Priya Ranganathan, C. S. Pramesh[1], Marc Buyse[2,3]

Department of Anaesthesiology, Tata Memorial Centre, [1]Department of Surgical Oncology, Division of Thoracic Surgery, Tata Memorial Centre, Mumbai, Maharashtra, India, [2]International Drug Development Institute, San Francisco, California, USA, [3]Department of Biostatistics, Hasselt University, Hasselt, Belgium

Address for correspondence:
Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Ernest Borges Road, Parel, Mumbai - 400 012, Maharashtra, India.
E-mail: drpriyaranganathan@gmail.com

Common pitfalls in statistical analysis: The perils of multiple testing

Abstract

Multiple testing refers to situations where a dataset is subjected to statistical testing multiple times - either at multiple time-points or through multiple subgroups or for multiple end-points. This amplifies the probability of a false-positive finding. In this article, we look at the consequences of multiple testing and explore various methods to deal with this issue.

Key words: Biostatistics, data interpretation, multiplicity, statistical significance

INTRODUCTION

In a previous article, we discussed the alpha error rate (or false-positive error rate), which is the probability of falsely rejecting the null hypothesis.[1] In any study, when two or more groups are compared, there is always a chance of finding a difference between them just by chance. This is known as a Type 1 error, in contrast to a Type 2 error, which consists of failing to detect a difference that truly exists. Conventionally, the alpha error is set at 5% or less which ensures that when we do find a difference between the groups, we can be at least 95% confident that this is a true difference and not a chance finding.

The 5% limit for alpha, known as the significance level of the study, is set for a single comparison between groups. When we compare treatment groups multiple times, the probability of finding a difference just by chance increases depending on the number of times, we perform the comparison. In many clinical trials, a number of interim analyses are planned to occur during the course of the trial, with the final analysis taking place when all patients have been accrued and followed up for a minimum period. If all these interim (and final) analyses were performed at the 5% significance level, the overall probability of a Type 1 error would exceed the prespecified limit of 5%. It can be calculated that if two groups are compared 5 times, the probability of a false positive finding is as high as 23%; if they are compared 20 times, the probability of finding a significant difference just by chance increases to 64%.[2,3] Fortunately, much statistical research has been devoted to

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/2229-3485.179436

How to cite this article: Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: The perils of multiple testing. Perspect Clin Res 2016;7:106-7.

this problem, and "group sequential designs" have been proposed to control the Type 1 error rate when the data of a trial need to be analyzed multiple times.

Another, more challenging type, of multiple testing occurs when authors try to salvage a negative study. If the primary endpoint does not show statistical significance, looking at multiple other (less important) comparisons quite often produce a "positive" result, especially if there are many such comparisons. Investigators can try to analyze different endpoints, among different subsets of patients, using different statistical tests, and so on, so the opportunity for multiplicity can be substantial.[2,4] One case in point is subset analyses when the treatments are compared among subsets of patients defined using prognostic features such as their gender, age, tumor location, stage, histology, and grade If there were only three such binary factors, $8 = 2^3$ subsets could be formed. If we were to compare the treatments among these 8 subsets, we would have one chance in three (33% probability) to observe a statistically significant ($P <= 0.05$) treatment effect in one of them even if there was no true difference between the treatments. Worse still, if there was an overall statistically significant benefit ($P <= 0.05$) in favor of one of treatments, we would have a nine in ten chance (90% probability) to observe a benefit in favor of the other treatment in one of the subsets!

It is to avoid these serious problems that all intended comparisons should be fully prespecified in the research protocol, with appropriate adjustments for multiple testing. However, for retrospective studies, it is difficult to ascertain with certainty whether the analyses performed were actually thought of when the research idea was conceived or whether the performed analyses were mere data dredging.

HOW ARE ADJUSTMENTS MADE FOR MULTIPLE TESTING?

Two main techniques have been described for controlling the overall alpha error:

1. The family–wise error rate: This approach attempts to control the overall false-positive rate for all comparisons. "Family" is defined as a set of tests related to the same hypothesis.[2] Various approaches for correcting the alpha error include the Bonferroni, Tukey, Hochberg and Holm's step-down methods. The Bonferroni correction consists of simply dividing the overall alpha level by the number of comparisons. For example, if 20 comparisons are being made, then the alpha level for significance for each comparison would be 0.05/20 = 0.0025. However, while this is simple to do (and understand), it has been criticized as being far too conservative, especially when the various tests being performed are highly correlated[3]
2. The false discovery rate: This approach attempts to control the fraction of "false significant results" among the significant results only. The Benjamini and Hochberg procedure has been described for this approach.[5]

IS ADJUSTMENT OR COMMON SENSE NEEDED FOR MULTIPLE TESTING?

Many statisticians feel that alpha-adjustment for multiple comparisons reduces the significance value to very stringent levels and increases the chances of a Type 2 error (false negative error; falsely accepting the null hypothesis).[2] It has also been argued that an obsessive reliance on alpha-adjustment may be counterproductive.[6]

The following simple strategies have been suggested to handle multiple comparisons:[2,7]

- Readers should evaluate the quality of the study and the actual effect size instead of focusing only on statistical significance
- Results from single studies should not be used to make treatment decisions; instead, one should look for scientific plausibility and supporting data from other studies which can validate the results of the original study
- Authors should try to limit comparisons between groups and identify a single primary endpoint; using a composite endpoint or global assessment tool is also an acceptable alternative to using multiple endpoints.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: "P" values, statistical significance and confidence intervals. Perspect Clin Res 2015;6:116-7.
2. Drachman D. Adjusting for multiple comparisons. J Clin Res Best Pract 2012;8:1-3. Available from: https://www.firstclinical.com/journal/2012/1207_Multiple.pdf. [Last cited on 2016 Mar 21].
3. Goldman M. Why is multiple testing a problem? 2008. Available from: http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf. [Last cited on 2016 Mar 21].
4. Cabral HJ. Multiple comparisons procedures. Circulation 2008;117:698-701.
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B 1995;57:289-300.
6. Buyse M, Hurvitz SA, Andre F, Jiang Z, Burris HA, Toi M, *et al.* Statistical controversies in clinical research: Statistical significance-too much of a good thing. Ann Oncol 2016. pii: Mdw047. [Epub ahead of print].
7. Feise RJ. Do multiple outcome measures require p value adjustment? BMC Med Res Methodol 2002;2:8.