Anesthesiology 2010; 113:779–81

Equivalence and Noninferiority Testing in Anesthesiology Research

"Absence of proof is not proof of absence."
—William Cowper, English poet.

CLINICAL studies frequently test whether one treatment or intervention is superior to another. When a test for superiority is statistically significant, we happily conclude that one method is better than the other, especially if it is in the expected direction! But what can it mean if the test is not significant? Can we conclude that the treatments are "similar" or "equivalent"? Actually, no, because although it could be that no clinically meaningful population difference exists, it is also possible that one does exist—but either the study was underpowered to detect it (sample size was too small) or we just got unlucky (false negative result). So from a nonsignificant test for superiority, we can really conclude only that no population difference was detected, and not equivalence, even if the observed means are very similar!

Fortunately, accepted methods of assessing and claiming equivalence between two randomized interventions do exist. They require an *a priori* definition of clinical "equivalence" between interventions, in the form of limits within which treatments are considered to be effectively the same. Equivalence is then claimed if the observed confidence interval for the difference between groups falls within the *a priori* defined equivalency region. Perhaps more commonly, however, the goal is to demonstrate that a new treatment is "as good as" or "not worse than" a standard. In such cases equivalence is not expected and superiority not needed; a third alternative, called a "noninferiority design," is best, because it allows one to claim noninferiority by refuting the null hypothesis that the preferred intervention is worse than the comparator.[1,2]

Suppose an intervention is known to have favorable intraoperative properties on certain key parameters but is suspected of adversely affecting other parameters. For example, Bala *et al.*[3] were interested in demonstrating that dexmedetomidine (*vs.* placebo) did *not* have a clinically important effect on evoked potentials in patients undergoing complex spinal surgery, because the drug had other known benefits. Because no effect (*i.e.*, no difference between groups) was the desired outcome, testing for superiority or noninferiority would not have addressed the research question. Rather, an equivalency trial was conducted, in which a clinically acceptable difference between groups on the primary outcome was specified *a priori*, and testing was performed to assess whether the true difference was within the equivalence region.

In an equivalence trial with a continuous outcome, we test the *null* hypothesis (H0) that the true difference between means is outside of the prespecified equivalency region, either below $-\delta$ or above $+\delta$ (*i.e.*, "not equivalent"), as H_0: $M_T - M_S \leq -\delta$ *or* $M_T - M_S \geq +\delta$, where M_T and M_S are the population means for *test* and *standard* treatments, respectively. The alternative hypothesis (H_1), which one wishes to conclude, is that the true difference lies between the specified limits (*i.e.*, "equivalent"), as H_1: $M_T - M_S > -\delta$ *and* $M_T - M_S < +\delta$.

Equivalence is claimed *only* if the treatment difference is concluded to be both significantly above the lower limit ($-\delta$) and significantly below the upper limit ($+\delta$) using a traditional one-sided test against a constant for each of the two components of H_1 (usually *t* tests if the outcome is continuous). Equivalence testing is thus referred to as "two one-sided tests" (or TOST).[4] If both tests are significant, the observed confidence interval (CI) for the difference will correspondingly fall within the equivalency region. It is then concluded that the true difference lies between $-\delta$ and $+\delta$, as in the first equivalency example (E) in figure 1. For the second equivalency trial result (F) in figure 1, the confidence interval is not within $\pm\delta$, and so the conclusion is that equivalence at the specified δ value cannot be claimed. In equivalence testing, we are given a bonus—no correction to the significance criterion for multiple comparisons is needed when performing the two one-sided tests because both must be significant (say, *P* less than 0.05) before equivalence can be claimed.

Sometimes it is more natural to express the equivalence region in terms of ratios of means than as an absolute value. For example, with the primary outcome of opioid consumption, researchers might have difficulty choosing an absolute number for an equivalency δ. Instead, they might *a priori* hypothesize δ to be, say, 0.90. Equivalence would imply that the true mean consumption for the *test* intervention was between 90 and 111% (1/90%) of the *standard* mean. When a ratio formulation is used, hypotheses are best specified on the log scale to create symmetric equivalency limits.[5,6] For a binary (yes/no) outcome, δ can be an absolute difference between proportions, relative risk, or odds ratio.[7]

Now suppose a new warming device can be made for one third the cost of an existing device, is known to have a better

Accepted for publication June 3, 2010. The author is not supported by, nor maintains any financial interest in, any commercial activity that may be associated with the topic of this article.

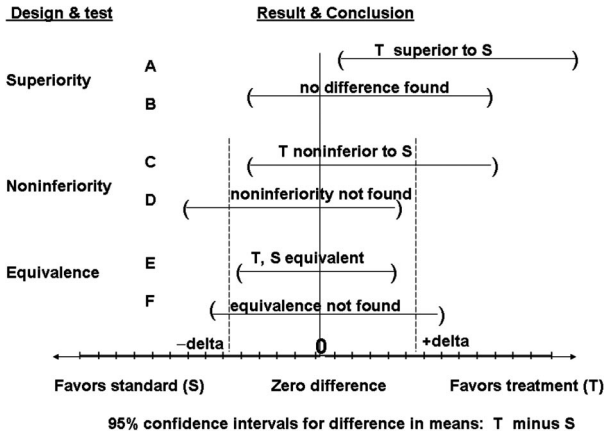


Fig. 1. Six clinical trial designs, confidence interval results and the corresponding conclusions, including two designed for superiority (*A* and *B*), two for noninferiority (*C* and *D*), and two for equivalence (*E* and *F*). The depicted conclusions can only be drawn if the study was designed to test the corresponding hypothesis (superiority, noninferiority, equivalence). An exception is that a significant test for noninferiority (as in trial *C*) can also be tested for superiority, because noninferiority includes the potential for superiority.

safety profile, and/or is much easier to use. Demonstrating superiority on rate of rewarming or intraoperative temperature of the new device over the existing one would be a luxury, as it would suffice to show that the new device was at least not worse than the existing device (*i.e.*, noninferior) on the main efficacy measure. Noninferiority designs are useful when the goal is to show that a new treatment is at least as effective as the standard (*i.e.*, equivalent or superior to), particularly when the new treatment is more favorable in other ways. They are essentially one-sided equivalency designs that test the null hypothesis that the preferred treatment is worse than the comparator by at least "δ," against the alternative (which a significant *P* value would conclude) that the preferred is "not more than δ worse than" or "at least as effective as" the comparator.[2] When higher values of the outcome are favorable, the null and alternative hypotheses are the same as the first components of the H_0 and H_1 statements, above, respectively, as H_0: $M_T - M_S \leq -\delta$ *versus* H_1: $M_T - M_S > -\delta$, and testing is the same as for the lower limit of an equivalence trial. When the noninferiority test is significant (*e.g.*, $P < 0.05$ for $\alpha = 0.05$), the lower limit of the CI for the difference between means correspondingly lies above $-\delta$, as in the first noninferiority example (C) in figure 1 (or below $+\delta$ if lower values are favorable). The second noninferiority example (D) in figure 1 shows a nonsignificant result because the lower limit is below $-\delta$.

For superiority designs, a two-sided superiority test significant in either direction corresponds to a CI that does not contain zero, as in the first example (A) in figure 1, where the test treatment is found to be superior to standard. The second superiority CI (B) contains zero, so the test must be nonsignificant, and we conclude that no difference was found. Equivalence cannot be claimed here because it was not tested and because no definition of equivalence is specified in the design of a superiority trial. Now notice that the second superiority CI (B) is identical to the first and significant, noninferiority CI (C). This prompts a question: in a trial designed for superiority, can researchers test for noninferiority after a nonsignificant test of superiority? No, a nonsignificant superiority test ends the testing, because further testing would increase the chance of type I error for which the trial was designed. However, it would be acceptable to assess superiority in a noninferiority trial after noninferiority had been established, because a significant noninferiority result implies potential superiority. In other words, additional testing to refine a statistically significant result is appropriate, but changing hypotheses to find statistical significance is not.

Choosing the equivalency δ is an integral part of study design and is best based on both clinical and statistical grounds. δ should be small enough to be of little clinical consequence, well within the range of background variability, and smaller than differences expected in superiority trials of an active treatment *versus* placebo. Too large a δ risks a claim of equivalence based on a clinically misleading δ, whereas too small a δ can waste sample size resources and make claiming equivalence unnecessarily difficult.[8–9]

Sample size calculation for an equivalence or noninferiority design is the same as for a one-tailed superiority trial powered to detect a difference equal to the chosen equivalency δ. However, because the equivalency δ is usually smaller than the superiority difference, a larger sample size is often needed. Often the sample size is calculated by using the postulated equivalency δ and assuming that the population difference is truly zero. If there are prior data and good intuition suggesting that the underlying difference is nonzero, the sample size may be calculated assuming a nonzero effect. For a noninferiority trial in which the underlying difference favored the preferred treatment, the sample size would be decreased.

In reporting results, studies designed to assess equivalency and noninferiority[10,11] should be clearly labeled as such. Choice of δ should be determined *a priori* and should be justified clinically; confidence intervals for the treatment difference should be presented in relation to δ.[12] In addition, treatments should be labeled "comparable" or "equivalent" only if formal tests for equivalence were done. Incorporating these readily available and widely accepted methods for assessing equivalency and noninferiority will strengthen clinical trial design and reporting.

Edward J. Mascha, Ph.D., Departments of Quantitative Health Sciences and Outcomes Research, Cleveland Clinic, Cleveland, Ohio. maschae@ccf.org

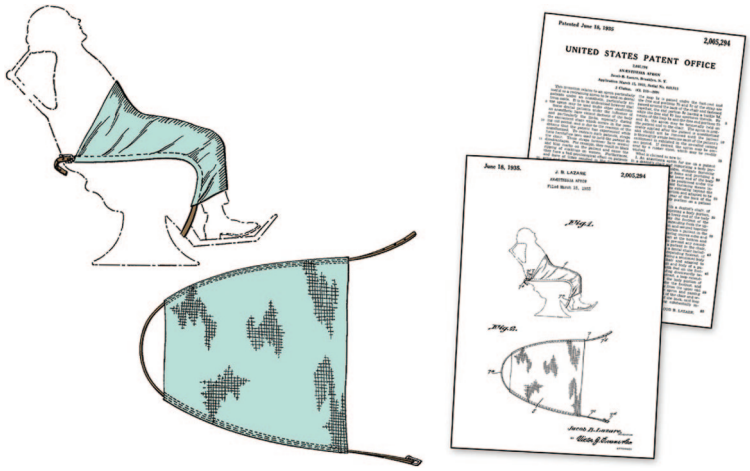
References

1. Tunes da Silva G, Logan BR, Klein JP: Methods for equivalence and noninferiority testing. Biol Blood Marrow Transplant 2009; 15:120–7
2. Blackwelder WC: "Proving the null hypothesis" in clinical trials. Control Clin Trials 1982; 3:345–53

3. Bala E, Sessler DI, Nair DR, McLain R, Dalton JE, Farag E: Motor and somatosensory evoked potentials are well maintained in patients given dexmedetomidine during spine surgery. ANESTHESIOLOGY 2008; 109:417–25
4. Schuirmann DJ: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 1987; 15:657–80
5. Berger RL, Hsu JC: Bioequivalence trials, intersection-union tests, and equivalence confidence sets (with discussion). Stat Sci 1996; 11:283–319
6. Laster LL, Johnson MF: Non-inferiority trials: The 'at least as good as' criterion. Stat Med 2003; 22:187–200
7. Laster LL, Johnson MF, Kotler ML: Non-inferiority trials: The 'at least as good as' criterion with dichotomous data. Stat Med 2006; 25:1115–30
8. Wiens BL: Choosing an equivalence limit for noninferiority or equivalence studies. Control Clin Trials 2002; 23: 2–14
9. Hou Y, Wu XY, Li K: Issues on the selection of noninferiority margin in clinical trials. Chin Med J (Engl) 2009; 122:466–70
10. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, for the CONSORT Group: Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. JAMA 2006; 295:1152–60
11. International conference on harmonisation; guidance on statistical principles for clinical trials; availability–FDA. Notice. Fed Regist 1998; 63:49583–98
12. Le Henanff A, Giraudeau B, Baron G, Ravaud P: Quality of reporting of noninferiority and equivalence randomized trials. JAMA 2006; 295:1147–51

ANESTHESIOLOGY REFLECTIONS

The Lazare Anaesthesia Apron



On the Ides of March in 1933, Jacob B. Lazare of Brooklyn, New York, filed for a U.S. Patent for his "Anaesthesia Apron" designed to restrain patients emerging from the excitement phase of nitrous oxide anesthetics in the dental chair. His invention featured a foot-rest loop attached to sturdy upholstering cloth with convenient straps (as featured above). According to U.S. Patent No. 2,005,294, the Lazare Anaesthetic Apron prevented many of the contusions, bone fractures, and even psychological trauma associated with prior efforts that had leather-strapped patients to dental chairs. A bonus, according to Lazare, was that his lady patients appreciated that his "Apron" prevented "the ripping of stockings." (Copyright © the American Society of Anesthesiologists, Inc. This image appears in color in the *Anesthesiology Reflections* online collection available at www.anesthesiology.org.)

George S. Bause, M.D., M.P.H., Honorary Curator, ASA's Wood Library-Museum of Anesthesiology, Park Ridge, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.