Defending the Rationale for the Two Tailed Test in Clinical Research

Lemuel A. Moyé, M.D., Ph.D.

Alan T. N. Tita, M.D., M.P.H.

From The University of Texas Houston Health Science Center, School of Public Health (LAM, ATNT) and Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston (ATNT)

Address Correspondence to

Lemuel A. Moyé, M.D., Ph.D.

RAS Building E815

1200 Herman Pressler

Houston, Texas 77030

Telephone 713-500-9518

Fax 713-500-9530

Email lmoye@utsph.sph.uth.tmc.edu

Abstract

The issue of test sidedness in hypothesis testing for clinical trial analyses has been the subject of debate in the medical and statistical literature. This design consideration represents a sharp line in the research community, dividing the investigators' deep seated beliefs in therapy effectiveness from their obligatory prime concern for patient welfare. A recent commentary advances the thesis that for individual trials, especially those evaluating new interventions not previously studied, a one-sided hypothesis test seems sensible from each of an ethics and cost efficiency perspective. However, we argue here that two-tailed testing should be routinely used in health care research from an ethics and cost effectiveness rationale, especially in randomized trials in which the investigator controls the intervention. Rather than reflect the investigators' *a priori* intuition, the type I error should reflect the uncertainty of the research effort's unpredictable and sometimes surprising conclusions. This is critical in a field in which health care practitioners, and health care researchers inadvertently do harm to their patients.

Key Words: statistics, hypothesis testing, sample size

Abstract

The issue of test sidedness in hypothesis testing for clinical trial analyses has been the subject of debate in the medical and statistical literature. This design consideration represents a sharp line in the research community, dividing the investigators' deep seated beliefs in therapy effectiveness from their obligatory prime concern for patient welfare. A recent commentary advances the thesis that for individual trials, especially those evaluating new interventions not previously studied, a one-sided hypothesis test seems sensible from each of an ethics and cost efficiency perspective. However, we argue here that two-tailed testing should be routinely used in health care research from an ethics and cost effectiveness rationale, especially in randomized trials in which the investigator controls the intervention. Rather than reflect the investigators' *a priori* intuition, the type I error should reflect the uncertainty of the research effort's unpredictable and sometimes surprising conclusions. This is critical in a field in which health care practitioners, and health care researchers inadvertently do harm to their patients.
Key Words: statistics, hypothesis testing, sample size

Introduction

The issue of one vs. two sided hypothesis testing in clinical trial analyses has been the subject of debate in the medical and statistical literature [1-4]. This design consideration represents a sharp line in the research community, dividing the investigators' deep-seated beliefs in therapy effectiveness from their obligatory prime concern for patient welfare. JA Knottnerus and LM Bouter in a recent commentary advance the thesis that, for individual trials evaluating new interventions not previously studied, a one-sided hypothesis test seems sensible from each

of an ethics and cost efficiency perspective[5]. Certainly, to many health care researchers, the temptation of one tailed hypothesis testing, in which the location of type I error 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 1 error, the allure of the smaller sample size associated with one tailed testing further strengthens its attraction to some researchers. However, we argue here that one tailed testing should be avoided in health care research from an ethics and cost effectiveness rationale, especially in randomized trials in which the investigator controls the intervention. Rather than reflect 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 health care practitioners, and health care researchers inadvertently do harm to their patients.

One sided thinking and ethical restraints

In intervention clinical trials, we as investigators wish to demonstrate that the tested intervention produces benefit. As clinical researchers, we do not like to harbor the notion that the interventions we have developed for the benefit of our patients can produce harm. Nevertheless, harm is often the result, as demonstrated by the use in the past of bleedings and potent purgatives for diseases which practitioners believed they understood. These now debunked medical procedures were applied by physicians who 1) took the same medical oath for patient protection, as we do, 2) acted in the best interest of their patients, as we do and 3) believed the therapy was appropriate and beneficial, again, as we do. Health care practitioners and researchers must be ever vigilant for the hazard of patient harm because patient harm is often the consequence of our

good intent. The more strongly we believe in the benefit of a therapy, the more observant we must become for the unsuspected occurrence of harm. The two-sided test shines bright, direct light on the health researcher's darkest fear - that she, despite 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 harmful the intervention might be.

The genesis of a health care research idea is often the observation of practicing physicians. We as physicians have a particular burden here, since 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 often must appeal to the better nature of patients who are uncertain in their decisions. Physicians have learned to use tact, firmness, prestige and character together to recommend and 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 health care providers, again vehemently expressed. Thus, physicians can bring strong beliefs about the effect of therapy to the research design table.

The force behind vehement investigator 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 which involves a sponsoring agency, recruiting centers, and hundreds of health care workers. The proponents of the intervention must persuade these colleagues of theirs that the experiment is worthy of their time and labor. The investigators must convince sponsors (private or public) that the experiment should be executed, and their argument

often includes a forcefully delivered thesis on the prospects for the trial's success. This is necessary, since financial sponsors, who often must choose from among 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 persistent force pushing her toward an overwhelming belief 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 vs. Sample Size Effectiveness

An argument raised in defense of one-sided testing is sample size efficiency. Others [5] correctly point out that the one tailed test produces a reduction in the minimum research sample size, since the one sided test focuses on only one tail of the effect probability distribution. However, although the savings are apparent for a given experimental alpha, they do not occur at the level one might expect. Figure 1 depicts the relationship between the fraction of observations needed in a two-tailed test that are required in a one-tailed test in a randomized clinical experiment whose goal is to demonstrate a 20 percent reduction in clinical event rates from a cumulative control group event rate of 25 percent with 80 percent power. If we would expect that 50 percent 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 y = 0.50 for the different levels of acceptable type 1 error (alpha). The curve in figure 1 demonstrates something quite different. For example, for an alpha level of 0.05, 79 percent of the observations required in the two-tailed test are needed for the one-sided test. At any level of alpha examined, the 50

percent value is not achieved. Although smaller under one tailed testing than under 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 looking for benefit only comes at the price of being unable to draw appropriate conclusions if the investigators are wrong and the study demonstrates harm. The medical community requires assurance that the finding of harm 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, since the p value is untrustworthy when no type I error is allocated prospectively [6]. Therefore the medical community does not receive its desired assurance, placing the investigator in an uncomfortable and untenable situation. This study's finding of harm may lead to the defensible believe that the research's replication is unethical. However, since *p* value estimates are unreliable in this setting, the findings may not represent the results in the population, and therefore require a second study 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 would 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 two separate one sided studies as inferred from figure 1, everything else being equal. Thus, although less efficient, the experiment designed for a two tailed hypothesis test is more effective by removing the necessity of repetition (with its attendant ethical dilemma) when the findings of harm and not benefit are produced from the study.

The one tailed test designed 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 increasing common usage of multiple endpoints in clinical studies. Assume that a one tailed test for benefit is carried out for one primary endpoint and one secondary endpoint in a clinical trial, each of which was prospectively identified in a concordantly executed study [7]. What is the correct conclusion to be drawn for the population if the null hypothesis is not rejected for the one tailed test for the primary endpoint, but is rejected for the secondary endpoint? Since the one tailed test does not differentiate 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 obviated by two sided testing with only a small marginal increase in sample size.

It must also be pointed out that, although some authors [5] point out 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 which can terminate the study prematurely in light of early strong evidence of benefit.

Knowledge vs. faith

Furthermore, In a strictly mathematical sense, and from strictly an optimality perspective, uniformly most powerful tests are available from the family of one tailed tests[8]. 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 one 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 yields the hypothesized beneficial outcome. So ideally if clinical research is designed to have the same level of evidence for the expected outcome, the argument that one-sided testing involves a smaller sample would be baseless. However, the 2 sided test is potentially more informative when faced with an unexpected outcome.

Physicians treat their patients based on what they believe. However, the best experimental designs 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 do not *know* that effect. We may have accepted the idea of the intervention's beneficial effect because of what we have seen in practice, however our view is not objective, but skewed. Admitting the necessity of the research effort to study the intervention is a first important acknowledgment 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 in carrying out a 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 intuition 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 may be more effective. Researchers eventually carried out a large scale clinical trial to assess the effect of these antiarrhythmic agents. However, 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 fifty percent 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, since the trial was terminated quickly due to an unanticipated, mortal effect of the drug. In a trial designed by the investigators to demonstrate only a survival benefit of antiarrhythmic therapy, this therapy was discovered to be almost four 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"[10].

The investigators reacted to these devastating findings with shock and disbelieve. They had embraced the arrhythmia suppression hypothesis, to the point where they had excluded all possibility of identifying a harmful effect. Yet the findings of the experiment proved them wrong. There has been much debate on the implications of CAST for the development of antiarrhythmic therapy[11]. However, an important lesson is that health care researchers must exert the greatest possible care in forming conclusions about population effects by extrapolating their own beliefs.

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 [12] similar to the approach used by the CAST advisory board. We uphold it is inappropriate to systematically apply such retrospective levels as associated p-values are inherently uninterpretable. 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 p-values only marginal, it becomes futile to attempt to dissociate unexpected harm from true null.

The CAST experience demonstrates this sense of invulnerability to harm can ambush well-meaning investigators, delivering them over to stupefaction and confusion as they struggle to assimilate the unexpected, devastating results of their efforts. We health care researchers don't 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 to often the case. 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.

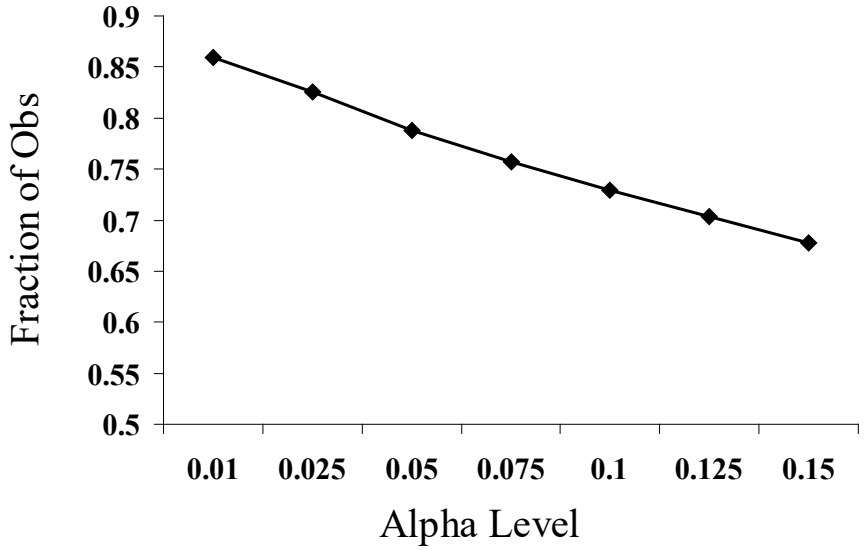
This concern, expressed early and formally in the trial's design, can be very naturally translated into effective steps taken during the course of the experiment. In circumstances where the pre-design clinical intuition is overwhelmingly in favor of a finding of benefit, the investigators should exert 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 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 experimental design, apparently 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 committee member's perspective. Journal of Biopharmaceutical Statistics 1991; 1(1):151-156,.
2. Enkin MW. One and two sided tests of significance. British Medical Journal 1994; 309: 874.
3. Dunnett CW Gent M. An alternative to the use of two-sided tests in clinical trials. Stat in Med 1996; 15:1729-1738.
4. Bland JM, Altman DG. Statistics notes:one and two sided tests of significance. BMJ1994; 309:248.
5. Knottnerus JA, Bouter LM. Commentary The ethics of sample size: two-sided testing and one-sided thinking. J Clin Epi 2001; 54:109-110.
6. Moyé LA. Random Research. *Circulation*. 2001; 103:3150-3..
7. Moyé LA. P-Value Interpretation and Alpha Allocation in Clinical Trials. *Ann Epidemiol* 1998;8:351-357.
8. Bickel PJ, Doksum KA. Mathematical Statistics: Basic Ideas and Selected Topics. San Francisco: Holden-Day, 1977: 312-332.
9. Thomas Moore. Deadly Medicine. New York: Simon and Schuster, 1995.
10. The CAST Investigators. Preliminary Report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction". N Engl J Med. 1989; 3212:406-412.

11. Pratt CM, Moyé LA. The Cardiac Arrhythmia Suppression Trial. Casting Ventricular Arrhythmia Suppression in a New Light. Circulation. 1995; 91(1): pps 245-247.
12. Koch GG. One-sided and two-sided tests and p-values. Journal of Biopharmaceutical Statistics 1991; 1(1):161-170.

Fig. 1 Fraction of Observations in a Two Tailed Test Required for a One Tailed Test



This graph depicts the fraction of observations in a two sided test required for a one tailed test. Assumes the trial is designed to detect a 20% reduction in endpoint events from a control group cumulative event rate of 25% with 80% power