Defending the Rationale for the Two-Tailed Test in Clinical Research

Lemuel A. Moyé, MD, PhD; Alan T.N. Tita, MD, MPH

The issue of one-sided versus two-sided hypothesis testing in clinical trial analyses has been the subject of debate in the medical and statistical literature. This design consideration represents a sharp divide in the research community, separating the investigators’ deep-seated beliefs in therapy effectiveness from their obligatory primary concern for patient welfare. In a recent commentary, Knottnerus and Bouter advanced the theory that, for individual trials evaluating new interventions not previously studied, a one-sided hypothesis test seems sensible from both an ethical and cost-efficiency perspective. Certainly, to many healthcare researchers, the temptation of one-tailed hypothesis testing, in which the location of type I errors coincides exactly with the investigator’s prospective intuition (based on available but sometimes misleading information) about the research result, can be difficult to resist. For a specified experimental probability of type I error, the allure of the smaller sample size associated with one-tailed testing further strengthens its attraction for some researchers. We argue here, however, that one-tailed testing should be avoided in healthcare research for ethical and cost-efficiency reasons, especially in randomized trials in which the investigator controls the intervention. Rather than reflecting the investigators’ a priori intuition, the type I error should reflect the uncertainty of the research effort’s future conclusions. This is critical in a field in which healthcare practitioners and healthcare researchers can inadvertently do harm to their patients.

One-Sided Thinking and Ethical Restraints

In clinical intervention trials, we as investigators wish to demonstrate that the tested intervention is beneficial. As clinical researchers, we do not like to harbor the notion that the interventions we have developed for the benefit of our patients can cause harm. Nevertheless, harm is often the result, as demonstrated by the use in the past of bleedings and potent purgatives for diseases that practitioners believed they understood. These now debunked medical procedures were applied by physicians who took the same medical oath for patient protection that we have, acted in the best interest of their patients, as we do, and believed the therapy was appropriate and beneficial, again, as we do. Healthcare practitioners and researchers must be ever vigilant for the hazard of patient injury, as it is often a consequence of our good intentions. The more strongly we believe in the benefit of a therapy, the more observant we must become for the unsuspected damaging effects. The two-sided test shines a bright, direct light on the health researcher’s darkest fear, that he or she, despite his or her best efforts, might do harm. This essential illumination provides an objective view of the effect of the studied intervention, regardless of how beneficial or how detrimental the intervention might be.

The genesis of a healthcare research idea is often an observation of practicing physicians. We as physicians have a particular burden here, because the persuasive power we bring to bear when discussing therapy options with patients can more deeply embed a one-sided view of therapy efficacy. We as physicians find ourselves in the position of advocating therapy choices for patients who rely heavily on our recommendations and opinions. We must often appeal to the better nature of patients who are uncertain in their decisions. Physicians have learned to use tact, firmness, prestige, and character to convince patients of our belief in the best approach in managing their health problems. Although the patient may choose to obtain a second opinion, these opinions are those of other healthcare providers, also vehemently expressed. Thus, physicians can bring strong beliefs about the effect of therapy to the research design table.

The force behind an investigator’s vehement opinion can be magnified by the additional energy required to initiate and drive a study program. In research, enthusiasm is required to carry forward a joint research effort that involves a sponsoring agency, recruiting centers, and hundreds of healthcare workers. The proponents of the intervention must persuade their colleagues that the experiment is worthy of their time and labor. The investigators must convince sponsors (private or public) that the experiment should be carried out, and their argument often includes a forcefully delivered thesis on the prospects for the trial’s success. This is necessary because financial sponsors, who often must choose from a collection of proposed experiments competing for funding, are understandably more willing to underwrite trials with a greater perceived chance of demonstrating that the intervention is...
beneficial. In this environment, the principal investigator must resist the urge for persistent blind faith in the intervention’s untested beneficial effect. The two-tailed hypothesis test appropriately reasserts the possibility that the investigator’s belief system about an effect of therapy might be wrong.

Sample Size Efficiency Versus Sample Size Effectiveness

An argument raised in defense of one-sided testing is sample size efficiency. Others correctly point out that the one-tailed test produces a reduction in the minimum research sample size, because the one-sided test focuses on only one tail of the effect probability distribution. Although the savings are apparent for a given experimental alpha, however, they do not occur at the level one might expect. The Figure depicts the fraction of observations needed in a two-tailed test that are required in a one-tailed test in a randomized clinical experiment where the goal is to demonstrate a 20% reduction in clinical event rates from a cumulative control group event rate of 25% with 80% power. If we would expect that 50% of the observations required for a two-sided test were needed for a one-tailed significance test in this example, then the curve would reveal a flat line at $\alpha=0.50$ for the different levels of acceptable type I error (alpha). The curve in the Figure demonstrates something quite different. For example, for an alpha level of 0.05, 79% of the observations required in the two-tailed test are needed for the one-sided test. At any level of alpha examined, the 50% value is not achieved. Although smaller in one-tailed testing than in two-tailed testing, the reduction in sample size is a modest one.

However, this apparent reduction in sample size in a clinical experiment produced by carrying out a one-sided hypothesis test comes at the price of being unable to draw appropriate conclusions if the investigators are wrong and the study demonstrates detrimental effects. The medical community requires assurance that the finding of injury is not due to sampling error. This assurance would typically be conveyed by the measure of type I error, but what type I error is associated with this finding of harm in a one-tailed test designed to find benefit? In fact, there is no good measure of type I error in this setting, because the $P$ value is untrustworthy when no type I error is allocated prospectively. Therefore, the medical community does not receive its desired assurance, placing the investigator in an uncomfortable and untenable situation. The study’s finding of harm may lead to the defensible belief that the research’s replication is unethical. Because $P$ value estimates are unreliable in this setting, however, the findings may not represent the results in the population, and therefore a second study is required for confirmation. Here, the finding of harm in a one-tailed test designed to find benefit makes it ethically unacceptable but scientifically necessary to reproduce the result. This conundrum causes confusion in the medical community and could have been completely avoided by carrying out a two-tailed test from the beginning. Such a two-tailed study requires only 63% of the requisite total sample size for 2 separate one-sided studies, everything else being equal, as inferred from the Figure. Thus, although less efficient, the experiment designed for a two-tailed hypothesis test is more effective because it removes the necessity of repetition with its attendant ethical dilemma when harm, rather than benefit, is shown.

A one-tailed test designed exclusively to find benefit does not permit the assessment of the role of sampling error in producing harm, a dangerous omission for a profession whose fundamental tenet is to first do no harm. This deficit is amplified by the increasingly common use of multiple endpoints in clinical studies. Assume that a clinical, one-tailed trial testing for beneficial effects is carried out for one primary endpoint and one secondary endpoint, each of which was prospectively identified in a concordantly executed study. What is the correct conclusion if the null hypothesis is not rejected for the primary endpoint, but is rejected for the secondary endpoint? Because the one-tailed test does not differentiate a harmful effect from a null effect, how can the investigator assure the medical community that the population will be spared from harm on the primary endpoint of the study? No trustworthy measure of type I error level is available in this setting, which may require the study to be reproduced, a replication that would be obviated by two-sided testing with only a small marginal increase in sample size.

It must also be noted that although some authors argue that more patients may be exposed to the control therapy and perhaps receive the inferior treatment in a two-sided test, this criticism is blunted somewhat by the use of prospectively designed monitoring rules that can terminate the study prematurely in light of early strong evidence of benefit.

Knowledge Versus Faith

Furthermore, in a strictly mathematical sense and from strictly an optimality perspective, the most powerful tests are uniformly available from the family of one-tailed tests. The fact that the minimum sample size required for the one-tailed test is smaller than that required for two-tailed testing is, however, not a question of statistical optimality, but merely a demonstration that the one-tailed test requires less strength of evidence for a positive result than the two-tailed test. When comparing statistical findings, the comparison should ideally be based on level of evidence. Thus, a two-sided symmetric 0.05 test has a greater level of evidence than a one-sided 0.05 test, but the same level of evidence as a one-sided 0.025 test that demonstrates the hypothesized beneficial outcome. Thus,
if clinical research is ideally designed to have the same level of evidence for the expected outcome, the argument that one-sided testing involves a smaller sample would be irrelevant. However, the two-sided test is potentially more informative when faced with an unexpected outcome.

Physicians treat their patients on the basis of what they believe. The best experimental designs, however, have their basis in knowledge, not faith. Research design requires that we separate our beliefs from our knowledge about the therapy. Although we are convinced of the intervention’s effect as the study is designed, we must acknowledge that we cannot be certain of that effect. We may have accepted the idea of the intervention’s beneficial effect because of what we have seen in practice; however, that view is not objective, but skewed. Admitting the necessity of the research effort to study the intervention is an important acknowledgment showing that the investigator does not know what the outcome will be. Therefore, a critical requirement in the design of an experiment is that the investigators separate their beliefs from available uncertain information. The experiment should be designed based on knowledge of, rather than faith in, the therapy.

**One-Tailed Plans and Opposite Tail Results**

Evidently, there are important limitations to the one-tailed test in a clinical research effort. The major difficulty is that the one-sided testing philosophy reveals a potentially dangerous level of investigator consensus that there is no possibility of patient harm produced by the intervention being tested. The Cardiac Arrhythmia Suppression Trial (CAST)\(^9\) experience is perhaps most reflective of the difference between belief and reality. In the middle of the 20th century, an idea developed among cardiologists that disorders in heart rhythm did not all have the same prognosis, but instead depicted a spectrum with well-differentiated mortality prognoses. Drugs had been available to treat heart arrhythmias at the time, but many of these produced severe side effects and were difficult for patients to tolerate. Scientists were, however, developing a newer generation of drugs that they believed produced fewer side effects and were more effective. Researchers eventually carried out a large-scale clinical trial to assess the effect of these antiarrhythmic agents. Perhaps being influenced by presentations in the clinical trial literature that suggested it would be appropriate to present clinical trial results demonstrating harm as though they were simply not beneficial,\(^10\) they designed the trial as one-sided, anticipating that only therapy benefit would result from this research. The fact that the investigators designed the trial as one-sided reveals the degree to which they believed the therapy would reduce mortality.

Recruitment in this study was crippled by the refusal of many recruiting physicians to allow their patients to be randomized into a study in which there was a 50% chance that the patients would not receive the intervention. Fortunately, the Data Safety and Monitoring Board of CAST imposed an advisory 0.025 lower bound for the possibility of harm because the trial was terminated quickly because of an unanticipated, mortal effect of the drug. In a trial designed by the investigators to demonstrate only a survival benefit of antiarrhythmic therapy, the therapy was shown to be almost 4 times as likely to kill active group patients as placebo. In this one-tailed experiment, the \(P\) value was 0.0003 in the “other tail.”\(^11\)

The investigators reacted to these devastating findings with shock and disbelief. They had embraced the arrhythmia suppression hypothesis to the point where they had excluded all possibility of there being a harmful effect. The findings of the experiment, however, proved the researchers wrong. There has been much debate about the implications of CAST for the development of antiarrhythmic therapy.\(^12\) An important lesson is that healthcare researchers must exert the greatest possible care when extrapolating their own beliefs to form conclusions about population.

Some authors advocate a one-sided confirmatory test for the expected beneficial outcome and an exploratory, post-hoc or hypothesis-generating interpretation of an unexpected outcome\(^13\) similar to the approach used by the CAST advisory board. We maintain it is inappropriate to systematically apply such retrospective levels, as associated probability values are inherently not interpretable. In the case of CAST, the high relative risk for harm and the extremely small \(P\) value strongly but not conclusively suggested a harmful effect of antiarrhythmic therapy in the population. In situations where relative risks are modest and retrospective probability value is only marginal, it becomes futile to attempt to dissociate unexpected harm from true null.

The CAST experience demonstrates that this denial of potential harm can ambush well-meaning investigators, causing confusion as they struggle to assimilate the unexpected, devastating results of their efforts. We healthcare researchers do not like to accept the possibility that, well-meaning as we are, we may be hurting the patients we work so hard to help; however, a thoughtful consideration of our history persuades us that this is all too often the case.

After a first trial demonstrated in a somewhat irregular fashion that the compound vesnarinone could reduce the cumulative total mortality rate of patients with heart failure,\(^14\) a second study was executed to confirm this result. In this second evaluation, the vesnarinone trial termed VEST,\(^15\) 3833 patients with congestive heart failure of New York Heart Association functional class III or IV with a left ventricular ejection fraction \(\leq 30\%\) were randomized to either conventional therapy plus placebo, conventional therapy plus 30 mg of vesnarinone, or conventional therapy plus 60 mg of vesnarinone. In VEST, the primary endpoint was all cause mortality, and the goal of the study was to compare the 30 mg vesnarinone dose and the 60 mg vesnarinone dose to the placebo. The maximum follow-up period was 70 weeks, and it was anticipated that 232 deaths would be required to be able to demonstrate the beneficial effect of vesnarinone on all cause mortality. In this confirmatory study, however, the mortality rate observed was higher in the patients randomized to 30 mg of vesnarinone (21%) or 60 mg vesnarinone (22%) than in the placebo group (18.9%). In addition, the time to death was significantly shorter in the 60 mg vesnarinone group than in the placebo group (\(P=0.02\)). The first trial, which demonstrated a mortality benefit for a 60 mg dose of vesnarinone had its findings reversed by the second trial,
which demonstrated a mortality hazard of this same dose. The
investigators stated, “Examination of the patient populations in
the two trials reveals no differences that could reasonably
account for the opposite response to the daily administration
of 60 mg of vesnarinone.” Fortunately, the two-tailed test
preserved the reliability of the conclusion involving this
compound.

The intelligent application of the two-tailed test requires
deliberate, overt effort to consider the possibility of patient
harm during the design phase of any experiment. This
concern, expressed early and formally during the trial’s
design, can be very naturally translated into effective steps
taken during the course of the experiment. In circumstances
where the predesign clinical intuition is overwhelmingly in
favor of beneficial results, the investigators should exert the
required discipline to provide adequate ability to determine if
the intervention produces harm. It is fine to hope for the best,
as long as we also prepare for the worst. The prospective use
of a two-sided significance test is of utmost importance.
Although the two-sided hypothesis test can complicate ex-
perimental design, increasing sample size requirements, this
approach is ultimately more informative and potentially
prevents subsequent exposure of research participants and the
general population to harmful interventions.

References
1. Fisher LD. The use of one-sided tests in drug trials: an FDA advisory
2. Enkin MW. One and two-sided tests of significance: one-sided tests
should be used more often. BMJ. 1994;309:874.
3. Dunnett CW Gent M. An alternative to the use of two-sided tests in
4. Bland JM, Altman DG. Statistics notes: one and two-sided tests of
5. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing
7. Moyé LA. P-value interpretation, and alpha allocation in clinical trials.
8. Bickel P, Doksum KA. Mathematical Statistics: Basic Ideas and
10. Lan KK, Friedman L. Monitoring boundaries for adverse effects in
11. The CAST Investigators. Preliminary report: effect of encainide and
flecainide on mortality in a randomized trial of arrhythmia suppression
12. Pratt CM, Moyé LA. The Cardiac Arrhythmia Suppression Trial: casting
ventricular arrhythmia suppression in a new light. Circulation. 1995;91:
245–247.
morbidity and mortality in patients with heart failure. Vesnarinone Study
15. Cohn J, Goldstein SC, Feenheid S, et al. A dose-dependent increase in
mortality seen with vesnarinone among patients with severe heart failure.

Key Words: statistics • tests • trials
Defending the Rationale for the Two-Tailed Test in Clinical Research
Lemuel A. Moyé and Alan T.N. Tita

Circulation. 2002;105:3062-3065
doi: 10.1161/01.CIR.000018283.15527.97
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/105/25/3062

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at: http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at: http://circ.ahajournals.org/subscriptions/