# 8

# One-Sided Versus Two-Sided Testing

## 8.1 Introduction

The charge of unethical conduct stikes at the heart of every physician who takes his or her oath seriously. Ethical conduct does not play an important role in our work in healthcare— in fact it is preeminent. It is present in every conversation we have with patients, in every treatment plan we formulate, and in each of our research efforts. Decisions concerning test sidedness clearly define the battlelines where a researcher's deep-seated belief in therapy effectiveness collides with his obligatory prime concern for patient welfare and protection.

The issue of test sidedness in a clinical experiment may be a technical matter for mathematical statisticians, but for healthcare workers it is an ethical issue. Test sidedness goes to the heart of the patient and community protection responsibilities of physicians.

## 8.2 Attraction of One-Sided Testing

Students instantly gravitate to one-sided (benefit only) testing, and many, after concluding an introductory section in significance testing, come away with the idea that the one-sided (benefit only) testing is efficient, and that two-sided tests are wasteful.

Consider the example of a test statistic that falls on the benefit side of a two sided critical region. Many investigators argue that the evidence seems clear. The test statistic is positive — why continue to insist on placing alpha in the negative, (opposite) side of the distribution? Be adaptive and flexible, they would say. Respond to the persuasive nature of the data before you. Clearly the test statistic falls in the upper tail — the benefit tail — of the probability distribution. That is where the efficacy measure is. That is where the magnitude of the effect will be measured. Place all of your alpha, like a banner in a newly claimed land, with confidence.

## 8.3 Belief Versus Knowledge in Healthcare

Physicians may have many reasons for conducting research, but a central motivation is the desire to relieve the suffering of patients. We do not like to harbor the notion that the interventions we have developed for the benefit of our patients can do harm. Nevertheless, harm is often done. Well-meaning physicians, through ignorance, injure their patients commonly. The administration of cutting and bleeding in earlier centuries, and the use of strong purgatives in this century are only two of the most notorious examples. The administration of potent hormone replacement therapy for menstrual symptoms, and the use of surgical lavage and debridement as palliative therapy for osteroarthritic knees are only two of the most recent examples of the well-meaning community of physicians treating patients under mistaken and untested assumptions. What will they say 200 years from now about our current oncologic treatments?

While we must continue to treat our patients with the best combination of intellect, rigor, knowledge, imagination, and discipline, we must also continue to recognize what for us is so difficult to see and accept — the possibility that we may have harmed others.

Physicians often develop strongly held beliefs because of the forces of persuasion we must bring to bear when we discuss options with patients. We 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. We must be persuasive. We educate patients about their options, and then use a combination of tact, firmness, and presige to influence them to act in what we understand to be their best interest.

Of course, patients may select a second opinion, but they are the opinions of other physicians, again vehemently expressed. Perhaps the day will come when physicians will be dispassionate dispensers of information about therapy choices, but today is not that day.

However, history teaches us repeatedly that a physician's belief in a therapy does not make that therapy right. Making our belief vehement only makes us vehemently wrong. Physicians must therefore remain ever vigilant about patient harm: the more strongly we believe in the benefit of a therapy, the more we must protect our patients, their families, and our communities from the harm our actions may inadvertently cause. Strong vigilance must accompany strong belief. This is the research role of the two-sided test: to shine bright, directed light on our darkest unspoken fears — that we as physicians and healthcare workers, despite our best efforts, might do harm in research efforts.

## 8.4 Belief Systems and Research Design

The force behind vehement physician opinion can be magnified by the additional energy required to initiate and nurture an innovative and controversial research program. Relentless enthusiasm sometimes appears necessary to sustain a joint effort. 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 is worth doing, and their argument often incorportates a forcefully delivered thesis on the prospects for the trial's success. This is necessary, since sponsors, who often must choose among a collection of proposed experiments, are understandably more willing to underwrite trials with a greater chance of demonstrating benefit. It is difficult to lobby for commitments of hundreds of thousands, if not millions of dollars, to gain merely an objective appraisal of the intervention's effect. People invest dollars for success. In this high-stakes environment, the principal investigator must fight the persistent undertow to become the therapy's adamant advocate. The argument for one-sided testing is cogently made by Knotterus [1]

The one-sided (benefit only) test is all available for these strong advocates. The investigators believe the therapy will work, and sometimes they unconciously allow themselves to imperceptibly drift away from the real possibility of harm befalling their patients. The one-sided (benefit only) significance test has the allure of being consistent with their belief and providing some statistical efficiency. This is the great trap.

## 8.5 Statistical Versus Ethical Optimization

We have pointed out earlier that there is no mathematical preference for one level of type I error over another; the common selection of 0.05 is based merely on tradition. The level of alpha is arbitrary and the mathematical foundation for significance testing is not strengthened or vitiated with the choice of a particular type I error rate. However, the case is different for choosing the sidedness of a significance test. In a strictly mathematical sense, the justification for the one-sided (benefit only) test is slightly stronger than that for two-sided testing. This argument is based solely on the criteria statisticians use to choose from amonst a host of competing statistical tests.

For a given testing circumstance, there are many imaginable ways to construct a significance test for a null and an alternative hypothesis. How does one decide which is best? Statisticians address this important question by comparing the properties of different tests [2].[*] A uniformly most powerful (UMP) test is a hypothesis test that has superior power (for the same alpha level) than any other competitor test.

When does a UMP test exist? When working with commonly used distributions (e.g., the normal Poisson, binomial distribution), one-sided significance tests concerning the parameters of these distributions are UMP tests. The two-sided tests are not UMP precisely because the type I error has to be apportioned across the two tails. Allocating type I error in each tail when the test statistic from any one experiment can fall in only one tail engenders an inefficiency in the significance test. Thus, strictly from an optimality perspective, the one-sided test is superior.

However, are statistically optimal tests the best criterion in healthcare? Are they of necessity clinically ethical? This is an issue where the tension between the mathematics and the ethics of clinical research.

One-sided (benefit only) testing speaks to our intuition as researchers. We believe we know the tail in which the test statistic will fall. Why not put all of the type I error there? The one-sided test resonates with our own belief in the effective-

[*] This examination is actually very straightforward. Competing tests are compared by keeping alpha constant, and then computing the probability that each test will reject the null hypothesis when the alternative hypothesis is true. The test that has the greatest power is preferred. The theory is easy to understand, although sometimes the mathematics can be complicated. Lehman's text is one of the best elaborations of this work.

ness of the therapy. We have experience with the therapy, and can recognize its effectiveness in patients. We may even come to rely on the therapy in our practices. Why place alpha in the tail of the distribution in which the test statistic will not fall? We are convinced of the importance of managing type I error. Why choose to waste it now? Why should we allocate type I error for an event that we do not believe will occur?

The operative word here is "believe." 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. We are convinced of the intervention's efficacy, but we do not know the efficacy. We accept the therapy because of what we have seen in practice. However, our view is not objective, but skewed. We may have seen only those patients who would have improved regardless of the therapy. Those patients who did not respond to therapy may have involved themselves in additional therapy without our knowledge. As convincing as our platform is, it provides only a distorted view of the true effect of the intervention in the community.

We may believe strongly, but our vantage point as practicing physicians assures us that we have not seen clearly. Admitting the necessity of the trial is an important acknowledgment that the investigator does not know what the outcome will be. Therefore, an important requirement in the design of an experiment is that the investigators separate their beliefs from their information. The experiment should be designed based on knowledge of rather than faith in the therapy.

There are many remarkable examples of the surprises that await physician scientists who are not careful with the tails of the distribution. Two singularly notorious illustrations follow.

## 8.6 "Blinded by the Light": CAST

The "arrhythmia suppression" theory is an example of the difficulties that occur when decisions for patient therapy at the community level are based on untested physician belief rather than on tried and tested fact [3].*

In the middle of this century, cardiologists began to understand that heart arrhythmias were not uniform, but instead, occurred in a spectrum with well-differentiated mortality prognoses. Some of these unusual heart rhythms, such as premature atrial contractions and premature ventricular contractions, are in and of themselves benign. Others, such as ventricular tachycardia, are dangerous. Ventricular fibrillation, in which vigorous, coordinated ventricular contractions are reduced to non-contractile, uncoordinated ventricular muscle movement lead to immediate death. The appearance of these dangerous rhythms was often unpredictable; however, they were common in the presence of atherosclerotic cardiovascular disease and, more specifically, were often present after a myocardial infarction.

Drugs had been available to treat heart arrhythmias, but many of these (e.g., quinidine and procainamide) produced severe side effects and were difficult for patients to tolerate. However, scientists were developing a newer generation of drugs (e.g., ecanide, flecanide, and moritzacine) that produced fewer side effects.

* Much of the discussion is taken from Thomas Moore's book *Deadly Medicine* [3].

The effectiveness and safety of these newer drugs were examined in a collection of case series studies. As the number of patients placed on these drugs increased, the perception was that patients with dangerous arrhythmias taking the new medications were showing some reversal of these deadly heart rhythms and perhaps some clinical improvement.

Unfortunately, the absence of a control group makes case series studies notoriously difficult to assess. When patients survived, their survival was often attributed to drug therapy. However, patient deaths are often not counted against the therapy being tested. When the patient died, investigators suggested that the patient was just too sick to survive i.e., no drug could have helped the patient. Thus the drugs were credited for saving lives, but commonly were not blamed for patient deaths. Despite some debate, a consensus arose that patient survival was also improved.

The consensus that these newer drugs were having a beneficial impact on arrhythmia suppression, and perhaps on mortality gained momentum as prestigious physicians lent their imprimatur to the arrhythmia suppression hypothesis. The sponsors of these drugs, along with established experts in cardiology and cardiac rhythm disturbances, continued to present data to the federal Food and Drug Administration (FDA), pressuring the regulatory body to approve the drugs for use by practicing physicians. This body of evidence presented did not contain a randomized controlled clinical trial to test the efficacy of these compounds – only case series. However, after persistent lobbying, intense advisory committee deliberations, and extensive public and private discussions, the FDA relented and approved the new antiarrhythmic agents. As a consequence of this approval, physicians began to prescribe the drugs not just to patients with severe rhythm disturbances, but also to patients with very mild arrhythmias. This expanded use was consistent with the growing consensus that these drugs were also beneficial in blocking the progression from mild heart arrhythmias to more serious rhythm disturbances. The FDA was extremely uncomfortable with this untested expansion, but was powerless to limit the use of the drugs. Soon, many of the nation's physicians were using the drug to treat the relatively mild ventricular arrhythmia of premature beats.

However, there were researchers who were interested in putting the arrhythmia suppression hypothesis to the test. Working with the National Institute of Health (NIH), they designed an experiment called CAST (Cardiac Arrhythmia Suppression Trial) that would randomize almost 4,400 patients to one of these new antiarrhythmic agents or placebo and follow them over time. Clinical trial specialists with established, prolific experience in clinical trial methodology designed this study. This experiment would be double blind, so that neither the physicians administering the drug nor patients taking the drug would know whether the agent they consumed was active. These workers computed that 450 deaths would be required for them to establish the effect of the therapy. However, they implemented a one-sided trial design, anticipating that only therapy benefit would result from this research effort. As designed, the trial would not be stopped because the therapy was harmful, only for efficacy or the lack of promise of the trial to show efficacy. They had no interest in demonstrating that the therapy might be harmful. Thomas Moore in *Deadly Medicine* states (pps 203–204),

> The CAST investigators…wanted a structure that eliminated the possibility of ever proving that the drugs were harmful, even by accident. If the drugs were in fact harmful, convincing proof would not occur because the trial would be halted when the chances of proving benefit had become too remote.

The fact that the investigators implemented a one-sided trial design reveals the degree to which they believed the therapy would reduce mortality. However, the Data and Safety Monitoring Board[*] insisted on reviewing the data using a significance test at the $\alpha = 0.025$ level. This was tantamount to a two-sided hypothesis test for the interim data review. However, the board did not overtly contravene the investigators' desire that this should be a test for benefit.

The CAST investigators' attempts to recruit patients were commonly met by contempt from the medical community. Clinicians were already using these drugs to suppress premature ventricular contractions. Why was a clinical trial needed? Hadn't the FDA already approved these new antiarrhythmic agents? Had not numerous editorials published by distinguished scientists in well-respected journals clarified the benefits of this therapeutic approach? Why do an expensive study at this point? Furthermore, why would a physician who believed in this therapy agree to enter her patients into a clinical trial where there was a fifty-fifty chance that the patient would receive placebo therapy? Vibrant discussions and contentious exchanges occurred during meetings of physicians in which CAST representatives exhorted practicing physicians to recruit their patients into the study. At the conclusion of a presentation by a CAST recruiter, a physician, clearly angry, rose and said of the trial recruiter, "You are immoral," contending that it was improper to deny this drug to half of the randomized participants [1].

However, well before recruitment was scheduled to end, CAST scientists who were mandated to collect and study the data noted an unanticipated effect emerging. After one-third of the required number of patients were recruited for the study, the data analysis provided shocking results. Out of the 730 patients randomized to the active therapy, 56 died. However, of the 725 patients randomized to placebo there were only 22 deaths. In a trial designed to demonstrate only the benefit of antiarrhythmic therapy, active therapy was almost four times as likely to kill patients as placebo. In this one-sided experiment the *p*-value was 0.0003, in the "other tail" [4].

The investigators reacted to these devastating findings with shock and disbelief. Embracing the arrhythmia suppression hypothesis they had excluded all possibility of identifying a harmful effect. It was difficult to recruit patients to this study because so many practicing physicians believed in suppressing premature ventricular contractions. 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. However, an important lesson is that physicians

[*] The Data and Safety Monitoring Board is a group of external, scientists, separate and apart from the trial's investigators, who periodically review the data in an unblinded fashion. This board's responsibility is to determine if there is early evidence of safety or harm.

cannot form conclusions about population effects by extrapolating their own beliefs.

## 8.7 LRC Results

CAST exemplified the reaction of investgigators who were forced to confront the findings of harm when they held the full expectation of therapy benefit. The issue in the Lipid Research Clinic (LRC) study was one of efficacy. For these investigators, not even a one-sided test was sufficient to insure success.

As we have seen earlier, LCR studied the role of cholesterol reduction therapies in reducing the risk of clinical events. Designed in the 1970s by lipidologists working in concert with experienced clinical trial methodologists, the LRC trial set out to establish with some finality the importance of cholesterol level reduction in reducing the clinical sequelae of atherosclerotic cardiovascular disease. It was designed to randomize patients either to cholesterol reduction therapy or to no therapy, and then to follow these patients over time, counting the number of fatal and nonfatal myocardial infarctions that occurred. LRC required over 3,500 patients to be followed for seven years to reach its conclusion, incorporated into a prespecified hypothesis test. The final trial test statistic would be assessed using a prospectively declared one-sided hypothesis test at the 0.01 level. If the resulting *z*-score at the trial's conclusion was greater than 2.33, the investigators would conclude that the therapy was beneficial.

The investigators did not underestimate the importance of their work. They knew the field was contentious and that their study would be criticized regardless of its findings. These researchers therefore designed the effort with great care. Upon its completion, they converted their lucid protocol into a design manuscript, publishing it in the prestigious *Journal of Chronic Diseases* [5]. This was a praiseworthy effort. The investigators prospectively and publicly announced the goals of the research effort and, more importantly, disseminated the rules by which they would decide the success or failure of the experiment for all to review before the data were collected and tabulated. This is one of the best approaches to reducing experimental discordance.

In 1984, the study's conclusion was anticipated with great excitement. When published in the *Journal of the American Medical Association* [6], the researchers revealed that active therapy produced an 8.5% reduction in cholesterol. Furthermore, there were 19% fewer nonfatal myocardial infarctions and 24% fewer deaths from cardiovascular disease in the active group. The final *z*-score was 1.92. The *p*-value was less than 0.05. The investigators determined the trial to be positive.

However, the preannounced threshold was $z > 2.33$. Since the achieved *z*-score of 1.92 did not fall in this critical region, the study should have been judged as null, and therapy unsuccessful by the LRC investigators' own criteria. Yet the investigators concluded that the experiment was positive. Their rationale was that the significance test should now be interpreted not as a one-sided test with a significance level of 0.01 as was planned and announced, but as a one-sided test at the 0.05 level. This adjustment changed the critical region to be $z > 1.645$.

They changed the significance level of the test based on the findings of the trial! The fact that the investigators published a design manuscript prospectively, highlighting the rules by which the trial would be judged and the standards to which the trial should be held, makes the change in the required test significance level singularly ignoble. It is hard to find a clearer example of alpha corruption than this.

As we might expect, there are many plausible explanations for the diluted finding of efficacy for LRC. The cholesterol reduction therapy chosen for the trial, cholestyramine, was difficult for patients to tolerate. This difficulty led to fewer patients taking the medication, vitiating the measured effectiveness of the compound. It must be said that even with this weak cholesterol reduction effect, the investigators were able to identify a trend for a reduction in morbidity and mortality associated with cholestyramine. The study produced new information that would serve as a firm, scientifically based foundation for the next clinical experiment.

However, the investigators chose instead to fly in the face of their own prospective rules for assessing the strength of evidence in their study, resulting not in illumination, but in withering criticism from the scientific community. The LRC investigators had too little objective evidence to believe in the cholesterol reduction hypothesis as strongly as they did at the trial's inception. They believed in it only to the point of excluding a formal examination for the presence of a harmful effect of cholestyramine, choosing instead a one sided (benefit only) evaluation. This belief also led them to take the astounding step of corrupting their study when the type I error appeared larger than they anticipated. Rather than bolster the cholesterol reduction hypothesis, perhaps refining it for the next study, their alpha corrupting maneuver besmirched it. The one-sided test was a symptom of "strong belief disease."

In healthcare, time and again, well-meaning researchers identify an intervention they believe will be effective in alleviating suffering and perhaps prolonging life. This noble motivation is the raison d'ệtre of medical research. Unfortunately, all too often the fervor of this belief spills over, tainting the design and execution of an experiment in order to demonstrate efficacy. These experiments are often encumbrances to medical progress rather than bold steps forward [7] .

The one-sided test is not the disease, only the disease's most prominent sign. The disease is untested investigator belief masquerading as truth.

## 8.8 Sample Size Issues

Another argument raised in defense of one-sided testing is sample size efficiency. With concern for only one of the two tails of the probability distribution, one might naturally expect there to be a substantial reduction in the size of the sample since the one-sided test focuses on only one tail of the probability distribution of effect.

However, although the savings are apparent, they do not occur at the level one might expect. Figure 8.1 depicts the relationship between the fraction of observations needed in a two-sided test that are required in a one-sided test, from the design of 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.

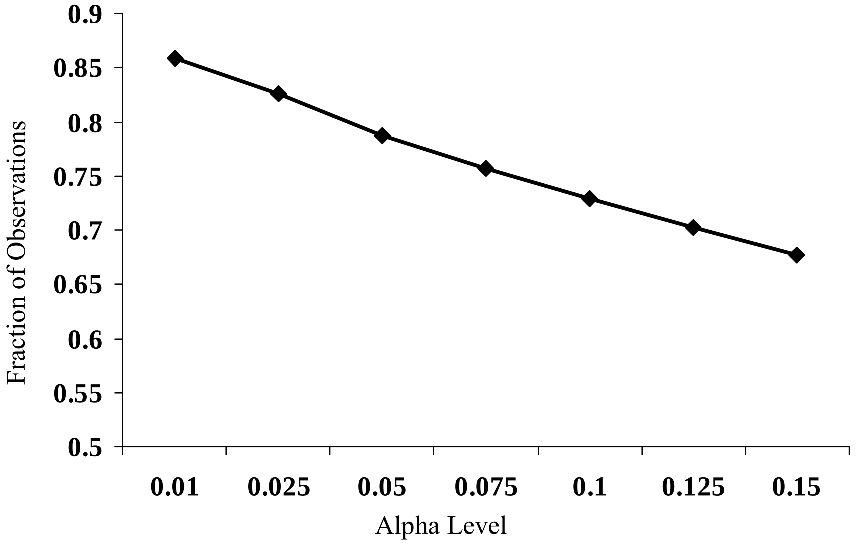

**Fig. 8.1**. Fraction of observations in a two-sided test required for a one-sided test.

From Figure 8.1, we would expect that if 50% of the observations required for a two-sided test were needed for a one-sided significance test in this example, then the curve would reveal a flat line at *y* = 0.50 for the different levels of alpha. The curve demonstrates something quite different. For example, for an alpha level of 0.075, 76% of the observations required in the two-sided test are needed for the one-sided test. At any level of alpha examined, the 50% value is not achieved. Thus, while some savings are evident, the number of required observations for a one-sided test are more than 50% of that necessary for the two-sided test. The sample size savings for the one tailed test are much smaller than one might naively expect.

## 8.9 Hoping for the Best, Preparing for the Worst

There are important limitations in carrying out a one-sided test in a clinical research effort. In my view, the major difficulty is that the one-sided testing philosophy reveals a potentially dangerous level of investigator consensus that there will be no possibility of patient harm.

As we have seen in CAST, 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 physicians 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 too often the case. The intelligent application of the two-sided test requires deliberate efforts to consider the possibility of patient harm during the design phase of the 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 predesign clinical intuition is overwhelmingly in favor of a finding of benefit, the investigators should go out of their way to assert a level of alpha that will provide community level protection from harm in the research program. It is fine to hope for the best as long as we prepare for the worst.

However, the use of a two-sided test does not in and of itself guarantee adequate community-level protection. In these circumstances, the investigators' mandate for alpha allocation extends beyond the simple stipulation that significance testing be two-sided.

## 8.10. Symmetrics versus Ethics

Once we have made the decision in the design phase of an experiment for a two-sided significance test, we very naturally and easily divide alpha into two equal components placing half in the tail signifying harm and half in the benefit tail of the probability distribution. This assumption of the symmetric allocation of alpha is perhaps the easiest to justify, but locks us into a reflexive choice involving a community protection issue. It would be useful to examine other options.

Consider, for example, an investigation of the treatment of diabetes mellitus. Diabetes, in addition to its serious renal, microvasculature, lens and retinal complications, is a well-known risk factor for atherosclerotic heart disease. Patients with diabetes mellitus have a greater risk for atherosclerotic disease than the general population. However, the debate over the control of adult onset (type II) diabetes mellitus has persisted for several decades. Should blood glucose be strictly monitored and confined to a narrow range (tight control), or should it be allowed to vary in accordance with a more liberal standard? Suppose an investigator is interested in determining if the tight control of adult onset diabetes mellitus can reduce the risk of death due to cardiovascular disease.

The investigator believes that the risk of cardiovascular disease in a community of type II diabetics can be reduced by 20% with the tight control of diabetes. He plans the following experiment. Select a random sample of patients suffering from adult-onset diabetes from the population at large, allocate them randomly to either standard control or tight control, and follow these patients for five years, counting the number of cardiovascular deaths in each group.

The investigator believes that the tight control group will experience fewer cardiovascular events, but he is also interested in protecting his community. He acknowledges that, despite his belief, the result might not be what he expects. Appropriately separating his belief from his information, he designs a two-sided test. A test statistic that appears in the extreme higher tail of the normal distribution would suggest benefit; one in the lower tail would suggest harm. He does not expect harm to occur as a result of the tight control, but he prepares for it.

This strategy is consistent with the points we have made earlier in this chapter, but is it sufficient? In order to decide in favor of benefit, the investigator needs a cardiovascular event rate to be 20% less in the tight control group than in the standard control group. However, must the tight control intervention be associated with 20% greater cardiovascular mortality before the investigator will conclude that tight control is harmful? Is this symmetric approach of strength of evidence in line with the oath to "first do no harm"? The symmetry argument ensures that the strength of evidence of harm is the same as that required to conclude benefit, but if the ethical concern is harm, perhaps less evidence of harm should be required.

The "first do no harm" principle identifies protection as the primary issue. If we are to protect the patients, we must be especially vigilant for the possibility of harm. We must stop the trial if harm occurs, at the acceptable risk of placing new therapy development on hold. Unfortunately, therapy used for the treatment of type II diabetes mellitus has been associated with harm in the past. Knowing this, the investigator must assure himself and his patients that he will remain vigilant so that the new method of diabetes control he is using will not harm his patients.

Consider the following alternative strategy for alpha allocation. Recognizing that his primary responsibility is to protect the community from a harmful effect of the tight control intervention, the investigator decides that a test statistic suggesting harm should not be as extreme as the test statistic he would accept as suggesting benefit. As a first approximation for the overall alpha for the experiment, the investigator chooses an alpha at 0.05. He apportions 0.03 of this alpha in the harm tail of the distribution and the remaining 0.02 in the benefit arm (Table 8.1).

**Table 8.1**. Alpha Table: Trial Conclusions

| Primary Endpoint | Alpha Allocation | *P*-value |
|---|---|---|
| Harm | 0.030 | 0.030 |
| Benefit | 0.020 | 0.015 |
| Total | 0.050 | 0.045 |

This leads to the following decision rule (*TS* is the test statistic).

Reject the null hypothesis in favor of harm if $TS \leq -1.88$.
Reject the null hypothesis in favor of benefit if $TS \geq 2.05$.

The investigator's greater concern for the possibility that he is harming his patients is explicitly expressed in this significance testing. The sample size is chosen to be adequate for the tail of the distribution with the smallest alpha allocation.

Now, at the conclusion of the concordantly executed trial, the test statistic is computed and is seen to be 2.18. What is the *p*-value for the study? It is the probability that a normal random variable is less than or equal to –1.88 plus the probability that a normal deviate is greater than 2.18 or $P[Z < -1.88] + P[Z > 2.18] = 0.03 + 0.015 = 0.045$. One would correctly conclude that this concordantly executed study is positive. Note that this *p*-value construction is different from the traditional *p*-value computation.

Traditionally with a test statistic of 2.18, the *p*-value would be computed as

$$\boldsymbol{P}\left[|Z|>2.18\right]=\boldsymbol{P}\left[|Z|<-2.18\right]+\boldsymbol{P}\left[|Z|>2.18\right]=2\boldsymbol{P}\left[|Z|>2.18\right]$$
$$=(2)(0.015)=0.030.$$

In this computation, the probability of a type I error is reduced from the pre-specified level of 0.05, not just on the benefit side of the distribution, but on the harm side as well. How is this justified? Why should we transmute the reduction of the possibility of type I error for benefit into a reduction in the possibility of the type I error for harm? The only justification for this reduction in the harm arm of alpha is the *a priori* notion of symmetry. If the investigator at the trial's inception constructs the two-sided test as symmetrical, he is arguing that the alpha error be channeled symmetrically at the end of the experiment. His best guess at the inception of the study was at the 0.05 level, so he apportions alpha symmetrically as 0.025. At the conclusion of the experiment, although the type I error in the benefit arm decreased from 0.025 to 0.015, the *a priori* assumption of symmetry remains in force. The finding of 0.015 in the benefit side of the distribution results in a total alpha expenditure of 0.030.

However, with the removal of the symmetry constraint *a priori*, there is no longer justification for the symmetric construction of the *p*-value. The possibility of a type I error for harm has not been updated by the experiment and remains unchanged at 0.035. The estimate of the magnitude of type I error in the benefit portion of the efficacy distribution is decreased from 0.035 to 0.015, and the *p*-value for the experiment is 0.030 + 0.015 = 0.045. It would be a mistake to view the results of this experiment as barely significant, as is often done in the reflexive 0.05

Let's return to the design phase of this argument to explore one other consideration. One could make a case that, given the strength of evidence so far suggesting the possibility of long-term harm of the treatment of diabetes, even more alpha should be evidence for harm (Table 8.2)

**Table 8.2.** Alternative Allocation and Trial Conclusion (Test Statistic = 2.18)

| Primary Endpoint | Alpha Allocation | *P*-value |
|---|---|---|
| Harm | 0.1000 | 0.1000 |
| Benefit | 0.0250 | 0.0125 |
| Total | 0.1250 | 0.1125 |

In this scheme, the overwhelming ethical concern is to identify the possibility of harm. The total alpha for the experiment is 0.125, but only 0.025 is identified for benefit. In this circumstance, the test statistic must be larger than 1.96 to demonstrate evidence of benefit, but only –1.28 to demonstrate evidence of harm.

Suppose now that the test statistic is observed as 2.18 (Table 8.2.). Then the experiment is positive and the total alpha expended

$$\boldsymbol{P}\left[Z < -1.28\right] + \boldsymbol{P}\left[Z > 2.18\right] = 0.10 + 0.015 = 0.115.$$

The study should be viewed as positive for benefit since the total alpha allocated is less than the alpha allocated at the trial's beginning. Also, adequate population protection was provided for harm. Critics of the study who would minimize the findings of this trial should be reminded that the alpha for significance for benefit = P[Z > 2.18] = 0.015, suggesting that sampling error is an unlikely explanation of the research findings. However, the overriding concern for population protection was the prospectively specified, asymmetric critical region.

In this case, the investigators have allocated alpha prospectively and intelligently, requiring sufficient strength of evidence for benefit while vigorously exercising their mandate for community protection. Nevertheless, many will feel uncomfortable about this nonstandard scheme for alpha allocation. We must keep in mind that the allocation the investigators have chosen stands on good ground, adhering to the following requirements, necessary and sufficient for its clear interpretation. First, alpha is allocated prospectively, in great detail. Secondly, the protocol involved the random selection of subjects from the population and the random allocation of therapy. Finally, the experiment was executed concordantly with no consequent alpha corruption. The investigators have wielded their mandate handsomely, providing adequate community protection while insisting on the same standard of efficacy required by the more traditional scientific community. The experimental findings would be reduced only in the eyes of those with a reflexive requirement of 0.05 with symmetric tails.

## 8.11 Conclusions

I believe one-sided (benefit only) testing reflects a mindset of physicians and healthcare researchers who believe their intervention can produce no harm, a philosophy that has been shown repeatedly to be faulty, and dangerous to patients and their families. Investigators who agree to the one-sided (benefit only) approach to significance testing in a clinical experiment have closed parts of their minds to the possibility of harm entering into an avoidable flirtation with danger. In this sense, the one-sided test is not the disease, it is only a symptom.

We as physicians and healthcare workers feel strongly about the treatment programs we advocate. This is required in our profession. However, these strong feelings often betray us since our day-to-day experience does not provide an objective view of the treatments. We need every tool we can find to help us gain that objective vantage point. The use of a two-sided significance test is of utmost importance. A forceful, intelligent argument for ethics will argue not only for a two-sided test, but asymmetrical allocation of alpha. Prospective identification of alpha is again critical here, and community protection predominates all other concerns. A sensitive, perceptive, ethical approach to alpha allocation for sidedness can complicate experimental design, but complexity in the name of ethics is no vice.

## References

1. Knottnerus JA, Bouter LM (2001) Commentary: The ethics of sample size: two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology* **54**:109–110.
2. Leyman EL (1986) *Testing Statistical Hypotheses.* New York. John Wiley and Sons.
3. Moore T (1995) *Deadly Medicine*. New York Simon and Schuster.
4. The CAST Investigators (1989) Preliminary Report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine*. **3212**:406–412.
5. The Lipid Research Clinic Investigators (1979) The Lipid Research Clinics Program: The Coronary Primary Prevention Trial; Design and implementation. *Journal of Chronic Diseases* **32**:609–631.
6. The Lipid Research Clinic Investigators.(1984)The Lipid Research Clinics Coronary Primary Prevention trial results. *Journal of the Amerian Medical Association* **251**: 351–74.
7. Moyé LA**,** Tita, A (2002) Defending the Rationale for the Two-Tailed Test in Clinical Research. *Circulation* **105**: 3062–3065.