Rethinking Statistics: Basing Efficacy Alpha Levels on Safety Data in Randomized Trials

Vance W. Berger, PhD; Ginny Hsieh, PhD

VB, GH: National Cancer Institute, University of Maryland Baltimore County
Biometry Research Group, National Cancer Institute

Abstract:

It has recently been proposed that the results of safety analyses be used to determine the alpha levels for the efficacy analyses in the same trial. The rationale for this proposal is that the consequences of various actions should be considered when making a decision and selecting one of these actions. The safety of a medical intervention is certainly relevant to the relative consequences of a Type I error versus a Type II error concerning the efficacy of this same medical intervention. The purpose of this paper is to provide some examples of how the safety data can be considered in determining the alpha level for the primary efficacy comparison.

MeSH Words: Randomized Trials, Epidemiology, Alpha Level,

Introduction

Randomized trials are considered the gold standard for clinical research [1]. In randomized trials, the alpha (α) level for tests of statistical significance is generally set to 0.05, either for each test (comparison) or overall (with adjustment for the multiplicity of tests) [2]. One example is a randomized trial of St. John's wort for major depression [3], which specified that "All statistical tests were two-tailed and significance was declared at the 0.05 level". Such a statement is hardly uncommon in published reports of randomized trials. But recently, the logic of setting the alpha level to be a constant, be it 0.05 or any other constant, without regard to the specifics of the trial in question, has been challenged [2]. Another issue that has not received sufficient attention in the published literature concerns who should determine the alpha level for hypothesis tests in randomized trials. These are the two issues to be considered in this paper. We do not consider the important issue of using a one-sided or two-sided

test, as this issue has been addressed elsewhere [2].

Who Should Determine Alpha?

In randomized trials, the current practice is for investigators to set the alpha levels for their own studies. One example was provided in the Introduction. This practice has many analogies in real life, which can help to shed light on the questionable basis for this practice. Imagine an athlete trying out for a team, and having to be timed in a sprint of 40 yards, as part of this try-out. Is it appropriate for the athlete to determine the time needed to make the team, as long as this threshold to define success is pre-specified? Imagine a candidate applying for a job, and having to give a lecture, which will be evaluated as part of the application process. Is it appropriate for the candidate to determine the extent to which she needs to impress her would-be colleagues in order to receive the job offer, as long as this threshold to define success is pre-specified? Imagine a student about to take a test in school. Is it appropriate for this student to determine the score necessary on a test to receive a grade of an A for the semester, as long as this threshold to define success is pre-specified?

It seems rather clear that in none of these cases should the party being evaluated also play the role of the party setting the standard by which the evaluation is measured. The current alpha practice is equally lacking of merit, because it essentially allows investigators to determine what the consumer requires to accept the results. Such a decision is better left to the purview of the consumer. Ideally, then, the producers of research should focus on quantifying the extent to which the data they produce are convincing (p-values), rather than the extent to which the data need to be convincing in order that a consumer can be convinced (alpha). The only exception to this policy would be sample size calculations, which generally will require an alpha level to be input, and so here it is the investigator who would specify this alpha level.

One could argue that regulatory agencies act on behalf of consumers, and hence are in a better position to select the alpha level than the sponsors of trials would be. This is probably true. In fact, a regulatory agency would have to pick an alpha level to reach its own decision, which would, of course, affect consumers. But this does not negate the right of each consumer (be it a patient contemplating an elective surgery or an over-the-counter medicine, a treating physician contemplating writing a given prescription, or an HMO contemplating adding a drug to a formulary) to select his or her own alpha level, based on whatever information seems to be most relevant to the decision at hand. Some consumers may be especially concerned about the safety of a new medical intervention, while others are concerned mostly with cost and ease of administration.

If the more efficacious of two interventions also happens to be the more toxic or more expensive or less convenient of the two, then there is a trade-off, and room for disagreement regarding how to proceed. How much extra efficacy will it take to tilt the balance in favor of the more efficacious and more toxic intervention? Clearly, this is a subjective decision that, to the extent possible, needs to be left to the consumer. But how can a sponsor simultaneously address all the potential consumers of their study, and the myriads of implied alpha levels? Would a sponsor need to present the decision to be reached for every conceivable alpha level? In fact, doing so is much simpler than it might first appear to be. A single p-value allows for the assessment of statistical significance at any alpha level. One has merely to compare one's preferred alpha level to the reported p-value to determine which is smaller. If the alpha level is smaller, then statistical significance has not been reached (at that alpha level). If, on the other hand, $p<\alpha$, then statistical significance has been reached (again, at that alpha level).

A New Paradigm for Determining Alpha Levels

Even if we can agree that the selection of the alpha level is in the purview of the consumer, this still does not offer any guidance for selecting alpha levels. A hint for how to do so can come from determining what can go wrong when the usual practice of choosing the standard 0.05 for the alpha level is used. This usual practice is predicated on the notion that every analysis, or comparison, is identical in terms of the consequences of a Type I error (false positive) and a Type II error (false negative). That is, in no study need the consequences of these two types of error be identical, or even comparable,

but the consequences of each type of error is identical across studies (and across comparisons). It is immediately clear that this is an unsound assumption.

To see the problem with this assumption, consider the analogy of false positive and false negative findings in diagnostic tests, which are used to either rule out further testing for and treatment of a disease or signal the need for further testing. A false positive means more testing than is necessary, whereas a false negative means that a presumably treatable disease is not being treated. The relative consequences of these two types of error will depend on both the nature of the follow-up testing and the natural course of the disease when left untreated, relative to when it is treated. If the follow-up testing is harmful and the treatment is only minimally effective, then a false positive could be much worse than a false negative. But if the follow-up testing is not harmful and the benefits of treatment are substantial, then a false negative could be much worse than a false positive. The specifics of the situation would dictate the relative likelihood of each type of error that one would tolerate. This is analogous to the distinction between requiring "a preponderance of evidence" (moderate alpha level) for a civil trial versus "beyond a reasonable doubt" (small alpha level) for a criminal trial.

The same concerns govern the weighing of the consequences of a Type I error in hypothesis testing relative to those of a Type II error in hypothesis testing. For instance, a Type I error (false positive) in the form of claiming that broccoli prevents arthritis when in fact it does not will only result in increased consumption of broccoli, which is both inexpensive and nutritious. The lack of any real harm coming from this Type I error signals the need for a relatively large alpha level in the evaluation of broccoli for the prevention of arthritis [2]. Conversely, in the case of testing whether a new, highly toxic, expensive chemotherapy might be effective in treating cancer, a Type I error has serious consequences. If the new treatment is found to be effective, then it will be widely adopted, whether or not it is actually effective. If it is not, then treatment recipients will experience unnecessary increased toxicity, at a high cost. A small alpha level may then be necessary to offset the associated harmful effects from a Type I error.

There is nothing particular to broccoli or toxic drugs in this formulation. Indeed, the salient points to consider are safety, cost, and convenience, which can be transformed, at least intuitively, into an acceptable range of alpha levels. Qualitatively, we see that evaluations of minimally invasive, safe, inexpensive, and convenient interventions (such as broccoli) deserve more alpha than do evaluations of toxic or expensive interventions. It is reasonable, then, to customize the alpha level based on these specific parameters, which serve as measures of the relative consequences of Type I and Type II errors. Such a paradigm can help to overcome one problem confronting the feasibility of some clinical research. Specifically, consider studies that focus on rare diseases, or for some other reason have limited patient populations, and hence have little chance of recruiting enough patients to allow for a demonstration of statistical significance at the customary 0.05 alpha level. If the intervention being studied is relatively safe, inexpensive, and convenient, then the new paradigm offers a way around having to demonstrate statistical significance at the customary 0.05 alpha level. A larger alpha level could be warranted, and if it is, then this would increase the statistical power, possibly to the point of making an otherwise infeasible study feasible.

How To Calculate Alpha from Safety Criteria

We have seen that it is reasonable to allow a larger chance for a Type I error when a Type I error is not a disaster, relative to when it is. The harm caused by a Type I error depends on the safety, cost, and convenience of the intervention being evaluated, so these should, ideally, all factor in to the determination of the alpha level for the efficacy comparisons. For simplicity, our illustrations will be based on safety only, but future work will detail the methodologies to incorporate all relevant information. Although the p-value may be an imperfect measure of the strength of evidence, it is a good starting point. Taken as a group, the safety p-values should be used to set the alpha level for the efficacy analyses. If these safety p-values are two-sided, then low values would not discriminate very safe compounds from very harmful ones. Hence, it is preferable to use one-sided p-values, testing the

alternative hypothesis that the experimental treatment is more toxic than the control agent, because with this formulation, lower p-values indicate greater toxicity and larger p-values indicate greater safety of the experimental intervention.

In general, there will be many of these safety p-values, and they can be combined in a variety of ways. For illustration purposes, we choose only the lowest p-value as our measure of safety. This minimum one-sided p-value will be such that p=0.00 indicates the maximal toxicity for the active group, and p=1.00 indicates the maximal safety for the experimental group. Now, this safety p-value is to be mapped into the unit interval, [0,1], to arrive at the alpha level for the efficacy comparison. This mapping needs to be monotonic, so that larger p-values earn more alpha and vice versa. Mathematically, the function $\alpha(p)$ satisfies the conditions $\alpha(0)=0$, $\alpha(1)=1$, and if $p1<p2$ then $\alpha(p1)<\alpha(p2)$. The simplest function $\alpha(p)$ that satisfies these conditions is the identity function, $\alpha(p)=p$. Such a function would set alpha to be exactly the value of the safety p-value.

The simplicity of this approach may be appealing, but it would entail using an alpha level of 0.50 if the safety p-value is 0.50, which it would be if the experimental agent appears to be exactly as safe as the control. In such a case, it might be deemed appropriate to use a somewhat lower alpha level, because the new intervention needs to win, and not tie, to supplant the standard of care. The specific alpha level to use when the active and control interventions are equally safe is somewhat arbitrary, but one could specify that in such a case the alpha level should be the conventional 0.05. If one specifies an exponential form for the function $\alpha(p)$, so that $\alpha(p)=p^k$ for some value of k, then the imposition of this condition, $\alpha(0.5)=0.05$, would allow one to arrive at a unique function $\alpha(p)=p^{4.32}$, since $0.5^{4.32}=0.05$. This is one possibility to consider.

If we use $\alpha(p)=p^{4.32}$, and if p=0.2000 for the safety comparison, then $\alpha=0.2^{4.32}=0.00096$ for the efficacy comparison. This alpha level may seem unnecessarily conservative, as a moderate p-value has been converted to a very small alpha level. Of course, one need not use an exponential function. The simplest function might be a step function that specifies a range of acceptable alpha levels, such as $\alpha(p)=0.01$ for $p<0.2$, $\alpha(p)=0.05$ for $0.2\leq p<0.4$, $\alpha(p)=0.10$ for $0.4\leq p<0.6$, $\alpha(p)=0.15$ for $0.6\leq p<0.8$, and $\alpha(p)=0.25$ for $p\geq 0.8$. There is nothing magical about this specific function, but its simplicity is appealing, and it satisfies the required conditions, so this is the one we will pursue in the examples. Regardless of the specific function, a large safety p-value would indicate that the experimental intervention is relatively safe, so that less evidence of efficacy would be required, leading to a higher corresponding efficacy alpha level.

Examples

In this section we offer four examples to illustrate the application of the alpha function defined in the previous section. Specifically, $\alpha(p)=0.01$ for $p<0.2$, $\alpha(p)=0.05$ for $0.2\leq p<0.4$, $\alpha(p)=0.10$ for $0.4\leq p<0.6$, $\alpha(p)=0.15$ for $0.6\leq p<0.8$, and $\alpha(p)=0.25$ for $p\geq 0.8$.

Example 1. First, consider a randomized trial of tricyclic antidepressant medication, stress management therapy, and their combination for chronic tension headache [4]. In this study "A modified Bonferroni procedure was used to control the family-wise Type I error rate for the five comparisons at 0.05". Also, several p-values were provided for safety analyses, and all are relevant. For simplicity we consider only the occurrence of any adverse event. In the analysis discussed, the rates of adverse events were 78/97 (80%) for antidepressant medication and 27/90 (30%) for placebo (p=0.001 as reported by the authors, but the one-sided Fisher exact test yields p=0.0000). There were more than just these two treatment groups, and, again, there were more safety comparisons than just this, but still we will proceed with p=0.0000 as the relevant safety p-value for the purposes of setting the alpha level for the efficacy comparisons, or at least those involving these two treatment groups. Based on the identity function, $\alpha(p)=p$, then the study will earn no α at all, but based on a step function, $\alpha(p)=0.01$ for $p<0.2$, this study could set the efficacy alpha level to 0.01. The efficacy p-values reported in the abstract were 0.006, 0.003, and 0.001. With no adjustment for multiplicity, all are significant.

Example 2. The second example is a randomized trial of St. John's wort for major depression [3]. In this study "All statistical tests

were two-tailed and significance was declared at the 0.05 level". Also, the only adverse event that was statistically significantly different across the treatment groups was headaches (39/95 or 41% in the St. John's wort group, 25/100 or 25% in the placebo group, p=0.02 as reported by the authors but p=0.0126 by the one-sided Fisher exact test). If other less significant safety p-values were combined with this one in some way, then this would result in less significance, and a larger p-value, but for simplicity we take 0.0126 to be the relevant safety p-value for the purposes of setting the alpha level for the efficacy comparisons. The step function allows for $\alpha(p)=0.01$ for $p<0.2$. The efficacy p-values reported in the abstract were $p<0.001$, p=0.16, and p=0.58. With an alpha level of 0.01, only the first of these efficacy p-values would be statistically significant.

Example 3. The third example is a randomized trial of atorvastatin for early recurrent ischemic events in acute coronary syndromes [5]. In this study "A significance level of p=0.001 was used for each interim analysis, with a significance level for the final analysis adjusted to p=0.049 to preserve to [sic] the overall Type I error rate at p=0.05. The testing of all secondary objectives was done at the two-sided p=0.05 level of significance". Also, abnormal liver transaminase levels occurred in 38 of the 1538 patients in the atorvastatin group and in 9 of the 1548 patients in the placebo group ($p<0.001$ as reported by the authors, but the one-sided Fisher exact test yields p=0.0000). Again, we take 0.0000 to be the relevant safety p-value for the purposes of setting the alpha level for the efficacy comparisons. Based on the reasoning that Type I error could have significant impact of having abnormal liver transaminase level, the alpha should be set accordingly at a conservatively low level, 0.01. The primary efficacy p-value reported in the abstract is p=0.048, which would not attain statistical significance.

Example 4. The fourth example is a randomized trial of glycine antagonist for patients with acute stroke [6]. In this study, "Two interim analyses and a final analysis were scheduled. These used a two-sided test with equal allocation of Type I error (0.025 in either direction), but with asymmetrical stopping boundaries". Also, Table 6 of [6] tabulates the safety data, including the data for agitation. It turned out that 77/819 patients in the gavestinel group (9%) and 93/786 patients in the placebo group (12%) experienced agitation (p=0.1233 by the two-sided Fisher exact test). Now the one-sided p-value is 0.0667, but this favors the active treatment, and so for the purposes of allocating alpha for the efficacy comparisons, the lower this p-value the better. The null probability of obtaining the value of the test statistic actually observed is 0.0186, so with this, we can calculate the reversed one-sided p-value (in favor of placebo) to be 0.9519. This is found by subtracting from one the difference 0.0667-0.0186, so 1-0.0481=0.9519, which we take to be the relevant safety p-value for the purposes of setting the efficacy alpha level. Logically, with treatment group being safer than the placebo, we would expect this study to earn the most alpha, at 0.9519 if the identity function is used, or at 0.25, if the step function is used. The three-month survival p-value was p=0.11, which the authors stated was not significant. At the 0.25 alpha level, however, it is.

Discussion

We believe that the appropriate alpha level is dependent on the study context, particularly on the relative adverse consequences of a Type I error versus a Type II error. To ensure the integrity of evaluations, the "one size fits all" approach should be replaced by a more flexible and rational system for allocating alpha. We have offered one particular flexible paradigm, and we have illustrated its use on four recent randomized trials. This approach transforms the safety measures achieved within the same study into specific alpha levels for the efficacy comparisons. Ultimately, however, each consumer of the research would need to reach his or her own decision regarding a reasonable alpha level to apply. In addition to the specific function, we provided also a framework to allow others to develop their own approaches to assigning these alpha levels.

References:

[1]. Berger VW, Bears JD. When can a clinical trial be called "randomized"? Vaccine 2003; 21;468-472.

[2]. Berger VW. On the generation and ownership of alpha in medical studies. Controlled Clinical Trials 2005, in press.

[3]. Shelton RC, Keller MB, Gelenberg A, et al. Effectiveness of St. John's Wort in Major Depression. *JAMA* 2001; 285(15):1978-1986.

[4]. Holroyd KA, O'Donnell FJ, Stensland M, et al. Management of Chronic Tension-Type Headache with Tricyclic Antidepressant Medication, Stress Management Therapy, and Their Combination. *JAMA* 2001; 285(17):2208-2215.

[5]. Schwartz GG, Olsson AG, Ezekowitz MD, et al. Effects of Atorvastatin on Early Recurrent Ischemic Events in Acute Coronary Syndromes. *JAMA* 2001; 285(13):1711-1718.

[6]. Sacco RL, DeRosa JT, Haley EC, et al. Glycine Antagonist in Neuroprotection for Patients with Acute Stroke. *JAMA* 2001; 285(13):1719-1728.

Competing interests: None Declared

Funding: None

Correspondence to:

Vance W. Berger, PhD
Executive Plaza North, Suite 3131
6130 Executive Boulevard, MSC 7354
Bethesda, MD 20892-7354
(301) 435-5303 (voice),
(301) 402-0816 (fax),
vb78c@nih.gov