Published in final edited form as:

J Clin Psychiatry. 2011 March ; 72(3): 331–340. doi:10.4088/JCP.10r06669.

Comparative Effectiveness Clinical Trials in Psychiatry: Superiority, Non-inferiority and the Role of Active Comparators

Andrew C. Leon, Ph.D.

Abstract

Background—The Agency for Healthcare Research and Quality 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.

Objective—This paper is written to educate the clinical community about the implications of an active comparator on clinical trial design.

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

Conclusion—Evidence of non-inferiority can only come from well designed and conducted non-inferiority CER. Sample sizes needed in non-inferiority trials, and in superiority trials that include an active comparator, are substantially larger than in trials that can utilize a placebo control in their scientific design. Because of this, 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.

Keywords

active comparator; clinical trial; comparative effectiveness; equivalence; non-inferiority

Introduction

The Agency for Healthcare Research and Quality (AHRQ) issued several recent Requests for Applications (RFA) that call for comparative effectiveness research (CER).¹ These were motivated primarily by the efforts for health care reform in the US 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

Corresponding Author: Andrew C. Leon, Ph.D. Weill Cornell Medical College Department of Psychiatry, Box 140 525 East 68th Street New York, NY 10065 telephone: (212) 746-3872 fax: (212) 746-8754 aleon@med.cornell.edu.

Presented, in part, at the Annual meeting of the International Society for CNS Clinical Trials and Methodology (ISCTM), March 3-5, 2009, Alexandria, VA.

Disclosures (past 12 months)

Data and Safety Monitoring Boards: AstraZeneca, Dainippon Sumitomo Pharma America, Pfizer, Vanda (2006-07)

Consultant/Advisor: Cyberonics, FDA, MedAvante, NIMH, Roche, and Schering Plough.

Equity: MedAvante

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 The definition conforms to those of both the Congressional Budget Office report on Research on Comparative Effectiveness of Medical Treatment³ and the Institute of Medicine report on Initial National Priorities on Comparative Effectiveness Research⁴. Much of the research supported by this initiative will involve randomized controlled clinical trials (RCT), yet meta-analyses and observational studies will also be used.

The AHRQ 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 are those that evaluate a treatment effect in a rarefied sample under ideal circumstances. This includes 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 NIMH 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 schizophrenia⁷, and Lithium Treatment -- Moderate dose Use Study (LiTMUS) for bipolar disorder⁸. Each was designed as a superiority CER trial; i.e., to test whether one intervention was *better than* another.

Here we consider the implications of including an active comparator on clinical trial design. Aspects of the superiority trial design will be contrasted with those of the non-inferiority design. The latter is seldom used in psychopharmacologic research for regulatory submissions, but could play a fundamental role in CER. For example, a study might seek to determine if an inexpensive generic intervention is inferior to a more costly, novel intervention; or if a brief psychotherapy is inferior to a 12 month psychotherapeutic intervention. The results of the CATIE schizophrenia trial⁷, for example, have been misinterpreted by some as if it were a non-inferiority trial; i.e., designed to show that one intervention is *not worse than* another.⁹ However, a superiority clinical trial result of “no statistically significant difference” does not imply “equivalence”. For a trial to make a claim of equivalence (or non-inferiority), the protocol and a priori hypothesis must indicate that it is not a superiority trial.¹⁰

The presentation will consider the choice of a non-inferiority margin, the required sample sizes, and an unintended consequence of including active 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 comparative 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 population treatment group difference ($\Delta = A - I$), d^* is a threshold of *clinical meaningfulness* and δ is the *non-inferiority margin*, which represents a margin of clinical indifference¹¹. I is clinically superior to A if $\Delta > d^*$ and A is clinically superior to I if $\Delta < -d^*$. I is “non-inferior” to A if $A - I < \delta$ and A is “non-inferior” to $A - I > -\delta$. In some situations, non-inferiority will include equivalence or clinical superiority (Figure 1). An investigator must choose among these hypotheses at the design stage of an RCT, prior to study initiation. It is unlikely that a trial would test non-inferiority in either direction, but instead the direction would be specified (e.g., 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 (page 9).

Superiority Trials

Consider an efficacy trial that includes three cells: I , 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 . Assay sensitivity of the trial is examined by contrasting A and P in a sample from that population. A third contrast compares 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 investigational 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 two 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 that both treatments are effective or neither treatment is effective, but the two 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 three arm efficacy trial, however, P provides a context to test assay sensitivity. That is, it will help inform the question of assay sensitivity, “Was the trial designed and implemented in such a way that differences between effective and ineffective interventions would be detected?” Assay insensitivity could stem from, among other things, an inadequate sample size, poor retention, or inappropriate inclusion criteria or dosing.

Sample Sizes for Superiority Trials

Returning to the design of a three 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 is designed to test that contrast. To understand the problematic nature of the I vs. A comparison, consider the basis for sample size determination in a clinical trial.

There are four components of statistical power analyses in a superiority RCT: the α -level (typically 0.05, although smaller in the case of with co-primaries¹²), statistical power (typically 80% or 90%), sample size, and the population effect size (e.g., Cohen's d , a standardized group difference). For simplicity assume that the trial is designed with equal cell sizes and the number of participants is based on that needed to provide statistical power of 80% with a two-tailed alpha level of .05. Further assume that a t-test will be used for two-group comparisons on pre-post change 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

vs. *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 Meibach¹⁴). For instance (based on Cohen¹⁵), 393 participants are needed per group to detect an effect size of $d=.20$ (where .20 sd units = 4 PANSS units), 64 participants are needed per group to detect an effect size of $d=.50$ (10 PANSS units), and 26 participants are needed per group to detect an effect size of $d=.80$ (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 schizophrenia¹⁶. It involved 7323 participants with a mean $d=0.51$, which can serve as a benchmark. About 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, say 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* vs. *P* will have substantially less power to detect the smaller *I* vs. *A* difference. That is, with 100/group there is 80% statistical power for $d=.40$ (8 PANSS units), but only 69% power for $d=.35$ (7 PANSS units), 56% for $d=.30$ (6 PANSS units), 42% for $d=.25$ (5 PANSS units), and 29% for $d=.20$ (4 PANSS units).

In contrast, if the goal of the *I* vs. *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 under-powered, does not provide scientific evidence of equivalence or non-inferiority. This is because a null hypothesis can be *rejected* based on a superiority trial; a null hypothesis cannot be *accepted*. Bias in superiority trials typically favors the null hypothesis. A non-significant result could stem from a poorly designed or implemented study (e.g., inadequate sample size, inappropriate dosing, or excessive attrition). If a three arm superiority trial will examine *I* vs. *A*, it must be designed with adequate power for that contrast in particular. However, if the objective is to demonstrate non-inferiority, a superiority trial design cannot be used.

Non-Inferiority Trials

An active comparator could be included in an efficacy trial to examine assay sensitivity or in CER to evaluate the relative efficacy or safety of the investigational intervention. Alternatively, in a non-inferiority trial *A* could be included to address the question, “Is *I* worse than *A*?”. To test this, the protocol must define a non-inferiority margin (δ), preferably one that is accepted by the clinical community. The non-inferiority design can be implemented for one of two reasons. In a two-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 available. Alternatively, a non-inferiority design could 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 could very well differ in those two settings.

Why is this referred to as a non-inferiority trial and not as an equivalence trial? The objective of an equivalence trial is to show that two interventions differ by no more than a specified amount, in either direction (i.e., $\pm \delta$); not better and not worse, perhaps stemming from studies of bioequivalence. A two-sided confidence interval (CI) is used to examine equivalence. In contrast, the objective of a non-inferiority trial is to show that *I* is not worse than *A* by more than a pre-specified amount. This involves a one-sided CI (Figure 1).

The null hypothesis in a non-inferiority trial states that *A* is superior to *I*

$$H_0: A - I \geq \delta.$$

The alternative hypothesis states that *I* is not inferior to *A*

$$H_1: A - I < \delta.$$

Non-inferiority is supported (i.e., rejection of H_0) whether, based on the data, *I* is superior, equivalent, or non-inferior to *A*. The inferential errors in non-inferiority tests almost appear to be inverted relative to those of superiority trials. A false positive result (Type I error) is seen when one incorrectly concludes non-inferiority. A false negative result (Type II error) occurs if one fails to conclude “non-inferiority” when, in fact, the treatments are similar¹⁸ Unlike a superiority trial in which poor study design and implementation favor the null hypothesis, much of the imprudence in a non-inferiority trial favors conclusion of non-inferiority. For example, inappropriate inclusion criteria, inadequate dosing, small sample sizes, unreliable assessment procedures, non-compliance, and excessive attrition all can contribute to a finding of “non-inferiority”.^{20, 21} For that reason, it is preferable that each aspect of the design mirror that of completed superiority trials for the same intervention and indication.

As stated earlier, the non-inferiority trial protocol must explicitly state the magnitude of the non-inferiority margin, and, of course, this must be done prior to the start of the trial. δ represents the largest *clinically acceptable* 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* vs. *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 $sd = 20$, $d = 0.51$ represents about 10 PANSS units. It would be convenient to simply reduce that effect size by 50% to define δ such that $\delta = 0.25 = 5$ PANSS units. Yet that would only be a reasonable choice if there were a consensus among clinical researchers that five points in PANSS change is, in fact, a clinically acceptable difference; that is, five points does not represent a notable difference.

The recent FDA Guidance Document on non-inferiority 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* vs. *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* vs. *P* effect that must be preserved by *I* (in the *A* vs. *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, two efficacious interventions are compared.

Sample Sizes for Non-Inferiority Trials

The required sample size for a non-inferiority trial is substantially larger than that of a superiority trial. This 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 $d^* = 0.40$ was used to define a clinically meaningful difference, I

would not accept $\delta=.39$ as a margin of indifference (instead perhaps .15 or .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 non-inferiority sample size must be fourfold higher.^{20, 21} This general guideline applies to both binary and continuous outcomes. However, despite the convenience of this axiom, fifty percent 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 (e.g., PANSS units). Examples of sample sizes required *per group* for various non-inferiority margins for continuous outcomes are: 6280 ($\delta=.05$), 1570 ($\delta=.10$), and 698 ($\delta=.15$)²². (These estimates assume an alpha level of .025 for the one-sided non-inferiority confidence interval and 80% power.)

An Unintended Consequence of Including an Active Comparator

It is conceivable that an active comparator is included in an efficacy trial strictly to examine assay sensitivity of the trial (*A vs. P*) and that the study is not powered for another contrast. Nevertheless, the data collected for a test of assay sensitivity provide what is needed for a comparison of investigational and active interventions. It is unlikely that reviewers would ignore the opportunity to examine it. Consider, for example, the Vanda Pharmaceuticals regulatory submission as a brief case study of one problem introduced by including an active comparator. Vanda conducted a randomized, double-blind RCT to evaluate the efficacy, safety and tolerability of iloperidone as an atypical antipsychotic for schizophrenia. Participants were randomized to receive 28 days of iloperidone, ziprasidone or placebo.^{23, 24} This trial followed what was an unsuccessful attempt by Novartis to assemble sufficient empirical evidence to gain regulatory approval for iloperidone. An active comparator (haloperidol in one trial or risperidone in the other two trials) was included to examine assay sensitivity in the negative Novartis trials. The US FDA reviewed the totality of the iloperidone evidence, including the Novartis experience, and in 2008 issued a non-approvable letter to Vanda Pharmaceuticals.²⁵ The decision was based on the results of at least two RCTs in which iloperidone was superior to placebo, yet at the same time, the active comparator was significantly better than iloperidone. Although this latter contrast was included for examining assay sensitivity (i.e., *A vs. P*), regulators used those data to determine if iloperidone is worse (i.e., less helpful or more harmful) than each active comparator. The FDA concern was based on several considerations; apparently, in part on safety issues. That is, 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 that pose a public health risk, particularly when dealing with a serious and persistent chronic illness such as schizophrenia? However, federal regulations do not require evidence of superiority or non-inferiority to an active comparator, but instead superiority (typically to placebo). 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 two and three arm trials as examples. An efficacy trial typically includes a superiority contrast (*I vs. P*) and must be designed to detect a clinically meaningful difference (d^*). If an active comparator is included, an efficacy trial could also test assay sensitivity, contrasting the *A vs. P*. A third contrast, which may or may not be a component of an efficacy trial design, but is fundamental in CER, is that which compares *I vs. A*. It could examine either superiority or non-inferiority, but the choice between the two must be stated a priori. A CER design, for example, might use a non-inferiority design to

compare a costly brand name medication with a generic. For this aspect of the trial, the design must pre-specify a clinically acceptable difference (δ), which operationalizes “not worse than”.

The results of the CATIE trial have been interpreted by some as evidence of non-inferiority (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 non-significant superiority contrast, as seen in CATIE, does not demonstrate equivalence or non-inferiority. The primary aim of the CATIE schizophrenia trial did not posit either a non-inferiority hypothesis nor a non-inferiority margin.

Evidence of non-inferiority can only come from well-designed and well-conducted non-inferiority CER. Poor design and implementation in a non-inferiority trial could, in fact, favor non-inferiority. Pocock described the ideal circumstances for a non-inferiority trial.²⁰ First, the investigators must select an active comparator that has well-documented and convincing evidence of superiority over placebo. Second, the non-inferiority 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 non-inferiority margin must be small enough that clinicians and researchers will be convinced that the investigational intervention has therapeutic value.

Fulfilling a non-inferiority trial objective is especially challenging when evaluating interventions for depression and anxiety disorders, for which only about 50% of RCTs for A (with known efficacy) are positive.¹⁹ It leaves open the very real possibility that a non-inferiority finding is based on comparison with an A that would not have actually separated from placebo, had it been included in the trial. Trial idiosyncrasies (e.g., the sites, sample characteristics, or assessors) can have more influence on results than we would like. A non-inferiority finding is more plausible if evidence of assay sensitivity is provided. Such evidence can only come from within the trial, which 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 active comparators tend to require more resources, both financial and human. This because a larger sample size is necessary to detect the smaller differences expected in trials that do not have a placebo control. As a result, trials with active comparators require more time for recruitment and expose more participants to the risks of an experiment than in two-arm placebo-controlled efficacy trials. This applies to both superiority and non-inferiority trials, albeit more so in the latter because a non-inferiority 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 investigational intervention. 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. The Psychiatric Drugs Division of the US FDA, in contrast, does not currently require that trials have an active comparator, and therefore the vast majority of efficacy trials do not include one. The European Medicines Agency (EMA) 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 to examine the assay sensitivity of the trial. That test is particularly valuable with indications for which approved medications have high failure rates.¹⁹ However, once data from the

active comparator cell are available in a three arm trial, it leaves the investigational intervention vulnerable to failure in a test of comparative efficacy or in a test of non-inferiority. 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.

Acknowledgments

This research was supported, in part, by grants from the National Institute of Mental Health (MH060447 and MH068638). The author gratefully acknowledges Lori L. Davis, M.D., Barbara Milrod, M.D., Nina Schooler, Ph.D., and an anonymous reviewer for valuable comments on an earlier draft of this manuscript.

References

1. Agency for Healthcare Research and Quality Funding Announcements. <<http://www.ahrq.gov/fund/grantix.htm#RFA>>.
2. Federal Coordinating Council for Comparative Effectiveness Research: Report to the President and the Congress. DHHS, ed. 2009
3. Research on Comparative Effectiveness of Medical Treatment: Issues and Options for an Expanded Federal Role: Congressional Budget Office. 2007
4. Initial National Priorities on Comparative Effectiveness Research. Institute of Medicine; Washington, DC: 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*. Sep 24; 2003 290(12):1624–1632. [PubMed: 14506122]
6. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. Feb; 2004 25(1):119–142. [PubMed: 15061154]
7. Lieberman JA, Stroup TS, McEvoy JP, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med*. Sep 22; 2005 353(12):1209–1223. [PubMed: 16172203]
8. Nierenberg AA, Sylvia LG, Leon AC, et al. Lithium treatment -- moderate dose use study (LiTMUS) for bipolar disorder: rationale and design. *Clin Trials*. Dec; 2009 6(6):637–648. [PubMed: 19933719]
9. Kraemer HC, Glick ID, Klein DF. Clinical trials design lessons from the CATIE study. *Am J Psychiatry*. Nov; 2009 166(11):1222–1228. [PubMed: 19797435]
10. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA*. Mar 8; 2006 295(10):1152–1160. [PubMed: 16522836]
11. Cook, TD.; DeMets, DL. Introduction to statistical methods for clinical trials. Chapman & Hall/CRC; Boca Raton: 2008.
12. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *J Clin Psychiatry*. Nov; 2004 65(11):1511–1514. [PubMed: 15554764]
13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987; 13(2):261–276. [PubMed: 3616518]
14. Marder SR, Meibach RC. Risperidone in the treatment of schizophrenia. *Am J Psychiatry*. Jun; 1994 151(6):825–835. [PubMed: 7514366]
15. Cohen J. A Power Primer. *American Psychologist*. 1992; 112:155–159.
16. Leucht S, Arbter D, Engel RR, Kissling W, Davis JM. How effective are second-generation antipsychotic drugs? A meta-analysis of placebo-controlled trials. *Mol Psychiatry*. Apr; 2009 14(4):429–447. [PubMed: 18180760]
17. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med*. Jun 15; 1992 11(8):1099–1102. [PubMed: 1496197]

18. Leon AC, Davis LL. Enhancing clinical trial design of interventions for posttraumatic stress disorder. *J Trauma Stress*. Dec; 2009 22(6):603–611. [PubMed: 19902462]
19. Center for Drug Evaluation and Research; Center for Biologics Evaluation and Research. US Food and Drug Administration; Rockville, MD: Guidance for Industry: Non-inferiority clinical trials (Draft guidance)..
20. Pocock SJ. The pros and cons of noninferiority trials. *Fundam Clin Pharmacol*. Aug; 2003 17(4): 483–490. [PubMed: 12914552]
21. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. Jul 6; 1996 313(7048):36–39. [PubMed: 8664772]
22. Julious SA. Sample sizes for clinical trials with normal data. *Stat Med*. Jun 30; 2004 23(12):1921–1986. [PubMed: 15195324]
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>>.
24. Cutler AJ, Kalali AH, Weiden PJ, Hamilton J, Wolfgang CD. Four-week, double-blind, placebo- and ziprasidone-controlled trial of iloperidone in patients with acute exacerbations of schizophrenia. *J Clin Psychopharmacol*. Apr; 2008 28(2 Suppl 1):S20–28. [PubMed: 18334909]
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>.
26. FDA Approves Fanapt to Treat Schizophrenia. <http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm149578.htm>.
27. Insel TR. Psychiatrists' relationships with pharmaceutical companies: part of the problem or part of the solution? *JAMA*. Mar 24; 303(12):1192–1193. [PubMed: 20332407]
28. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments - Part 1: Ethical and scientific issues. *Annals of Internal Medicine*. Sep 19; 2000 133(6):455–463. [PubMed: 10975964]
29. Note for Guidance on Clinical Investigations of Medicinal Products in the Treatment of Depression: Committee for Proprietary Medicinal Products, The European Agency for the Evaluation of Medicinal Products. 2002
30. Note for Guidance on Clinical Investigations of Medicinal Products in the Treatment of Schizophrenia: Committee for Proprietary Medicinal Products, The European Agency for the Evaluation of Medicinal Products. 2002

Figure 1.

Confidence Intervals for Differences between *Investigational Intervention* and *Active Comparator* in Hypothetical Non-Inferiority Trials

Notes:

Non-inferiority of Investigational Intervention requires that the CI falls entirely to the left of $+\delta$, the non-inferiority margin. (e.g., #'s 3, 4, 5 and 6).

Superiority of the Investigational Intervention requires that the CI falls entirely to the left of 0 (e.g., #'s 4 and 6). If such a result is seen in a non-inferiority trial and a two-stage testing procedure was specified (first test non-inferiority, then test superiority), the evidence would support both non-inferiority and superiority.¹⁹

The Indeterminate result (#2) stems from the CI that overlaps both 0 and $+\delta$, the non-inferiority margin.

Assuming that the equivalence margin ranges from $-\delta$ to $+\delta$, #'s 3 and 4 not only provides evidence of non-inferiority, but also equivalence.