Published in final edited form as:
J Trauma Stress. 2008 October ; 21(5): 433–439. doi:10.1002/jts.20367.

Noninferiority and Equivalence Designs: Issues and Implications for Mental Health Research

Carolyn J. Greene[1], **Leslie A. Morland**[1], **Valerie L. Durkalski**[2], and **B. Christopher Frueh**[3]

1 *National Center for PTSD- Pacific Islands Division, Department of Veterans Affairs Pacific Islands Healthcare System, Honolulu, Hawaii*

2 *Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, South Carolina*

3 *The Menninger Clinic and Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas*

Abstract

"Noninferiority" and "equivalence" are often used interchangeably to refer to trials in which the primary objective is to show that a novel intervention is as effective as the standard intervention. The use of these designs is becoming increasingly relevant to mental health research. Despite the fundamental importance of these designs, they are often poorly understood, improperly applied, and subsequently misinterpreted. This paper explains the noninferiority and equivalence designs and key methodological and statistical considerations. Decision points in using these designs are discussed, such as choice of control condition, determination of the noninferiority margin, and calculation of sample size and power. With increasing utilization of these designs, it is critical that researchers understand the methodological issues, advantages, disadvantages and related challenges.

In biomedical and mental health research, by far the most common conceptual approach is traditional significance testing in between-group designs, useful for determining if group outcomes are clinically and statistically different from each other and if so, in which direction. This is the standard approach used in most treatment outcome studies, where it is hypothesized that an experimental treatment is superior to a comparison treatment. However, it has become increasingly important in research to evaluate a slightly different question: determining if the effects of two treatments are *not* clinically and statistically different from each other. This is important when, for example, we want to know if a novel treatment or service delivery mode, that might be less costly, safer, and/or more convenient is at least as effective as another more well established treatment with proven effectiveness. Traditional significance testing designs and analyses cannot sufficiently address this latter question. Thus, there is a need for development and implementation of alternative conceptual research models to address research questions that involve hypotheses of "no meaningful differences" between examined outcomes.

Noninferiority and equivalence designs are becoming increasingly common and relevant in randomized controlled trials (RCTs) in the fields of medicine and mental health. Noninferiority designs are a one-sided test used to determine if a novel intervention is no worse than a standard intervention. Equivalence designs, a two-sided test, pose a similar question, but also allow for

Correspondence: Carolyn Greene, National Center for PTSD- Pacific Islands Division, Department of Veterans Affairs Pacific Islands Healthcare System; 3375 Koapaka St. Ste. I-560 Honolulu, HI 96819; (808)566−1651 (telephone); (808) 566−1885 (fax); E-mail: Carolyn.Greene2@va.gov..

the possibility that the novel intervention is no better than the standard one. Issues related to the comparability between novel interventions and better established ones are of great importance to improving health service delivery across a wide range of settings and contexts. However, noninferiority and equivalence designs are often poorly understood, improperly applied, and subsequently misinterpreted. These designs differ from traditional "superiority" trials in significant conceptual and statistical ways and pose a unique set of challenges to investigators and consumers of research. In this paper, we identify: design concepts, issues, and special considerations of noninferiority and equivalence trials; how they differ from standard superiority trials; when the designs are appropriate for use; and key statistical considerations. We also discuss the importance of these designs to mental health research, review common misapplication of the designs in published studies and provide recommendations for proper application.

The first papers utilizing noninferiority or equivalence designs began appearing in the late 1970s (e.g., Dunnett & Gent, 1977). Early equivalence and noninferiority studies were primarily designed to test the bioequivalence of generic forms of established medications. New, less expensive generic formulations of medications were compared to the standard, brand-name pharmaceuticals. In these studies, if bioequivalence was demonstrated, therapeutic equivalence was indirectly demonstrated (Garrett, 2003). In other words, there was no need to conduct a full evaluation of all the therapeutic properties of the medication as long as key biological properties were comparable to those of the standard treatment. Although bioequivalence studies are still widely conducted, noninferiority and equivalence designs have evolved and are being extensively used in certain areas of medical research such as cardiology, dentistry, oncology, and health care service delivery to directly demonstrate therapeutic equivalence. These applications often have more complex parameters of comparison including ease of dosing, side-effect profile, degree of discomfort, recovery time, accessibility of treatment, and tolerability.

Description/Explanation of Noninferiority or Equivalence Designs

"Noninferiority" and "equivalence" are often used interchangeably to refer to a trial in which the primary objective is to show that the response to the novel intervention is as good as the response to the standard intervention. In actuality, there is an important distinction between the two terms. Part of the confusion regarding these designs stems from ambiguous terminology. The International Conference on Harmonisation (ICH) defines an equivalence trial as a trial designed to show that two interventions do not differ in either direction by more than a pre-specified unimportant or insignificant amount (i.e., a two-sided test), whereas a noninferiority trial is designed to show that the novel treatment is no less than a certain amount from the standard intervention (i.e., a one-sided test) (ICH, 1998). In both cases, that small amount of allowable difference is the margin that defines the "zone of indifference" within which the interventions are considered equivalent or noninferior, respectively (Blackwelder, 1982). Figure 1 provides an illustration of this zone. Equivalency designs are rarely used in therapeutic trials evaluating effectiveness since the study objective often is to show that a new treatment is not inferior to a standard, which corresponds to a noninferiority design, as opposed to showing that the novel treatment is neither inferior nor superior to the standard, which corresponds to the equivalency design. Because noninferiority trials are much more common in mental health research, this article focuses on that design.

Noninferiority designs differ from standard superiority trials in several fundamental ways. The null hypothesis of standard superiority trials is that there is no true difference between the interventions. In other words, unless strong evidence is found indicating the superiority of one intervention over the other, the default conclusion is that no treatment difference exists. A Type I error is erroneously concluding a treatment difference that is unlikely to actually exist (thus

erroneously rejecting the null hypothesis). A Type II error is a failure to reject the null hypothesis when in fact the alternative is true (thus erroneously missing an actual difference in treatments). In contrast, noninferiority trials have a null hypothesis that the experimental treatment is inferior to the standard treatment by at least a certain prespecified amount (δ): H0 : μStandard Treatment - μExperimental Treatment ≥δ. The alternative hypothesis to be proven is that the experimental treatment is inferior to the standard treatment by less than that same amount (δ): H1 : μStandard Treatment - μExperimental Treatment <δ. Thus, the definitions of the type I and type II errors are reversed. The type I error for a noninferiority trial is the probability of erroneously concluding noninferiority when evidence is lacking that the novel treatment truly is inferior to the control and a type II error is the probability of erroneously failing to reject the null hypothesis when the novel treatment is truly noninferior.

When to Use the Designs

Noninferiority designs can be of value when a novel treatment has been developed that is easier to use, has fewer side effects, is less costly, and/or provides increased access to care compared to the standard treatment. If the new treatment offers these more appealing characteristics, then the research goal may be to show that its effectiveness is not much less than that of the standard treatment. A key methodological consideration that we elaborate upon later is the choice of the standard treatment. Neither an equivalence nor a noninferiority design can be appropriately used without a consistently effective treatment to serve as the active control.

Although the existence of a standard treatment is necessary for the use of these designs, it is not sufficient justification. There must be a genuine question as to the equivalence or noninferiority of the new treatment. If the new treatment is expected to be better than the standard treatment, a superiority trial is called for.

Importance to Mental Health Research

Within recent years, there has been a significant increase in the number of effective, well validated treatments for individuals suffering from traumatic stress disorders. Noninferiority designs can be used to evaluate modifications to these interventions, alternate delivery modalities, and applications to special populations. A noninferiority conceptual approach is also relevant to treatment dismantling studies, comparative outcome studies, cost studies, and efforts to understand mechanisms of action in mental health treatments. These designs may play a critical role in expanding access to evidence based treatment of traumatic stress.

Collectively, the authors are working on a number of randomized clinical trials for mental health and medical interventions that incorporate noninferiority designs. One example is an ongoing noninferiority trial examining the clinical effectiveness of using a telemental health service delivery mode to provide a cognitive-behavioral therapy for treating combat related PTSD (Morland et al., 2008). For purposes of illustration, we will describe and discuss elements of the project as one example of the appropriate application of the noninferiority design.

The Department of Defense's office of Congressionally Directed Medical Research Programs has recently funded this 4-year prospective clinical trial evaluating the effectiveness of a telemental health modality, specifically videoteleconferencing, as a means of providing a well validated PTSD specific treatment intervention, Cognitive Processing Therapy (CPT; Resick & Schnicke, 1992) to veterans in remote locations, as compared to the traditional face-to-face modality. In this ongoing study we hypothesize that using a telemental health modality will be non-inferior to the traditional mode of service delivery (face-to-face), on both clinical and process measures, for providing CPT in a group format for veterans with PTSD. In this study we are conducting 9 cohorts of CPT groups over a 3-year period. Each cohort includes a telemental health condition and a traditional face-to-face condition. The noninferiority design

was selected as the best way to determine if the novel, less expensive telemental health modality is as good as the traditional face-to-face delivery mode for the delivery of CPT. The prespecified noninferiority margin or "clinically meaningful difference" was based on both clinical and statistical justifications. Both confidence interval and hypothesis testing approaches will be used in the final analysis. If our final data analyses show that the difference in clinical effectiveness (significant change in PTSD symptoms on the Clinician Administered PTSD Scale [CAPS; Weathers, Keane, & Davidson, 2001] between the two conditions) is less than the prespecified noninferiority margin, we can conclude that telemental health delivery is not inferior to in-person deliver of CPT (i.e., we can reject the null hypothesis of inferiority). With a finding of noninferiority, one could be confident in recommending the novel, more easily accessible telemental health treatment modality in lieu of the standard face-to-face treatment without any significant loss of clinical benefit. It is expected that results from this project can be applied to other sites where specialized PTSD clinical services are needed but unavailable due to geographic barriers.

Key Methodological Issues

The methodological challenges in conducting a noninferiority study are significant; especially since poor trial design and execution can erroneously suggest a similarity. Challenges include (a) choice of an active control treatment, (b) choice of a noninferiority margin, (c) sample size estimation, and (d) statistical analysis.

As previously mentioned, these designs cannot be used without a well-established standard treatment to use as the active control (ICH, 2001). There must be convincing prior evidence of the effectiveness of the active control compared with placebo; its effectiveness must be consistently demonstrated (Blackwelder, 2004*).* It must be clear that the active control is effective in the specific application, ideally with the specific population, used in the current study. The conditions of the trial (e.g., setting, dose, duration) should not unfairly favor one treatment over another (Hwang & Morikawa, 1999). Also, it must be truly unknown if one treatment is inferior to the other (Djulbegovic & Clarke, 2001). This can be a difficult standard to meet in studies with underserved populations in which the actual standard care is little or insufficient care (e.g., telemental health trials with rural patients). For example, in our ongoing trial the true "standard of care" in many of our rural locations is minimal to no care. However for the purpose of the trial, we needed to choose an active control with a well established evidence base. Therefore we selected the face-to-face delivery of CPT services as our active control.

The choice of noninferiority margin is another critical decision. Unfortunately, there is no gold standard criterion for determining an appropriate margin (W.L. Greene, Concato, & Feinstein, 2000). In fact, the only consistent recommendations from regulators are that the margin is determined in advance and that it should not be greater than the smallest effect size the active drug would be reliably expected to have compared with a placebo (Hwang & Morikawa, 1999; ICH, 2001). If the margin is too large, rejecting the null hypothesis is meaningless; but if the margin is too small, power to detect noninferiority is dramatically reduced (Wiens, 2002). Some researchers prefer to derive the margin based on statistical properties. That approach often leads to noninferiority margins that are relative to effect size. Typically it is a fraction, usually one half or less, of the historical effect size of the standard intervention (Temple & Ellenberg, 2000). It can also be a percentage of the effect of the standard treatment in the current trial. For example, the novel intervention must be at least 80% as effective as the standard intervention. Perhaps the most common approach in treatment outcome studies is to set a margin based on what is considered "clinically unimportant."

Conceptually, if the experimental treatment is almost as good as the standard treatment and there is only a trivial or unimportant difference between the two, then that margin has to be smaller than an amount that would make a difference clinically. This is quantified by taking the smallest value that would be clinically meaningful and using this as the margin of noninferiority. For example, in our ongoing CPT study with a population of combat veterans with PTSD, our noninferiority margin or the minimum clinical meaningful difference between the two conditions of interest on the primary clinical outcome measure is a decrease of 10 points on the CAPS. Thus, a difference smaller than this magnitude would not be enough to be considered clinically meaningful. This prespecified difference is both clinically and statistically established (Schnurr, et al., 2003). Therefore, our margin is set such that the mean score reduction of the experimental treatment could not be more than 10 points lower than that of the standard treatment. The most rigorous approach is to utilize both clinical and statistical judgment (Wiens, 2002). Whatever criterion is chosen, the margin and its derivation should be explicitly reported in the manuscript (Piaggio, Elbourne, Altman, Pocock, & Evans, 2006).

It is generally not enough for a noninferiority study to merely show that the novel treatment is not inferior to a standard treatment. Rather, it is preferable that the study be able to demonstrate that both of the treatments are actually effective. Even when the standard treatment has a strong history of effectiveness, there are many trial-specific factors that could make the standard treatment ineffective such as choice of study population, treatment setting and choice of primary outcome. The most methodologically rigorous approach is to include a placebo or wait list condition to confirm that the actual control (i.e., standard treatment) and novel treatment are both superior to placebo (Hwang & Morikawa, 1999). When a placebo is not feasible, as is frequently the case in psychotherapy research, every other aspect of the study design should be as similar as possible to the previous trials establishing the effectiveness of the active control (ICH, 2001; Temple & Ellenburg, 2000).

As with superiority trials, the required sample size depends upon the specific analyses and variables to be used in the trial. Power calculations must take into account the noninferiority margin and Type I and II errors (Julious, 2004). Regulatory authorities recommend the use of a 95% confidence interval, which corresponds to a Type I error value of 0.025 (Lewis, Jones, & Rohmel, 1995). In order to achieve power no less than 80%, Type II error can be no less than 0.20. We have frequently heard researchers voice a misconception that equivalence and noninferiority studies necessitate extremely large sample sizes and as a result are virtually impossible to conduct. However, investigators should not be discouraged from pursuing an equivalence or noninferiority design based on this erroneous assumption. Although these studies may require large sample sizes, the size of any trial (superiority or other) is dependent upon the primary objective, the choice of error rates, and the differences one is expecting to detect. For our ongoing study, we considered that in using a noninferiority design the consequence of a Type II error is the same as the consequence of a Type I error in traditional studies and adjusted values accordingly, in order to have a resultant power of 0.90.

The statistical analysis of noninferiority can be conducted by either using confidence intervals or by applying variations to the analytic strategy of null hypothesis testing. Under conventional hypothesis testing in a comparative study of two interventions, the goal is to reject the null in favor of a difference between the two interventions. If this approach is extended to a noninferiority study, then can noninferiority be claimed when the test fails to reject the null hypothesis of no difference? Some argue that it is acceptable assuming that a strict level of Type II error (failing to reject the null when it is false) is provided (Jennison & Turnbull, 2000; Ng, 1995). However, others argue that it is logically impossible to conclude noninferiority on the basis of failing to reject the null hypothesis (Blackwelder, 1982; Dunnett & Gent, 1977; Jones, Jarvis, Lewis, & Ebbutt, 1996). Rather, failure to reject the null means

that there is not sufficient evidence to accept the alternative hypothesis. Alternately, if the null hypothesis states that the true difference is greater than or equal to the pre-specified noninferiority margin, then failing to reject it can be interpreted as insufficient evidence to conclude that the difference of the two procedures is less than the pre-specified noninferiority margin.

The statistical analysis plan for our ongoing trial is to employ a multilevel (also called hierarchical or random effects) modeling procedure for a non-inferiority analysis on our primary outcome measure (change in PTSD symptoms on the CAPS). The magnitude of the difference in the means (effect sizes), as estimated by the confidence intervals, will provide useful clinical information and will allow a clinical judgment relative to the clinical non-inferiority of the two modes of delivery. In a multilevel design such as this one, power is influenced not only by the number of participants, the effect size, and the α-level, but also by the number of clusters (i.e., the number of cohorts) and the effect size variability (measured in our study as the effect size variance across cohorts). High values of the effect size variability would result in reduced power, but because care will be taken to ensure that all treatment sessions will undergo the same procedures, there is no reason to expect high effect size variability.

To avoid misinterpretation of null hypothesis testing, some investigators, including those who contributed to the CONSORT guidelines, favor the confidence interval approach to show noninferiority of two treatments (Durrleman & Simon, 1990; Jones et al., 1996; Makuch & Simon 1978; Piaggio et al., 2006). The width of the interval signifies the extent of noninferiority, which is a favorable characteristic of this approach. If the confidence interval for the difference between two interventions lies to the right of the noninferiority margin (Figure 1) then noninferiority can be concluded. If the interval crosses the boundary (contains the value of the margin) then noninferiority cannot be claimed. Some investigators prefer the confidence interval approach to examine the precision of noninferiority. Others prefer hypothesis testing. The authors recommend that both confidence interval and *p*-values be provided to allow the audience to interpret the extent of the findings.

ICH E10 (2001) and CONSORT (Piaggio et al., 2006) guidelines recommend that noninferiority trials conduct an intent-to-treat (ITT) analysis including all participants who were randomized into groups, regardless of their actual participation in treatment. It should be noted that although the importance of ITT analysis in a traditional superiority trial is well established, the role of the ITT population in a non-inferiority trial is not equivalent to that of a superiority trial (Brittain & Lin, 2005). An ITT analysis in a superiority trial tends to reduce the treatment effect, minimizing the difference between groups—in essence favoring the null hypothesis of no difference. This is a conservative way to view the results. However, in the non-inferiority setting, because the null and alternative hypotheses are reversed, a dilution of the treatment effect actually favors the alternative hypothesis, making it more likely that true inferiority is masked. An alternative approach is to use a per-protocol population, defined as only participants who comply with the protocol. The per-protocol analysis can potentially have bias as well. Because non-completers are not included, it can distort the reported effectiveness of a treatment. If, for example, a significant percentage of participants dropped out of the experimental treatment because they felt it was ineffective, the per-protocol analysis would not adequately capture that. Another approach is to use a modified ITT which excludes participants who never actually received the treatment, but includes non-compliant participants who started the treatment but did not fully complete it. The most rigorous approach is to use both a per-protocol analysis and an ITT analysis with the aim of demonstrating noninferiority with both populations (Jones et al., 1996). Regardless of the approach used, the reported findings should specify and fully describe the type analysis population.

Use of the Noninferiority and Equivalence Designs in Published Studies

Noninferiority conclusions are frequently drawn in studies that did not properly utilize an appropriate design. In 2000, W.L. Greene and colleagues conducted a literature review of human research papers across a wide range of disciplines and topics claiming equivalence. Eighty-eight eligible papers were selected. Thirty-five came from Abridged Index Medicus journals such as Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association, The Lancet, and New England Journal of Medicine. The papers were evaluated for 5 core methodological attributes: (1) statement of research aim, (2) magnitude of reported difference, (3) choice of quantitative boundary, (4) method of statistical testing, and (5) calculation of sample size. The authors' findings revealed major problems with the misuse of these designs. Only 22% of the 88 papers evaluated met all 5 criteria. Only 51% of the 88 studies were specifically designed to test equivalence. Only 51% of the 88 studies were specifically designed to test equivalence. Other studies either tested superiority or did not state the research aim. The quantitative designation of equivalence ranged from 0−21% for absolute differences and from 0−76% for relative differences. Only 23% of the reports set an equivalence boundary and confirmed it with an appropriate statistical test. Sixty-seven percent of the reports declared equivalence after a failed test for superiority. In 10% of the reports, the claim of equivalence was not statistically evaluated. The sample size needed to confirm results had been calculated in advance for only 33% of the reports (n = 20 patients per group or fewer in 25% of reports).

Le Henanff, Giraudeau, Baron, and Ravaud (2006) conducted a similar review after the Greene study and the issuance of ICH and other guidelines. They evaluated 162 studies (116 noninferiority and 46 equivalence) from a diverse array of fields using 4 specific requirements: (1) noninferiority or equivalence margins defined, (2) sample size calculation taking the margin into account, (3) conducted both ITT (or modified ITT) and per-protocol analyses, and (4) used confidence intervals to report results. They found that the overall quality of reporting had improved; however, there were still major problems. Only 20.3% fulfilled all 4 of the requirements identified. Among the subgroup of reports that fulfilled all of the requirements, 12.1% had misleading conclusions in which they claimed equivalence or noninferiority although the results were inconclusive.

Conducting a comprehensive critical review of mental health studies that have used an equivalence or noninferiority design is outside the scope of this paper. As but one of numerous examples, Frueh et al. (2007) reported findings from a randomized trial, concluding that results provided "preliminary support for the hypothesis that PTSD treatment delivered via telepsychiatry is as efficacious as traditional care (p. 145)." Cautious though their conclusions were (e.g., "preliminary support"), this study was not powered or designed specifically to conduct equivalence or non-inferiority analyses, relying instead on traditional analyses for significance testing. Thus, results were not optimal to support conclusions of equivalence between the two treatment conditions.

In the interest of assisting readers' understanding of these designs, we reviewed the published literature and identified four studies that serve as excellent examples. The studies contained explicit explanations of the methodology used, presented the necessary sample size parameters, and conducted appropriate analyses. One study evaluated a telephone administered cognitive behavior therapy for treatment of OCD in comparison to the face-to-face delivery (Lovell et al., 2006). Another study tested the noninferiority of telepsychiatry consultations and follow-up compared to those conducted in-person (O'Reilly et al., 2007). A third study compared Repetitive Transcranial Magnetic Stimulation to Electroconvulsive Therapy for treatment of severe depression (Eranti, et al., 2007). The fourth investigated the effectiveness of standard

treatment of depression in primary care clinics with versus without antidepressant medication (Hermens et al., 2007).

It should be noted that even among these strong studies there were some weaknesses. The use of terminology, particularly use of “noninferiority” versus “equivalence,” was inconsistent. In some cases, the terms seemed to be used interchangeably. Also, in the choice of an active control treatment, some studies did not clearly establish that the standard treatment had previously been well validated as effective. Consistent documentation of obtained power was not always provided either.

Conclusions

With the ongoing development of new approaches to administering standard treatments, noninferiority and equivalence designs will continue to become more common in mental health research. The increasing utilization of these designs to expand access to evidence based treatment of traumatic stress necessitates improved understanding of the methodological challenges and issues. It is important for both researchers and consumers of research to be aware of how these designs are implemented and what conclusions may be appropriately drawn from findings. As a researcher, one must pay careful attention to methodological rigor including choice of active control, noninferiority margin, trial conduct, and analytic methods. The study methods must be pre-determined through a systematic process of decision making and analysis. As a consumer of research, it is important to assess the accuracy of claims of noninferiority or equivalence.

Acknowledgements

This work was supported in part by grants Tel 03−080−3 and IIR-04−421−3 from the Department of Veterans Affairs, Office of Research and Development, Health Services R&D Service and MH074468 from the National Institute of Mental Health. Support was also provided by VA Pacific Islands Health Care System, Spark M. Matsunaga Medical Center and VA National Center for PTSD. All views and opinions expressed herein are those of the authors and do not necessarily reflect those of our respective institutions or the Department of Veterans Affairs. We acknowledge the valuable comments provided by Rebecca G. Knapp on an earlier draft of this essay.

References

Blackwelder WC. Current issues in clinical equivalence trials. Journal of Dental Research 2004;83(Spec Iss C):C113–C115. [PubMed: 15286135]

Blackwelder WC. “Proving the null hypothesis” in clinical trials. Controlled Clinical Trials 1982;3:345–353. [PubMed: 7160191]

Brittain E, Lin D. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. Statistics in Medicine 2005;24:1–10. [PubMed: 15532089]

Djulbegovic B, Clarke M. Scientific and ethical issues in equivalence trials. Journal of the American Medical Association 2001;285:1206–1208. [PubMed: 11231752]

Dunnett CW, Gent M. Significance testing to establish equivalence between treatments with special reference to data in the form of 2 × 2 tables. Biometrics 1977;33:593–602. [PubMed: 588654]

Durrleman S, Simon R. Planning and monitoring of equivalence studies. Biometrics 1990;46:329–336. [PubMed: 2194579]

Eranti S, Mogg A, Pluck G, Landau S, Purvis R, Brown RG, et al. A Randomized, controlled trial with 6-month follow-up of repetitive transcranial magnetic stimulation and electroconvulsive therapy for severe depression. American Journal of Psychiatry 2007;164:73–81. [PubMed: 17202547]

Frueh BC, Monnier J, Yim E, Grubaugh AL, Hamner MB, Knapp RG. A randomized trial of telepsychiatry for post-traumatic stress disorder. Journal of Telemedicine and Telecare 2007;13:142–147. [PubMed: 17519056]

Garrett AD. Therapeutic equivalence: fallacies and falsification. Statistics in Medicine 2003;22:741–762. [PubMed: 12587103]

Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: Are they supported by the evidence? Annals of Internal Medicine 2000;132:715–722. [PubMed: 10787365]

Hermens M, van Hout HPJ, Terluin B, J Adèr H, Penninx BWJH, van Marwijk HWJ, et al. Clinical effectiveness of usual care with or without antidepressant medication for primary care patients with minor or mild-major depression: A randomized equivalence trial. Biomed Central Medicine 2007;5:36.

Hwang IK, Morikawa T. Design issues in noninferiority/equivalence trials. Drug Information Journal 1999;33:1205–1218.

International Conference on Harmonisation. ICH Topic E9: Statistical principles for clinical trials. 1998 [March 8, 2006]. from, http://www.emea.europa.eu/pdfs/human/ ich/036396en.pdf

International Conference on Harmonisation. ICH Topic E10: Choice of control group and related issues in clinical trials. 2001 [March 8, 2006]. from, http://www.emea.europ a. eu/pdfs/human/ich/036496en.pdf

Jennison, C.; Turnbull, BW. Group sequential methods with applications to clinical trials. Chapman and Hall/CRC; Boca Raton, FL: 2000.

Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: The importance of rigorous methods. British Medical Journal 1996;313:36–39. [PubMed: 8664772]

Julious SA. Tutorial in biostatistics: Sample sizes for clinical trials with normal data. Statistics in Medicine 2004;23:1921–1986. [PubMed: 15195324]

Le Henanff AL, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. Journal of the American Medical Association 2006;295:1147–1151. [PubMed: 16522835]

Lewis JA, Jones DR, Rohmel J. Biostatistical methodology in clinical trials—A European guideline. Statistics in Medicine 1995;14:1655–1657. [PubMed: 7481201]

Lovell K, Cox C, Haddock G, Jones C, Raines D, Garvey R, et al. Telephone administered cognitive behaviour therapy for treatment of obsessive compulsive disorder: Randomised controlled non-inferiority trial. British Medical Journal 2006;333:883. [PubMed: 16935946]

Makuch RW, Simon RM. Sample size requirements for evaluating a conservative therapy. Cancer Treatment Reports 1978;62:1037–1040. [PubMed: 688245]

Morland LA, Greene CJ, Frueh B, Rosen C, Mauldin P, Chard K, et al. Telemental Health and Cognitive Processing Therapy for Rural Combat Veterans with PTSD. PT074516 −20063 funded by Congressionally Directed Medical Research Programs (CDMRP). 2008

Ng T. Conventional null hypothesis testing in active control equivalence studies. Controlled Clinical Trials 1995;16:356–358. [PubMed: 8582153]

O'Reilly R, Bishop J, Maddox J, Hutchinson L, Fisman M, Takhar J. Is telepsychiatry equivalent to face-to-face psychiatry? Results from a randomized controlled equivalence trial. Psychiatric Services 2007;58:836–843. [PubMed: 17535945]

Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. Journal of the American Medical Association 2006;295:1152–1160. [PubMed: 16522836]

Resick PA, Schnicke MK. Cognitive processing therapy for sexual assault victims. Journal of Consulting and Clinical Psychology 1992;60:748–756. [PubMed: 1401390]

Ruskin PE, Silver-Aylaian M, Kling MA, Reed SA, Bradham DD, Hebel JR, et al. Treatment outcomes in depression: Comparison of remote treatment through telepsychiatry to in-person treatment. American Journal of Psychiatry 2004;161:1471–1476. [PubMed: 15285975]

Schnurr P, Friedman M, Foy D, Shea M, Hsieh F, Lavori P, et al. Randomized trial of trauma-focused group therapy for posttraumatic stress disorder. Archives of General Psychiatry 2003;60:481–489. [PubMed: 12742869]

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 2000;133:455–463. [PubMed: 10975964]

Weathers FW, Keane TM, Davidson JRT. Clinician-Administered PTSD Scale: A review of the first ten years of research. Depression and Anxiety 2001;13:132–156. [PubMed: 11387733]

Wiens BL. Choosing an equivalence limit for noninferiority and equivalence studies. Controlled Clinical Trials 2002;23:2–14. [PubMed: 11852160]

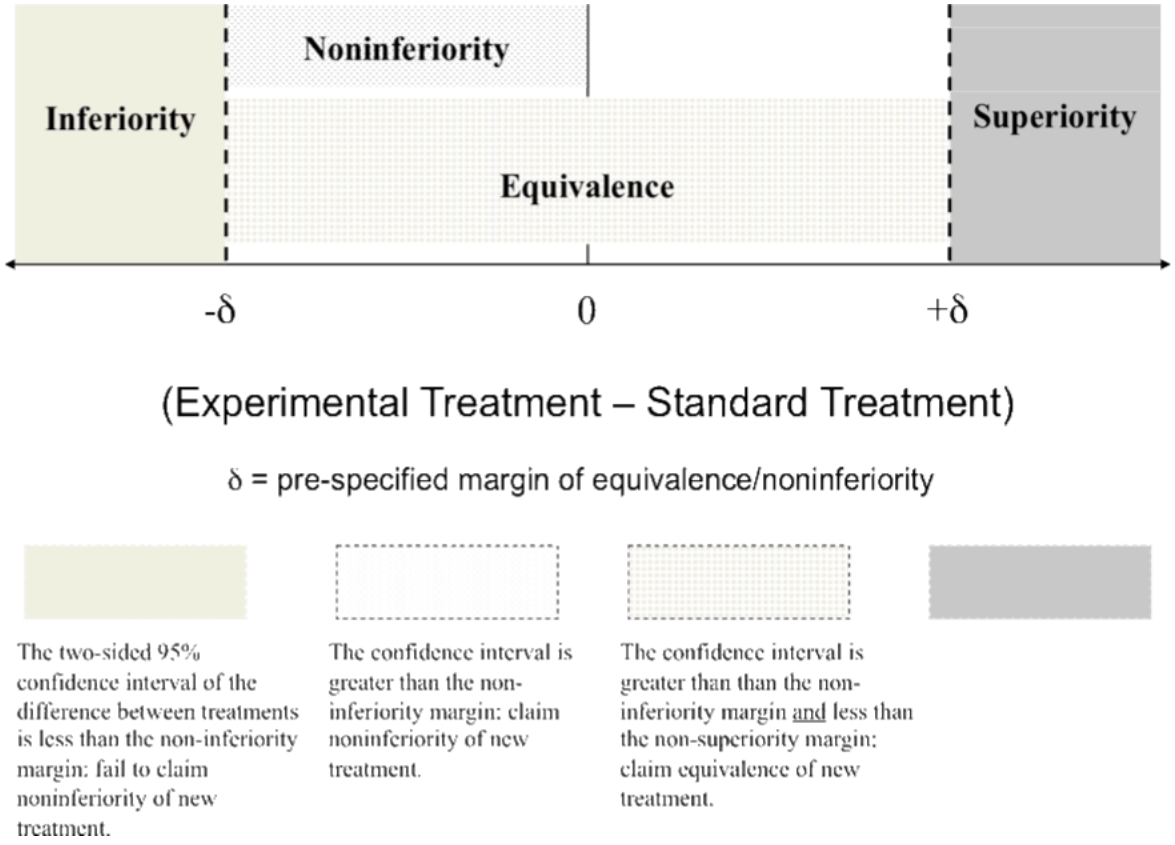

Figure 1.
Boundaries of equivalence and noninferiority for a 2-sided 95% confidence interval of the difference (Experimental Treatment – Standard Treatment).