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 equivalence 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, “not worse than” a standard. In such cases equivalence is not the goal is to demonstrate that a new treatment is “as good as” or equivalency region. Perhaps more commonly, however, the difference between groups falls within the

Equivalence is claimed only if the treatment difference is concluded to be both significantly above the lower limit (−δ) and significantly below the upper limit (+δ) using a traditional one-sided test against a constant for each of the two components of H1 (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 equivalence region. It is then concluded that the true difference lies between −δ and +δ, as in the first equivalence example (E) in figure 1. For the second equivalence trial result (F) in figure 1, the confidence interval is not within ±δ, 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 equivalence 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. For a binary (yes/no) outcome, δ 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...
Choosing the equivalency $\delta$ is an integral part of study design and is best based on both clinical and statistical grounds. $\delta$ 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 $\delta$ risks a claim of equivalence based on a clinically misleading $\delta$, whereas too small a $\delta$ can waste sample size resources and make claiming equivalence unnecessarily difficult.

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 equivalence $\delta$. However, because the equivalence $\delta$ 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 $\delta$ 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 should be clearly labeled as such. Choice of $\delta$ should be determined a priori and should be justified clinically; confidence intervals for the treatment difference should be presented in relation to $\delta$. 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.

**References**


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.