

# Using optimal alpha rather than traditional NHST thresholds reverses conclusions for 14-34% of cancer clinical trials.

**Abstract:** Clinical trials are conducted to test treatment efficacy and safety and to provide high quality data for healthcare decision making. Most phase III cancer clinical trials continue to use traditional Null Hypothesis Significance Testing (NHST), which has been criticized for using an arbitrary rejection threshold, sensitivity to sample size, and ignoring Type II error or the effect size. Optimal alpha is a method for setting the rejection threshold that is explicitly designed to address the key limitations of NHST. This study re-analyzed 2,197 statistical tests from published phase III cancer clinical trials using optimal  $\alpha$ , and compared conclusions that were originally reached using traditional NHST thresholds to those reached using optimal  $\alpha$ . Our results show that in 23.6% of the tests, using optimal alpha would have resulted in a different conclusion than was reached using traditional thresholds.

**One Sentence Summary:** We conclude that adopting optimal  $\alpha$  for setting rejection thresholds would improve our ability to make correct inferences from cancer clinical trials.

## INTRODUCTION

Clinical trials are the ‘gold standard’ in healthcare decision-making (1) with four phases. If there is preliminary evidence in phases I and II that the treatment offers a benefit, Phase III trials broaden the population that receives the treatment, and are used to identify side effects and compare results to existing treatments (1-2). Phase III clinical trials have endpoints that are more clinically important; such as survival, or absence of clinical events (e.g. stroke), and are considered a definitive phase before the sponsor submits their drug effectiveness claim to the regulatory agency (1,3).

Investigators evaluate clinical trials by considering characteristics such as: the context in which the trial is being conducted, the severity of the disease, the study design and research methods, strength of findings, and the benefits of the treatment compared to its costs (1). However, despite the careful design and evaluation of trials, final conclusions and recommendations can be incorrect.

Darrow et al (4) suggest that FDA is often criticized for either slowing patients’ access to promising therapies, or approving treatments that may be “ineffective, unsafe, or both”. The analysis of 222 new treatments approved by the FDA from 2001 through 2010, showed 32% had post-market safety events (5). Further, Das (6) showed that more than half of randomized clinical trials (59/105) supported FDA approval of anticancer drugs that subsequently failed to yield clinically meaningful benefits (7). In addition, Begley et al. (8) reviewed 53 cancer clinical trials and found that only 11% of the findings could be confirmed while Prinz et al.

(9) found that only 20-25% of findings of 67 trials were reproducible. Such failures have raised many questions about the efficacy of clinical trials in supporting FDA decision-making (7).

A contributing factor to this failure is the focus on reporting the statistical significance of a treatment (P-value) rather than considering the biological significance and clinical importance of observed effect sizes (10-13). In a review of 100 randomized clinical trials, P-values were reported for 100% of the trials, but 25% of the trials had effects that were considered clinically unimportant despite being statistically significant (11). One recommendation is to shift our clinical trials’ questions from “is there a difference between groups?” to “how large is the effect, if real?” (10,12). Although communicating relative effect sizes in randomized clinical trials has improved, investigators often ignore absolute effect sizes perhaps because they are “unimpressively small in magnitude” (11).

In response to the concerns raised above, FDA has (1) required registration of clinical trials on publicly available databases such as [clinicaltrials.gov](http://clinicaltrials.gov) (14) and (2) released an act that requires sponsors to submit all evidence of clinical benefit for a treatment before approval (15). Further, the National Institute of Health has collaborated with journal editors to provide principles and guidelines for clinical trialists to report their research; such as requiring the submission of detailed reporting of study design, inclusion/exclusion criteria, and data archiving. Although these solutions have likely mitigated some of the concerns, none

of these changes address a core problem - using Null Hypothesis Significance Testing (NHST) as the “default” method for making inferences.

Null Hypothesis Significance Testing is widely used in medical research although it has often been criticized (1). The American Statistical Association recently expressed many concerns about the use of null hypothesis testing (16). They emphasized that the binary decision (i.e. yes/no) associated with P-values, is not appropriate for making claims about scientific findings. Yet, this method is still entrenched in all types of clinical research.

Criticisms of NHST and P-values include: (1) Traditional NHST often uses a ‘straw man’ null hypothesis, (2) NHST is based on arbitrary thresholds, (3) p-values are commonly misinterpreted as being the probability of a null and/or alternative hypothesis being true, (4) p-values are sensitive to sample size, and (5) p-values do not measure effect sizes. The International Committee of Medical Journal Editors even recommends avoiding the reliance on NHST in most biomedical research as it fails to provide enough information about the effect size (17). Problems with traditional NHST testing may contribute to reduced reproducibility of clinical trials (18,19).

Despite these problems and criticisms, clinical trials are often analyzed with traditional statistical hypothesis testing and P-values. Following FDA guidelines, most phase III trials continue to use Null Hypothesis Significance Testing, P-values, and the conventional 1%, 5% or 10% statistical thresholds (20-21). Thus, most clinical trials are primarily concerned with

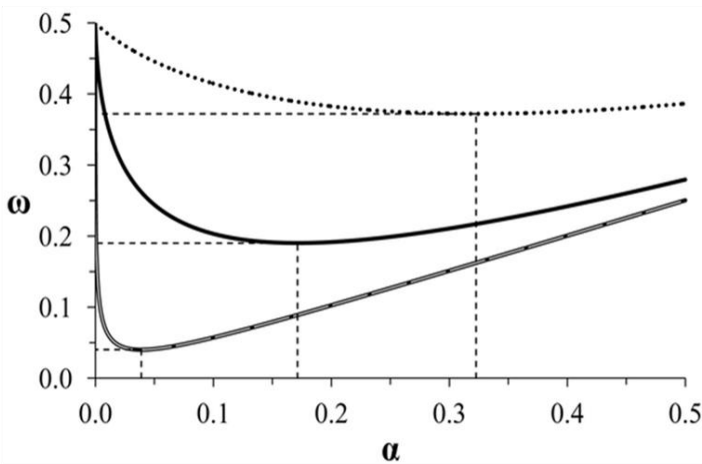
setting the maximum acceptable Type I error rate. That is, investigators and regulators aim to ensure that ineffective drugs are not incorrectly approved. However, Type II error (i.e. rejecting an effective therapy) also require consideration. Investigators try to manage this unavoidable trade-off by designing adequately powerful trials with large sample sizes (22), but it is not always feasible to have a powerful test, especially in urgent cases where the risk of unmet medical need is greater than the risk of approving an ineffective treatment (22). In those cases, FDA may order accelerated approval and a “P < 0.05 in one study would be sufficient” (20). The flaws of hypothesis testing method and the arbitrary nature of its “conventional” thresholds raise many questions about how clinical trials are evaluated. If there are useful alternatives, they should be considered.

Optimal alpha ( $\alpha$ ) is an alternative to traditional NHST that provides a study-specific statistical threshold and presents conclusions with explicit consideration of effect size and the relative costs of Type I and Type II errors (23).

Optimal  $\alpha$ : There is a non-linear negative relationship between the probability of making a Type I error under the null hypothesis ( $\alpha$ ) and the probability of making a Type II error under the alternate hypothesis ( $\beta$ ) (23). That is, all other things being equal, setting  $\alpha$  at a low value to reduce the probability of making a Type I error will inflate the probability of making a Type II error and vice versa. Optimal  $\alpha$  is designed to account for this trade-off by setting a threshold that minimizes the combined probability of making any error – the average

of  $\alpha$  and  $\beta$ :  $(\alpha+\beta/2)$  assuming equal costs of

Figure 2. Determination of optimal  $\alpha$  from the a priori combined probabilities of Type I and Type II error.



Mudge JF, Baker LF, Edge CB, Houlihan JE (2012) Setting an Optimal  $\alpha$  That Minimizes Errors in Null Hypothesis Significance Tests. PLOS ONE 7(2): e32734. <https://doi.org/10.1371/journal.pone.0032734>  
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032734>

**PLOS ONE**

Type I and II errors ( Figure 1, 23).

**Fig. 1.** Taken from Mudge et al (2012): “Determination of optimal alpha from a priori combined probabilities of Type I and Type II error.  $\alpha$  and  $\omega$  (the average of Type I and Type II error) for independent, 2-tailed, 2-sample t-tests ( $n_1 = n_2$ ). Data are for 3 (dotted line), 10 (solid line), and 30 (double line) samples per group, with critical effect sizes of 1 SD of either group. Drop lines indicate the minimum average of Type I and Type II error and its associated value of  $\alpha$ .”

Effect size has been identified as one of the most important indicators of clinical significance – it reflects the magnitude of the difference in outcomes between treatment and control groups (1). A “critical” effect size is the magnitude of the relationship between the dependent and independent variables that the investigator would consider important to detect if it exists. Optimal  $\alpha$  requires that

researchers identify the effect size they consider to be important and presents conclusions directly related to that specific effect size. Setting a meaningful critical effect size is nontrivial because there is rarely a completely objective rationale for choosing it (23-24). For example, a 3% decline in stroke events resulting from developing new drug might not be considered big enough by representatives of pharmaceutical industries but affected patients might consider such a decline to be important and beneficial. However, when there is not a clear rationale for a single critical effect size, multiple critical effect sizes representing “small”, “medium”, and “large” effects could be presented, where optimal  $\alpha$  would allow explicit conclusions for each of the critical effect sizes (23).

Optimal  $\alpha$  is an improvement of traditional NHST because it is explicitly designed to address the key problems in NHST. It recognizes the difference between “statistical” significance and “practical” significance and the importance of reporting effect sizes. Optimal  $\alpha$  minimizes the combined probabilities of making Type I and Type II errors without changing the experimental design. However, it is not clear that using optimal  $\alpha$  would regularly result in different decisions. So, the question remains: is optimal  $\alpha$  worth the effort? How often would we reach different conclusions using optimal  $\alpha$  instead of traditional  $\alpha$  of 0.05 in phase III cancer clinical trials? Our objective is to review the clinical trials literature, select tests from published papers, calculate optimal  $\alpha$  for each test and compare the conclusions reached by optimal  $\alpha$  and traditional NHST.

## MATERIALS AND METHODS

This research requires 5 steps - Step 1: Select published cancer clinical trials; Step 2: Extract observed P-value(s) and all information required to estimate optimal  $\alpha$ ; Step 3: Quality assessment and quality control of extracted data; Step 4: Use extracted information to calculate optimal  $\alpha$  for small, medium and large effect sizes; Step 5: Summarize results of comparison between optimal  $\alpha$  and traditional NHST thresholds.

### Clinical trial selection

We used the PubMed database to select all clinical trials used. We included a clinical trial if it: (1) was a phase III cancer clinical trial that used a two-group comparison method, (2) used a null hypothesis significance testing technique and an explicit  $\alpha$  significance threshold, (3) reported a P-value as a discrete number; (4) was written in English; and (5) was available online.

Further, methods for calculating optimal  $\alpha$  have been developed for seven statistical tests - t-test, regression, ANOVA, chi-square test, log-rank, cox proportional hazard ratio test, and logistic regression. We only included clinical trials that used one of these statistical tests. A study was excluded if it did not contain data required to calculate optimal  $\alpha$  (Table 1) or if the data were not provided in a clear and unambiguous form.

The PubMed database was searched from May through August 2019 using the Clinical Trial, Phase III 'article type' filter and the following search terms: "cancer" AND "log-rank" or "logrank" or "log rank"

or "cox regression" or "cox" or "chi square" or "chi-square" or " $\chi^2$ " or "logistic regression" or "ANOVA" or "t-test" or "regression". All search results were then reviewed, and studies selected based on the criteria described above. Two investigators (M.A. and S.Y.) reviewed 2,015 articles and retrieved a total of 829 articles identified as meeting the selection criteria.

### Data extraction

The 829 selected trials were divided such that the first half of the search results was analyzed by M.A., and the other half by S.Y..

We extracted the values required to calculate optimal  $\alpha$  from each article (Table 1), and the observed P-value for each test. In addition, we extracted data that described key characteristics of the clinical trials including the type of statistical test, year, location, type of intervention, primary and secondary endpoints or primary outcome, and the statistical threshold.

**Table 1 To calculate optimal  $\alpha$  for the different statistical tests**

Test	Input
Regression	Total sample size ( $N$ )
t-test	Sample sizes ( $N_1$ and $N_2$ ), type of test
ANOVA	Degrees of freedom
Chi square	Total sample size ( $N$ ), degrees of freedom
Log-rank/Cox regression	Sample sizes ( $N_1$ and $N_2$ ), proportion of subjects in control group
Logistic regression	Sample sizes ( $N_1$ and $N_2$ ), proportion of subjects in treatment group, the average incidence rate

To calculate optimal  $\alpha$ , we must use power calculations for each of log-rank, Cox regression, chi-square, logistic regression, ANOVA, Regression/correlation, and t-test. See Table 1 for a description of the inputs required for power calculations for each of the different tests. For logistic regression, the average incidence rate of the endpoint across both treatment and control groups was underreported in most clinical trials. Hence, for each logistic regression test, we used average incidence rates of 0.1 (i.e. small) and 0.5 (i.e. large) in calculations.

In cases where authors explicitly stated their primary endpoint, we extracted data related to that primary endpoint. If authors did not explicitly state their primary outcome, we had to infer it from either the title of the article or after looking at the main study question or study hypothesis in the introduction section. After data related to the primary endpoint/outcome had been extracted, we searched for other secondary

endpoint tests in tables, graphs, and/or in-text. We did not extract data for all secondary endpoint tests due to time and effort constraints. In trials where we selected only a subset of all P-values for secondary endpoints, we chose two P-values that were close to the study's significance threshold and two P-values that were far from the threshold. We defined P-values as close to threshold if they were within  $\pm 0.04$  of the threshold and far from the threshold if they were beyond  $\pm 0.04$  of the statistical threshold.

### Quality assurance/ Quality control

The data were extracted independently by two different individuals to ensure that data were extracted accurately (42).

In the second stage of data extraction, M.A. extracted inputs for all articles originally examined by S.Y. and S.Y. extracted inputs for all articles initially examined by M.A., without either being aware of what the other had extracted. Then, data extraction results were compared and where there were inconsistencies, the articles were re-examined by M.A. In this stage, we had a total of 2,911 individual tests. Of these, there were 632 tests that M.A. and S.Y. disagreed on: 82 of these disagreements were resolved. However, the other 550 disagreements could not be resolved and were excluded. Common examples of these disagreements include disagreements on degrees of freedom for Chi square tests since they are usually unreported in the journal articles, and disagreements on sample sizes especially in secondary analyses when there were cases of patient withdrawals or deaths. Note that in this stage, optimal  $\alpha$  thresholds

had not been calculated yet, thus tests were excluded only if there was an unresolvable source of ambiguity in the data.

In the final stage of quality assurance and control a third investigator (C.Z.) independently extracted required data for a subset of articles examined by both M.A. and S.Y. Results were compared for consistency across all three investigators. Disagreements were resolved by discussion among the investigators and where inconsistencies could not be resolved the test was discarded. In the last stage, we had an additional 164 unresolved disagreements that were also excluded. Finally, a total of 2,197 individual tests from 718 articles were included in the study. Of these tests, 1,112 were examined by all three investigators and 1085 by M.A. and S.Y. with unanimous agreement on the data extracted.

### **Calculating optimal $\alpha$**

The objective of optimal  $\alpha$  is to minimize the combined probability of making Type I ( $\alpha$ ) and Type II ( $\beta$ ) errors in a study. For a given statistical test, sample size, critical effect size, and alpha value there is a single beta value. Thus, for a specific test, sample size and critical effect size, the value of  $(\alpha + \beta)/2$  can be estimated across all possible  $\alpha$  values and optimal  $\alpha$  identified as the value that minimizes  $(\alpha + \beta)/2$  (23). For each optimal  $\alpha$  there is an associated optimal  $\beta$  (i.e. the probability of making a Type II error when the statistical threshold is set at optimal  $\alpha$ ) and an overall probability of error (i.e. the probability of making either error  $(\alpha + \beta)/2$  when the statistical threshold is set at optimal  $\alpha$ ). Note that sample sizes will be a

characteristic of a particular clinical trial, but the critical effect size(s) must be defined, a priori, for all tests to calculate optimal  $\alpha$ . Codes for calculating optimal  $\alpha$  for each test can be found in Appendix A.

### **Identifying a critical effect size**

A “critical” effect size is the effect size that is considered clinically important (23). Since there is no completely objective rationale for identifying a meaningful critical effect size, we chose to calculate optimal  $\alpha$  levels for multiple critical effect sizes; large, medium, and small, using published estimates of appropriate effect sizes (Table 2; 43-45). For example, in a clinical trial that uses control and treatment groups and applies a t-test, Cohen’s  $d$  represents the difference between the treatment and control means divided by the common standard deviation (46). Cohen (43) identified 1.0, 0.5 and 0.2 as large, medium and small effects sizes for Cohen’s  $d$ , respectively. These effect sizes imply that the difference in means between the control and treatment groups are 1.0, 0.5 and 0.2 standard deviations, respectively.

Some papers included explicit mention of ‘target’ effect sizes so, for a small subset of analyses we have estimated optimal alpha using the effect size targeted in the paper. We will focus on the generic ‘critical’ effect sizes but present the results on the subset for comparison.

**Table 2 Critical effect sizes suggested by Cohen that corresponds to the different statistical tests.**

Test	Critical effect size		
	Small	Medium	Large
Regression	$r=0.2$	$r=0.5$	$r=0.8$
t-test	$d=0.2$	$d=0.5$	$d=1.0$
ANOVA	$f^2=0.04$	$f^2=0.25$	$f^2=0.65$
Chi square	$W=0.1$	$W=0.3$	$W=0.5$
Log-rank/Cox regression	$HR=1.3$	$HR=1.9$	$HR=2.8$
Logistic regression	$OR=1.68$	$OR=3.47$	$OR=6.71$

## Assumptions

When calculating optimal  $\alpha$ , we have made two assumptions - that the prior probabilities of the null and alternative hypotheses are equal (i.e. 0.5) and the cost of making Type I and Type II errors are also equal (23). It is not necessary to make these assumptions because optimal  $\alpha$  can incorporate unequal priors and unequal cost assumptions but, these are the reasonable default assumptions when no explicit quantitative alternatives have been identified.

## Data analysis

After extracting the necessary inputs, we calculated optimal  $\alpha$  (i.e. the new null hypothesis rejection threshold) under the pre-specified assumptions: equal prior probabilities of null and alternate hypotheses, and equal costs of Type I and Type II error rates at each of the pre-defined critical effect

sizes (Table 2). It is important to note here that our choices for generic critical effect sizes could be different than the effect sizes authors considered important to detect, if specified. This means that it is not possible to comment on whether conclusions of individual trials would have been different, but rather whether the conclusions would have been different if they had used these specific assumptions. Our choice of three critical effect sizes resulted in three different optimal  $\alpha$ 's, one for each of the critical effect sizes; small, medium, and large, with the exception of logistic regression cases where we calculated six different optimal  $\alpha$ 's – three for the small defined incidence rate, and three for the large defined incidence rate. In this thesis, only logistic regression cases with the large incidence rate will be presented and included in the analysis.

We compared inferences about rejecting or accepting the null hypothesis made using the traditional  $\alpha$  and optimal  $\alpha$  for each of the tests, and we present the differences in inferences for the full data set and for the subset that was examined by all three investigators. Then, we found the proportion of tests where  $P < 0.05$  and optimal  $\alpha$  disagreed (i.e. one rejected and the other accepted the null hypothesis). We also compared how disagreement rates and the direction of disagreement (i.e. optimal  $\alpha$  rejecting the null when traditional NHST accepts versus optimal  $\alpha$  failing to reject when traditional NHST rejected) varied depending on the critical effect size.

## RESULTS

### Data Description

Seven hundred and eighteen articles (containing 2,197 statistical tests) met the inclusion criteria for detailed data abstraction. Articles were published between the years 1993 and 2019. Most selected clinical trials were conducted in Europe, North America, or on an international platform. Only 11.1% of the trials were conducted in other locations such as Asia, Australia, South America, and Africa.

### Optimal $\alpha$ values

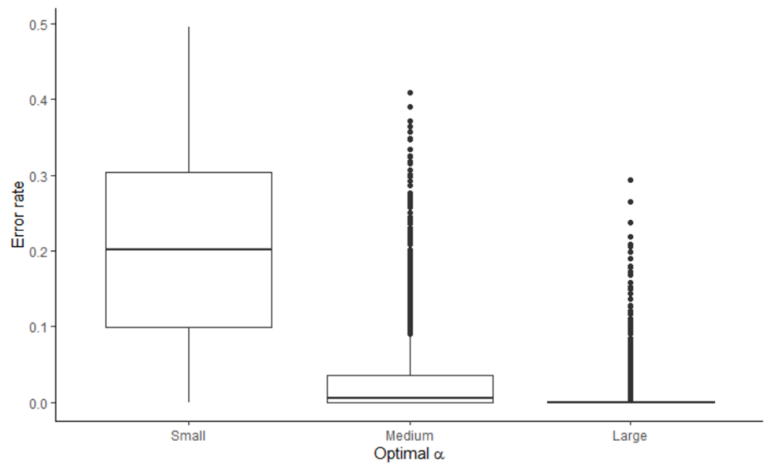
Optimal  $\alpha$  values vary negatively with effect size because, all other things being equal, tests have more power to detect large effect sizes than small. The median optimal  $\alpha$  for detecting small effects was 0.166, for medium effects was 0.00543 and large effects was 0.0000148. Thus, the evidence required to conclude that there is a large effect is greater than the evidence required to conclude that there is a medium or small effect. For none of the effect sizes was the median optimal  $\alpha$  close to the traditional threshold of 0.05.

### Optimal $\beta$ values

Optimal  $\beta$  values follow the same trend as optimal  $\alpha$  values – they vary negatively with effect size because, all other things being equal, tests have more power to detect large effect sizes. The median optimal  $\beta$  for detecting small effects was 0.235, for medium effects was 0.00594 and large effects was 0.0000153.

### Overall error probability

For pre-defined small critical effect sizes half of the studies had a 20% chance or greater of making an error when they rejected or accepted the null (i.e. the median (optimal  $\alpha$  + optimal  $\beta$ )/2 was 0.2). However, for medium and large critical effect size, half of the studies only had 0.57% and 0.0015% chance or greater of making an error, respectively (i.e. the median (optimal  $\alpha$  + optimal  $\beta$ )/2 were 0.0057 and 0.000015, respectively) (Figure 2).



**Fig. 2.** Distribution of overall error probability associated with each optimal  $\alpha$  category.

### Optimal $\alpha$ versus $\alpha \leq 0.05$

Inferences made using optimal  $\alpha$  were inconsistent with the ones made using traditional  $\alpha$  in 23.6% of tests. Optimal  $\alpha$  disagreed with traditional  $\alpha$  in 14.01% of tests for the small effect size, 24.48% of the tests for a medium effect size, and 32.31% of tests for the large effect size (Table 3). Disagreement between optimal  $\alpha$  and traditional  $\alpha$  implies either that the null would have been rejected using optimal  $\alpha$  and accepted using a traditional threshold or vice versa. The results were similar when we

only included the primary endpoint/primary outcome (N=460). Similar results were obtained when using only the data set that was examined by the three investigators (1,112 tests).

**Table 3.** Number of times optimal  $\alpha$  reached consistent and inconsistent conclusions compared to traditional  $\alpha$  at three different effect sizes (N= 2,197)

Effect Size	Comparison between traditional and optimal $\alpha$	
	Disagree	Agree
$\alpha_1$ small	308 (14.01%)	1889 (85.99%)
$\alpha_2$ medium	538 (24.48%)	1659 (75.52%)
$\alpha_3$ large	710 (32.31%)	1487 (67.69%)

### Proportion of tests where traditional $\alpha$ and optimal $\alpha$ disagreed

Ninety-three percent of the disagreements between optimal  $\alpha$  and traditional  $\alpha$  for the medium critical effect size and 98% of the disagreements for the large critical effect size were when authors rejected the null and concluded that there was an effect when they would have failed to reject it if they had used optimal  $\alpha$ . On the other hand, optimal  $\alpha$  was more aggressive than conventional statistical thresholds for the small critical effect size where 81% of

the disagreements were in favor of rejecting the null using optimal  $\alpha$  when the authors had concluded the opposite (Table 4).

**Table 4.** A detailed breakdown of the total disagreements and agreements between optimal  $\alpha$  and traditional  $\alpha$  for the full data set (N= 2,197).

Disagreements	Small critical effect size	Medium critical effect size	Large critical effect size
Traditional $\alpha$ rejected when optimal $\alpha$ accepted	57	502	699
Traditional $\alpha$ accepted when optimal $\alpha$ rejected	251	36	11
Total	308	538	710
Agreements			
Both traditional $\alpha$ and optimal $\alpha$ rejected	755	310	113
Both traditional $\alpha$ and optimal $\alpha$ accepted	1134	1349	1374
Total	1889	1659	1487

Investigators are more likely to conclude that therapies will have medium, or large effects when they will not, and that therapies would have no effect when they would have a small effect on patients' health.

## DISCUSSION

We show that inferences made using optimal  $\alpha$  were inconsistent with those made using traditional  $\alpha$  in 23.6% of tests. Optimal  $\alpha$  disagreed with traditional  $\alpha$  in 14.01% of tests for the small effect size, 24.48% of the tests for a medium effect size, and 32.31% of tests for the large effect size. We found that that using traditional thresholds increased the likelihood of (1) missing a ‘true’ small effect and (2) detecting ‘false’ medium and large effects. The estimated overall error rate when trying to detect a small critical effect size was often very high (i.e.. 40-50%), which indicates that these trials rarely had enough power to consistently detect small effects.

### **NHST and optimal $\alpha$**

Traditional NHST has been criticized for weighing Type I error rate more than Type II error, for using arbitrary thresholds that have no scientific meaning (such as  $\alpha = 0.05$ ), failing to provide information on effect sizes, ignoring prior probabilities, and for being very sensitive to sample size (16,17,25,26). Thus, using traditional NHST may reduce the reliability of inferences attained from clinical trials (27).

Optimal  $\alpha$  mitigates many of the problems associated with traditional NHST. Optimal  $\alpha$  provides a study-specific threshold that minimizes the combined probabilities of making Type I and Type II errors and explicitly identifies the critical effect size(s) (23). Hence, the sensitivity to sample size and the potential for inflated Type II error rates associated with NHST are both minimized. Optimal  $\alpha$  requires researchers to identify the effect size they consider to be important and

present conclusions directly related to that specific effect size. Our results suggest that if the goal in health-care research is to make the fewest mistakes, the traditional threshold often leads to wrong conclusions, and investigators using NHST should apply optimal  $\alpha$  rather than the traditional 0.05, 0.01 or 0.1 thresholds.

### **Type of disagreements and critical effect sizes**

We found that many clinical trials are poorly designed to detect small effects. By contrast, there has been a growing consensus that clinical trials may show statistically significant effects that have limited clinical value (28). Unfortunately, there is no general agreement on specific thresholds for assessing clinically important effects - clinical significance can vary among patients, clinicians, policy-makers and regulatory agencies (30-32, 7). For example, the use of erlotinib in patients with pancreatic cancer showed 18% reduction in the risk of death ( $P = 0.038$ ), however, this translated into a less than 2-week improvement in median overall survival (32,31). Is this effect clinically significant?

Lawrence et al (28) suggest that an ideal hypothesized clinical effect size should consider the benefits of a treatment in relation to its toxicity, side-effects, price, and the severity of the disease. For example, expensive or highly toxic treatments may require larger benefits to warrant the side effects or cost. On the other hand, smaller benefits may be justified for cheap or less toxic treatments (28). The American Society of Clinical Oncology recommended a HR of

$\leq 0.80$  (which corresponds to an improvement in median OS within a range of 2.5 to 6 months) as a reasonable criterion for assessing clinical benefit in phase III trials (29). (Note:  $HR \leq 0.80$  is equivalent to  $HR \geq 1.25$  if the treatment is intended to increase the incidence of the endpoint. For example, if the endpoint is survival rather than mortality). Yet 48 drugs approved by FDA for metastatic solid tumors between the years 2002 and 2012 showed only small median gains of 2.16 months in OS (33). We identified an HR of 1.3 as small, which is larger than what has often been suggested but our results show that few clinical trials are designed to detect such small effects.

### **Optimal $\alpha$ is a measure of strength of experimental design**

Strong experimental design assures adequate power to detect predefined clinically important effects (34), low probabilities of making Type I or II errors and reliable inferences from clinical trials (35, 34). Optimal  $\alpha$  is a measure of strength of experimental design because it is dependent on the power of the test – when the power is high, the overall probability of making an error will be low, therefore, the rejection threshold that minimizes that error probability (i.e. optimal  $\alpha$ ) will also be low. On the other hand, when the power of a test is low (i.e. when the sample size is too small to detect the critical effect size), the overall probability of error will be large. Consequently, optimal  $\alpha$  will be large. Optimal  $\alpha$  is designed to minimize the probability of making an error but when the experimental design for a targeted effect size is poor, the minimum probability of error may still be quite large and the associated

optimal  $\alpha$  also quite large. Our study showed that more than half of the clinical trials in this study had greater than 20% probability – four times the probability we usually accept for Type I error (5%) - of making a mistake when trying to detect a small effect size. This implies that most of the trials are poorly designed to detect small effect sizes. If optimal  $\alpha$  was used in clinical trials, researchers would provide the optimal  $\alpha(s)$  for their critical effect size(s) and it would be clear how strong the evidence for or against a critical effect size was, simply based on its optimal  $\alpha$ .

### **Optimal $\alpha$ and Type I and Type II errors**

Our results show that investigators who used traditional NHST thresholds compared to optimal thresholds were more likely to make a Type I error when detecting medium and large effects, and a Type II error when detecting small effect sizes. This begs the question, which of the two errors is worse?

Clinical trials that use NHST with its conventional thresholds usually put more weight on avoiding Type I rather than a Type II error (22), and regulators are usually more concerned with Type I error (36). However, being too conservative with Type I error, increases the probability of rejecting effective therapies. Regulatory agencies acknowledge this trade-off, which can be seen in the approval trends. For example, Wong et al (37) examined 400,000 entries of 185,994 clinical trials to estimate the probability of success of a clinical investigation and found that FDA approval rates were dependent on disease type – higher approval rates were seen in vaccines for

infectious diseases: 33.4%, while much lower approval rates (3.4%) were seen in drugs that treat cancer (39). The need to balance Type I and Type II errors has long been recognized by clinicians and regulatory agencies (38,22). Yet, traditional NHST does not provide a formal framework to assess the potential risks and benefits of a treatment. Because

optimal  $\alpha$  can incorporate the relative costs of Type I and II errors in setting the rejection threshold, adopting optimal  $\alpha$  for clinical trials might encourage clinicians and regulatory agencies to more explicitly consider the relative costs of Type I and II errors.

## REFERENCES

1. L. M. Friedman, C. Furberg, D. L. DeMets, D. M. Reboussin, C. B. Granger, *Fundamentals of clinical trials* (Vol. 4). New York: Springer (2010).
2. FDA Office of the Commissioner Step 3: Clinical Research (2018). Retrieved January 2020, from <https://www.fda.gov/patients/drug-development-process/step-3-clinical-research>
3. C. A. Umscheid, D. J. Margolis, C. E. Grossman, Key concepts of clinical trials: a narrative review. *Postgraduate medicine*, **123**, 194–204 (2011).
4. J. J. Darrow, J. Avorn, A. S. Kesselheim, New FDA breakthrough-drug category—implications for patients. *New Engl. J. Med.* **2014**, 370:1252-1258 (2014).
5. N. S. Downing, N. D. Shah, J. A. Aminawung, A. M. Pease, J. D. Zeitoun, H. M. Krumholz, J. S. Ross, Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010. *JAMA*, **317**, 1854-1863 (2017).
6. M. Das, Many FDA-approved cancer drugs might lack clinical benefit. *The Lancet Oncology*. **19**, E82 (2018).
7. A. Tibau, C. Molto, A. Ocana, A. J. Templeton, L. P. Del Carpio, J. C. Del Paggio, A. Barnadas, C. M. Booth, E. Amir, Magnitude of clinical benefit of cancer drugs approved by the US Food and Drug Administration. *JNCI* **110**, 486-492. (2018).
8. C. G. Begley, L. M. Ellis, Raise standards for preclinical cancer research. *Nature*, **483**, 531-533 (2012).
9. F. Prinz, T. Schlange, K. Asadullah, Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, **10**, 712-712. (2011).
10. R. E. Kirk, Promoting good statistical practices: Some suggestions. *Educ. Psych. Meas.*, **61**, 213-218. (2001).
11. G. A. Diamond, S. Kaul, On reporting of effect size in randomized clinical trials. *Am. J. Cardio.*, **111**, 613-617. (2013).
12. J. F. Piccirillo, Improving the quality of the reporting of research results. *JAMA O-H & N Surg.*, **142**, 937-939. (2016).
13. A. Bakker, J. Cai, L. English, G. Kaiser, V. Mesa, W. Van Dooren, Beyond small, medium, or large: points of consideration when interpreting effect sizes. *Educ. Stud. Math.*, **102**, 1-8. (2019).

14. Food and Drug Administration Act, "Amendments Act" 121 C.F.R. § 801 (2007).
15. Adequate and well-controlled studies, 21 C.F.R. § 314.126 (2016).
16. R. L. Wasserstein, N. A. Lazar, The ASA Statement on p-Values: Context, Process, and Purpose. *Am. Stat.*, **70**, 129-133 (2016).
17. International Committee of Medical Journal Editors, Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals (2019). Retrieved January 2020, from <http://www.icmje.org/icmje-recommendations.pdf>
18. D. D. Boos, L. A. Stefanski, P-value precision and reproducibility. *Am. Stat.*, **65**, 213-221. (2011).
19. G. Cumming, The new statistics: Why and how. *Psych. Sci.*, **25**, 7-29. (2014).
20. R. Temple, How FDA currently makes decisions on clinical studies. *Clinical Trials*, **2**, 276-281 (2005).
21. L. Kennedy-Shaffer, When the Alpha is the Omega: P-Values, "Substantial Evidence," and the 0.05 Standard at FDA. *Food and Drug Law Journal*, **72**, 595–635, (2017).
22. L. Isakov, A. W. Lo, V. Montazerhodjat, Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design. *J. Econometrics*, **211**, 117-136. (2019).
23. J. F. Mudge, L. F. Baker, C. B. Edge, J. E. Houlihan, Setting an optimal  $\alpha$  that minimizes errors in null hypothesis significance tests. *PLoS one*, **7**, e32734. (2012).
24. A. Banerjee, U. B. Chitnis, S. L. Jadhav, S. J. S. Bhawalkar, S. Chaudhury, Hypothesis testing, type I and type II errors. *Indust. Psych. J.*, **18**, 127 (2009).
25. J. Woodcock, FDA introductory comments: clinical studies design and evaluation issues. *Clinical Trials*, **2**, 273-275. (2005).
26. K. S. Lee, When the Alpha is the Omega: P-Values, "Substantial Evidence," and the 0.05 Standard at FDA. *Food and Drug Law Journal*, **72**, 595. (2017).
27. L. Pedro-Roig, C. H. Emmerich, The reproducibility crisis in preclinical research—lessons to learn from clinical research. *Medical Writing*, **26**, 28-32 (2017).
28. N. J. Lawrence, F. Roncolato, A. Martin, R. J. Simes, M. R. Stockler, M. R. Effect Sizes Hypothesized and Observed in Contemporary Phase III Trials of Targeted and Immunological Therapies for Advanced Cancer. *JNCI cancer spectrum*, **2**, pky037 (2018).
29. L. M. Ellis, et al., American Society of Clinical Oncology perspective: raising the bar for clinical trials by defining clinically meaningful outcomes. *J. Clin. Oncol.*, **32**, 1277-1280. (2014).
30. N. I. Cherny et al., A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). *Annals of Oncology*, **26**, 1547-1573. (2015).
31. Wilson, M. K., Karakasis, K., & Oza, A. M. (2015). Outcomes and endpoints in trials of cancer treatment: the past, present, and future. *The Lancet Oncology*, **16**(1), e32-e42.

32. M. J. Moore et al., Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *Journal of clinical oncology*, **25**, 1960-1966. (2007).
33. A. T. Fojo, A. Noonan, A. Why RECIST works and why it should stay—counterpoint. *Cancer Research*, **72**, 5151-5157 (2012).
34. K. P. Suresh, S. Chandrashekara, Sample size estimation and power analysis for clinical research studies. *J. Human Rep. Sci.*, **5**, 7 (2012).
35. R. H. Breau, T. A. Carnat, I. Gaboury, Inadequate statistical power of negative clinical trials in urological literature. *J. Urology*, **176**, 263-266 (2006).
36. U.S. Food & Drug Admin., Guidance for Industry: E9 Statistical Principles for Clinical Trials (1998).
37. C. H. Wong, K. W. Siah, A. W. Lo, Estimation of clinical trial success rates and related parameters. *Biostatistics*, **20**, 273-286. (2019).
38. D. A. Berry, Interim analysis in clinical trials: The role of the likelihood principle. *Am. Stat.*, **41**, 117-122. (1987).
39. M. J. Bown, A. J. Sutton, Quality control in systematic reviews and meta-analyses. *Eur. J. Vasc. Endovasc. Surg.*, **40**, 669-677 (2010).