Paper AS09

Across the Multiverse…Statistically Speaking

Cralen Davis, Cytel, Massachusetts, U.S.A.
Trevano Dean, Cytel, Massachusetts, U.S.A.

Abstract

In a clinical or other controlled trial, the goal is often to demonstrate the validity of a claim or effectiveness of a product to a governing or regulatory body using specific measures from collected data, generally referred to as endpoints. It is very common for a trial to have multiple endpoints or multiple tests involving 1 or more endpoints. However, formal testing of multiple hypotheses within a trial may result in multiplicity, the increased probability of falsely concluding that a result is statistically significant. Guidance from both the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) states that special consideration should be given to adjusting the alpha level, or probability of observing a false positive, to account for multiplicity. This paper discusses some of the methods and approaches that can be used to address multiplicity.

Keywords: multiplicity, type I error, hypotheses, adjustment, p-value, critical value

Introduction

There is an old expression that says, curiosity killed the cat. If that saying were literally true, researchers would be the greatest committers of animal-related atrocities in all history! Fortunately, it is just a saying and the kitties are fine even as new research trials start daily and curiosity abounds. So much curiosity abounds that it is not unusual to evaluate multiple statistical hypotheses in a clinical trial or study. However, the evaluation of multiple hypotheses can often lead to an increase in the type I error rate (α), the probability of falsely concluding that a result is statistically significant (Table 1). This multiple hypothesis testing within the same study and potential increase in the type I error rate, also referred to as alpha, is known as multiplicity.

Specifically, multiplicity occurs when 2 or more endpoints are used to determine significance of an important outcome or outcomes from a study or trial. Examples of multiple endpoints can be in terms of: pair-wise comparisons over more than 1 measure (ex. group 1 vs. group 2 on hospitalizations, falls, lipid levels, BMI, etc.); 1 measure and multiple pair-wise comparisons (ex. weight loss for group 1 vs. group 2, group 1 vs. group 3, etc.); or a mixture of the two. While not all cases of multiplicity are problematic (ex. a trial with two endpoints where both have to be significant to conclude effectiveness), adjustment should be taken into consideration more times than not, as both the Food and Drug Administration (FDA) and European Medicines Agency (EMA) recommend acknowledging and accounting for actual, or even potential, multiplicity.

Table 1.
Types of Conclusions in Statistical Hypothesis Testing

Error types		**Actual fact**	
		H_0 true	**H_0 false**
Statistical inference	H_0 true	Correct	Type II error (β)
	H_0 false	Type I error (α)	Correct (Power)

When to Consider Adjustment for Multiplicity

Any case where a trial can be deemed successful if fewer endpoints than the total number of endpoints considered for making such a determination show statistical significance, adjustment for multiplicity is most likely needed. For example, consider a clinical study with 2 primary endpoints where the trial could be deemed successful if statistical significance is achieved by at least one of the endpoints. Adjustment for multiplicity would be needed in this example because the risk of coming to a false positive conclusion is increased. This can be shown using the following 2-endpoint example. If we assume that each endpoint will be tested at a 1-sided alpha of 0.025 and the tests are independent, then each test has a 0.975 probability of giving the true conclusion under the null hypothesis. However, in this example only 1 null hypothesis has to be rejected to conclude overall success. This means that under the overall null hypothesis for the study, that both tests are non-significant, the probability of getting the correct

conclusion is 0.975 x 0.975 = 0.951, which results in the probability for reaching the wrong conclusion as being 1 – 0.951 = 0.049. This is almost a 5% type I error rate, nearly double that 2.5% rate for 1 test and illustrates the mistake that can be made when adjustment for multiplicity has not been taken into consideration.

Controlling Family-Wise Error Rate

In the example above, the initial alpha was set as 0.025 but in actuality, it was nearly double that due to testing multiple endpoints. The collection of these tests are often referred to as a family and the alpha related to that collection, referred to as the family-wise error rate (FWER). The example showed how the alpha can be inflated due to multiplicity; however, by adjusting for multiplicity the desired alpha level can be preserved. This is also known as controlling the family-wise error rate. There are several methods for doing this and some of the more common approaches will be presented in this document.

Methods: FWER

Most methods that adjust for multiplicity involve changing the alpha-level-based critical value or adjusting p-values; however, there are some approaches that do not change the alpha, such as prespecifying the order in which endpoints will be evaluated. With the latter approach, each test can be performed at the original, unchanged alpha level where a subsequent test can only be performed if the preceding test was below the prespecified alpha level. For illustration, consider two studies that have alpha=0.05: one with 4 ordered endpoints (A1, A2, A3, A4); and one with 5 ordered endpoints (B1, B2, B3, B4 B5). In trial A: A1, A2, and A3 are significant at 0.05 but A4 is not, yet A4 was able to get tested; for trial B: B1 and B2 are significant but not B3. This means B3 was evaluated but B4 and B5 were not because the process stopped at the non-significant B3 endpoint.

Examples of methods covered here that call for a change to the alpha-level-based critical value or p-value adjustment include Bonferroni, Holm, Hochberg, and minP. While this is by no means an exhaustive list, it does provide some known and fairly straight-forward methods that can be implemented without a great degree of complexity.

Bonferroni

The Bonferroni method of adjusting for multiplicity is a simple approach where all endpoints are evaluated using the same alpha-level-based critical value/significance level, which is based on the number of endpoints. The significance level is calculated by dividing the original, desired alpha level by the number of endpoints. For example, if the desired FWER is 0.05 and there are 5 endpoints, then the Bonferroni adjusted alpha is 0.05/5 = 0.01. Each endpoint would be evaluated at the 0.01 level. The Bonferroni method is a quick and easy approach that will address multiplicity; however, it is one of the most conservative approaches for doing so and can significantly reduce power. The Bonferroni method controls FWER under the null hypothesis and assumes the evaluations are independent.

Holm

The Holm method, also referred to as Bonferroni-Holm/Holm-Bonferroni method of adjusting for multiplicity, is an approach that requires the p-values for each endpoint be ordered from smallest-to-largest before being evaluated in that order. This approach is often referred to as a step-down approach due to the sequential evaluation of p-values from the smallest-to-largest and potentially stepping from the lowest p-value, down to the next lowest p-value. P-values are evaluated one at a time and the evaluation is stopped at the first instance of a p-value being greater than its corresponding significance level, which is adjusted in a Bonferroni-like manor. The significance level is determined by dividing the FWER by the result of the following: number of endpoints plus 1 minus the order rank of the p-value. Consider an example with 3 endpoints: A, B, and C; which have corresponding p-values: pA, pB, and pC. Let the order of p-values from smallest to largest be pB with rank 1, pA with rank 2, and pC with rank 3. Further, assume a FWER of 0.05. In the example, pB would be evaluated at the significance level of 0.05/(3 + 1 – 1) or 0.01667. In the rank 1 case, it is essentially the Bonferroni method. If pB ≤ 0.01667, then pA would be evaluated at the significance level of 0.05/(3 + 1 – 2) or 0.025. If pA ≤ 0.025 then pC would be evaluated at 0.05; however, if pA > 0.025, then the process would stop, endpoints A and C would be deemed non-significant and pC would not be evaluated. The Holm method controls FWER under the null hypothesis and is less conservative than the Bonferroni method, which in turn makes it more powerful, more likely to reject the null hypothesis when it is false (Table 1). The Holm method assumes the hypotheses are independent.

Hochberg

The Hochberg method of adjusting for multiplicity is similar to the Holm method. in that the P-values are ordered and evaluated sequentially; however, the p-values are ordered from largest to smallest. This approach is often referred to as a step-up approach due to the sequential evaluation of p-values from largest to smallest. Unlike the Holm method, the evaluation for Hochberg is stopped at the first instance of a p-value being less than its corresponding significance level. The significance level in this method is also adjusted in a Bonferroni-like approach; it is determined by dividing the FWER by the order rank of the p-value. Consider the example from above with endpoints: A; B; and C; and p-values: pA; pB; and pC. The Hochberg order calls for a reverse of the Holm ordering. For the Hochberg approach, p-values are ordered as follows, from largest to smallest: pC with rank 1, pA with rank 2, and pB with rank 3. The FWER is 0.05. In this example, pC would be evaluated at the significance level of 0.05/1 or 0.05. If pC < 0.05, the process would stop and endpoints C, A, and B would be deemed significant despite the values of pA or pB; pA and

pB would not be evaluated. If pC ≥ 0.05, then pA would be evaluated at the significance level of 0.05/2 or 0.025. If pA < 0.025 then endpoints A and B would be deemed significant despite the value of pB, which would not be evaluated. The Hochberg method controls FWER under the null hypothesis and it is slightly more powerful than the Holm method. The Hochberg method assumes the hypotheses are independent.

MinP

The Bonferroni, Holm, and Hochberg methods of adjusting for multiplicity all require an assumption that the testing or evaluations of the endpoints are independent. The minP method, is not limited by that assumption and can be used when the endpoint evaluations are related. This non-independence feature is due to the empirical nature by which the alpha-level-based critical values are obtained and the bootstrap sampling approach employed by the minP method. With bootstrap sampling, or random sampling with replacement, samples are taken from the original data to make a collection of representative datasets, then test statistics are calculated for each resampled dataset. Critical values corresponding to the FWER can then be selected from the distributions created by combining all the resampled data-based test statistics. These critical values are used to evaluate the test statistics from the original dataset. Alternatively, p-values can be obtained by computing the proportion of critical values that are more extreme than the original data-based test statistics.

The nature of bootstrap sampling, which is used in the minP method, allows the datasets to retain any endpoint correlation that is present in the original data. Since that correlation is already accounted for in the collection of datasets and the alpha-level-based critical values or p-values ultimately come from the collection, the requirement of independent hypotheses is no longer needed with the minP approach.

Implementation in SAS®: FWER

SAS® can be used to implement the FWER controlling methods mentioned above. Implementation can be done in multiple ways such as data-step-level programming, using macros, etc. SAS® also has a procedure called PROC MULTTEST that implements these methods. Rather than providing adjusted critical values based on a specific method, the procedure generally operates in reverse by providing method-based adjusted p-values. As an example, assume we use the Bonferroni method to evaluate three endpoints at an FWER of 0.05. Based on the calculations described above, each p-value would be compared to 0.05/3 = 0.0167. PROC MULTTEST, on the other hand, would multiply each p-value by 3 and compare it to the FWER. In a three endpoint Holm method example with an FWER of 0.05, where the resulting ordered significance levels or critical values would be 0.05/3, 0.05/2, and 0.05/1 respectively; if this same example were conducted using PROC MULTTEST, it would return the following ordered adjusted p-values of 3 x p1; 2 x p2; and 1 x p3. Note that in both cases, the sequential evaluation of the p-values and stopping rule implementation would have to be done manually.

Although the use of SAS® was discussed here, there are many different ways to implement FWER controlling methods that do not require the use of complex software.

Controlling False Discovery Rate

The importance of controlling the family-wise error rate to address type I error rate inflation that can occur due to multiplicity was discussed earlier in this paper. While FWER is about preserving an acceptable rate of not rejecting the null hypothesis when it is true, given multiplicity; there may be times when the choice is more about making sure the null hypothesis is rejected at least every time it is false. The false discovery rate (FDR) is something that can be used in those situation.

Methods: FDR

When multiplicity is present and the desire is to determine each endpoint or test that may be important, one may use an FDR approach. This type of approach is typically more lenient to having greater type I errors than in an FWER approach but lower type II errors (Table 1). Consideration for controlling FDR usually involves determining if doing more testing or evaluating more endpoints to identify important signals or measures is of greater value than having false positives. If an FDR approach is desired and identifying all significant or potentially significant results are more valuable than having false positives, then a higher FDR of say 0.10 or 0.20 may be reasonable. If an FDR approach is desired with more of a balance or if having multiple type I errors is a concern, then a more conservative FDR such as 0.05 may be preferred.

In general, FDR methods are more powerful than FWER methods, given it is a more liberal approach for determining significance. The Benjamini-Hochberg method is an example of a common FDR approach.

Benjamini-Hochberg

The Benjamini-Hochberg method of adjusting for multiplicity has some similarities to the Holm method such as it calls for the ordering of p-values from smallest to largest value and the use of a Bonferroni type of adjustment. However, it is considered a step-up rather than a step-down method and the p-values are not processed sequentially like in Holm. Instead, each p-value is evaluated individually against its corresponding critical value. The critical value or

significance level for each ranked p-value is determined by dividing the FDR by the following result: number of endpoints plus 1 minus the order rank of the p-value. Consider the earlier example with 3 endpoints: A; B; and C; which have corresponding p-values: pA; pB; and pC. Let the smallest-to-largest order be pB with rank 1, pA with rank 2, and pC with rank 3. Assume an FDR of 0.15. Using the Benjamini-Hochberg method, pB would be evaluated at the significance level of 0.15/(3 + 1 – 1) or 0.05; pA at the significance level 0.15/2 = 0.075; and pC at the significance level 0.15. If all p-values are greater than their significance level, all would be non-significant, and all would be significant if they were all less than or equal to their significance level. Consider the following two scenarios: pB < 0.05, pA > 0.075, and pC > 0.15; and pB > 0.05, pA > 0.075, and pC < 0.15. This FDR approach returns pB only as significant and pB, pA, and pC as significant for each respective scenario. This is due to the Benjamini-Hochberg method defining significant results by identifying the largest p-value that is less than its corresponding critical value and including all p-values equal to or smaller than that largest p-value as significant, even if those p-values are not less than their corresponding critical value. The Benjamini-Hochberg method assumes the hypotheses are independent.

Implementation in SAS®: FDR

The SAS® procedure, PROC MULTTEST, can also be used to implement the FDR controlling method mentioned above. More information about the procedure can be found at: https://documentation.sas.com/doc/pl/pgmsascdc/v_037/statug/statug_multtest_toc.htm.

Conclusion

Multiplicity can be a real concern for inflating the type I error rate when multiple statistical hypotheses are being evaluated simultaneously. Fortunately, there are several methods and approaches available to account for the potential type I inflation, whether controlling the FWER or the FDR, and whether the situation calls for an option that is simple or complex. Although all but one (minP) of the methods covered in this paper require an assumption that the hypotheses are independent, multiple approaches are available for situations when that assumption may not be valid. As this paper demonstrates, there is great flexibility in how to address multiplicity; however, understanding the trial, the data, and the hypotheses to be evaluated a priori is key to determining if multiplicity is a concern that needs to be addressed or not, as well as identifying appropriate ways to address it if needed. Please choose wisely!

References

Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017 Jun;9(6):1725-1729.

European Agency for the Evaluation of Medicinal Products. (2002, September). Points to consider on multiplicity issues in clinical trials. https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials-scientific-guideline

Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesth. 2018;71(5):353-360.

SAS/STAT® 15.2 User's Guide. SAS Institute Inc., Cary, NC.

Vickerstaff, V., Omar, R. & Ambler, G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. BMC Med Res Methodol. 2019;19(129):1-13.

U.S. Food and Drug Administration. (2022, October). Multiple endpoints in clinical trials guidance for industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials

Acknowledgments

The authors would like to acknowledge and thank Cytel Leadership for affording the opportunity and time to complete this work, as well as the Cytel Methodology Workstream Committee who helped develop this work. The authors would also like to thank PHUSE 2026 Leadership for allowing this work to be published and presented. Last, but certainly not least, thank you to Everyone who reads this paper or attended the presentation.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Cralen Davis
Company: Cytel, Inc.
Email: cralen.davis@cytel.com