

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

When busy clinicians bump into a new treatment, they ask themselves 2 questions. Firstly, is it *better* than (“superior to”) what they are using now? Secondly, if it’s not superior, is it *as good as* what they are using now (“non-inferior”) and preferable for some other reason (eg, fewer side effects or more affordable)? Moreover, they want answers to these questions right away. *Evidence-Based Medicine* 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 centre reduced mortality compared with first contact thrombolysis”<sup>1</sup> (<http://ebm.bmjournals.com/cgi/content/9/2/42>) and “Ximelagatran was non-inferior to warfarin in preventing stroke and systemic embolism in atrial fibrillation.”<sup>2</sup> (<http://ebm.bmjournals.com/cgi/content/9/2/43>) The latter of these 2 studies 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 non-significant” results of these 2-tailed tests. Rather than recognising 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 importantly, 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. Two-sided p values tell us the probability that the results are compatible with that null hypothesis. When that probability is small (say, <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. Firstly, by performing “2-sided” tests of statistical significance, investigators turn their backs on the “1-sided” clinical questions of superiority and non-inferiority. Secondly, they often fail to recognise 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. Thirdly, 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 recognising 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.”<sup>3</sup>

A solution to both this incompatibility (between 1-sided clinical reasoning and 2-sided statistical testing) and confusion (about the clinical interpretation of statistically non-significant 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, and others have also contributed to its development<sup>4</sup> (although the latter sometimes refer to “non-inferiority” as “equivalence,” a term whose common usage fails to distinguish 1-sided from 2-sided thinking). I’ll illustrate the contribution of Charles Dunnett and Michael Gent 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.<sup>5</sup>

Thirty years ago, a group of us performed a randomised controlled trial (RCT) of nurse practitioners as providers of primary care.<sup>6</sup> 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 realise 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 non-inferiority question: Were the outcomes of patients cared for by nurse practitioners non-inferior to those of patients cared for by general practitioners? Mike then helped us recognise 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 randomised to our nurse practitioners as we observed among patients randomised to our general practitioners. As it happened, our 1-sided analysis revealed that the probability that our nurse practitioners’ patients were worse off (by  $\geq 5\%$ ) 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.<sup>7</sup> To the disappointment of many, we failed to show a statistically significant superiority of surgery for preventing subsequent fatal and non-fatal 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 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 non-inferiority 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 non-inferiority and superiority trials have now appeared in the *New England Journal of Medicine*,<sup>8</sup> *Lancet*,<sup>9</sup> and *JAMA*,<sup>10</sup> and this issue of *Evidence-Based Medicine* includes another *Lancet* article (<http://ebm.bmjournals.com/cgi/content/9/2/43>).<sup>2</sup>

An essential prerequisite to doing 1-sided testing is the specification of the exact non-inferiority 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 (<http://www.biomedcentral.com>).

## Patients at the centre: in our practice, and in our use of language

Dear editors,

In their editorial about patients sharing decision making responsibilities with physicians, Guyatt *et al*<sup>1</sup> review some of the evidence that patients favour involvement in their own care. They neglect a more complex body of evidence, however, that suggests that while patients like the idea of greater choice, they often find it to be less satisfying in practice.<sup>2</sup> Just as the authors suggest that physicians incorrectly “assume that physiological outcomes will lead to improvements in mortality and quality of life,” they themselves assume that more choice will lead to greater patient satisfaction.

As suggested by Barry Schwartz, a psychologist at Swarthmore College, the opposite may be true.<sup>3</sup> Schwartz has identified a number of problems that arise as choices proliferate.<sup>4,5</sup> Firstly, reliable information is difficult to obtain. The advent of the internet has increased access to information, as Guyatt *et al* contend, but this access may be to as much unreliable as trustworthy information. Obtaining to additional opinions and performing research are potentially costly and time consuming.

Another problem is that as options multiply, patients’ standards for what is an acceptable outcome mount, with a loss of perspective of what may reasonably be achieved. Schwartz also notes that people may fear making the wrong choice and allow that fear to guide their choices. Or, post hoc, they may come to believe that an unacceptable result is their fault, and that their choice of therapy was mistaken. As the saying goes, good judgment is the result of experience, and experience is the result of bad judgment.

In one example of a study that stands in contrast to those cited by Guyatt *et al*, people were asked to describe the role they would select in choosing their treatment plan if they had

I hope that this editorial will help free frontline 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*

- 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 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, *et al*. 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; GUSTO V Investigators. 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.

cancer.<sup>6</sup> One hundred fifty women, newly diagnosed with breast cancer, were compared to 200 women with benign breast disease. The majority of the women with breast cancer preferred a passive role, leaving the decision making responsibility to their physicians; the benign disease control group preferred a collaborative role with decisions made jointly between the patient and the physician.

Schwartz has written that “Indeed, there may be a point when choice tyrannizes people more than it liberates them.”<sup>7</sup> Like the participants in the study cited above, my own mother, an intelligent and strong willed professional, was faced with the choice between lumpectomy and mastectomy for breast cancer. She sought opinions from 2 oncologists, an internist, a gynaecologist, and her son, a nephrologist. I thoroughly reviewed with her the medical literature of the day. Finally, after many weeks, her unwavering decision was a non-decision, to let me make the choice, without her offering any real guidance. For the rest of her life, she was grateful to me for having “saved her breast.”

DAVID S GOLDFARB M.D.  
*New York Harbor VA Medical Center  
NYU School of Medicine  
New York, USA*

- 1 Guyatt G, Montori V, Devereaux PJ, *et al*. *Evidence-Based Medicine* 2004;**9**.
- 2 Schneider CE. The practice of autonomy: patients, doctors, and medical decisions. Oxford: Oxford University Press, 1999.
- 3 Schwartz B. The paradox of choice: why more is less. New York: Ecco, 2004.
- 4 Schwartz B. *Am Psychol* 2000;**55**:79–88.
- 5 Schwartz B, Ward A, Monterosso J, *et al*. *J Pers Soc Psychol* 2002;**83**:1178–97.
- 6 Beaver K, Luker KA, Owens RG, *et al*. *Cancer Nurs* 1996;**19**:8–19.
- 7 Schwartz B. Nation of second guesses. *New York Times*, January 22, 2004. <http://www.nytimes.com/2004/01/22/opinion/22SCHW.html>.

## Patients at the centre: in our practice, and in our use of language

In response:

Dr Goldfarb is correct that patients may legitimately choose not to participate in decision making. Indeed, at the outset of our article, we described how in the past most patients may have preferred to leave decisions to their clinician.

We argue for the importance of patient autonomy. That autonomy begins with a preference, on the patient's part, about the extent to which the patient wants information, wants to participate in deliberations, and wants to take responsibility for the decision. We believe, as does Dr Goldfarb, that clinicians can harm patients by forcing them into a decision making role that they do not want. Some patients, however, may choose not to participate in the decision making process because they believe physicians will not present relevant information in a manner the patients can easily understand.

Dr Goldfarb seems to believe that a study in which the degree of enthusiasm for active participation in decision making varied with the underlying condition somehow undercuts our point. Preferences will vary between patients and between contexts. Our contention is that evidence suggests that patients today are more inclined toward active participation in decision making than were those of several decades ago.

Will patients be better off if they choose to participate more actively in decision making? This is an empirical question best answered by randomised trials of alternative decision making approaches. A systematic review of 34 randomised trials of use of decision aids, a process that promotes active patient participation, showed that in general these approaches reduce decisional conflict (weighted mean difference 19 on a 100-point scale, 95% CI 13 to 24).<sup>1</sup> Furthermore, results of randomised trials have shown that coaching patients to become more involved in making decisions can improve patient wellbeing and some disease specific outcomes.<sup>2,3</sup>

In our article, we stated that "Shared decision-making—like evidence-based medicine when in its initial phases, and perhaps even today—also faces risks of misunderstanding." Dr Goldfarb's letter vividly illustrates our point. In his final paragraph, Dr Goldfarb tells the story of his mother's choice about lumpectomy or mastectomy for breast cancer. The patient sought opinions from 5 physicians. She ultimately reviewed the evidence with her son. In the end, having heard several opinions, and with an understanding of the evidence,

she chose to leave the decision taking to her son. It appears that Dr Goldfarb's mother was interested in obtaining the best understanding she could of the tradeoffs around her decision. Presumably, had there been a clear choice, she would have made it herself. In the end, it appears she concluded that the decision was closely balanced. Thus, she was most comfortable leaving the final choice to someone whom she trusted would have an in-depth knowledge of the benefits and risks of the alternatives, and an equally deep understanding of her values and preferences.

To the extent that this characterisation of events is accurate, Dr Goldfarb's story is completely consistent with the shared decision making that our article advocates. Dr Goldfarb's mother's choice to have her son make the decision proved the right one because he understood her values and preferences—this is no doubt why she chose him, rather than the other 4 physicians she consulted, as the final decision-maker. Outcomes might be less happy if decision makers do not have a deep understanding of patients' values and preferences—had, for instance, Dr Goldfarb's mother relied on a physician who selected mastectomy as the best option for her cancer.

Among the skills required by clinicians aspiring to evidence-based decision making are an appraisal of the complexity of the decision, the informational needs, and the decision style of the patient and the likelihood that patients with different values and preferences will make different choices after receiving the same information. In preparing information, clinicians must avoid framing bias and attend to uncertainty in knowledge about likely outcomes. Ultimately, whatever the degree of involvement the patient chooses, the clinician must focus not on clinically relevant but rather on patient important outcomes.

VICTOR MONTORI, MD  
P J DEVEREAUX, MD  
HOLGER SCHÜNEMANN, MD  
MOHIT BHANDARI, MD  
GORDON GUYATT, MD  
*McMaster University  
Hamilton, Ontario, Canada*

- 1 O'Connor AM, Stacey D, Entwistle V, et al. *Cochrane Database Syst Rev* 2003;(2):CD001431.
- 2 van Dam HA, van der Horst F, van den Borne B, et al. *Patient Educ Couns* 2003 sep;**51**:17–28.
- 3 Michie S, Miles J, Weinman J. *Patient Educ Couns* 2003;**51**:197–206.