

## The State of Being Noninferior

David C. Musch, PhD, MPH - Ann Arbor, Michigan

Brenda W. Gillespie, PhD - Ann Arbor, Michigan

This issue presents the results of a study by Diestelhorst et al<sup>1</sup> that was designed to demonstrate that intraocular pressure (IOP) reduction from a fixed combination of latanoprost and timolol given once daily is “noninferior” to the 2 drugs administered separately. A noninferiority design also is being used in a large, ongoing phase II clinical trial comparing anecortave acetate with photodynamic therapy in 530 patients with age-related macular degeneration (AMD). If you are somewhat confused about what is meant by *noninferior*, rest assured that you are in the large majority. The purpose of this editorial is to bring some clarity to this increasingly prevalent trial design.

In the absence of a treatment that has been proven effective, clinical trials of proposed treatments have made use of the placebo-controlled randomized clinical trial, the hallmark design of drug development for decades. Such studies are designed to show that the treatment being tested is superior to the placebo control. Today’s world, thankfully, often presents a different scenario, with effective treatments available for many diseases. Drug development has now shifted to more frequent use of active-controlled randomized clinical trials, wherein the goal is either to improve upon an already effective treatment (which invokes the “superior to” hypothesis) or to demonstrate that, compared with the current treatment, the new treatment is equivalent (i.e., about the same) or not inferior (i.e., possibly better, but definitely no worse). Equivalence and noninferiority trials are attractive because the new treatment can be declared a success even if it does not beat the standard treatment. Such trials are used to test new treatments that are expected to have an efficacy similar to that of the standard treatment, but may have an advantage such as fewer side effects, lower cost, or easier administration.

In a noninferiority trial, a margin of noninferiority ( $\Delta$ ) must be specified such that a new treatment that is either better than the standard treatment or worse by no more than  $\Delta$  is considered acceptable. For example, Diestelhorst et al specified a noninferiority margin of 1.5 mmHg in mean daily IOP reduction—that is, they wished to test if the once-a-day combined treatment would result in a mean reduction in IOP between baseline and week 12 no more than 1.5 mmHg less, on average, than the twice-a-day (standard) treatment. They concluded that the once-a-day treatment was noninferior to the standard treatment because, although the IOP reduction with the once-a-day treatment was 0.3 mmHg less than the reduction with the standard treatment, the confidence interval (CI) for the difference in IOP reduction (−0.1 to 0.7) did not include 1.5. In general, a conclusion of noninferiority is made when the CI for the treatment effect lies entirely in the noninferiority region.

Noninferiority trials require statistical advice on how to frame the study hypotheses. Practically speaking, noninferiority trials require high statistical power to detect small differences. In general, they require sample sizes as large as or larger than superiority trials. For example, the Diestelhorst et al study includes 502 patients in the intent-to-treat (ITT) analysis. This well-powered study has a sample size that yields 89% probability of the CI remaining within the 1.5-mmHg noninferiority margin even if the fixed combination treatment truly has an average of 0.5 mmHg less IOP reduction than the standard treatment. If the sample size had been 200 patients, the probability of the CI remaining within the 1.5-mmHg noninferiority margin would have been only 52%. It is important to note that in the context of a superiority design, one can often find no significant difference by using a small sample size with limited statistical power,<sup>2,3</sup> but such a result is not sufficient to declare noninferiority.

Guidance on issues to consider in designing a noninferiority trial is provided by both the European Agency for the Evaluation of Medicinal Products<sup>4</sup> and the United States Food and Drug Administration (FDA).<sup>5</sup> The FDA lists 4 key criteria that must be met in a noninferiority trial. First, there must be solid evidence that the active control is efficacious. Without such evidence, it would be easy to show that any treatment is noninferior to, essentially, no treatment. There is no doubt that Diestelhorst et al’s active control, timolol and latanoprost administered separately, is effective in reducing IOP, with a mean reduction over 12 weeks of approximately 9 mmHg. Second, there must be an acceptable noninferiority margin—that is, the extent to which the test treatment could be worse in its effect than the standard treatment and yet be considered essentially no different. The noninferiority margin must be much smaller than the known minimal effect of the active control treatment.<sup>6</sup> This practice will ensure that we do not include “no effect” in our noninferiority region. The noninferiority margin should also be chosen to reflect a range of acceptable efficacies that is indistinguishable from a clinical perspective. Clinical judgment plays an important role in this decision. Diestelhorst et al’s choice of a 1.5-mmHg noninferiority margin is well below the effect of the active treatment (~9 mmHg), and many would agree that it is an acceptable difference in 2 effective treatments. Note that the outcome measure must also be carefully chosen. Mean diurnal IOP was the selected outcome, but the averaging over time could obscure time-specific variation in effect and, thereby, contribute to a possibly erroneous noninferiority conclusion. This possibility is countered, however, by the provision of time-specific IOP data in the article.

The third FDA criterion states that details of the trial design (study population, end points, etc.) must adhere

closely to the trials that led to evidence of efficacy for the active control. This condition insures that we are comparing apples with apples, not oranges—that is, that we can expect the results for the active control to be similar to the results obtained in the original efficacy trials. By anchoring the active control to previous trial results (with identical trial designs), we can interpret the results from the new intervention with confidence. We would need to know the reference literature to verify that Diestelhorst et al met this criterion.

The fourth FDA criterion states that the conduct of the trial must be of high quality. This criterion requires careful consideration, particularly because clinical trials that are not designed or conducted well are more likely to show noninferiority. For example, poor treatment compliance in both study arms will push the results of both treatments closer together. Diestelhorst et al do not mention compliance in their report. Furthermore, conventions that are conservative in some settings, such as using “last observation carried forward” in the case of missing data, also tend to decrease the difference between treatments. In any use of last observation carried forward, the number of subjects for which it was used should be given; this information was lacking in the report. Even the ITT analysis, the gold standard of superiority trials by which subjects are analyzed in the treatment group to which they were randomized (despite any noncompliance or crossover to the other treatment), leads to well-known conservatism or reduced ability to detect significant differences.<sup>6</sup> It is worthy to note that Diestelhorst et al exclude 15 patients from the ITT analysis, some because they did not receive study medication. Such patients should not be excluded from an ITT analysis. Many authors, including Diestelhorst et al, supplement the ITT analysis with a per protocol analysis, by which subjects are analyzed in the treatment group for the treatment they actually received, irrespective of randomized assignment. Although this analysis removes the conservatism of the ITT analysis, it can introduce bias of unknown magnitude and direction due to the nonrandom nature of patients who do not comply. Thus, a superiority trial has a safe (at worst, conservative) analysis, whereby if a trial shows a significant result, we are fairly certain that the treatments are truly different. In contrast, a noninferiority trial has no safe analysis, and most flaws in trial conduct will bias the result toward a conclusion of noninferiority.

Although a methodologist should be involved in designing a noninferiority study, the expert judgment of a clinician is also critical to its design. Key to deciding upon the acceptability of a noninferiority conclusion is how certain one is of the active control’s effectiveness, the size of this effect, and the margin of acceptable deviation from this effect. If evidence from prior studies for any of these considerations is weak or lacking in consistency across multiple studies, the foundation of a noninferiority study is shaky. Moreover, of course, there is no substitute for soundness in study design, methodology, and analytical aspects.

Given approved treatments for conditions like neovascular AMD, glaucoma, and uveitis, we can expect to see more randomized clinical trials designed to demonstrate noninferiority. And so, unlike Garrison Keillor’s statement that children in fictitious Lake Wobegon are all above average, the most one could conclude from a noninferiority trial is that the children are no worse than average. Nevertheless, showing that a treatment has an effect that is no worse than average is not necessarily bad, when “average” is based on a known, efficacious treatment. A sound noninferiority trial can validate the efficacy of a new beneficial treatment—with its accompanying benefits of easier delivery, fewer side effects, and/or reduced cost.

## References

1. Diestelhorst M, Larsson L-I, European-Canadian Latanoprost Fixed Combination Study Group. A 12-week, randomized, double-masked, multicenter study of the fixed combination of latanoprost and timolol in the evening versus the individual components. *Ophthalmology* 2006;113:70–6.
2. Bourne WM. “No statistically significant difference.” So what? *Arch Ophthalmol* 1987;105:40–1.
3. Javitt JC. When does the failure to find a difference mean that there is none? *Arch Ophthalmol* 1989;107:1034–40.
4. European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products. Points to consider on switching between superiority and non-inferiority. CMP/EWP/482/99. 2000. Available at: <http://www.emea.eu.int/pdfs/human/ewp/048299en.pdf>.
5. Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for industry. E 10 choice of control group and related issues in clinical trials. 2001. Available at: <http://www.fda.gov/cder/guidance/4155fnl.pdf>.
6. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med* 2000;1:19–21.