# Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis

When busy clinicians bump into a new treatment, they ask themselves 2 questions. First, is it *better* than ("superior to") what they are using now? Second, if it's not superior, is it *as good as* what they are using now ("noninferior") and preferable for some other reason (e.g., fewer side effects or more affordable)? Moreover, they want answers to these questions right away. *ACP Journal Club* and its related evidence-based journals do their best to answer these questions in their "more informative titles." That's why this issue contains titles such as: "Angioplasty at an invasive-treatment center reduced mortality compared with first-contact thrombolysis" (1) and "Ximelagatran was noninferior to warfarin in preventing stroke and systemic embolism in atrial fibrillation" (2). The latter study prompted this editorial.

Progress toward this "more informative" goal has been slow because we have been prisoners of traditional statistical concepts that call for 2-sided tests of statistical significance and require rejection of the null hypothesis. We have further imprisoned ourselves by misinterpreting "statistically nonsignificant" results of these 2-tailed tests. Rather than recognizing such results as "indeterminate" (uncertain), we conclude that they are "negative" (certain, providing proof of no difference between treatments). This editorial will address the problems created by these ways of thinking and, more important, their clinically relevant solutions.

At the root of our problem is the "null hypothesis," which decrees that the difference between a new and standard treatment ought to be zero. 2-sided *P* values tell us the probability that the results are compatible with that null hypothesis. When the probability is small (say, less than 5%), we "reject" the null hypothesis and "accept" the "alternative hypothesis" that the difference we've observed is not zero. In doing so, however, we make no distinction between the new treatment being better, on the one hand, or worse, on the other, than the standard treatment.

There are 3 consequences of this faulty reasoning. First, by performing "2-sided" tests of statistical significance, investigators turn their backs on the "1-sided" clinical questions of superiority and noninferiority. Second, they often fail to recognize that the results of these 2-sided tests, especially in small trials, can be "statistically nonsignificant" even when their confidence intervals include clinically important benefit or harm. Third, investigators (abetted by editors) frequently misinterpret this failure to reject the null hypothesis (based on 2-sided *P* values > 5%, or 95% confidence intervals that include zero). Rather than recognizing their results as uncertain ("indeterminate"), they report them as "negative" and conclude that there is "no difference" between the treatments. By doing so, authors and editors and readers regularly fall into the trap of concluding that the "absence of proof of a difference" between 2 treatments constitutes "proof of an absence of a difference" between them. This mistake was forcefully pointed out by Phil Alderson and Iain Chalmers: "It is never correct to claim that treatments have no effect or that there is no difference in the effects of treatments. It is impossible to prove ... that two treatments have the same effect. There will always be some uncertainty surrounding estimates of treatment effects, and a small difference can never be excluded" (3).

A solution to both this incompatibility (between 1-sided clinical reasoning and 2-sided statistical testing) and confusion (about the clinical interpretation of statistically nonsignificant results) has been around for decades but is just now gaining widespread recognition and application. I assign most of the credit to a pair of biostatisticians, Charles Dunnett and Michael Gent (others have also contributed to its development [4], although they sometimes refer to "noninferiority" as "equivalence," a term whose common usage fails to distinguish 1-sided from 2-sided thinking). I'll illustrate Charlie Dunnett's and Mike Gent's contribution with a pair of trials in which their thinking helped clinical colleagues escape from the prison of 2-sided null hypothesis testing and, by doing so, prevented the misinterpretation of statistically nonsignificant results (5).

Thirty years ago, a group of us performed a randomized controlled trial (RCT) of nurse practitioners as providers of primary care (6). We wanted to know if patients fared as well under their care as under the care of general practitioners. Guided by Mike Gent, we came to realize that a 2-sided analysis that produced an "indeterminate," statistically nonsignificant difference in patient outcomes could confuse rather than clarify matters. We therefore abandoned our initial 2-sided null hypothesis and decided that we'd ask a noninferiority question: Were the outcomes of patients cared for by nurse practitioners noninferior to those of patients cared for by general practitioners? Mike Gent then helped us recognize the need to specify our limit of acceptable "inferiority" in terms of these outcomes. With his prodding, we decided that we would tolerate no worse than 5% lower physical, social, or emotional function at the end of the trial among patients randomized to our nurse practitioners as we observed among patients randomized to our general practitioners. As it happened, our 1-sided analysis revealed that the probability that our nurse practitioners' patients were worse off (by ≥ 5%) than our general practitioners' patients was as small as 0.008. We had established that nurse practitioners were not inferior to general practitioners as providers of primary care.

Twenty years ago, a group of us performed an RCT of superficial temporal artery–middle cerebral artery anastomosis ("EC–IC bypass") for patients with threatened stroke (7). To the disappointment of many, we failed to show a statistically significant superiority of surgery for preventing subsequent fatal and nonfatal stroke. It became important to overcome the ambiguity of this "indeterminate" result. We therefore asked the 1-sided question: What degree of surgical benefit could we rule out? That 1-sided analysis, which calculated the upper end of a 90% (rather than a 95%) confidence interval, excluded a surgical benefit as small as 3%. When news of this 1-sided result got around, performance of this operation rapidly declined.

Thanks to statisticians like Charlie Dunnett and Mike Gent, we now know how to translate rational, 1-sided clinical reasoning

  

into sensible, 1-sided statistical analysis. Moreover, this modern strategy of asking 1-sided noninferiority and superiority questions in RCTs is gathering momentum. The CONSORT statement on recommendations for reporting RCTs omits any requirement for 2-sided significance testing. Even some journal editors are getting the message, for 1-sided noninferiority and superiority trials have now appeared in the *New England Journal of Medicine* (8), *Lancet* (9), and *JAMA* (10), and this issue of *ACP Journal Club* includes another *Lancet* article (2) (see Ximelagatran was nonin-ferior to warfarin in preventing stroke and systemic embolism in atrial fibrillation).

An essential prerequisite to doing 1-sided testing is the specification of the exact noninferiority and superiority questions before the RCT begins. As with unannounced subgroup analyses, readers can and should be suspicious of authors who apply 1-sided analyses without previous planning and notice. Have they been slipped in only after a peek at the data revealed that conventional 2-sided tests generated indeterminate results? This need for prior specification of 1-sided analyses provides yet another argument for registering RCTs in their design stages and for publishing their protocols in open-access journals such as Biomed Central (www.biomedcentral.com).

I hope that this editorial will help free front-line clinicians, investigators, and editors from the 2-sided null-hypothesis prison. If any traditional, 2-sided biostatisticians happen upon it, they may object. If their objections are relevant to this journal's readers, they might appear in these pages.

*David L. Sackett, MD*
*Trout Research and Education Centre at Irish Lake*
*Markdale, Ontario, Canada*

***References***

1. **Andersen HR, Nielsen TT, Rasmussen K, et al.** A comparison of coronary angioplasty with fibrinolytic therapy in acute myocardial infarction. N Engl J Med. 2003;349:733-42.
2. **Olsson SB.** Stroke prevention with the oral direct thrombin inhibitor ximelagatran compared with warfarin in patients with non-valvular atrial fibrillation (SPORTIF III): randomised controlled trial. Lancet. 2003;362:1691-8.
3. **Alderson P, Chalmers I.** Survey of claims of no effect in abstracts of Cochrane reviews. BMJ. 2003;326:475.
4. **Ware JH, Antman EM.** Equivalence trials. N Engl J Med. 1997;337:1159-61.
5. **Dunnett CW, Gent M.** An alternative to the use of two-sided tests in clinical trials. Stat Med. 1996;15:1729-38.
6. **Sackett DL, Spitzer WO, Gent M, Roberts RS.** The Burlington randomized trial of the nurse practitioner: health outcomes of patients. Ann Intern Med. 1974;80:137-42.
7. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an International randomized trial. The EC/IC Bypass Study Group. N Engl J Med. 1985;313:1191-200.
8. **Zeuzem S, Feinman SV, Rasenack J, et al.** Peginterferon alfa-2a in patients with chronic hepatitis C. N Engl J Med. 2000;343:1666-72.
9. **Topol EJ.** Reperfusion therapy for acute myocardial infarction with fibrinolytic therapy or combination reduced fibrinolytic therapy and platelet glycoprotein IIb/IIIa inhibition: the GUSTO V randomised trial. Lancet. 2001;357:1905-14.
10. **Parienti JJ, Thibon P, Heller R, et al.** Hand-rubbing with an aqueous alcoholic solution vs traditional surgical hand-scrubbing and 30-day surgical site infection rates: a randomized equivalence study. JAMA. 2002;288:722-7.