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.

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. 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 \( (\leq -\delta) \) 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). 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 equivalence 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 \( \delta \) 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 \( \delta \). Instead, they might a priori hypothesize \( \delta \) 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.

For a binary (yes/no) outcome, \( \delta \) can be an absolute difference between proportions, relative risk, or odds ratio.

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.

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.

References
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.