial or other relationship relevant to the subject of this article.

***Previous presentation:*** Presented, in part, at the 5th Annual Scientific Meeting of the International Society for CNS Clinical Trials and Methodology; March 3–5, 2009; Alexandria, Virginia.

## REFERENCES

1. Agency for Healthcare Research and Quality Funding Announcements. http://www.ahrq.gov/fund/grantix.htm#RFA. Accessed January 28, 2010.
2. Department of Health and Human Services. Federal Coordinating Council for Comparative Effectiveness Research: report to the president and the congress. http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf. Updated June 30, 2009. Accessed October 8, 2010.
3. Congressional Budget Office. Research on comparative effectiveness of medical treatment. issues and options for an expanded federal role.

© Copyright 2011 Physicians Postgraduate Press, Inc. © Copyright 2011 Physicians Postgraduate Press, Inc.

http://www.cbo.gov/ftpdocs/88xx/doc8891/Frontmatter.1.2.shtml. Updated December 2007. Accessed October 8, 2010.
4. *Initial National Priorities on Comparative Effectiveness Research*. Washington, DC: Institute of Medicine; 2009.
5. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*. 2003;290(12):1624–1632.
6. Rush AJ, Fava M, Wisniewski SR, et al; STAR*D Investigators Group. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. 2004;25(1):119–142.
7. Lieberman JA, Stroup TS, McEvoy JP, et al; Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med*. 2005;353(12):1209–1223.
8. Nierenberg AA, Sylvia LG, Leon AC, et al; LiTMUS Study Group. Lithium Treatment—Moderate Dose Use Study (LiTMUS) for bipolar disorder: rationale and design. *Clin Trials*. 2009;6(6):637–648.
9. Kraemer HC, Glick ID, Klein DF. Clinical trials design lessons from the CATIE study. *Am J Psychiatry*. 2009;166(11):1222–1228.
10. Piaggio G, Elbourne DR, Altman DG, et al; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA*. 2006;295(10):1152–1160.
11. Cook TD, DeMets DL. *Introduction to Statistical Methods for Clinical Trials*. Boca Raton, LA: Chapman & Hall/CRC; 2008.
12. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *J Clin Psychiatry*. 2004;65(11):1511–1514.
13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987;13(2):261–276.
14. Marder SR, Meibach RC. Risperidone in the treatment of schizophrenia. *Am J Psychiatry*. 1994;151(6):825–835.
15. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
16. Leucht S, Arbter D, Engel RR, et al. How effective are second-generation antipsychotic drugs? a meta-analysis of placebo-controlled trials. *Mol Psychiatry*. 2009;14(4):429–447.
17. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med*. 1992;11(8):1099–1102.
18. Leon AC, Davis LL. Enhancing clinical trial design of interventions for posttraumatic stress disorder. *J Trauma Stress*. 2009;22(6):603–611.
19. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research. *Guidance for Industry: Non-Inferiority Clinical Trials (draft guidance)*. Rockville, MD: Food and Drug Administration; 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Updated March 2010. Accessed October 8, 2010.
20. Pocock SJ. The pros and cons of noninferiority trials. *Fundam Clin Pharmacol*. 2003;17(4):483–490.
21. Jones B, Jarvis P, Lewis JA, et al. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313(7048):36–39.
22. Julious SA. Sample sizes for clinical trials with normal data. *Stat Med*. 2004;23(12):1921–1986.
23. ClinicalTrials.gov. A randomized, double-blind, placebo- and active-controlled, multicenter study to evaluate the efficacy, safety and tolerability of iloperidone in schizophrenic patients in acute exacerbation followed by a long-term treatment phase. http://clinicaltrials.gov/ct2/show/NCT00254202. Accessed January 28, 2010.
24. Cutler AJ, Kalali AH, Weiden PJ, et al. Four-week, double-blind, placebo- and ziprasidone-controlled trial of iloperidone in patients with acute exacerbations of schizophrenia. *J Clin Psychopharmacol*. 2008;28(suppl 1):S20–S28.
25. Vanda Pharmaceuticals announces receipt of not approvable letter From FDA for iloperidone. http://phx.corporate-ir.net/phoenix.zhtml?c=196233&p=irol-newsArticle&ID=1179851&highlight. Accessed January 9, 2010.
26. FDA approves Fanapt to treat schizophrenia http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm149578.htm. Accessed January 9, 2010.
27. Insel TR. Psychiatrists' relationships with pharmaceutical companies: part of the problem or part of the solution? *JAMA*. 2010;303(12):1192–1193.
28. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments: part 1: ethical and scientific issues. *Ann Intern Med*. 2000;133(6):455–463.
29. The European Agency for the Evaluation of Medicinal Products: Committee for Proprietary Medicinal Products. Note for Guidance on Clinical Investigations of Medicinal Products in the Treatment of Depression.. http://www.tga.gov.au/docs/pdf/euguide/ewp/051897en.pdf. Updated April 25, 2002. Accessed October 8, 2010.
30. The European Agency for the Evaluation of Medicinal Products: Committee for Proprietary Medicinal Products. Note for Guidance on Clinical Investigations of Medicinal Products in the Treatment of Schizophrenia. http://www.tga.gov.au/docs/pdf/euguide/ewp/055995en.pdf. Updated February 26, 1998. Accessed October 8, 2010.

© Copyright 2011 Physicians Postgraduate Press, Inc. © Copyright 2011 Physicians Postgraduate Press, Inc.