THE ORTHOPAEDIC FORUM

"Not Statistically Different" Does Not Necessarily Mean "the Same": The Important but Underappreciated Distinction Between Difference and Equivalence Studies

Alex H.S. Harris, MS, PhD, Sara Fernandes-Taylor, PhD, and Nicholas Giori, MD, PhD

Researchers often want to evaluate whether a new medical or surgical treatment is equivalent to an existing treatment. The new treatment may be preferred if its results are equivalent to those of the existing approach in terms of complications or outcomes but it is superior in terms of ease of use, safety, or cost. However, many researchers are unaware that the equivalence of two interventions cannot be established by failing to find a statistical difference between them. This somewhat subtle detail of research design and statistical analysis has very important clinical implications. The goal of this brief paper is to explain the distinctions between the familiar difference trial (or superiority trial) study design and the often more appropriate but much less familiar equivalence study design. These designs have different underlying hypotheses, power calculations, statistical analyses, and conclusions. Claiming the equivalence of two interventions on the basis of a nonsignificant difference in the results of a difference trial and analysis, as is unfortunately common, may lead to incorrect conclusions and inappropriate changes in clinical practice[1,2]. Erroneous equivalence claims in the medical literature have been reviewed by Greene et al.[3].

The examples presented in the present paper hinge on two fundamental statistical concepts that warrant description, specifically type-I and type-II errors[4]. A type-I error, or false positive, is an error in which the null hypothesis is rejected when, in fact, the null hypothesis is true. For example, a type-I error has occurred if researchers declare that two treatments produce different outcomes when, in reality, no difference exists. The probability of a type-I error is denoted by α. A type-II error, or false negative, occurs when researchers fail to reject the null hypothesis when, in fact, the null hypothesis is false. For example, a type-II error has occurred if researchers declare that two treatments have no difference in outcome when, in reality, a difference does exist. The probability of a type-II error is denoted by β.

Difference Trials

The difference trial (or superiority trial) is most familiar to researchers and serves as the basis for the explanation of type-I and type-II errors outlined above. The purpose of a difference trial is to examine whether one intervention is different from (usually better than) a comparison or control intervention. The null hypothesis of the difference trial is that the mean outcome for patients exposed to intervention 1 is the same as the mean outcome for patients exposed to intervention

Disclosure: None of the authors received payments or services, either directly or indirectly (i.e., via his or her institution), from a third party in support of any aspect of this work. One or more of the authors, or his or her institution, has had a financial relationship, in the thirty-six months prior to submission of this work, with an entity in the biomedical arena that could be perceived to influence or have the potential to influence what is written in this work. No author has had any other relationships, or has engaged in any other activities, that could be perceived to influence or have the potential to influence what is written in this work. The complete **Disclosures of Potential Conflicts of Interest** submitted by authors are always provided with the online version of the article.

Disclaimer: The views expressed herein are those of the authors and are not necessarily those of the Department of Veterans Affairs.

TABLE I Distinctions Between Difference and Equivalence Trials

	Difference Trial	Equivalence Trial
Null hypothesis	$Mean_1 - Mean_2 = 0$	$\lvert Mean_1 - Mean_2 \rvert \geq d$, where d is the prespecified threshold for "clinically meaningless"
Alternative hypothesis	$Mean_1 \neq Mean_2$	$\lvert Mean_1 - Mean_2 \rvert < d$
Basis of the power analysis	Smallest interesting difference	Biggest difference that would be clinically meaningless
Simple analysis method	Independent-sample t test	Confidence interval method or TOST (two one-sided tests)

2 (H0: $Mean_1 - Mean_2 = 0$). The alternative hypothesis of the difference trial is that the mean outcome for patients exposed to intervention 1 is different from the mean outcome for patients exposed to intervention 2 (HA: $Mean_1 - Mean_2 \neq 0$). In other words, $Mean_1 - Mean_2$ is statistically different from zero. To conduct a power analysis, the researcher needs to specify the smallest difference between $Mean_1$ and $Mean_2$ (the effect size) that would be of clinical interest. The smaller the effect that the researcher wants to be able to detect, the larger the required sample size[5].

In the simplest case, patients are randomized to the two interventions and the outcomes can be compared with use of an independent-sample t test. If the difference between the group means is statistically different from zero (at some prespecified α, e.g., 0.05), then the researcher claims this as evidence against the null hypothesis and in support of the alternative hypothesis. However, if the means are not statistically different from each other (e.g., $p > 0.05$), the researcher can claim only that no evidence of a difference was found, not that the interventions are the same. Failure to find a difference is not the same as establishing equivalence[6,7].

Equivalence Trials

The equivalence trial is different from the difference trial. First, the goal of the equivalence trial is often to establish that one treatment is clinically equivalent to another treatment in terms of a particular outcome (e.g., complications, postoperative patient function, or mortality). Evidence that the interventions are equivalent might be meaningful because the new intervention might possess additional benefits such as lower cost, improved safety, or greater ease of use. Often, the one-sided version of this design (the noninferiority trial) is used to assess whether the new intervention is "at least as good as" the old intervention.

The distinctions between the difference trial and the equivalence trial are summarized in Table I. For the equivalence trial, the null hypothesis is that $\lvert Mean_1 - Mean_2 \rvert \geq d$, where d is a prespecified threshold equal to the largest difference that is still considered to be "clinically meaningless." Deciding on this threshold (d) is a difficult and critically important aspect of the equivalence design. It serves as the operational definition of "the same as." The alternative hypothesis is that $\lvert Mean_1 - Mean_2 \rvert < d$, or that the confidence interval for the difference between the means falls within the prespecified threshold.

The power analysis for the equivalence trial is based on d, the prespecified threshold for what is considered to be "clinically meaningless." The smaller the value of d, the larger the sample size will need to be. In the simplest case, patients are randomized to the two interventions and the outcomes can be compared with use of a confidence interval method or an equivalence test with a prespecified value for α (e.g., 0.05).

There are two basic methods for evaluating the equivalence hypothesis. The first and more complicated method is a "two one-sided tests" (TOST) procedure in which the two null hypotheses are $H0_1$: $Mean_1 - Mean_2 \leq -d$ and $H0_2$: $Mean_1 - Mean_2 \geq d$, and the alternative hypotheses are HA_1: $Mean_1 - Mean_2 > -d$ and HA_2: $Mean_1 - Mean_2 < d$. If each of the one-sided null hypotheses is rejected with use of a one-sided t test at the prespecified α, then we have support for the equivalence hypothesis that $-d < Mean_1 - Mean_2 < d$. Procedures for testing equivalence with use of a TOST procedure have been implemented in some statistical software packages such as SAS (SAS Institute, Cary, North Carolina) and R (R Foundation, Vienna, Austria).

A substantially simpler, and therefore more commonly used, method[8] involves constructing the confidence interval for the difference between the means and checking whether it falls completely within the interval from –d to d. A properly constructed confidence interval for the parameter of interest allows the researcher to perform the hypothesis testing merely by inspection and gives much richer information regarding the range of plausible values for the difference. However, one must understand how to choose the level for the confidence interval, and misinformation abounds. Here is the simple rule, which is outlined in more detail elsewhere[9]: When the underlying hypotheses involve one-sided tests, as is true for equivalence and noninferiority designs, then the corresponding confidence intervals should be at the $(1 - 2\alpha)$ level. Therefore, if you want to be 95% sure not to commit a type-I error (i.e., $\alpha = 0.05$), then a 90% confidence interval should be used in the analyses. If you want to be 97.5% sure not to commit a type-I error (i.e., $\alpha = 0.025$), then a 95% confidence interval should be used. When the underlying hypotheses involve two-sided tests, as is true for most difference designs (e.g., H0: $Mean_1 - Mean_2 = 0$) but not for equivalence designs, then the corresponding confidence intervals should be at the $(1 - \alpha)$ level.

Regardless of the α level that is chosen, if the relevant confidence interval for the difference between the means falls entirely within the zone of equivalence (–d to d), the null hypothesis is rejected and the researcher claims this as evidence in support of the alternative hypothesis, which states that the

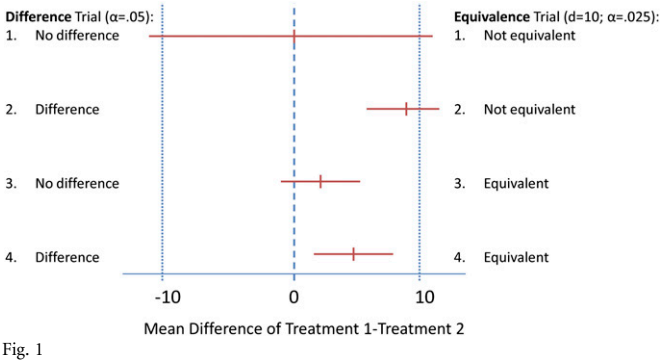

Fig. 1

Interpretation of four confidence intervals for a difference trial (H0: $M_1 - M_2 = 0$; HA: $M_1 - M_2 \neq 0$) and for an equivalence trial (H0: $|M_1 - M_2| > 10$; HA: $|M_1 - M_2| \leq 10$). H0 is the null hypothesis, HA is the alternate hypothesis, M_1 and M_2 are the means for the two interventions, α is the significance level, and d is the equivalence threshold.

means are equivalent within the threshold d. If the null hypothesis is not rejected, however, the researcher has failed to find evidence of equivalence. This is different from finding that the interventions are significantly different.

Figure 1 helps to illustrate these distinctions. It shows four 95% confidence intervals for the difference between the means of two interventions.

The top 95% confidence interval represents a comparison in which a researcher running a difference trial (with $\alpha = 0.05$) and analysis would conclude that there is no evidence of a difference between the means because the interval includes zero. However, a researcher running an equivalence trial (with $\alpha = 0.025$) and analysis with an equivalence threshold of $d = 10$ would conclude that there is no evidence of equivalence because the interval extends beyond –10 and 10. This scenario exemplifies the problem with failing to find a statistically significant difference with use of a t test and claiming equivalence; an interval that contains both zero (i.e., no difference) and values that represent a clinically meaningful difference (i.e., no equivalence) provides evidence for neither equivalence nor difference. Many small-sample studies in the orthopaedics literature produce similar results.

The second 95% confidence interval represents a comparison in which a researcher running a difference trial and analysis would reject the null hypothesis and claim evidence of a difference between the means because the interval does not include zero. However, a researcher running an equivalence trial and analysis with an equivalence threshold of $d = 10$ would conclude that there is no evidence of equivalence because the interval extends beyond 10.

For the third 95% confidence interval, a researcher running a difference trial and analysis would fail to reject the null hypothesis because the interval includes zero. A researcher running an equivalence trial and analysis with an equivalence threshold of $d = 10$ would reject the null hypothesis and claim evidence of equivalence because the interval lies entirely between –10 and 10. Under this scenario, researchers who have used a difference trial analysis, found no evidence of a difference, and erroneously claimed equivalence have just happened to reach the correct conclusions by using the wrong analysis, including the unintentional application of a more stringent α of 0.025 rather than 0.05. If their confidence interval had a larger range that extended above 10 (e.g., the first interval in the figure), their conclusion would have been wrong.

The bottom 95% confidence interval does not contain zero and lies between –10 and 10. The null hypotheses for both the difference and equivalence trials would therefore be rejected. This scenario shows that an effect can be statistically different from zero even though the interventions are clinically equivalent.

Noninferiority Trials ("Good Enough")

The goal of a noninferiority study is to evaluate whether the result of a new intervention ($Mean_1$) is at least as good as the result of another intervention ($Mean_2$); assuming that larger values are clinically better (e.g., survival, quality of life), the null hypothesis is H0: $Mean_1 - Mean_2 \leq -d$, and the alternative hypothesis is HA: $Mean_1 - Mean_2 > -d$. If α is set to 0.05, this analysis can be done by constructing a 90% confidence interval and checking that its lower limit is greater than –d.

Brief Example of an Equivalence Trial

In the following example of an equivalence trial, researchers compare a novel approach for treating a medical condition with a traditional approach. Since the novel approach has advantages over the traditional approach in terms of cost savings, the goal of the investigators is to determine whether the novel approach is equivalent to the traditional approach in terms of clinical efficacy. The outcome in this example is the patient score on a self-reported health-related quality of life measure with a scale of 0 to 100. Patients with the particular condition being studied are randomized to receive either (1) the novel treatment or (2) the traditional treatment. Clinical outcomes are assessed at a prescribed time point. The researchers decide that the largest difference that would still be clinically meaningless is 10 points on the 100-point quality of life measure (i.e., $d = 10$). The researchers also use published estimates of population-level scores on the outcome measure to estimate that the standard deviation of scores on the quality of life measure for this patient group should be 10.

The sample size for each group is determined so that the power, or $1 - \beta$, of the study is 0.90; i.e., the probability that the two treatments will be deemed equivalent if they are, in fact, equivalent is 0.90. The sample size calculations can be made with use of a Z-table and a calculator, as shown in Figure 2, or with use of statistical software, such as PASS 2008 (NCSS, Kaysville, Utah) or SAS, that contains routines for power analysis of equivalence tests. The required sample size is determined to be twenty-one per group or forty-two total.

The null hypothesis for the trial is H0: $|Mean_{novel} - Mean_{traditional}| \geq 10$. The alternative hypothesis is HA: $|Mean_{novel} - Mean_{traditional}| < 10$. The trial is conducted and yields findings of $Mean_{novel} = 53.8$ (standard error = 2.5) and $Mean_{traditional} = 56.3$ (standard error = 2.5). The difference between the means is therefore –2.5 (i.e., 53.8 – 56.3), and the standard error of the difference is 3.54 (i.e., the square root

To calculate the necessary sample size per group given:

Alpha (α) = .05

Desired Power (1-β) = .90

Outcome Standard Deviation (SD) = 10

Equivalence Threshold (d) = 3

The following formula can be used: $[Z(1-\alpha/2)+ Z(1-\beta)]^2 * 2 * [SD^2/ d^2]$, where Z(x) is the cumulative normal distribution function.

Thus,

$Z(1-\alpha/2)=z(0.975)= 1.959964$

$Z(1-\beta)=z(0.90)= 1.281552$

$N = (1.96 + 1.28)^2 * 2 * (10^2 / 3^2) = 233.50$

A sample size needed for each group in the trial is 234 (468 total).

Fig. 2

Calculation of the sample size requirement for an equivalence trial[10].

of $[2.5^2 + 2.5^2]$). The 90% confidence interval of the difference between the means equals the difference plus and minus 1.65 times the standard error of the difference. In this example, the confidence interval is 2.5 ± 5.83, or –8.33 to 3.33. The null hypothesis is rejected and the alternative equivalence hypothesis is supported because the interval is contained within –10 to 10 and therefore satisfies the prespecified definition of "same."

Some statistical programs have built-in TOST procedures that test the equivalence of means and calculate a p value. In our example, the associated p value is 0.02. (Data for this example and syntax for the TOST procedure in the R software package are available from the authors.)

Summary

Researchers need to be clear about their goals and hypotheses, recognizing that difference and equivalence are not mutually exclusive statistical opposites that can be inferred from each other. Some researchers may still wonder whether they can claim equivalence if they have failed to find a difference in a study that was powered to detect very small (clinically meaningless) effects. However, equivalence still cannot be claimed because the definition of equivalence is not directly used in the evaluation of the study hypothesis. Proper analyses and conclusions should flow from a thorough understanding of research aims. Failure to distinguish difference trials and equivalence designs is common, is potentially misleading, and may support inappropriate clinical practices. The understanding of these distinctions and their importance is the largest hurdle. Once researchers are clear about their goals and hypotheses, then the power analyses, statistical comparisons, and proper conclusions outlined here are simple to apply. As William W. Watt aptly stated, "Do not put your faith in what statistics say until you have carefully considered what they do not say."[11]

Alex H.S. Harris, MS, PhD
Sara Fernandes-Taylor, PhD
Nicholas Giori, MD, PhD
Center for Health Care Evaluation,
VA Palo Alto Health Care System,
795 Willow Road (MPD-152),
Menlo Park, CA 94025.
E-mail address for A.H.S. Harris: Alexander.Harris2@va.gov

References

1. Ware JH, Antman EM. Equivalence trials. N Engl J Med. 1997;337:1159-61.
2. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36-9.
3. Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? Ann Intern Med. 2000;132:715-22.
4. Petrie A. Statistics in orthopaedic papers. J Bone Joint Surg Br. 2006;88:1121-36.
5. Petrie A. Statistical power in testing a hypothesis. J Bone Joint Surg Br. 2010;92:1192-4.
6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;74:311.
7. Wellek S. Testing statistical hypotheses of equivalence and noninferiority. 2nd ed. New York: CRC Press; 2010.
8. Moseley JB, O'Malley K, Petersen NJ, Menke TJ, Brody BA, Kuykendall DH, Hollingsworth JC, Ashton CM, Wray NP. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med. 2002;347:81-8.
9. Steiger JH. Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychol Methods. 2004;9:164-82.
10. Julious SA. Sample sizes for clinical trials with normal data. Stat Med. 2004;23:1921-86.
11. Quotations about statistics. http://www.quotegarden.com/statistics.html. Accessed 09 Dec 2011.