STATISTICS FOR THE PEOPLE

INTERNATIONAL JOURNAL OF
RADIATION ONCOLOGY • BIOLOGY • PHYSICS

www.redjournal.org

Finding Multiple Signals in the Noise: Handling Multiplicity in Clinical Trials

Amit K. Chowdhry, MD, PhD,[*,†] John Park, MD,[‡,§] John Kang, MD, PhD,[‖] Gukan Sakthivel, MD,[*] and Stephanie Pugh, PhD[¶]

[*] *Department of Radiation Oncology, University of Rochester Medical Center, Rochester, New York;* [†] *Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry, Rochester, New York;* [‡] *Department of Radiation Oncology, Kansas City VA Medical Center, Kansas City, Missouri;* [§] *Department of Radiology, University of Missouri Kansas City School of Medicine, Kansas City, Missouri;* [‖] *Department of Radiation Oncology, University of Washington, Seattle, Washington; and* [¶] *NRG Oncology Statistics and Data Management Center, American College of Radiology, Philadelphia, Pennsylvania*

Received Nov 21, 2023; Accepted for publication Dec 3, 2023

Case Vignette

Lassman et al[1] investigated the efficacy of depatuxizumab mafodotin in the treatment of newly diagnosed glioblastoma with epidermal growth factor receptor gene amplification (EGFR-amp). There were 639 patients randomized to receive radiation, temozolomide, and either depatuxizumab mafodotin or placebo. The primary endpoint was overall survival with secondary endpoints including progression-free survival, overall survival and progression-free survival subgroup analyses, quality of life, and neurocognitive functioning. If positive, this trial would have been submitted to the Food and Drug Administration (FDA) for drug approval in the treatment of patients with newly diagnosed glioblastoma with EGFR-amp. Unfortunately, the trial was deemed futile at an interim analysis and results were subsequently reported. Given the FDA's guidance on multiplicity in clinical trials,[2] the trial used multiplicity adjustment and will serve as an example here.

Introduction

Classically in clinical trials, there is a single primary endpoint and multiple secondary endpoints (although more recently, coprimary endpoints have been introduced). Unless one corrects for the fact that multiple endpoints are being tested, and, in the presence of interim analyses, that there are tests being conducted throughout the trial, an erroneous conclusion of significance may be drawn by chance alone at a rate higher than represented in the associated P values (or Bayesian alternatives). The more one tests, the more one is going to find a "significant" result by pure chance. Multiplicity refers to a problem which occurs when multiple statistical tests are conducted in each study, such that the probability of finding false positives goes up as the number of tests increases.

A simple solution to multiplicity would be to allow only one hypothesis test per trial. Although that would mollify the statisticians' concerns, as Stephen Senn correctly points out in *Statistical Issues in Drug Development*, it would be a large waste of resources, and even ethically questionable, to only look at one hypothesis per trial; therefore, we must deal with the problem of multiplicity.[3]

In nearly every modern trial, investigators evaluate a variety of outcomes such as patient-reported outcomes (PROs) collected at multiple time points (eg, the many components of the Expanded Prostate Cancer Index Composite score in prostate cancer treatment trials). Consider the

Corresponding author: Amit K. Chowdhry, MD, PhD; E-mail: amit_chowdhry@urmc.rochester.edu

Disclosures: This work was supported by a National Cancer Institute NRG Oncology Statistical and Data Management Center grant (U10CA180822).

Acknowledgments—Thank you to Dr David Sher for invaluable feedback throughout the writing of this article. Thank you to Dr Brian Hobbs and Dr Steven Lin for their feedback on an earlier version of this article.

Int J Radiation Oncol Biol Phys, Vol. 119, No. 3, pp. 750−755, 2024

https://doi.org/10.1016/j.ijrobp.2023.12.007

situation in which one is interested in testing 20 statistical hypotheses (such as subgroup analyses), using a false positive rate (eg, α or type I error) of .05. Assuming the statistical model is correctly chosen and assuming there is no true difference between the treatments, then, on average, one of the tests would be falsely positive (ie, one divided by 20). Thus, for any study there may be well over 20 statistical tests, with more than one false positive. This conundrum may explain some trials (although most definitely not all) with incomprehensible results, and it emphasizes the importance of understanding multiplicity.

It is important that clinical trials are designed a priori with multiplicity in mind, including the choice of primary and secondary endpoints as well as the methods used for handling multiplicity. If proper considerations are not made, the interpretation of the study may be limited. Therefore, it is helpful that consumers of trial literature understand the key design issues that go into these trials.

One of the fundamental goals of using statistical methods to design and analyze trials is to ensure that the results are due to effects of treatment and not chance. We hope the reader will gain an appreciation for the underlying problem of multiplicity and understand different approaches for correcting for these multiple comparisons. Perhaps even more importantly, it will help readers and reviewers understand the limitations of research that does not consider the effect of multiplicity on the results and conclusions.

Frequentist Perspective on Multiplicity

Basics of frequentist inference

For the case of a simple superiority randomized trial comparing 2 treatments (eg, radiation therapy vs lobectomy for lung cancer), a P value for a 2-sided test is roughly defined as the probability of seeing a difference as or more extreme than what was observed under the null hypothesis (eg, there is no difference between treatments). Type I error (α) is most commonly the probability of finding a statistically significant difference between groups when they are in fact not different (eg, finding that surgery and radiation have different efficacy when there is not a difference). A common value of α used in the literature is .05. It is one measure of a false positive probability. Type II error (β) is most commonly the probability of not finding a statistically significant difference between groups when they are in fact different. The statistical power, or $1 -$ Type II error, is the probability of finding a statistically significant difference when a difference exists and is a measure of true positive probability (eg, finding that radiation therapy is superior to surgery or surgery is superior to radiation therapy for lung cancer).

Misuse of P values has been an extensively discussed topic. Much controversy was started when Ronald Wasserstein, executive director of the American Statistical Association, wrote an editorial that argued it is "time to stop using the term 'statistically significant' entirely. Nor should variants such as 'significantly different,' '$P < 0.05$,' and 'nonsignificant' survive."[4] One of the great contributors to the scorn of the misuse of P values comes from what is commonly called p-hacking, HARKing (hypothesizing after the results are known), data dredging, cherry picking, or fishing expeditions and occurs when any study showing a $P < .05$ is associated with a statistically significant finding.[5] Data sets can be easily manipulated to show $P < .05$ and thus may be a spurious finding. Although this problem is common in retrospective studies, it can also be seen in secondary analyses of clinical trials.

When designing clinical trials, one aims to reduce the frequency of false positive trials (type I error, trials that show a treatment is better or worse when this is not the correct conclusion) and false negative trials (type II error, trials that are unable to show a treatment is better or worse when in fact there is a difference). These concepts are important regardless of statistical philosophy: we do not want to make incorrect conclusions from our studies.

Frequentists generally believe that one should formally adjust for this multiplicity (eg, by lowering the acceptable type I error rate per test, or by performing hierarchical tests, eg, testing for subgroups only if the initial analysis is positive). There is consensus that formal adjustment is needed for the primary endpoint if there is more than one. It is more controversial for how to handle secondary endpoints. Some researchers argue that secondary endpoints should only be tested if the primary endpoint is positive. This argument makes sense if one would only accept a particular treatment if the trial was positive on the primary endpoint, and that the secondary endpoints are merely to provide additional information. It does provide some degree of control of type I error. For example, in the depatuxizumab mafodotin trial,[1] hierarchical testing was used to control the type I error due to the inclusion of multiple secondary endpoints. The FDA requires multiplicity adjustment for primary and secondary endpoints but not exploratory endpoints. Per FDA guidance, the primary endpoint(s) are tested initially, and if an effect is shown, the secondary endpoints can be formally tested.

Other authors argue the other extreme: that no adjustment for secondary endpoints is needed if researchers analyze and interpret the studies in an unbiased manner (if outcomes are independent, the argument is that it is just as if multiple studies have been conducted, adjustment is not necessary because we do not adjust for independent data from different studies). Unbiased researchers can evaluate each claim individually, and the false positive risk for each claim would have a probability of .05 or less. Others argue that for secondary and exploratory analyses, while adjustment for multiplicity can be done, if they are not done it must be understood that these analyses are hypothesis-generating in nature and should in general not be used to change practice. It is the view of the authors (who have differing positions) that regardless of which position is taken, readers should understand analysis approaches and the potential multiplicity issues involved in an analysis.

Simple approaches: Composite endpoints

One approach for handling multiplicity is to use composite endpoints. For example, a composite endpoint may use a weighted average of toxicity: one takes a number of toxicity outcomes and comes up with a single number, which can be tested, thus removing the problem of multiple testing in the event different toxicity outcomes were tested simultaneously.

Another simple way of dealing with multiple primary endpoints is the *all-or-none procedure*.[6] For the all-or-none procedure, consider the scenario in which there are 2 coprimary endpoints. For this procedure, both endpoints need to have a P value less than the prespecified type I error rate (.05 is commonly used) in order for each of them, and the trial, to be declared statistically significant. There are more advanced statistical procedures that allow for consideration of multiple hypotheses without compromising statistical rigor, which we will discuss in the following.

P value adjustment with Bonferroni and similar procedures including Holm/Hochberg and hierarchical testing

The Bonferroni procedure is probably the most well-known approach for adjusting for multiple comparisons. To perform the Bonferroni procedure, the α level (type I error probability) is divided by the number of tests being performed. The Bonferroni adjustment is straightforward, is easy to implement in both a prospective and retrospective analysis, and does not have any assumptions.

A limitation of this correction method is that it is overly conservative and does not account for any potential correlation between the outcomes being tested. For example, because scores across a PRO measure are likely to be correlated to some degree, the Bonferroni approach would reduce the statistical power to detect a difference if one exists.

An alternative approach is hierarchical testing. This approach can be used to develop methods with better properties than the Bonferroni procedure. Consider the situation of a study in which multiple doses of a treatment are being considered. Let us assume that with increasing dose, there is an increasing probability of efficacy on a particular outcome (eg, dose of radiation therapy for local control). For ordered null hypotheses, such as this situation, one may use a form of hierarchical testing, where we assume a priori that the higher doses are more likely to have better local control than lower doses.

For this next part, let us consider a concept called the family-wise type I error rate. The family-wise type I error rate of a group of tests (or "family") is the probability of making at least one type I error (false positive) among the group of tests. In clinical trials, including for the FDA, the family-wise type I error rate is the quantity that investigators try to keep under a prespecified value. This is the probability that making at least one type I error is equal to α (eg, .05).[7]

For example, consider a family-wise type I error (probability) of .05. Then one may design a test where if the highest dose is significant with $P < .05$, the second highest dose can then be tested against α until one gets to a dose which is not significant. With this approach, one can test multiple hypotheses without an adjustment because the testing is being done hierarchically. This approach is perhaps one of the simplest examples of a hierarchical test, but one can imagine more complicated versions of this test. A hierarchical test is a process in statistical testing, in which hypotheses are tested sequentially. In this study, if the first comparison is significant at an α level, the second test can be tested at that same α level without any compromise of type I error. Intuitively, one can imagine that if we require that a prior test is significant before considering the next hypothesis then we are not considering all hypotheses as equally important. Whether a hypothesis is tested is dependent on the result of the prior hypothesis and corresponding test.

The Lassman et al[1] trial in the vignette used hierarchical testing to control the type I error for secondary endpoints. The primary endpoint was overall survival and the first 2 secondary endpoints were progression-free survival and overall survival in the O-6-methylguanine DNA methyltransferase (MGMT) unmethylated subgroup. If overall survival were positive, then progression-free survival would be tested at the same significance level. If positive, then overall survival in the MGMT unmethylated subgroup would be tested also at the same significance level, and so on. If an endpoint is tested and not significant, then the remaining endpoints would not be tested. This type of adjustment is attractive when conducting a registration-intent trial because it maintains the type I error for the individual tests as well as overall. However, it does require prioritizing endpoints and understanding that some may not be tested.

There are also other more advanced versions of type I error level or P value adjustment, some of which are always equally good or better (ie, same type I error, same or more power) than the Bonferroni (such as the Holm step-down procedure). The concept behind Holm is relatively simple. For Holm, one orders the P values by magnitude. You compare the smallest P values to α divided by the number of tests (m). If the first test is significant, you reject the null hypothesis for that test, and then you start over with the remaining tests, with the number of tests being $m - 1$ (and comparing the next smallest P value to α divided by $m - 1$). This process is continued. Based on this process, one can see how it is easier to find significant tests, and it can be mathematically proven that this test controls type I error well and has more statistical power than the Bonferroni procedure.[8]

Here is a simple example of the Bonferroni and Holm procedures. Consider a study with 3 endpoints: (1) a toxicity measure, (2) a tumor size difference, and (3) overall survival.

Table 1 Benefits and limitations of common methods for multiplicity adjustment within frequentist statistical testing

Multiplicity adjustment procedure	How it is performed	Benefits	Limitations
Bonferroni procedure	Take the α level/type I error (eg, .05) and divide it by the number of comparisons	Minimal assumptions Very easy to implement	Overly conservative Does not account for any correlation between outcomes (eg, many PROs are correlated), thus resulting in a reduced ability to detect a difference
Holm step-down procedure	Rank *P* values for comparisons. Compare the smallest *P* value to α/m (where m is the number of comparisons). If the smallest *P* value is less than α/m, then check if the second smallest *P* value is less than $\alpha/(m-1)$. Continue this procedure for the third smallest *P* value (comparing with $\alpha/[m-2]$). Stop the procedure when the *i*th smallest *P* value is not smaller than $\alpha/(m-i)$. The differences are significant if the corresponding *P* value is less than the comparison.	Minimal assumptions At least as powerful as Bonferroni Same assumptions as Bonferroni	Although easy to implement with statistical software, not as simple as Bonferroni
Hochberg procedure	Similar to Holm, but instead of starting with the smallest *P* value and going to bigger values, we start with the largest *P* value and compare with $\alpha/1$, the second largest and compare with $\alpha/2$, the third with $\alpha/3, \ldots$, until α/m (if not stopped earlier). We stop when $P_i \geq \alpha/i$. All comparisons for which $P < \alpha/i$ for each comparison are considered significant.	More power than Holm step-down procedure	Requires independence and positively correlated values for control of the family-wise error rate More conservative than Hommel (ie, will have lower power)
Hommel procedure	Details are described elsewhere (beyond the scope of the article)[8]	More power than the Hochberg procedure	Requires independence and positively correlated values for control of the family-wise error rate More difficult to understand than Hochberg

Abbreviation: PRO = patient-reported outcome.

The 3 *P* values were .04, .02, and .005. If the α level was set to .05, by the Bonferroni procedure, the *P* values would be compared with $\alpha/3$ = .017, and only 1 out of the 3 tests would be significant. For Holm, the .005 would be compared with $\alpha/3$ = .017 and would still be significant, the .02 would then be compared with $\alpha/2$ = .025 and would now be significant, and the .04 would be compared with $\alpha/1$ = .05 and would also be significant.

Therefore, we recommend that the Holm step-down procedure should always be used in situations in which one would otherwise plan to use the Bonferroni procedure. There are a number of other common procedures, many of which are listed in Table 1. Having a treatment be superior to another in one outcome makes it likely that it is also superior in other outcomes, such as different domains of quality of life. Bonferroni ignores this fact by treating all tests as independent while other procedures, such as the Holm step-down procedure, take this into account.

For those interested, the following is a proof of why the Bonferroni procedure gives too strict control of Type I error (feel free to skip to the next section if the proof is not of interest). Consider 2 statistical tests in a clinical trial for which we perform the Bonferroni correction. For the next part, let *P*(X) denote the probability of some event X occurring. Based on probability theory, it is known that *P*(either test A or test B is a false positive) = *P*(test A is false positive) + *P*(test B is a false positive) − *P*(both A and B are false positives) = type I error of test A + type I error of test B. One can see from this formula that the type I error when the Bonferroni procedure is performed is always the same or larger than the true type I error because it does not subtract out *P*(both A and B are false positives).

Adjusting for multiplicity when there are interim analyses

Consider a clinical trial with multiple interim analyses in which one rejects the null hypothesis if $P < .05$ or finds that the probability of effectiveness is enough to stop the trial and declare the treatment or control superior.

Any time there are multiple tests being performed and decisions are made based on statistical tests, the probability of a false positive (ie, family-wise type I error rate) of the multiple tests taken together is higher than each test alone. Consider a trial with 6 interim analyses (and 1 final analysis) with the following P values for each test: .99, .52, .043, .34, .41, .37, .38. If one had stopped the trial whenever $P < .05$, then one would have found that one treatment had a different efficacy than another. However, as accrual to the trial and follow-up continued, the difference disappeared.

There are a variety of solutions to this problem, but perhaps the best way to understand the problem conceptually is that of the α spending function.[9] With an α spending function, one splits the type I error rate to be used at each look of the data. For example, the simplest approach (but generally thought to not be the best approach) is to split the type I error evenly across the time points. In the aforementioned example, the .05 could be split into .05/7 = .0071. Therefore, we would not reject the null hypothesis unless one of the time points had a P value less than .0071. However, in practice, one would not want to "spend" too much of the type I error on the early endpoints because these would have low power to detect a difference, as the sample size is smaller. Therefore, we can use increasing functions to allow for a larger cutoff later in the study. The benefit of this approach is that it allows one to declare a winner for dramatically successful or harmful treatments without compromising too much on power.

Indeed, this approach allows one to declare significance for very small P values early on and more moderate P values later in the study. One popular method that does this is called the O'Brien-Fleming (which interestingly predates the concept of an α spending function). For the simple normal case, with a z statistic, the O'Brien-Fleming method rejects the null hypothesis when the z statistic is greater than the z critical value times $\sqrt{K/k}$, where K is the total number of potential interim analyses and k is the interim analysis number.[10] Consider a simple study with a normally distributed outcome variable (such as a difference in tumor size changes over time between 2 treatments) with 3 interim analyses using a 2-sided z test. Let's assume the z statistic for the difference in tumor size is 2.0 at the first interim analysis (A test statistic is a value that is computed for the purpose of conducting a statistical test. For the change in tumor size example, a z statistic would be the sample mean difference in tumor size between treatment and control groups minus the mean for the null hypothesis divided by the true standard deviation.). If no adjustment for multiplicity were done, then this would be greater than the critical value of 1.96 (for example), and we would declare one treatment superior to the other. However, using the O'Brien-Fleming method, we would multiply 1.96 by $\sqrt{3/1}$, which is $1.96 \times 1.73 = 3.39$, as this test is being done for the first of the 3 interim analysis. Therefore, because 2 is less than 3.39, we do not reject the null hypothesis, and the study continues.

Bayesian Perspectives on Multiplicity

The Bayesian perspective and the longstanding debate about its place relative to frequentist statistics is described in detail in an earlier *Statistics for the People* article by Fornacon-Wood et al and subsequent discussions.[11-13] From the Bayesian perspective, its approach is a strength with respect to multiplicity, as when appropriate priors are used there is some degree of control of multiplicity, without requiring formal adjustment.

There is controversy among Bayesians as to whether one should adjust for multiplicity. Some Bayesians would argue that, if you use an appropriate prior, perhaps a skeptical prior (a prior that assumes large treatment effects are unlikely) for a proposed treatment, no adjustment is necessary even for unlimited looks at the data in interim analyses. This would hold because Bayesian probabilities represent the current state of knowledge and thus new information only needs to be added to the prior and no adjustments are needed other than to update the posterior probability to reflect the new current state of knowledge after looking at the interim data. Others say that you can have unlimited looks at the data in interim analyses, and it should not change Bayesian inference unless there are multiple primary outcomes, any one of which could be used to declare the study successful.

By contrast, many Bayesian clinical trial statisticians say that frequentist type I error control is of interest, as most new treatments are ineffective, so the type I error probability is a measure of the false positive probability of a study design. There are philosophical Bayesian arguments against this view. To understand the controversy fully, one needs a deep understanding of Bayesian statistics. For interested readers, the details of the debate are discussed in the following paragraph (which can be skipped for those not interested in the debate).

Bayesian statisticians who support controlling type I error may give the example of a randomized trial of a new systemic therapy agent for which there are 10 interim analyses. If one makes a decision at each analysis regarding whether to stop the trial based on whether one treatment is superior, there is a higher likelihood of finding a treatment effect even when it does not exist (which is the most common situation in trials for new systemic therapy agents). The response to this from strict Bayesian statisticians is that for the classical definition of type I error in frequentist statistics, at each "look" one asks the question of whether there is a treatment effect after assuming there is no treatment effect. One must then control the probability that at that “look” there is a possibility of drawing a false conclusion; by contrast, from a Bayesian perspective one can evaluate the probability of no effect given the available and accumulating data.

Therefore, there is not a single Bayesian approach to handling multiplicity. A deeper discussion of Bayesian approaches to handling multiplicity are beyond the scope of the article.

Conclusion

Multiplicity is a critically important problem in analyzing prospective trials and must be managed for proper inference. Regardless of the perspective taken, understanding multiplicity can help readers of journal articles understand whether statistical findings are likely to be due to chance alone and have a risk of being spurious.

References

1. Lassman AB, Pugh SL, Wang TJC, et al. Depatuxizumab mafodotin in EGFR-amplified newly diagnosed glioblastoma: A phase III randomized clinical trial. *Neuro Oncol* 2022;25:339-350.
2. Food and Drug Administration. Multiple endpoints clinical trials guidance for industry. Available at: https://www.federalregister.gov/documents/2022/10/21/2022-22882/multiple-endpoints-in-clinical-trials-guidance-for-industry-availability. Accessed February 6, 2023. .
3. Senn SS. *Statistical Issues in Drug Development*. Wiley; 2021.
4. Wasserstein RL, Schirm AL, Lazar NA. *Moving to a World Beyond “P < 0.05.”* Taylor & Francis; 2019:1-19.
5. Andrade C. HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *J Clin Psychiatry* 2021;82:20f13804.
6. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. CRC Press; 2009.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57:289-300.
8. Cui X, Dickhaus T, Ding Y, Hsu JC. *Handbook of Multiple Comparisons*. CRC Press; 2021.
9. Demets DL, Lan KG. Interim analysis: The alpha spending function approach. *Stat Med* 1994;13:1341-1352.
10. O’Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549-556.
11. Fornacon-Wood I, Mistry H, Johnson-Hart C, et al. Understanding the differences between Bayesian and frequentist statistics. *Int J Radiat Oncol Biol Phys* 2022;112:1076-1082.
12. Fornacon-Wood I, Mistry H, Price GJ, Faivre-Finn C, O’Connor JP. In Reply to Chowdhry et al. *Int J Radiat Oncol Biol Phys* 2023;115:250-251.
13. Chowdhry AK, Mayo D, Pugh SL, Park J, Fuller CD, Kang J. In regard to Fornacon-Wood et al. *Int J Radiat Oncol Biol Phys* 2023;115:249-250.