BMC Medical Research Methodology

Debate

Do multiple outcome measures require p-value adjustment?

Ronald J Feise

Address: Institute of Evidence-Based Chiropractic 6252 Rookery Road, Fort Collins, Colorado 80528

E-mail: rjf@chiroevidence.com

Published: 17 June 2002

BMC Medical Research Methodology 2002, **2**:8

Received: 15 March 2002
Accepted: 17 June 2002

This article is available from: http://www.biomedcentral.com/1471-2288/2/8

Abstract

Background: Readers may question the interpretation of findings in clinical trials when multiple outcome measures are used without adjustment of the p-value. This question arises because of the increased risk of Type I errors (findings of false "significance") when multiple simultaneous hypotheses are tested at set p-values. The primary aim of this study was to estimate the need to make appropriate p-value adjustments in clinical trials to compensate for a possible increased risk in committing Type I errors when multiple outcome measures are used.

Discussion: The classicists believe that the chance of finding at least one test statistically significant due to chance and incorrectly declaring a difference increases as the number of comparisons increases. The rationalists have the following objections to that theory: 1) P-value adjustments are calculated based on how many tests are to be considered, and that number has been defined arbitrarily and variably; 2) P-value adjustments reduce the chance of making type I errors, but they increase the chance of making type II errors or needing to increase the sample size.

Summary: Readers should balance a study's statistical significance with the magnitude of effect, the quality of the study and with findings from other studies. Researchers facing multiple outcome measures might want to either select a primary outcome measure or use a global assessment measure, rather than adjusting the p-value.

Background

Clinical trials often require a number of outcomes to be calculated and a number of hypotheses to be tested. Such testing involves comparing treatments using multiple outcome measures (MOMs) with univariate statistical methods. Studies with MOMs occur frequently within medical research [1]. Some researchers recommend adjusting the p-values when clinical trials use MOMs so as to prevent the findings from falsely claiming "statistical significance" [2]. Other researchers have disagreed with this strategy, because it is inappropriate and may cause incorrect conclusions from the study [3]. The examination of this issue is important to both researchers and readers. Researchers are concerned about p-values and their effect upon power and sample size. Both readers and researchers are concerned about accepting erroneous studies and rejecting beneficial interventions. The primary aim of this study was to evaluate the need to adjust p-values in clinical trials when MOMs are used.

Discussion

Classical view

Classicists believe that if multiple measures are tested in a given study, the p-value should be adjusted upward to reduce the chance of incorrectly declaring a statistical significance [4–7]. This view is based on the theory that if you test long enough, you will inevitably find something statistically significant – false-positives due to random varia-

bility, even if no real effects exist [4–7]. This has been called the multiple testing problem or the problem of multiplicity [8].

Adjustments to p-value are founded on the following logic: If a null hypothesis is true, a significant difference may still be observed by chance. Rarely can you have absolute proof as to which of the two hypotheses (null or alternative) is true, because you are only looking at a sample, not the whole population. Thus, you must estimate the sampling error. The chance to incorrectly declare an effect because of random error in the sample is called type I error. Standard scientific practice, which is entirely arbitrary, commonly establishes a cutoff point to distinguish statistical significance from non-significance at 0.05. By definition, this means that one test in 20 will appear to be significant when it is really coincidental. When more than one test is used, the chance of finding at least one test statistically significant due to chance and incorrectly declaring a difference increases. When 10 statistically independent tests are performed, the chance of at least one test being significant is no longer 0.05, but 0.40. To accommodate for this, the p-value of each individual test is adjusted upward to ensure that the overall risk or family-wise error rate for all tests remains 0.05. Thus, even if more than one test is done, the risk of finding a difference incorrectly significant continues to be 0.05, or one in twenty [4–7].

Those who advocate multiple comparison adjustments argue that the control for false-positives is imperative, and any study that collects information on a large number of outcomes has a high probability of producing a wild goose chase and thereby consuming resources. Thus, the main benefit of adjusting p-value is the weeding out of false positives [4–7,9]. Although Bonferroni is the classical method of adjusting p-value, it is often considered to be overly conservative. A variety of alternative methods have been developed, but no gold standard method exists [10–21].

Original intent

An examination of the need for p-value adjustments should begin by asking why adjustments for MOMs were developed in the first place. Neyman and Pearson's original statistical test theory in the 1920s was a theory of multiple tests, and it was used to aid decisions in repetitive industrial circumstances, not to appraise evidence in studies [22,23]. Neyman and Pearson were solving problems surrounding rates of defective materials and rejection of lots where there were multiple samples within each lot – a situation which clearly does require a p-value adjustment.

Rational analysis

The opponents of p-value adjustments raise several practical objections. One objection to p-value adjustments is that the significance of each test will be interpreted according to how many outcome measures are considered in the family-wise hypothesis, which has been defined ambiguously, arbitrarily and inconsistently by its advocates. Hochberg and Tamhane define family-wise error rate as any collection of inferences, including potential inferences, for which it is meaningful to take into account some combined measure of errors [17]. It is unclear how wide the operative term "family" should be. Thus, the use of a finite number of comparisons is problematic. Does "family" include tests that were performed, but not published? Does it include a meta-analysis upon those tests? Should future papers on the same data set be accounted for in the first publication? Should each researcher have a career-wise adjusted p-value, or should there be a discipline-wise adjusted p-value? Should we publish an issue-wise adjusted p-value and a year-end-journal-wise adjusted p-value? Should our studies examine only one association at a time, thereby wasting valuable resources? No statistical theory provides answers for these practical issues, because it is impossible to formally account for an infinite number of potential inferences [23–26].

An additional objection to p-value adjustments is that if you reduce the chance of making a type I error, you increase the chance of making a type II error [23,24,27,28]. Type II errors can be no less important than type I errors, and by reducing for individual tests the chance of type I errors (the chance of introducing ineffective treatments), you increase the chance of type II errors (the chance that effective treatments are not discovered). Thus, the consequences of both Type I and Type II errors need to be considered, and the relation between them established on the basis of their severity. Additionally, if you lower the alpha level and maintain the beta level in the design phase of a study, you will need to increase the sample size, thereby increasing the financial burden of the study.

The debate over the need for p-value adjustments focuses upon our ability to make distinctions between different results – to judge the quality of science. Obviously, no scientist wants coincidence to determine the efficacy of an intervention. But MOMs have produced a tension between reason and the classical technology of statistical testing [29,30]. The issue cannot be sidestepped by using confidence intervals (which are preferred by most major medical journals), because it applies equally to statistical testing and confidence intervals. Moreover, the use of multivariate tests in place of univariate tests does not solve the dilemma, because multivariate tests present their own shortfalls, including interpretation problems (if there is a difference between experimental groups, multi-

variate tests do not tell us which variable might differ as a result of treatment, and univariate testing may still be needed). Thus, we need to confront the uncomfortable and subjective nature of the most critical scientific activity – assessing the quality of our findings. Ideally, we should be able to recognize the well-grounded and dismiss the contrived. But we might have to admit that there is no one correct or absolute way to do this.

Conscientious readers of research should consider whether a given study needs to be statistically analyzed at all. We must be careful to focus not only upon statistical significance (adjusted or not), but also upon the quality of the research within the study and the magnitude of improvement. Effect size and the quality of the research are as important as significance testing! Does it really matter whether there is a statistical difference between two treatments if the difference is not clinically worthwhile or if the research is marred by bias?

An astute reader of research knows that statistical significance is a statistical statement of how likely or unlikely it is that an outcome has occurred by chance. If a p-value is .05, there is a rather large chance (1/20) that the finding is in doubt. However, if a p-value is .0001, the chance of error is significantly less (1/10000).

Multiple comparisons strategies

To date, the issues that separate these two statistical camps remain unresolved. Moreover, other strategies may be used in lieu of p-value adjustment. Some authors have suggested the use of a composite endpoint or global assessment measure consisting of a combination of endpoints [31–34]. For example, in chronic fatigue syndrome there are multiple manifestations that tend to affect different people differently. Because no manifestation dominates, there is no way to select a primary endpoint. Use of a composite endpoint provides efficacy of "nonspecific" benefits and is valuable in testing multiple endpoints that are suitable for combining.

Zhang has advocated the selection of a primary endpoint and several secondary endpoints as a possible method to maintain the overall type I error rate [34]. For example, in chronic low back pain, although there are numerous measurements that can be used, a researcher might focus the study on symptoms while using a pain instrument as the key outcome and other measures (such as function, cost, patient satisfaction, etc.) as secondary outcomes. Even though selecting a single endpoint is not always easy because of the multifarious sphere of conditions, it is a practical approach. The selection of a primary outcome measure or composite endpoint is also necessary in the planning stages of any experimental trial to estimate the study's power and sample size. Additionally, ethical review boards, funding agencies and journals need a rationale for handling the statistical conundrum of MOMs. The selection of a primary outcome measure or a composite endpoint provides such a rationale.

Reader strategies

The following strategies should enable the reader to reach a reasonable conclusion, regardless of p-value adjustments [23,25,27,28,35–39]:

1. Evaluate the quality of the of the study and the amplitude (effect size) of the finding before interpreting statistical significance.

2. Regard all findings as tentative until they are corroborated. A single study is most often not conclusive, no matter how statistically significant its findings. Each test should be considered in the context of all the data before reaching conclusions, and perhaps the only place where "significance" should be declared is in systematic reviews. Beware of serendipitous findings of fishing expeditions or biologically implausible theories.

Author strategies

The following strategies are for the consideration of the author-researcher when faced with MOMs [31–34]:

1. Select a primary endpoint or global assessment measure, as appropriate.

2. Communicate to your readers the roles of both Type I and Type II errors and their potential consequences.

Summary

Statistical analysis is an important tool in clinical research. Disagreements over the use of various approaches should not cause us to waver from our aim to produce valid and reliable research findings. There are no "royal" roads to good research [40], because in science we are never absolutely sure of anything.

Competing interests

None declared

Acknowledgments

I gratefully acknowledge Doug Garant, PhD. for his helpful comments on the manuscript.

References

1. Godfrey K: **Statistics in practice. Comparing the means of several groups.** *N Engl J Med* 1985, **313:**1450-1456
2. Feise RJ: **Behavioral-graded activity compared with usual care after first-time disk surgery: Considerations of the design of a randomized clinical trial (Letter).** *J Manipulative Physiol Ther* 2001, **24:**67-68
3. Ostelo RW, de Vet HC: **Behavioral-graded activity compared with usual care after first-time disk surgery: Considerations of the design of a randomized clinical trial (Letter).** *J Manipulative Physiol Ther* 2001, **24:**68

4. Tukey JW: **Some thoughts on clinical trials, especially problems of multiplicity.** *Science* 1977, **198:**679-684
5. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method.** *BMJ* 1995, **310:**170
6. Greenhalgh T: **Statistics for the non-statistician. I. Different types of data need different statistical tests.** *BMJ* 1997, **315:**364-366
7. Ludbrook J: **Multiple comparison procedures updated.** *Clin Exp Pharmacol Physiol* 1998, **25:**1032-1037
8. Ahlbom A: **Biostatistics for Epidemiologists.** *Boca Raton (FL), Lewis Publishers* 1993, 52-53
9. Steenland K, Bray I, Greenland S, Boffetta P: **Empirical bayes adjustments for multiple results in hypothesis-generating or surveillance studies.** *Cancer Epidemiol Biomarkers Prev* 2000, **9:**895-903
10. Sidak Z: **Rectangular confidence regions for the means of multivariate normal distribution.** *J Am Statist Assoc* 1967, **62:**626-633
11. Williams DA: **A test for differences between treatment means when several dose levels are compared with a zero dose control.** *Biometrics* 1971, **27:**103-117
12. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Statis* 1979, **6:**65-70
13. Mantel N: **Assessing laboratory evidence for neoplastic activity.** *Biometrics* 1980, **36:**381-399
14. Stoline MR: **The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs.** *Am Stat* 1981, **35:**134-141
15. Tukey JW, Ciminera JL, Heyse JF: **Testing the statistical certainty of a response to increasing doses of a drug.** *Biometrics* 1985, **41:**295-301
16. Shaffer JP: **Modified sequentially rejective multiple test procedures.** *J Amer Stat Assn* 1986, **81:**826-831
17. Hochberg Y, Tamhane AC: **Multiple comparison procedures.** *New York, John Wiley* 1987
18. Hommel G: **A stepwise rejective multiple test procedure based on a modified Bonferroni test.** *Biometrika* 1988, **75:**383-386
19. Westfall PH, Young SS: **p-Value adjustments for multiple tests in multivariate binomial models.** *J Amer Stat Assn* 1989, **84:**780-786
20. Tarone RE: **A modified Bonferroni method for discrete data.** *Biometrics* 1990, **46:**515-522
21. Turkheimer F, Pettigrew K, Sokoloff L, Smith CB, Schmidt K: **Selection of an adaptive test statistic for use with multiple comparison analyses of neuroimaging data.** *Neuroimage* 2000, **12:**219-229
22. Neyman J, Pearson ES: **On the use and interpretation of certain test criteria for purposes of statistical inference.** *Biometrika* 1928, **20A:**175-240
23. Perneger TV: **What's wrong with Bonferroni adjustments.** *BMJ* 1998, **316:**1236-1238
24. Rothman KJ: **No adjustments are needed for multiple comparisons.** *Epidemiology* 1990, **1:**43-46
25. Savitz DA, Olshan AF: **Multiple comparisons and related issues in the interpretation of epidemiologic data.** *Am J Epidemiol* 1995, **142:**904-908
26. Thompson JR: **Invited commentary: Re: "Multiple comparisons and related issues in the interpretation of epidemiologic data".** *Am J Epidemiol* 1998, **147:**801-806
27. Cole P: **The evolving case-control study.** *J Chronic Dis* 1979, **32:**15-27
28. Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG: **The problem of multiple inference in studies designed to generate hypotheses.** *Am J Epidemiol* 1985, **122:**1080-1095
29. Aickin M, Gensler H: **Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods.** *Am J Public Health* 1996, **86:**726-728
30. Manor O, Peritz E: **Re: "Multiple comparisons and related issues in the interpretation of epidemiologic data".** *Am J Epidemiol* 1997, **145:**84-85
31. O'Brien PC: **Procedures for comparing samples with multiple endpoints.** *Biometrics* 1984, **40:**1079-1087
32. Simes RJ: **An improved Bonferroni procedure for multiple tests of significance.** *Biometrika* 1988, **73:**751-754
33. Goldsmith CH, Smythe HA, Helewa A: **Interpretation and power of a pooled index.** *J Rheumatol* 1993, **20:**575-578
34. Zhang J, Quan H, Ng J, Stepanavage ME: **Some statistical methods for multiple endpoints in clinical trials.** *Control Clin Trials* 1997, **18:**204-221
35. Walker AM: **Reporting the results of epidemiological studies.** *Am J Public Health* 1986, **76:**556-558
36. deGruy F: **Significance of multiple inferential tests.** *J Fam Pract* 1990, **30:**15-16
37. Hart AA: **The interpretation of multiple P-values.** *Radiother Oncol* 1994, **33:**177-178
38. Voss S, George S: **Multiple significance tests.** *BMJ* 1995, **310:**1073
39. Goodman SN: **Multiple comparisons, explained.** *Am J Epidemiol* 1998, **147:**807-815
40. Small RD, Schor SS: **Bayesian and non-Bayesian methods of inference.** *Ann Intern Med* 1983, **99:**857-859

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/2/8/prepub