

# Optimal $\alpha$ 2.0: A new approach to interpreting p-values in clinical trials

**Abstract:** Null hypothesis significance testing (NHST) is widely used in clinical trials despite widespread understanding of the weaknesses associated with NHST including that (1) p-values provide little information about effect sizes and (2) traditional rejection thresholds are arbitrary. Optimal  $\alpha$  1.0, introduced in 2012, was an improved approach to approach to setting rejection thresholds because it had an explicit and defensible rationale, minimizing the probability of making Type I and II errors. However, it still used observed p-values to make binary decisions (reject or fail to reject the null) and it required the arbitrary selection of 'critical' effect sizes. Here we describe an extension of optimal  $\alpha$  1.0, optimal  $\alpha$  2.0, that mitigates many of the weaknesses associated with optimal  $\alpha$  1.0. Optimal  $\alpha$  2.0 uses observed p-values to estimate the maximum effect size for which the data provide evidence and the range of effect sizes against which the data provide evidence. Using optimal  $\alpha$  2.0, p-values are no longer used to reject or fail to reject the null but are used instead to make inferences about the range of effects sizes for which there is evidence. Here, I show that relationship between observed p-values and the hazard ratios for which the data provide evidence, for log-rank analyses of simulated clinical trials data. I also apply optimal  $\alpha$  2.0 to three clinical trials that used a traditional NHST approach and demonstrate how the results obtained using optimal  $\alpha$  2.0, allow more nuanced and useful interpretation than the traditional NHST interpretation.

**Keywords:** Null hypothesis significance testing, Type I errors, Type II errors, log-rank tests, hazard ratios, effect size

## INTRODUCTION

Historically, null hypothesis significance testing (NHST) including associated null rejection thresholds – usually 0.05 although occasionally shifted lower or higher for relatively arbitrary reasons – have played an important role in interpreting cancer clinical trials results<sup>1,2</sup>. However, cancer clinicians and researchers have discouraged over-reliance on p-values and encouraged increased emphasis on effect sizes and power analyses to assist in the interpretation and design of clinical trials<sup>3,4</sup>.

This concern about NHST and an overreliance on p-values in data interpretation extends far beyond medical research<sup>5,6</sup> and has led to a recent recommendation by the American Statistical Association<sup>7</sup> to stop using p-values as a tool for accepting or rejecting the null.

Most criticisms of NHST rest on three fundamental properties – (1) that effect size is not an explicit consideration so that meaningful effects can be overlooked at small sample sizes and trivial effects can be determined to be statistically significant at large sample sizes, (2) that Type II errors are not considered and (3) that p-values don't provide what researchers want - error probability statistics (i.e. estimates of the probability of making a mistake when the null is rejected or not)<sup>8,9,10</sup>.

To combat problems with p-values, medical biostatisticians and clinical trials

researchers have proposed alternative analytical approaches such as Bayesian statistics<sup>11</sup> and alternative inferential approaches such as likelihood ratios<sup>12</sup>, confidence intervals<sup>13</sup> or Akaike's Information Criteria<sup>14</sup>. However, p-values are still widely used in cancer clinical trials to make inferences from data<sup>10</sup>.

Mudge et al.<sup>15</sup> introduced a new approach to setting rejection thresholds, optimal  $\alpha$  (1.0), which resolves or mitigates many of the problems associated with traditional NHST. For a given statistical test, sample size and critical effect size, optimal  $\alpha$  1.0 allows researchers to identify the rejection threshold that will minimize the probability of making a mistake when rejecting or failing to reject the null hypothesis. Because, for a given statistical test, sample size and critical effect size, there is a mathematical relationship between  $\alpha$  and  $\beta$  (i.e. the probability of Type I and II errors, respectively) it is possible to identify  $\alpha$  so as to minimise the probability of  $\alpha + \beta$  (see Methods for technical details). However, optimal  $\alpha$  1.0 requires researchers to identify a critical effect size (or a range of critical effect sizes) and there is rarely consensus about the effect sizes that are clinically meaningful. The original recommendation was to identify three effect sizes, (1) the smallest effect size anybody would likely care about, (2) the effect size so large it would be considered clinically meaningful by most people and (3) a value intermediate between (1) and (2).

This is a simple way to implement optimal  $\alpha$  1.0, but the arbitrary nature of threshold selection and the binary interpretation of results (i.e. reject or fail to reject the null hypothesis) isn't desirable.

I have developed an extension of the earlier optimal  $\alpha$  (optimal  $\alpha$  2.0) that doesn't pre-select arbitrary rejection thresholds and allows a non-binary interpretation of p-values. Optimal  $\alpha$  2.0 is used to estimate the maximum effect size for which the data provide evidence and the effect sizes for which the data provide evidence against, by identifying the effect size for which the observed p-value is the optimal  $\alpha$ . Thus, rather than selecting an effect size, identifying the rejection threshold that would minimize the probability of making an error and comparing the observed p-value with the rejection threshold and choosing to reject or fail to reject the null, I identify the effect size (I will call this the 'optimal  $\alpha$  effect size') for which the observed p-value is the rejection threshold that would minimize the probability of making an error (i.e. optimal  $\alpha$ ). This would imply that there is more evidence for any effect size up to the 'optimal  $\alpha$  effect size' than for the null but more evidence for the null than any effect size larger than the 'optimal  $\alpha$  effect size'. Further, the observed p-value is an index of the relative strength of the evidence for these conclusions. This is because optimal  $\alpha$  (and the associated optimal  $\beta$ ) are estimates of the conditional probabilities of making either a Type I error when rejecting the null or a Type II error when failing to reject a

null. When those probabilities are very small it implies there is a very small probability of making an error about our effect size conclusions and therefore the evidence is strong. By contrast a large p-value implies weak evidence for the conclusions because the error probabilities are high.

#### *Clinical Trial Example*

Clinical trial: A comparison of the 5-year survival rates when patients are treated with a combination of drugs (Drug A + Drug B) versus only receiving Drug A. The trial concludes with 586 participants, split evenly between the control and treatment arms and we observe a p-value of 0.075 using a log-rank test. The traditional NHST approach with a rejection threshold of  $p < 0.05$  would fail to reject the null and conclude we had no evidence that the combination of drugs was more effective than Drug A alone. Using optimal  $\alpha$  1.0 we set small, medium, and large 'critical' effect sizes, at 1.3, 1.65 and 2 respectively. The 'optimal alpha' rejection thresholds for these three effect sizes were  $OA_{\text{small}} = 0.07109$ ;  $OA_{\text{medium}} = 0.00168$ ;  $OA_{\text{large}} = 0.00002$ . I would fail to reject the null and conclude there is no evidence of an effect for any of the three effect sizes. But I might include a statement describing the p-value as 'near-significant' for a small effect size.

Using optimal  $\alpha$  2.0, I estimate the critical effect size for which an observed p-value of 0.075 would be the optimal  $\alpha$ . In this case, it is a hazard ratio of 1.32. I conclude that these data provide evidence of a hazard

ratio greater than 0 but less than or equal to 1.32 and evidence against a hazard ratio larger than 1.29. We would consider the evidence to be moderate because there is a 7.5% chance of observing data as or more extreme than the observed data if the data came from a population where there is no difference in 5-year survival rates between the treatment and control regimes (i.e. optimal  $\alpha = 0.075$ ) and there is a 9.1% probability of observing data less extreme than these data if the effect size was greater than 1.32. Note that the conclusions about effect size are asymmetric. That is, optimal  $\alpha$  2.0 provides evidence that the true effect size is NOT larger than the optimal  $\alpha$  effect size, but it doesn't provide evidence of what the true effect size IS – rather it provides evidence of an effect size greater than zero but less than or equal to the optimal  $\alpha$  effect size. So, the true effect size could be as large as the optimal  $\alpha$  effect size but in most cases it will be smaller. This is because the optimal  $\alpha$  effect size is at the outer edge of the distribution of effects sizes that are more probable than the null given the data, so there are many effects sizes that are more probable than the optimal  $\alpha$  effect size, but the optimal  $\alpha$  effect size is more probable than the null. By contrast, the null is more probable than any effect size larger than the optimal  $\alpha$  effect size.

Thus, optimal  $\alpha$  2.0 addresses one of the key concerns associated with using p-values – that effect sizes are ignored in an NHST context or, if effect sizes addressed, it is in an ad hoc and inconsistent fashion.

Optimal  $\alpha$  2.0 provides a rigorous and quantitative integration of p-values and estimated effect sizes. The objectives of this paper are to (1) describe the methodological extension to optimal alpha, (2) run simulations describing the distribution of optimal  $\alpha$  effect sizes over different experimental designs and p-values and (3) apply the extended optimal  $\alpha$  technique to three published studies that used the traditional NHST approach.

## METHODS

### ***Optimal alpha 1.0***

For any statistical test,  $\beta$  (i.e. the probability of Type II error) is a function of (i) sample size (total or by group), (ii) critical effect size, (iii) variability, and (iv)  $\alpha$  (i.e. the rejection threshold). Thus, for a given test, sample size, critical effect size and variability there will be a unique value of  $\beta$  for each value of  $\alpha$  and a

deterministic non-linear relationship between  $\alpha$  and  $\beta$  (Figure 1). This implies that there is also a deterministic non-linear relationship between  $\alpha$  and  $\alpha+\beta$  and therefore, a single  $\alpha$  value that minimizes  $\alpha+\beta$  (Figure 2). If the objective of setting a rejection threshold is to make the fewest errors possible when rejecting or failing to reject the null, then the  $\alpha$  that minimises the combined probability of making either Type I or II errors (i.e.  $\alpha+\beta$ ) is the “optimal  $\alpha$ ”.

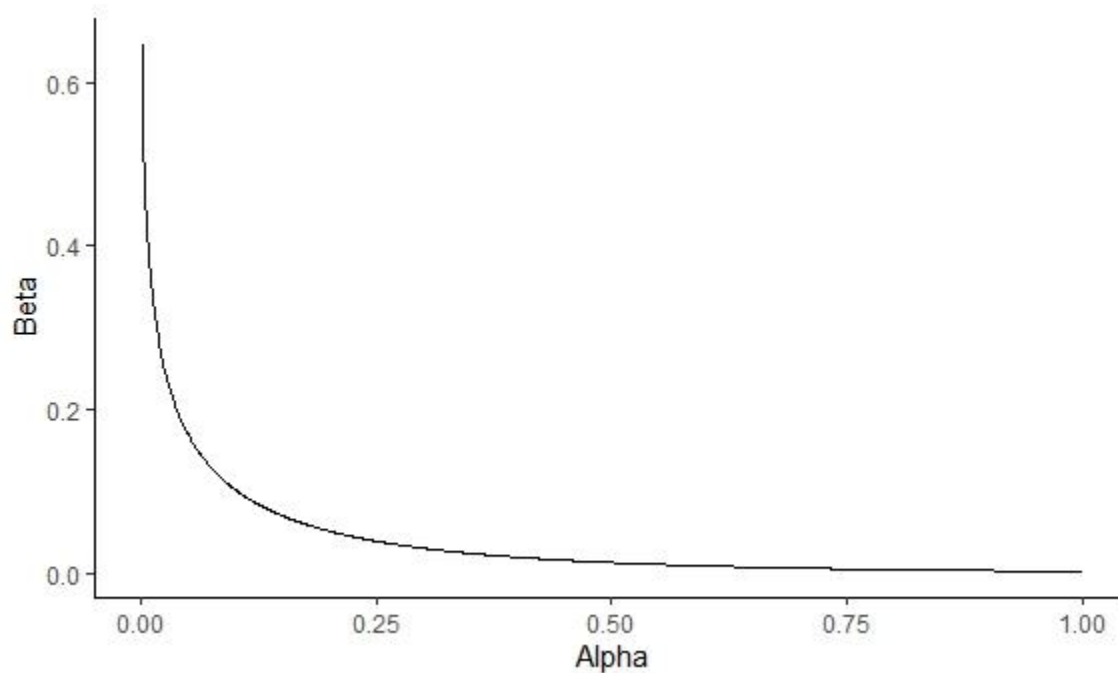


Figure 1: Relationship between  $\alpha$  and  $\beta$  for a log-rank test with 500 subjects split evenly between control and treatment groups and a ‘critical’ hazard ratio of 1.3.

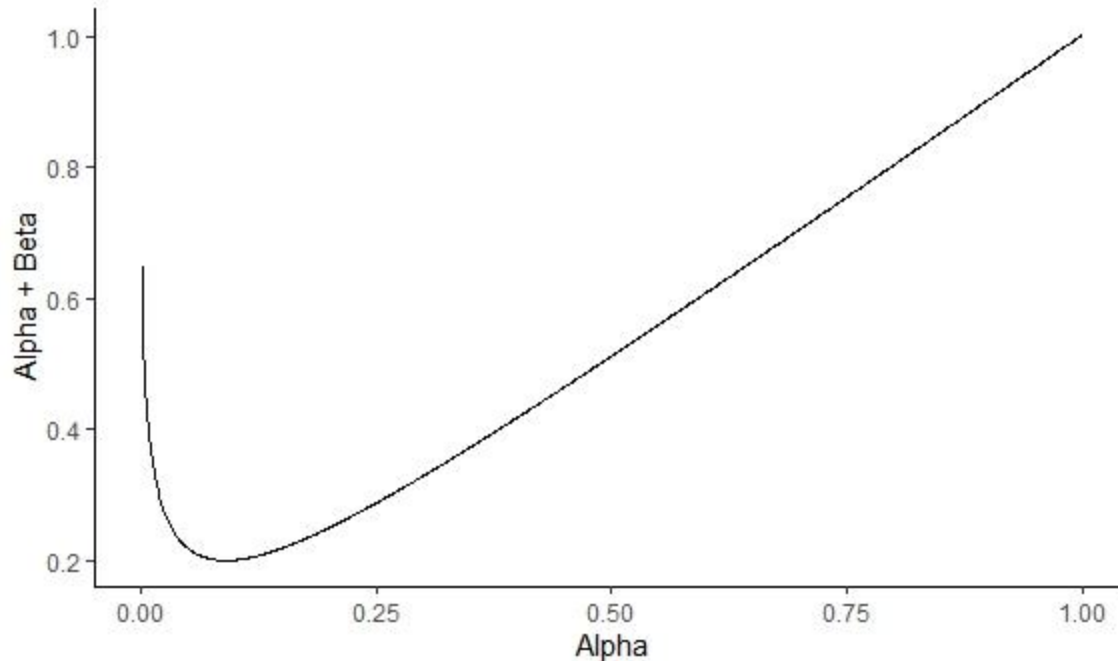


Figure 2: Relationship between  $\alpha$  and  $\alpha + \beta$  for a log-rank test with 500 subjects split evenly between control and treatment groups and a 'critical' hazard ratio of 1.3.

### ***Optimal alpha and prior probabilities***

To estimate the probability of making Type I or II errors using p-values, it is necessary to establish prior probabilities of the null and alternate hypotheses ( $P_0$  and  $P_A$ , respectively). These can easily be incorporated into optimal  $\alpha$  estimations by minimizing  $[(\alpha * P_0) + (\beta * P_A)]$  rather than  $\alpha + \beta$ . However, we propose that when there is no information about the prior probability of the null and alternate hypotheses, minimizing  $\alpha + \beta$  is still a rational approach for setting the rejection threshold because it implies the assumption that the probabilities of the null and alternate hypotheses are equal<sup>16,17</sup>. In this paper, I make no a priori assumptions about prior probabilities of  $H_0$  or  $H_A$  and my objective will be to minimize  $\alpha + \beta$ .

### ***Optimal alpha and critical effect sizes***

The rejection threshold that will minimize the probability of making an inferential error (i.e. optimal  $\alpha$ ) varies with the statistical test, sample size and the critical effect size. I define critical effect size as 'the smallest effect that would be considered biologically significant'. Thus, any effect size larger than the critical effect size must, by definition, also be biologically significant and any effect size smaller than the critical effect size is equivalent to the absence of an effect. To estimate the optimal rejection threshold using optimal  $\alpha$  1.0, a single critical effect size or a set of critical effect sizes must be chosen because the relationship between  $\alpha$  and  $\alpha + \beta$  varies with the chosen critical effect size. Thus, for a single 'critical' effect size, the subjective opinion of the researchers about what constitutes a critical

effect size, will have a large effect on the inferences they make. To mitigate this subjectivity, we proposed setting three different critical effect sizes<sup>15</sup> covering the plausible range of critical effect sizes, resulting in three different rejection thresholds and a separate inference for each of the three effect sizes. However, there is still a great deal of potential subjectivity in defining the 'plausible' range of critical effect sizes. Further, it doesn't allow any inferences to be made for effect sizes that lie in the intervals between each of the three chosen critical effect sizes.

#### ***Description of extended optimal alpha***

Optimal  $\alpha$  2.0 incorporates the concept that it is possible to characterize the distribution of optimal  $\alpha$ 's across a very wide range of critical effect sizes, rather than estimating optimal  $\alpha$  for only three critical effect sizes (Figure 3). Given the distribution of optimal  $\alpha$ 's across a wide range of critical effects sizes, the critical effect size for which the observed p-value would be optimal  $\alpha$  can be identified (Figure 4).

This link between the observed p-value and a critical effect size allows for inferences about (1) the range of critical effects for which there is more evidence than there is for an effect size equal to zero and (2) the range of effect sizes for which there is less evidence than there is for an effect size equal to zero.

#### ***Worked example***

To apply optimal  $\alpha$  2.0, one requires four pieces of information – (1) the type of statistical test used, (2) the total number of subjects in the clinical trial, (3) the proportion of subjects in the control group and (4) the observed p-value. For the example provided in the introduction, these are (1) log rank test, (2) 500 subjects, (3) 0.5 and (4) 0.075. I can estimate optimal  $\alpha$  for multiple critical effects sizes. I can extend the concept of estimating optimal  $\alpha$  for multiple critical effect sizes to estimate the distribution of optimal  $\alpha$  across the complete range of plausible critical effect sizes (Figure 4). Lastly, I find the effect size associated with the observed p-value (Figure 5).

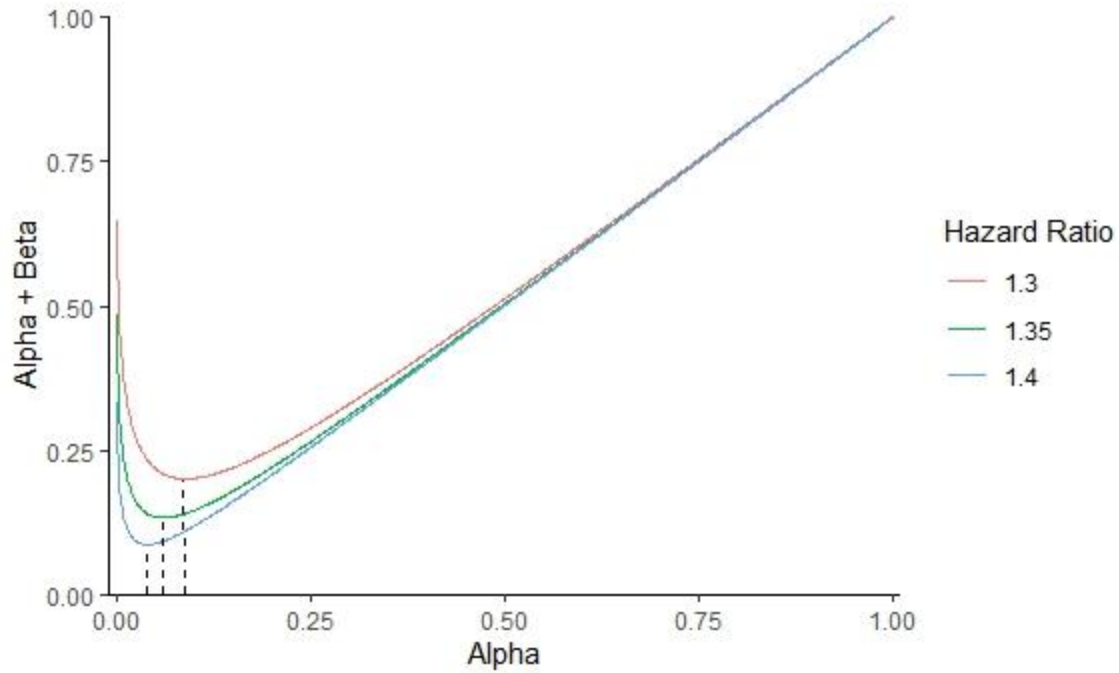


Figure 3: Relationship between  $\alpha$  and  $\alpha + \beta$  for a log-rank test with 500 subjects split evenly between control and treatment groups and 'critical' hazard ratios of 1.3 (red), 1.35 (green) and 1.4 (blue).

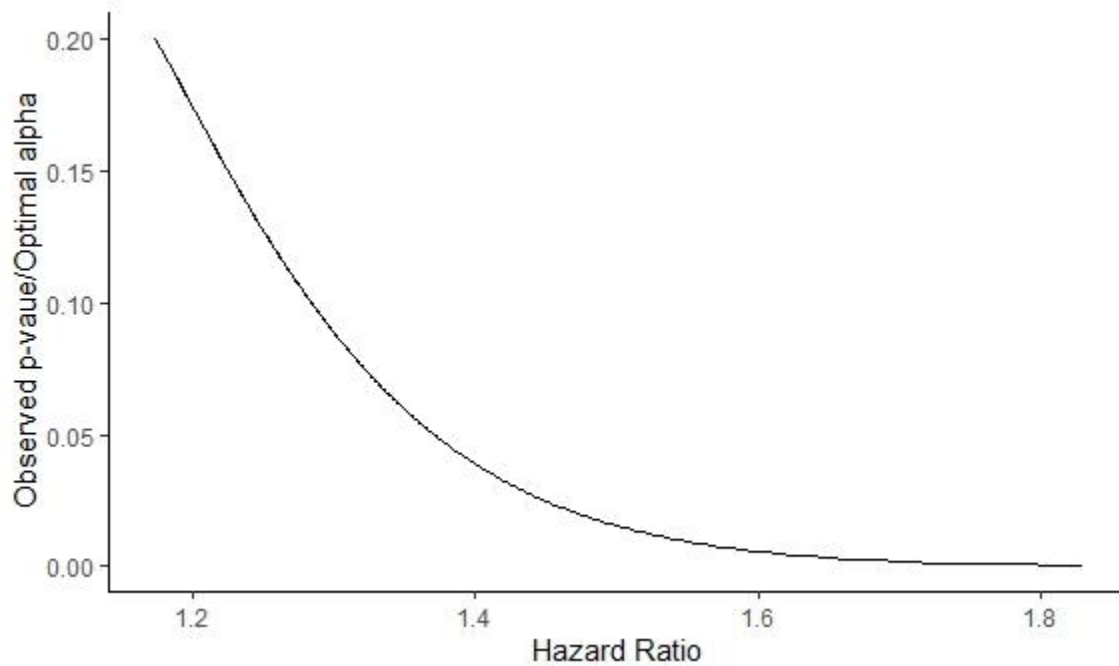


Figure 4: Relationship between the maximum hazard ratio for which there is evidence and the observed p-value for a log-rank test with 500 subjects split evenly between control and treatment groups.

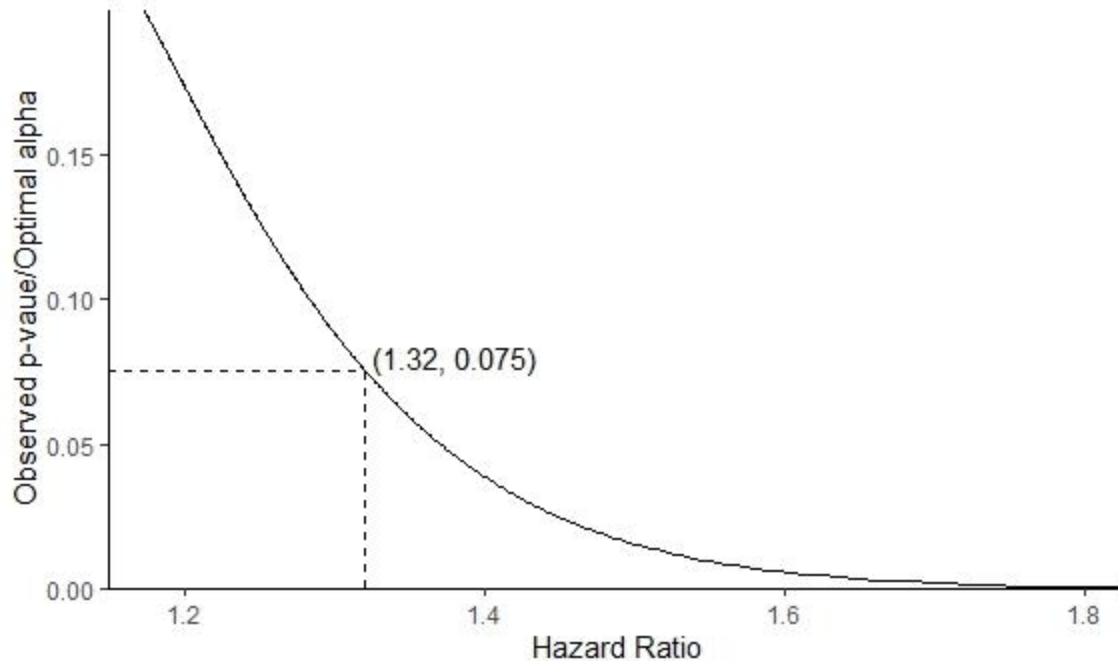


Figure 5: Relationship between the maximum hazard ratio for which there is evidence and the observed p-value for a log-rank test with 500 subjects split evenly between control and treatment groups and an observed p-value of 0.075.

### ***Description of simulations***

I simulated clinical trials that varied in number of subjects and the proportion in the control group and examined the relationship between observed p-value and critical effect size. In the first set of simulation, I set the control proportion = 0.5 and used sample sizes of 100, 200, 300, 400, 800, 1200, 1500, 2000 and 2500. For the second set of simulations, I set the samples sizes at 100, 500 and 1000 and for each sample size used control proportions of 0.1, 0.3 and 0.5. Note that the effect of control proportion on the relationship between observed p-value and critical effect size is symmetrical so that the relationships for control proportions of 0.1, 0.3 and 0.5 are the same as for control proportions of 0.9, 0.7 and 0.5, respectively.

### ***Selection of published studies for case study examples***

We have recently completed a retrospective analysis of ~ 1000 cancer clinical trials completed between 1993 and 2018, applying optimal  $\alpha$  1.0 to each of ~ 2000 statistical tests to estimate how often using optimal  $\alpha$  1.0 instead of traditional NHST thresholds would have resulted in a different inference. I chose tests from three separate clinical trials from our database. The only criteria were that (1) the researchers used a log rank test, (2) that the test(s) should address the primary question of the clinical trial, (3) that there was some variation in the proportion if subjects in the control treatment and (4) that we assessed a range of observed p-values from very small to greater than 0.05. The case studies I chose were: Case 1- three tests from Sandra-Petrescu

et al.<sup>18</sup>; Case 2 – four tests from de Boer et al<sup>19</sup>; Case 3 – four tests from Reck et al.<sup>20</sup>. See the results section for a more detailed description of the case study clinical trials.

## RESULTS

### *Simulation results*

The relationship between observed p-value and the largest effect size for which there is evidence is negative, monotonic, non-linear, and similar across simulated trials of different sample sizes (Figures 6 A-C). The slope of the relationship is much steeper over p-values less than 0.01 than above. For example, p-values of 0.01, 0.05 and 0.10 in a clinical trial with 100 subjects split evenly between control and treatment groups are evidence for a hazard ratio as large as 2.65, 2.02 and 1.75, respectively. The shape of the relationship between p-values and largest effect size for which there is evidence is essentially identical across all sample sizes but the largest effect size for which there is evidence is smaller for the same p-value as sample size increases (Figures 6 A-C). For an observed p-value of 0.01 in clinical trials with 100, 1000 and 2000 subjects split evenly between control and treatment, the largest effect size for which there is evidence is 2.65, 1.36 and 1.24, respectively. The observed p-value necessary to demonstrate a hazard ratio as large as 2.0 in a clinical trial with 1000 subjects and a control proportion = 0.5 is 0.00000003. For a clinical trial with 2000 subjects an observed p-value of 0.00000000000006 is required to provide evidence of a hazard ratio as great as 2.0. For a clinical trial with only 200 subjects, an observed p-value of 0.01 is evidence of a hazard ratio as large as 2.0.

The effect of proportion of subjects in the control group on the relationship between observed p-value and the largest effect size for which there is evidence is similar across control proportions to what we saw across sample sizes – that is, the relationship is negative,

monotonic, and non-linear and very similar across simulated trials of different control proportions (Figures 7 A-C). Again, the slope of the relationship is much steeper over small observed p-values. But all else being equal, as the relative proportion of subjects in the control and treatment arms moves further from an equal split, the same observed p-value provides

evidence of a larger effect size (Figures 7 A-C). For example, an observed p-value of 0.01 in three clinical trials each with 100 subjects but the proportion of subjects in the control arm being 0.1, 0.3 and 0.5 is evidence of a hazard ratio as large as 5.06, 3.38 and 2.65, respectively.

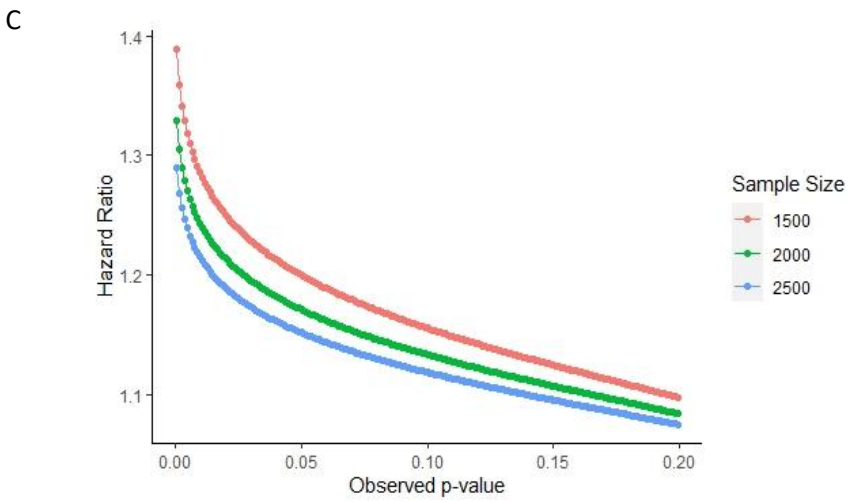
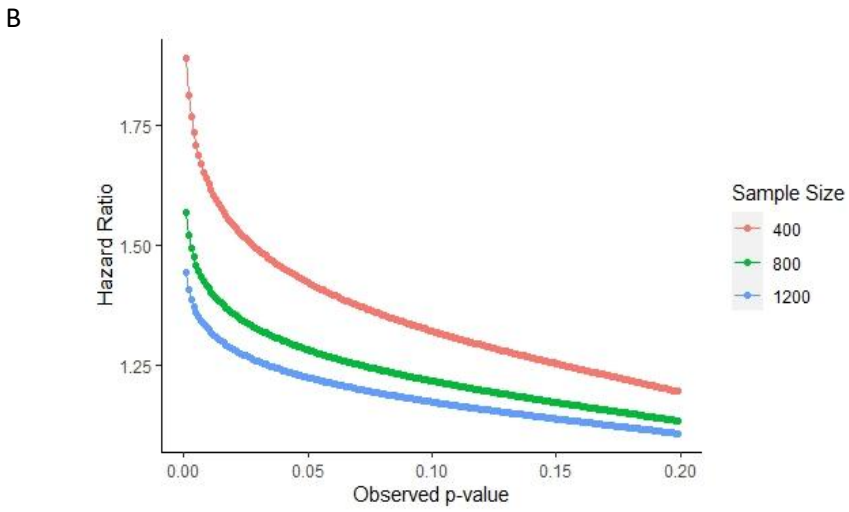
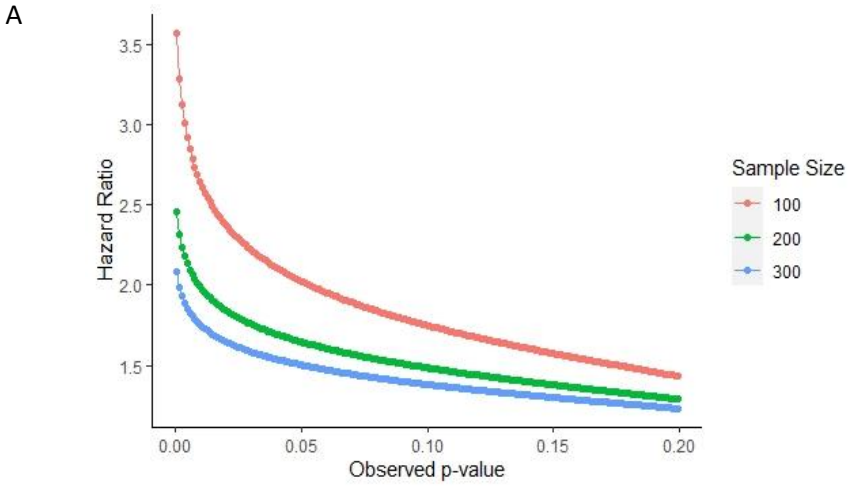


Figure 6: Relationship between the observed p-value and the ‘maximum’ hazard ratio for a log-rank test with A. 100 subjects, B. 500 subjects and C. 1000 subjects.

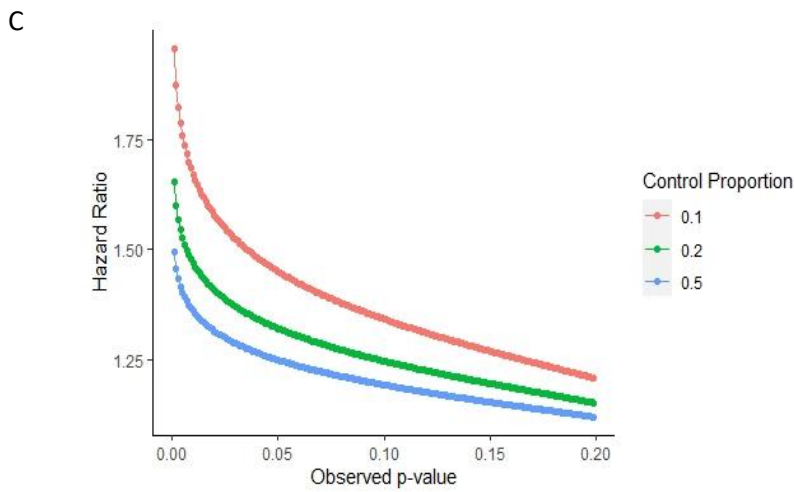
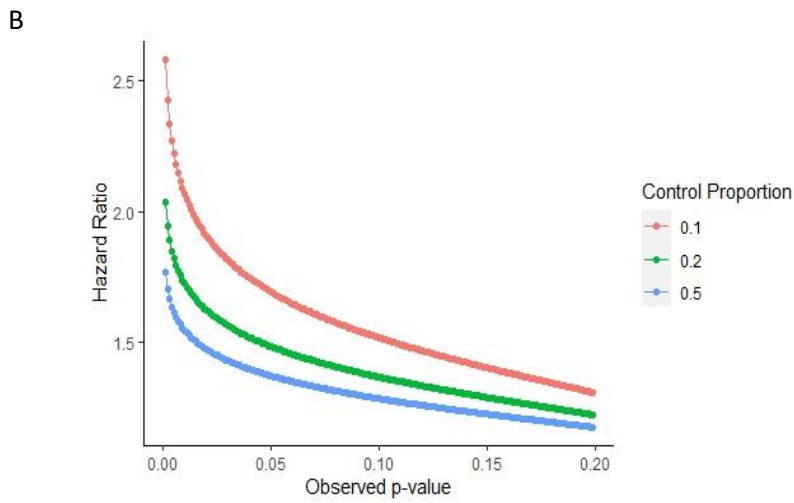
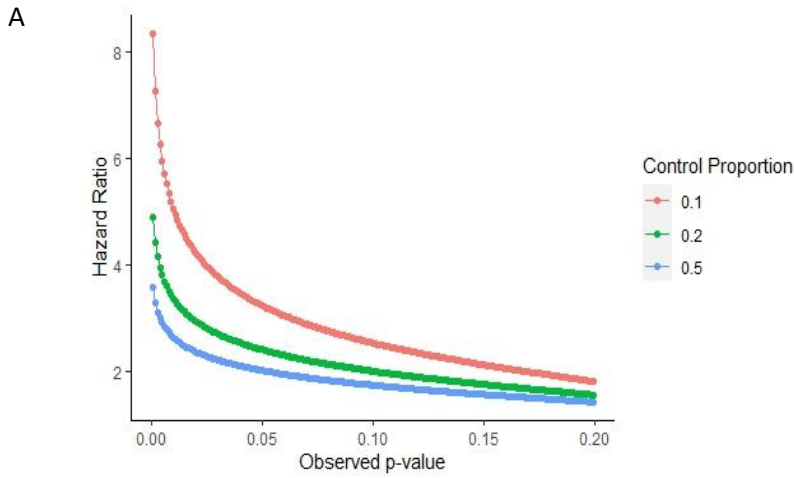


Figure 7: Relationship between the observed p-value and the 'maximum' hazard ratio for a log-rank test with subjects split evenly between control and treatment groups - A. 100, 200, 300 subjects, , B. 400, 800, 1200 subjects and C. 1500, 2000, 2500 subjects.

## Case studies

**Case 1:** Sandra-Petrescu et al. <sup>18</sup> did a post-hoc analysis of a clinical trial originally designed to compare the relative efficacy of capecitabine versus fluorouracil as treatments for rectal cancer. The post-hoc analysis examined how receiving complete versus incomplete cycles of chemotherapy affected overall survival and disease-free survival. They concluded that (1) overall survival for five years was better for subjects who received complete versus incomplete cycles (Treatment N = 251, Control N = 141; observed p-value < 0.0001), (2) in the subgroup of subjects that survived at least 6 months, overall survival was better for subjects who received complete versus incomplete cycles (Treatment N = 89, Control N = 21; observed p-value = 0.0073), and (3) that there was improved 5-year disease-free survival in the complete cycle group but it was not statistically significant (Treatment N = 251, Control N = 141; observed p-value = 0.0646).

Using the extended optimal  $\alpha$  approach, we would conclude that –

- (1) It is difficult to provide an effect size estimate because the authors did not provide a precise p-value. For a p-value = 0.0001 we would have strong evidence of an increased probability of overall survival after 5 years in the complete cycle group of up to 122% (i.e. HR = 2.22) and strong evidence against a hazard ratio larger than 2.22. However, the observed p-value is less than 0.0001 so the precise observed value is unknown though it implies evidence for a larger effect size, but we don't know how much larger.
- (2) For the subgroup that survived beyond 6 months, there was evidence of an

increased probability of overall survival after 5 years in the complete cycle group of up to 244% (i.e. HR = 3.44). There was evidence against an improvement of greater than 244%. The evidence for these conclusions is strong (p=0.0073).

- (3) There was evidence of an increased probability of disease-free survival after 5 years in the complete cycle group of up to 41% (i.e. HR = 1.41). There was evidence against an improvement of greater than 41%. But the evidence for these conclusions is moderate (p=0.0646).

**Case 2:** de Boer et al. <sup>19</sup> led a phase-3 clinical trial examining the relative efficacy of radiotherapy alone versus a combination of radiotherapy and chemotherapy (during and after radiotherapy) in treating women with high-risk endometrial cancer. They concluded that (1) there was improved 5-year overall survival in subjects receiving radiotherapy + chemotherapy relative to those receiving radiotherapy only but the difference was not statistically significant (Treatment N = 330, Control N = 330; observed p-value = 0.109), (2) there was a statistically significant improvement in the 5-year failure-free survival for subjects receiving chemo-radiotherapy (Treatment N = 330, Control N = 330; observed p-value = 0.022), (3) adverse events were statistically significantly more likely to occur in subject receiving radiation + chemotherapy than radiation alone (Treatment N = 330, Control N = 330; observed p-value < 0.0001), and (4) neuropathy persisted significantly more often in subjects receiving radiation and chemotherapy than those receiving radiation alone (Treatment N = 330, Control N = 330; observed p-value < 0.0001).

Using the extended optimal  $\alpha$  approach, we would conclude that –

- (1) There was evidence of an increased probability of overall survival after 5 years in the radiotherapy + chemotherapy group of up to 21% (i.e. HR = 1.21). There was evidence against an improvement of greater than 21%. But the evidence for these conclusions is relatively weak ( $p=0.109$ ).
- (2) There was evidence of an increased probability of failure-free survival in the combined group after 5 years of up to 39% (i.e. HR = 1.39). There was evidence against an improvement of greater than 39%. The evidence for these conclusions is moderate ( $p=0.022$ ).
- (3) It is difficult to provide an effect size estimate for adverse effects or neuropathy because the authors did not provide a precise p-value. For a p-value = 0.0001 we would have strong evidence of an increased probability of 81% of adverse events and of neuropathy (i.e. HR = 1.81) and strong evidence against a hazard ratio for both adverse events and neuropathy larger than 1.81. However, the observed p-value is less than 0.0001 so the evidence is for a larger effect size for both adverse events and neuropathy, but we don't know how much larger.

**Case 3:** Reck et al.<sup>20</sup> examined the relative efficacy of ramucirumab plus docetaxel versus docetaxel plus placebo in the treatment of refractory patients with advanced non-small cell lung cancer. They concluded that including ramucirumab improved overall survival (Treatment N = 182, Control N = 178; observed p-value = 0.197), progression-free survival

(Treatment N = 182, Control N = 178; observed p-value = 0.022), objective response rate (Treatment N = 182, Control N = 178; observed p-value = 0.014), and tumour response (Treatment N = 182, Control N = 178; observed p-value = 0.049). They did not report p-values when stating their conclusions in the abstract but did conclude in the discussion that only progression-free survival, objective response rate and tumour response showed a 'significant' improvement when ramucirumab was combined with docetaxel.

Using the extended optimal  $\alpha$  approach, we would conclude that –

- (1) There was evidence of an increased probability of overall survival after 5 years in the ramucirumab + docetaxel group of up to 21% (i.e. HR = 1.21). There was evidence against an improvement of greater than 21%. But the evidence for these conclusions is very weak ( $p=0.197$ ).
- (2) There was evidence of an increased probability of progression-free survival in the combined group after 5 years of up to 87% (i.e. HR = 1.87). There was evidence against an improvement of greater than 87%. The evidence for these conclusions is strong ( $p=0.002$ ).
- (3) There was evidence of an increased probability of the objective response rate in the combined group after 5 years increasing by up to 63% (i.e. HR = 1.63). There was evidence against an improvement of greater than 63%. The evidence for these conclusions is moderate ( $p=0.014$ ).
- (4) There was evidence of an increased probability of the tumour response rate in the combined group after 5 years increasing by up to 45% (i.e. HR = 1.45).

There was evidence against an improvement of greater than 45%. The

evidence for these conclusions is moderate ( $p=0.049$ ).

## DISCUSSION

### **Results summary**

Optimal  $\alpha$  2.0 allows researchers to make inferences explicitly about the effect sizes for which there is evidence AND inferences about effect sizes against which there is evidence.

Small p-values in clinical trials with few subjects is evidence of larger effects than in clinical trials with many subjects. Evidence of hazard ratios as large as 2 in clinical trials with large sample sizes (i.e. > 1000) requires very low observed p-values.

Further, the same observed p-value in clinical trials with subjects equally distributed between control and treatment groups is evidence for a smaller effect than in clinical trials where subjects are more unevenly distributed among treatment and control groups.

The case studies demonstrate that optimal  $\alpha$  2.0 leads to qualitatively different inferences. In these case studies, optimal  $\alpha$  2.0 provided hazard ratio estimates even where the original decision had been to accept the null hypothesis of 'no effect'. The hazard ratio estimates ranged from relatively small (e.g. 1.21) to estimates that are certainly clinically significant (e.g. 2.22. and 3.44), if real.

### **Optimal $\alpha$ 2.0 versus traditional use of p-values**

Historically in clinical trials, p-values have been embedded in an NHST context <sup>21</sup>. That is, they have been used to make decisions about rejecting or failing to reject the null hypothesis. It was recognized quickly that this use of p-values ignored an essential aspect of any clinical trial – the size of the treatment effect <sup>1</sup>. There have been attempts to incorporate effect size estimates when interpreting clinical trials results but they have generally been ad hoc – either ignoring p-values if effect sizes were above or below some pre-defined threshold <sup>22</sup> or discussing 'near significant' results if the effect size estimates were above some arbitrary threshold and the p-value was within some arbitrary distance of the rejection threshold

<sup>23,24</sup>. Optimal  $\alpha$  2.0 integrates p-values and effect sizes, so that inferences explicitly identify the range of effect sizes for which there is evidence and the range of effect sizes against which there is evidence.

### **Optimal $\alpha$ 2.0: Interpreting effect size estimates**

However, as with traditional use of p-values, there is potential for misinterpreting optimal  $\alpha$  2.0 results. Optimal  $\alpha$  2.0 identifies an effect size threshold – that is, the largest effect size for which there is more evidence than an effect size of zero. Thus, it is more likely that the true effect size is zero than that it is larger than the effect size threshold. However, it does not mean that the true effect size is (1) as large as the effect size threshold or (2) at least as large

as the effect size threshold. In fact, it would be reasonable to infer that the true effect size is smaller than the effect size threshold because the effect size threshold is found at precisely the value where the data provide equal support for the value at the threshold AND for an effect size of zero.

The greatest potential for misinterpretation is to infer that the effect size threshold is the best estimate of the true effect size because, the effect size threshold will be, in most cases, an overestimate of the true effect size. There are a variety of approaches to calculating the 'best' effect size estimate<sup>25</sup>. However, Optimal  $\alpha$  2.0 can provide the range of effect sizes for which there is more evidence than an effect size of 0 and the range of effect sizes for which there is more evidence of an effect size of 0.

#### ***Optimal $\alpha$ 2.0: Interpreting p-values***

Optimal  $\alpha$  2.0 provides estimates of the Type I and II error probabilities (i.e.  $\alpha$  and  $\beta$ ) at the effect size threshold. The observed p-value is the Type I error probability if one concludes that the effect size is as large as the effect size threshold and the  $\beta$  associated with the observed p-value and the effect size threshold, is the Type II error probability if one concludes that the effect size is zero rather than a value as large as the effect size threshold. Thus,  $\alpha$  and  $\beta$  at the effect size threshold, provide estimates of one's confidence in one's conclusions about the effect size, because they provide conditional probabilities of making either error. If the probability of making Type I and II errors is very small, then one has a great deal of confidence in conclusions about evidence for an effect size up to the effect size threshold but against an effect size larger than the effect size threshold. By contrast, large values of  $\alpha$  and  $\beta$

imply less confidence in clinical trial conclusions.

The case study section of the results provides explicit examples of how Optimal  $\alpha$  2.0 results should be interpreted.

#### ***Optimal $\alpha$ 2.0: Prior probabilities***

Optimal  $\alpha$  2.0 can be integrated into a Bayesian framework by incorporating prior probabilities of the null and the alternate hypotheses ( $\text{Pr}H_0$  and  $\text{Pr}H_A$ , respectively)<sup>15</sup>. If we assume  $\text{Pr}H_0 = \text{Pr}H_A = 0.5$  optimal  $\alpha$  is identified by minimizing  $\alpha + \beta$ . If  $\text{Pr}H_0 \neq \text{Pr}H_A$ , then optimal  $\alpha$  is identified by minimising  $(\text{Pr}H_0 * \alpha) + (\text{Pr}H_A * \beta)$ . For example, if  $\text{Pr}H_0 = 0.3$  and  $\text{Pr}H_A = 0.7$ , then optimal  $\alpha$  is identified by minimizing  $(0.3\alpha + 0.7\beta)$ . Note that all estimates in this paper were made assuming  $\text{Pr}H_0 = \text{Pr}H_A = 0.5$ .

However, this simple approach to incorporating prior probabilities has at least one conceptual problem because when a critical effect size is selected, the definition of the null hypothesis shifts. That is, when we identify the smallest hazard ratio that should be considered biologically significant, we shift the definition of the null hypothesis from effect size equal to zero to any effect size between zero and the critical effect size. For example, if we identify any hazard ratio less than 1.1 as not biologically significant, it implies that a hazard ratio between 1 and 1.1 will be treated as if it is a hazard ratio of 1 (i.e. effect size equal to zero). Thus, the null hypothesis shifts from  $\text{Pr}H_0 \leq 1$  to  $\text{Pr}H_0 < 1.1$ . However,  $\alpha$  is still being estimated based on a null hypothesis  $\text{Pr}H_0 \leq 1$ .

#### ***Optimal $\alpha$ 2.0: Relative costs of Type I and II errors***

All estimates in this paper were made assuming that the relative costs of Type I and II errors (i.e.

$\$ \alpha$  and  $\$ \beta$ ) were equal. However, it is not difficult to imagine contexts where the costs of Type I and II errors differ <sup>26</sup>. A difference in relative costs of Type I and II errors can be integrated into optimal  $\alpha$  2.0 by minimizing  $(\$ \alpha * \alpha) + (\$ \beta * \beta)$  rather than  $\alpha + \beta$ .

However, how to interpret effect size thresholds is not clear when minimizing costs rather than probabilities. Incorporating relative costs of Type I and II error probabilities using optimal  $\alpha$  1.0 wasn't difficult because it simply implied choosing the rejection threshold for a particular effect size that would minimize the costs rather than the probability of an error. Conceptually, this fits very well into a framework where p-values are used to make binary decisions. However, optimal  $\alpha$  2.0 extends the use of optimal  $\alpha$  beyond binary NHST decision-making to explicit inferences about effect sizes. Thus, while it is possible technically to incorporate different relative costs of Type I and II errors into optimal  $\alpha$  2.0, it's not clear that doing so integrates well with the objectives of optimal  $\alpha$  2.0.

However, this issue of relative costs of Type I and II errors is an important one for making inferences from critical trials <sup>27,28</sup>.

### ***Optimal $\alpha$ 2.0: Outstanding issues***

Estimating optimal  $\alpha$  rests on the ability to estimate  $\beta$ , that is, on the ability to estimate power for any particular test, sample size and effect size. However, for complex clinical trial designs there is often disagreement about the best approach to power analysis <sup>29,30</sup>. If different methods for estimating power, result in different estimates of  $\beta$  they will also result in different estimated of effect size thresholds and therefore different inferences about the effect sizes for which there is evidence and against which there is evidence. This implies that a

different choice of power analysis could lead to different conclusions from a clinical trial. Further investigation is needed to assess how sensitive conclusions are to choice of power analysis.

Further, we have only developed code for a limited number of relatively simple statistical techniques used in analysing clinical trials (i.e. log rank, Cox regression, chi-square, and logistic regression). Currently, optimal  $\alpha$  2.0 couldn't be used for more complex techniques.

Lastly, we need further investigation of how best to incorporate prior probabilities and relative costs of Type I and II errors into estimations of optimal  $\alpha$ .

### ***Optimal $\alpha$ 2.0: Implications for design and interpretation of clinical trials***

Optimal  $\alpha$  2.0 provides the foundation for a fundamental change in how p-values are used in clinical trials. There have been growing calls to abandon the use of p-values as a test of 'statistical significance' entirely - in clinical trials <sup>31</sup> and statistical analyses in general <sup>32</sup>.

However, researchers are understandably reluctant to discard a tool that has been effective, if flawed. Optimal  $\alpha$  2.0 mitigates many of the problems associated with the use of p-values (i.e. that p-values are uninformative about effect sizes, that rejection thresholds are arbitrary and that they disregard Type II error probabilities) and would allow health researchers to exploit the information contained in p-values without the problems associated with NHST and p-value use.

Optimal  $\alpha$  2.0 fundamentally changes how p-values are interpreted. Currently p-values are used to make a binary decision – to conclude that there was either 'no effect' or 'some effect'. If there is evidence for 'some effect'

then there is often an ad hoc attempt to discuss what the size of that effect might be. Optimal  $\alpha$  2.0 provides an effect size threshold and any conclusion is with respect to a specific and effect size. Optimal  $\alpha$  2.0 makes it impossible to make inferences from data without explicitly stating the effect sizes for which there is evidence and the effect sizes against which there is evidence. Optimal  $\alpha$  2.0 also allows observed p-values to be used to identify how strong the evidence for or against particular effect sizes are. These are dramatic improvements over how researchers currently interpret p-values in an NHST context.

Though researchers have been struggling with the concept of 'clinically significant' effects, there has not been enough emphasis placed on the question of 'What is a clinically significant effect size?'. Optimal  $\alpha$  2.0 insists on an answer to that question. For example, if there is evidence for a hazard ratio as large as 1.1 but evidence against a hazard ratio larger than 1.1, is this an adequate rationale for approving a new drug or changing an established treatment protocol? Effective interpretation of clinical trial results will require a deeper understanding of clinically significant effects.

Further, how confident must we be in our conclusions? Must the probabilities of Type I or II errors be extremely small (e.g.  $p < 0.00001$ ) or would even 1 chance in ten of making a mistake be considered good enough to recommend drug approval or treatment protocol changes? Optimal  $\alpha$  2.0 explicitly uses observed p-values for  $\alpha$  and  $\beta$  as indices of 'strength of evidence' (note: that precise p-values should be presented rather than vaguer terms like  $p < 0.0001$ ). Currently, at the scale of individual studies, the question of 'quality of evidence' is either not addressed or addressed in an ad hoc fashion. Again, effective interpretation of

clinical trials will require more explicit consideration of the strength of evidence contained in the data and how the strength of evidence should be incorporated into decision-making.

In addition, because optimal  $\alpha$  can incorporate prior probabilities and relative costs of Type I and II errors, I hope that it will stimulate increased discussion and research in those two areas. Understanding the relative costs of Type I and II errors in health research seems particularly important given (1) that the stakes can be life and death and (2) the enormous financial investment most countries make in health care. This is a difficult and contentious area of research because we are forced to address the trade-off between human health/lives and financial costs but there is no way to make responsible decisions without addressing the question of relative costs.

## REFERENCES

1. Schober P and Schwarte LE. Statistical significance versus clinical importance of observed effect sizes: What do p values and confidence intervals represent? *Anesthes Analg.* 2018; 126: 1068-1072.
2. Palesch YY. Some common misperceptions about p values. *Stroke* 2014; 45: e344-e246.
3. Lamberink HJ, Otte WM, and Sinke MRT et al. Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *J. Clin. Epid.* 2018; 102: 123-128.
4. Huang Z, Muniz-Terrera G and Tom BDM. Power analysis to detect treatment effects in longitudinal clinical trials for Alzheimer's disease. *Alzheimer's & Dementia: Trans. Res. & Clin Interven.* 2017; 3: 360-366.
5. Hirschauer N and Gruner S et al. Twenty steps towards an adequate inferential interpretation of p-values in econometrics. *J Econ Stat.* 2019; 239: 703-721.
6. Ho J and Tumkaya T et al. Moving beyond p-values: data analysis with estimation graphics. *Nature Methods* 2019 ; 16:565-566.
7. Wasserstein RL and Lazar NL. ASA statement of statistical significance and p-values: Context, process and purpose. *Am Stat.* 2016; 70: 129-133.
8. Anderson DR, Burnham KP and Thompson WL. Null hypothesis testing: Problems, prevalence, and an alternative. *J. Wildl Manage.* 2000; 64: 912-923.
9. Gliner JA, Leech NL and Morgan GA. Problems with null hypothesis significance testing (NHST): What do the textbooks say? *J Expl Educ.* 2002; 71: 83-92.
10. Nordahl-Hansen A, Øien RA and Volkmar F et al. Enhancing the understanding of clinically meaningful results: A clinical research perspective. *Psych Res.* 2018; 270: 801-806.
11. Lee JJ and Chu CT. Bayesian clinical trials in action. *Statistics in Medicine* 2012; 31: 2955-2972.
12. Blume JD. Likelihood methods for measuring statistical evidence. *Stat Med.* 2002; 21: 2563-2599.
13. Borenstein M. The case for confidence intervals in controlled clinical trials. *Control Clin Trials.* 1994; 15: 411-428.
14. Nagashima K and Sato Y. Information criteria for Firth's penalized partial likelihood approach in Cox regression models. *Stat Med.* 2017; 36: 3422-3436.
15. Mudge JF et al. Setting an optimal  $\alpha$  that minimizes errors in null hypothesis significance testing. *PLoS one* 2012; 7: e32374.
16. Dupont P. Laplace and the indifference principle: Essai philosophique des probabilités. *Rend Sem Mat Univ Politec Torino* 1977; 36: 125-137.
17. Hawthorne J and Landes J et al. The principle principle implies the principle of indifference. *Br J Phil Sci.* 2017; 68: 123-131.
18. Sandra-Petrescu F, Herrle F and Burkholder I et al. Influence of complete administration of adjuvant chemotherapy cycles on overall and disease-free survival in locally advanced rectal cancer: post hoc analysis of a randomized, multicenter, non-inferiority, phase 3 trial. *BMC Cancer* 2018; 18: 369.

19. de Boer SM et al. Adjuvant chemoradiotherapy versus radiotherapy alone for women with high-risk endometrial cancer (PORTEC-3): final results of an international, open-label, multicentre, randomised, phase 3 trial. *Lancet Oncol.* 2018; 19: 295-309.
20. Reck M, et al. Outcomes in patients with aggressive or refractory disease from REVEL: a randomized phase III study of docetaxel with ramucirumab or placebo for second-line treatment of stage IV non-small-cell lung cancer. *Lung Cancer* 2017; 112: 181-187.
21. Hillary FG and Medaglia JD. What the replication crisis means for intervention science. *International J. Psychophysiol.* 2020; 154: 3-5.
22. Del Paggio JC, Azariah B and Sullivan R et al. Do contemporary randomized controlled trials meet EMSO thresholds for meaningful benefits. *Ann Oncol.* 2017; 28: 157-162.
23. Gibbs NM and Gibbs SV. Misuse of 'trend' to describe 'almost significant' differences in anesthesia research. *J Anesth.* 2015; 115: 337-339.
24. Otte WM, Vinkers CH and Habets P et al. Almost significant: trends and P values in the use of phrases describing marginally significant results in 567,758 randomized controlled trials published between 1990 and 2020. medRxiv 2021
25. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. *Prof Psychol – Res Pr.* 2009; 40: 532-538.
26. Ioannidis JPA, Tarone R and McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* 2011; 22: 450-456.
27. Van Ravenzwaaij D and Ioannidis JPA. True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *Med Res Meth.* 2019; 19: 218.
28. Rogatko A and Litwin S. Phase II studies: Which is worse, false positive or false negative? *JNCI - J Natl Cancer I.* 1996; 88: 462.
29. Li D, Zhang S and Cao J. Incorporating pragmatic features into power analysis for cluster randomized trials with a count outcome. *Stat Med.* 2020; 39: 4037-4050.
30. Jakobsen JC, Ovesen C and Winkel P et al. Power estimations for non-primary outcomes in randomized clinical trials. *BMJ Open* 2019; 9: e27092.
31. Leopold SS and Porcher R. The minimum clinically important difference – the least we can do. *Clin Orthop and Relat R.* 2017; 475: 929-932.
32. McShane BB, Gal D and Gelman A et al. Abandon statistical significance. *Am Stat.* 2019; 73: 235-245.