Facts and Fictions About Handling Multiple Comparisons

Gail M. Sullivan, MD, MPH
Richard S. Feinn, PhD

Most papers submitted to this journal are quantitative in nature: that is, they ask *how much* or compare different groups through numbers. Despite how common quantitative methods are—in outcomes-type research and everyday life—there are aspects of manipulating numbers that educators may have forgotten since their long-ago (or never-taken) statistics classes. One aspect concerns analyses using many comparisons. Educators and researchers who do not take into account multiple independent comparisons may receive reviewer comments such as: Where did you prespecify how many comparisons you planned to make? How did you adjust for these multiple comparisons? or How do the multiple comparisons affect your statistical inferences? Not considering multiple comparisons can raise questions of internal validity (ie, are these findings actually true?). It can also lower, in the minds of reviewers and authors, confidence in the authors: *Do these authors know what the heck they're doing?* As clinicians and educators we may be less aware of these issues and how they can doom a study or paper if not handled transparently and well.

When considering a quantitative paper, 3 questions immediately arise: (1) Does this paper apply to my setting or trainees (external validity, generalizability); (2) Are the findings likely due to chance or true for the overall population being studied (false vs true positive finding); and (3) How large or meaningful are the findings (effect size).[1] This editorial provides a brief introduction to the second issue, the holy grail for many authors: a significant *P* level.

Back to Basics

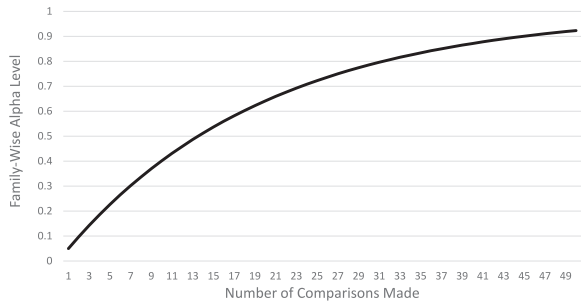
Why do we cherish *P* levels? Let's start with a single comparison, comparing 2 means. Suppose a group of internal medicine residents took an expensive board examination prep course and a similar group of residents did not, and we want to compare board score means between groups to determine if the course should be continued. If we assume that the null hypothesis is true (ie, there is no difference between the groups), the *P* value is the probability that our selection of residents—a random sample of *all* residents—produced a difference in the 2 board score means of at least the size found.

Type I error (alpha) is the error level deemed reasonable by the research team, who must select it before conducting the statistical test. It is the probability of committing a false positive error: in other words, of concluding that a difference between groups exists when there is truly no difference. If the *P* level from the statistical test is less than the selected error level, usually 5% (.05), we view the test difference as having only a 5% chance that the difference found is due to the selection of residents (as we cannot study the entire population) rather than the board prep course (ie, a 5% chance that the test score difference is due to the residents selected for our study, ie, by *chance alone*).

But what if we wish to look at additional factors that might be important to understanding who should be targeted for this expensive board prep course? For example: in-training examination scores, resident age and gender, US medical graduate vs international medical graduate, Milestones ratings during residency—or preferred breakfast drink? These issues can occur when we order many lab tests for a patient, too. If the alpha or type I error level remains at .05 for each comparison, the probability of at least one finding being "statistically significant" increases above 5% (see FIGURE). For example, for just 10 comparisons, the probability rises to 40% that you will find at least one "statistically significant" ($P < .05$) comparison that is due to the population of residents randomly selected rather than the factor under examination—that is, by chance. For 13 independent comparisons, the probability of finding a "significant" *P* level by chance increases to 50%.[2] These are called family-wise error rates, for a family of comparisons. You could erroneously conclude that residents who drink tea for breakfast are the best target for taking this board prep course—and create tortured explanations for this finding in your Discussion section.

Remember that statistical significance is determined by the level of error accepted (alpha or type I error) and reflects the likelihood that the sampled

DOI: http://dx.doi.org/10.4300/JGME-D-21-00599.1



FIGURE

Probability of at Least One Significant Comparison Found, With Increasing Number of Comparisons[3]

Note: This equation assumes that the comparisons are all independent: the chance of any 1 comparison having a significant *P* value is unrelated to the chance of another comparison having a significant *P* value. $y = 1-(1-0.05)^x$

population resembles the entire population (eg, that internal medicine residents in 2020–2021 at several institutions resemble *all* internal medicine residents). Note that this issue of multiple comparisons also pertains to 95% confidence intervals. If multiple comparisons are performed and a 95% confidence interval of the difference in means is created for each comparison, the probability that all the intervals will contain the true difference in means will be *less* than 95%.

As an extreme example, imagine if researchers conducting genome-wide association studies (GWAS) did not adjust for multiple testing. GWAS may test 100 000 different loci for an association with a disease. If an alpha level of .05 was used for each locus you can guarantee there would be numerous false positives.

Fishing Expeditions and P-Hacking

The terms *fishing expedition* or *P-hacking* refer to when researchers examine their data for every possible comparison of independent variables (eg, numerous demographic factors, postgraduate year levels, specialties, undergraduate locations, residency rotations) and/or dependent variables (eg, well-being index, burnout index, burnout subgroup elements, work-life balance index). The more comparisons, the more likely a P level of $< .05$ will be found for a comparison, and the null hypothesis (ie, no difference) may be rejected inaccurately. These terms are generally pejorative and reserved for when only the significant findings are reported, for example in the Abstract or Results section of a paper.

This problem may be inevitable when exploring entirely new questions with no expectation of where the interesting findings may lie. However, in medical education this is rarely true; we usually have hypotheses based on prior work or plausible theory. To avoid the appearance of "fishing," it is best to prespecify, based on the literature and theoretical framework for your approach, your planned comparisons in the Methods section. This fishing problem was found often enough in clinical trials that it is now mandatory for researchers to post the primary outcome(s) on a public site (clinicaltrials.gov) before the data are collected and analyzed. Resist the temptation to add additional analyses *after* you have seen the data!

When there are no plausible prior hypotheses, it can be acceptable to make many comparisons, report *all* of them with the associated P levels and/or confidence intervals, and state in your Methods section that these were exploratory hypotheses and that no adjustment for multiple comparisons was made for this reason. Be cautious in drawing inferences in these situations: as the number of tests expands, so does the family-wise error rate.

Why to Limit Comparisons and Pre-Plan Analyses

In preparing for a project, the first step is a deep dive into the literature: What methods did other researchers use? What theories may support different approaches? What gaps remain in our knowledge? Often prior work will provide you with specific directions or questions as next steps. This in turn will help you limit the collection of data as well as planned analyses of the data. If data were already collected (eg, Accreditation Council for Graduate Medical Education and national program director groups have enormous data collections to explore), choose carefully what you need to answer your question(s).

Here's the conundrum: If you don't correct for multiple comparisons, you risk finding "significant" results that are false positives and that will not be found by others in replication studies. If you do correct for multiple comparisons, you lose statistical power to find differences that actually exist (false negatives). Ergo, limit your comparisons to what fits your questions best.

Correction for a comparison may not be needed in some instances. For example, consider that you are looking at the effects of a new experiential orientation week on intern performance on aggregated professionalism milestones at 6 months, in current US psychiatry interns. Half of the interns receive the new week-long experiential orientation, and the other half receive a combination of large group and virtual orientation sessions. Those with the experiential orientation score significantly (and meaningfully) higher at the $P < .05$ level. You plan secondary analyses to look at subgroups: international medical

graduates vs US medical graduates, male vs female, older (> 30) vs younger (< 31 years), USMLE Step 1 quintile, and those at university-based vs non-university-based programs. In this example, it is not necessary to correct for the primary analysis, although the secondary analyses may require adjustment for multiple comparisons.

Strategies for Handling Multiple Comparisons

After considering the most important comparisons you plan to make, if you have more than a few, you should consider adjusting your analysis to reflect the multiple comparisons. (Remember that if you look at your data *before* deciding what comparisons to make, *you have already made multiple comparisons*. We suggest not to do this unless you are performing truly exploratory research.)

There are many methods to consider, and full texts as well as numerous articles describe them well. Of these, the Bonferroni correction is often used in medical education. The Bonferroni correction adjusts the alpha level (error) downward by dividing alpha by the planned number of comparisons. For 10 comparisons, with a type I error of 0.05, the corrected alpha level is .05/10 or .005. This is sometimes termed the *comparison-wise* error rate. The Bonferroni correction is easy to remember and thus popular, but it is overly conservative, especially if the associations are not in fact independent of each other. Thus, it can lead to a type II error (falsely accepting the null hypothesis of no association). There are many modifications of this general approach; some include using a less conservative adjustment (eg, Benjamini–Hochberg method), varying the alpha level for primary and secondary hypotheses, or switching to a lower alpha level for all tests (eg, .01 instead of .05).

But what if the various comparisons we want to make are not independent of each other? Or what if we are making a large number of comparisons, such as 25? There are methods for when independent and/or dependent variables are correlated and situations where numerous tests are performed. While beyond the scope of this introductory article, there are many good resources for readers to learn more about multiple comparisons and the various approaches that can support your methods (see Resources).

How to Discuss in Limitations

As you have seen, decisions must be made before examining your data—optimally before even collecting your data—that will inevitably affect the "truth" of your findings. Clearly present your reasoning in choice of comparisons and alpha error levels in the Methods section. Then, in the Discussion section, consider how your decisions may have affected your findings in either direction: false positives (differences observed that are actually due to chance) or false negatives (no difference found when one does exist). This latter problem more often occurs as a result of a type II error (beta), which we will save for another discussion. Laying out the potential effects of your methods' decisions in a transparent way enhances credibility in the eyes of reviewers, editors, and readers, and does not have to be lengthy. It's better to have "too much" transparency vs "too little," and any excess words can be trimmed away in the revision process.

Conclusions

This article barely scratches the surface of the topic of multiple comparisons in medical education research. We hope to raise awareness so that educators and researchers keep this issue in mind when reading articles, considering analyses, and writing up their work for presentations or publications. Most important:

1. Preplan your comparisons at the start. If you have not, but have examined the data before deciding which analyses to make, consider these post-hoc analyses as *all possible comparisons*.
2. Decide if your comparisons are likely independent of each other or if some may be related to each other.[4]
3. Consider adjusting your alpha level (error) for more than a few comparisons.
4. Present your decisions clearly in the Methods section.
5. Discuss how your methods may have affected your findings in the Discussion.
6. When in doubt, ask a friendly biostatistician.

Let us know if this article is helpful and whether you would like more *JGME* papers on this or related topics at www.jgme.org and on Twitter @JournalofGME.

References

1. Sullivan GM, Feinn R. Using effect size—or why the *P* value is not enough. *J Grad Med Educ*. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1

2. GraphPad. HJ The multiple comparison problem. https://www.graphpad.com/guides/prism/latest/statistics/beware_of_multiple_comparisons.htm Accessed May 21, 2021.
3. UC Berkely. Spring 2008—Stat C141/Bioeng C141—Statistics for Bioinformatics. https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf. Accessed May 21, 2021.
4. Derringer, Jaime. A Simple Correction for Non-Independent Tests. https://psyarxiv.com/f2tyw/. Accessed May 21, 2021.

Resources

- Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis*. 2017;9(6):1725–1729. doi:10.21037/jtd.2017.05.34
- Mutolsky, H. *Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 4th ed. Oxford, UK: Oxford University Press; 2017:203–223.
- TTB. Bonferroni Correction Method Explained. https://toptipbio.com/bonferroni-correction-method/. Accessed May 21, 2021.

Gail M. Sullivan, MD, MPH, is Editor-in-Chief, *Journal of Graduate Medical Education* (*JGME*), and Associate Director for Education, Center on Aging, and Professor of Medicine, University of Connecticut Health Center; and **Richard S. Feinn, PhD,** is Statistical Editor, *JGME*, and Associate Professor of Medical Sciences, Quinnipiac University.

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut Health Center, gsullivan@uchc.edu, Twitter @DrMedEd_itor