Limitations of Significance Testing in Clinical Research: A Review of Multiple Comparison Corrections and Effect Size Calculations with Correlated Measures

Terrie Vasilopoulos, PhD,* Timothy E. Morey, MD,* Ketan Dhatariya, MD, FRCP,† and Mark J. Rice, MD‡

Modern clinical research commonly uses complex designs with multiple related outcomes, including repeated-measures designs. While multiple comparison corrections and effect size calculations are needed to more accurately assess an intervention's significance and impact, understanding the limitations of these methods in the case of dependency and correlation is important. In this review, we outline methods for multiple comparison corrections and effect size calculations and considerations in cases of correlation and summarize relevant simulation studies to illustrate these concepts. (Anesth Analg 2016;122:825–30)

The appropriate use of multiple comparison corrections is an issue that is important and almost continually discussed across medicine.[1] We ask, first, why is correcting for multiple comparisons an important issue? To answer this, it is necessary to review the meaning of significance thresholds (α) and *P* values in relation to hypothesis testing. Researchers publishing in medical journals commonly set the significance threshold for their analyses at 5%. This metric is based on the probability of incorrectly rejecting the null hypothesis, also known as type 1 error (false positive). For a single statistical test, if $P < 0.05$, the null is rejected. For example, a researcher observes a difference between 2 means, and the test of this difference results in a *P* value <0.05. Provided assumptions hold, this means that there is a <5% chance that this mean difference would be observed if the null hypothesis were actually true. In other words, by setting this threshold to 5%, researchers accept a 5% chance that they will falsely conclude there is an effect. Conversely, type 2 error (β) refers to failing to reject the null when there is actually an effect (false negative; Table 1). The commonly used α level of 0.05 was first introduced by R. A. Fisher; however, its utility remains frequently questioned (discussed in later sections).[2]

Second, we ask, when are multiple comparison corrections needed? There is ongoing debate as to when and how to implement corrections for multiple comparisons.[3–5] The classical perspective posits that for any instance of repeated testing within a sample, the α (e.g., 0.05) or the *P* values themselves must be adjusted to reduce the probability of type 1 error.[4] However, critics of multiple comparison corrections argue that there is no consensus on what is considered a comparison. For example, does this include all performed tests (even exploratory) or just the ones that are published? Would corrections apply to different articles published from the same sample? Would a researcher who has worked on the same sample for many years need to report some type of "careerwise" error?[4] Others view multiple comparison corrections as unnecessary, with multiple comparison concerns being adequately addressed through different modeling approaches.[6] Although this review is more focused on how to correct for multiple comparisons as opposed to this debate, it is still important to acknowledge the concerns of researchers about the best way to report the most accurate results possible. Despite these differing opinions, some agreement has been achieved. First, researchers should strive to reduce the number of comparisons via thoughtful selection of end points, identification of primary versus secondary end points, and creation of global/summary measures, as appropriate.[4] Next, researchers should be transparent in both the consequences of type 1 and type 2 error with regard to their sample and the rationale for their approach (or absence of) for multiple comparison corrections.[4,7] Finally, multiple comparison corrections should be strongly considered for confirmatory analyses but are less needed for exploratory analyses (e.g., hypothesis-generating analyses).[3,4]

Many commonly used controls for type 1 error specifically aim to control for family-wise error (FWER), which is the probability of at least 1 false positive occurring (equation below). For example, with a significance threshold of $\alpha = 0.05$, the FWER for 10 tests would be 0.4 or 40% (Fig. 1). In other words, the chance for there being at least 1 false positive among 10 tests performed simultaneously is 40%.

FAMILY-WISE ERROR RATE

$$\text{FWER} = 1 - (1 - \alpha)^n$$

n = number of tests performed, α = significance threshold (typically 0.05).

From the *Department of Anesthesiology, University of Florida College of Medicine, Gainesville, Florida; †Elsie Bertram Diabetes Centre, Norfolk and Norwich University Hospitals, Norwich, United Kingdom; and ‡Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, Tennessee.

Accepted for publication October 15, 2015.

Funding: None.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

Address correspondence to Mark J. Rice, MD, Department of Anesthesiology, Vanderbilt University Medical Center, 1301 Medical Center Dr., Suite 4648, The Vanderbilt Clinic, Nashville, TN 37232. Address e-mail to mark.j.rice@vanderbilt.edu.

DOI: 10.1213/ANE.0000000000001107

Table 1. Type 1 and 2 Errors in Relation to Hypothesis Testing

	Truth of null hypothesis	
Decision	**Null hypothesis true**	**Null hypothesis false**
Reject null	Type 1 error (α) False positives	Correct decision
Fail to reject null	Correct decision	Type 2 error (β) False negatives

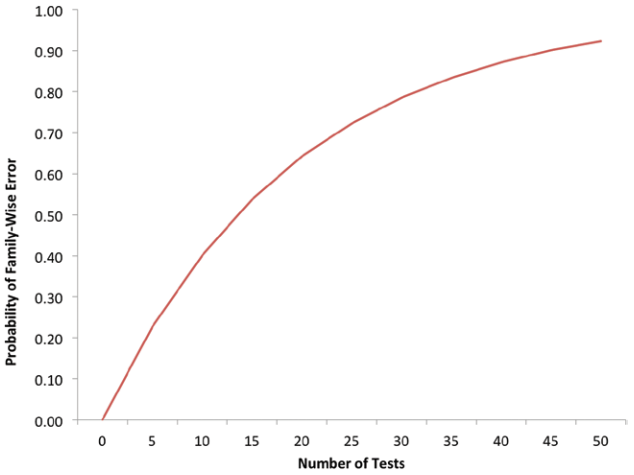

Figure 1. Probability (reported as a proportion) of family-wise error as a function of the number of tests performed, with family-wise error rate increasing with the addition of more comparisons.

The most common and simplest control of the FWER is the Bonferroni correction.[8] For this correction, the significance threshold is adjusted for the number of tests performed (Appendix 1). For example, if 10 tests are performed, the adjusted significant threshold would be from 0.05 to 0.005; thus, only tests with $P < 0.005$ would be considered statistically significant (i.e., null would be rejected).

However, the Bonferroni correction is also the most conservative and strict of the multiple testing correction approaches, and many researchers advocate alternatives under appropriate circumstances.[9,10] One popular alternative to the Bonferroni correction is the sequential (step-down) Bonferroni developed by Holm (the traditional Bonferroni correction is considered a single-step approach).[11] When using this step-down approach (also know as the Bonferroni-Holm method), *P* values of each single test are placed in rank order based increasing *P* values (Appendix 1). The smallest *P* value is compared with the standard Bonferroni-adjusted α. If it is not statistically significant in relation to the adjusted threshold, no adjustments to α are made. If it is less than the threshold, the second smallest *P* value is then compared with a significance threshold in which the original α (e.g., 0.05) is adjusted for the number of tests minus 1. This continues until no further *P* values are statistically significant according to the adjusted significance thresholds. This step-down procedure is considered less conservative and can better limit type 2 error compared with the standard Bonferroni correction; this method is commonly available in statistical software for users.[9] Both the above approaches are commonly used as upward adjustments to the α, the metric to which generated *P* values from all tests are compared. This adjusted α can also be used to calculate corrected confidence intervals (CIs). Of note, multiple comparison approaches can also be used to correct the estimated *P* values from the test performed instead of the α. In this case, the adjusted *P* values are compared with the a priori α (e.g., 0.05).

Although the Bonferroni-Holm procedure demonstrates increased power in comparison with the standard Bonferroni corrections; overall, controlling FWER is still a conservative approach to addressing the issue of type 1 error. By using strict corrections for multiple comparisons, researchers run the risk of reducing their power to detect real, existing effects (i.e., type 2 errors or false negatives, Table 1). Clinical journals have been advising researchers to move away from strict correction for multiple testing because of these (and other) concerns.[3,10,12] Alternatively, more powerful methods proposed include Bayesian methods, the use of likelihood ratios, and modified false discovery rate (FDR) procedures.[3,10,12]

FDR control is considered a less conservative approach to address false positives in contrast to FWER control methods.[10] FDR is the proportion of false positives among all rejected null hypotheses; specifically, FDR = number of false positives/(number of false positives + number of correct decisions to reject null) (Table 1). Benjamini and Hochberg[13] developed a straightforward approach to control for FDR. In this approach, *P* values are first ordered from smallest to largest, similar to the Bonferroni-Holm method. This *P* value is then compared with an adjusted threshold defined as the product of the maximum FDR threshold (typically 0.05) and the rank order of the *P* value divided by the number of tests.

Table 2 extends the example worked out by Glickman et al.[10] to compare the Bonferroni, step-down Bonferroni-Holm, and Benjamini and Hochberg FDR procedures. In this comparison, for the same set of tests, the Bonferroni and Bonferroni-Holm methods would reject the null hypothesis for 2 tests (i.e., the 2 tests were statistically significant), whereas the FDR method would reject the null hypotheses for 4 tests. Furthermore, in a simulation study comparing the approaches,[14] all 3 procedures completely controlled the number of type 1 errors across 50 tests ($\alpha = 0.05$). However, this simulation also modeled type 2 errors (false negatives), with the preset number of true alternatives to be $n = 15$. The FDR procedure resulted in fewer false negatives ($n = 10$) compared with Bonferroni and Bonferroni-Holm procedures (both $n = 14$). In other words, whereas the FDR correctly recognized 33% of true alternatives, the 2 FWER approaches only recognized 7% of true alternatives.

MULTIPLE COMPARISONS FOR CORRELATED OUTCOMES

One concern with the FWER and FDR methods is that there is an assumption that the tests are independent. These procedures may yield overly conservative adjustments in the case of dependence.[15] Modern clinical trials are increasingly more complex and often have multiple related outcomes; thus, procedures that take dependency into account should be considered.[16,17]

Resampling approaches, such as bootstrap methods, have been utilized to account for correlation in multiple comparisons.[15,18–21] Briefly, bootstrapping is used to estimate

Table 2. Comparisons Across Multiple Comparison Methods for 10 Tests

		Bonferroni		Bonferroni-Holm		False discovery rate	
Test no.	Estimated *P* value[a]	Corrected threshold[b]	Decision	Corrected threshold[b]	Decision	Corrected threshold[b]	Decision
1	0.0001	0.005	Significant	0.005	Significant	0.005	Significant
2	0.0002	0.005	Significant	0.0056	Significant	0.01	Significant
3	0.01	0.005	Not significant	0.0063	Not significant	0.015	Significant
4	0.013	0.005	Not significant	0.0071	Not significant	0.02	Significant
5	0.03	0.005	Not significant	0.0083	Not significant	0.025	Not significant
6	0.04	0.005	Not significant	0.01	Not significant	0.03	Not significant
7	0.07	0.005	Not significant	0.0125	Not significant	0.035	Not significant
8	0.15	0.005	Not significant	0.0167	Not significant	0.04	Not significant
9	0.26	0.005	Not significant	0.025	Not significant	0.045	Not significant
10	0.52	0.005	Not significant	0.05	Not significant	0.05	Not significant

[a]Estimated *P* value refers to *P* value produced by 1 of the multiple tests performed during analyses for a given study.
[b]Corrected threshold refers to the multiple comparison corrected α level, the metric to which *P* values are compared.

Table 3. Simulation (n = 1,000,000) Comparing Westfall-Young Bootstrap and Permutation *P* Value Adjustments to Those from Step-Down Bonferroni

Raw *P* value	Step-down Bonferroni-corrected *P* value	Bootstrap-corrected *P* value	Permutation-corrected *P* value
0.098	0.0982	0.0985	0.1007
0.0262	0.0525	0.0509	0.0521
0.0067	0.02	0.0187	0.0185

Westfall et al. (2011).[19]

the population distribution or a given test statistic or metric by using information from multiple random samples taken from the real sample data set.[22–24] Bootstrapping approaches have been incorporated into single-step and step-down FWER and FDR methods and implicitly account for the underlying correlational structure of data; overall, they should yield less conservative *P* value adjustments (Westfall-Young method).[19,20] However, bootstrapping methods can be criticized because of their approximate nature.Westfall et al.[19] also outline a permutation method for *P* value adjustment that yields similar results to their bootstrapping approach. One drawback to using resampling-based techniques is that they can be computationally intensive.

Follow-up studies have shown that Benjamini and Hochberg FDR is still robust when tests are positively correlated.[13] In cases where tests are negatively correlated or have a complex dependency structure, a modification to their original formula was developed.[13] This approach is less computationally intensive than the resampling approaches. Appendix 2 describes how these (and other) adjustments can be conducted in SAS.

Furthermore, it is important to understand how multiple comparison adjustments pertain to CIs because the reporting of CIs, sometimes in lieu of *P* values, is becoming more widely accepted. The limits of CIs are set by a priori α values, with the confidence limit equaling 1 – α. The most common CI of 95% corresponds to the common α level of 0.05. Thus, CIs can be adjusted using the same methods above, which decrease the α level to control for type 1 error. This adjustment would result in wider CIs, for example, α = 0.01 would correspond to 99% CIs.[25,26]

An example that highlights these potential issues arises in the recently published article from Murphy et al.[27] They report a corrected threshold of 0.01 (Bonferroni correction), which implies that the correction was applied to a group of 5 tests; however, the number of tests performed, and which tests were considered for this correction, was not clearly specified. Furthermore, due to the repeated nature of their study, the outcomes would most likely be correlated. Thus, a correction that considered correlations would have been more appropriate. This is important, considering the issue of balancing the correction of type 1 error versus inflating type 2 error. Furthermore, as noted previously in this journal, as well as others, the probability of reproducing results is as important as a significant result from a given individual study.[28–30] However, *P* values need to be rather small before achieving a satisfactory level of reproducibility in similar populations (with this journal recommending a threshold of $P < 0.0001$).[30] Thus, controlling for multiple comparisons is not only important in interpreting the results of a single study but also in evaluating how well a given study's results will predict future similar studies.

Overall, when correcting for multiple comparisons, a prudent approach would be for researchers to fully and clearly describe all testing performed in the study to justify the multiple comparison calculation used and to report all raw *P* values. If the authors believe that a flood of *P* values may exhaust readers or distract from central messages, an acceptable solution is to place the most important *P* values in the journal and then to place additional values in the journal's supplemental digital content section available online.

Simulation Studies

Table 3 illustrates the overall difference in *P* value adjustment between the common step-down Bonferroni with both step-down bootstrap and permutation approaches (Westfall-Young).[19] As noted above, these resampling techniques implicitly consider correlations. Overall, the adjusted *P* values using either the bootstrap or the permutation methods were lower than those from the step-down Bonferroni procedure.

Table 4 illustrates simulation results from Hutson[15] that demonstrate differences in corrected *P* values from the Bonferroni method and their semiparametric bootstrap approach from a simulated data set with 4 correlated variables (p1, p2, p3, and p4). While the standard Bonferroni

Table 4. Simulation Results for Semiparametric Bootstrap Approach (n = 1000) Compared with Bonferroni, Across Multiple Correlation Structures (for 4 Variables)

Pairwise correlations[a]									
ρ 12	ρ 13	ρ 14	ρ 23	ρ 24	ρ 34	Resampling-based correct factor	Bootstrap-corrected α[b]	Bonferroni-corrected α[b]	Bootstrap FWER
0	0	0	0	0	0	3.74	0.0134	0.0125	0.046
0.3	0.3	0.3	0.3	0.3	0.3	3.56	0.0140	0.0125	0.049
0.6	0.6	0.6	0.6	0.6	0.6	3.01	0.0166	0.0125	0.055
0.9	0.9	0.9	0.9	0.9	0.9	1.98	0.0253	0.0125	0.05
0.3	0.5	0.7	0.3	0.5	0.7	3.15	0.0159	0.0125	0.045
−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	2.54	0.0197	0.0125	0.059
0.5	0.5	0.5	0.5	0.5	0.5	3.24	0.0154	0.0125	0.05

Reproduced from Hutson (2004).[15]
FWER = Family-wiser error rate.
[a]All possible pairwise Pearson correlations (ρ) for each possible set of 2 variables among a 4-variable sample.
[b]Corrected α threshold is the metric to which P values are compared.

Table 5. Actual Versus Overestimated Effects Sizes Based on Correlation Between Measures

Actual effect size (Cohen's d)	r^a = 0	r = 0.2	r = 0.4	r = 0.6	r = 0.8
0.3	0.3	0.34	0.39	0.47	0.67
0.6	0.6	0.67	0.77	0.95	1.34
0.9	0.9	1.01	1.16	1.42	2.01
1.2	1.2	1.34	1.55	1.9	2.68

Abbreviated reproduction from Dunlap et al. (1996).[34]
[a]Absolute Pearson correlation.

Table 6. Simulation (n = 10,000) Comparing Uncorrected and Corrected Cohen's d Effect Sizes Across Difference Correlation Where the Actual Effect Size Is 1.0

Correlation[a] (n = 20)	Uncorrected Cohen's d[30]	Corrected Cohen's d[31]
0	1.024	1.023
0.1	1.023	1.021
0.3	1.023	1.017
0.5	1.029	1.018
0.7	1.027	1.011
0.9	1.036	1.012

Abbreviated reproduction from Dunlap et al. (1996).[34]
[a]Absolute Pearson's correlation.

formula corrects the study α (to which P values are compared) by dividing it by a correction factor equal to the number of tests (in this example n = 4, correct α = 0.05/4 = 0.0125), the approach by Hutson calculates a correction factor via a bootstrapping approach. Even in the case of no correlation, the bootstrap method is slightly less conservative than the Bonferroni method. Overall, as the correlation within the sample increases (or are negative), the bootstrap method becomes less conservative while maintaining a FWER near 0.05.

LIMITATIONS OF SIGNIFICANCE TESTING

While appropriately correcting for multiple comparisons can reduce type 1 error, researchers, reviewers, and readers should be cautious to interpret a nonstatistically significant finding as "no effect" because these 2 concepts can differ. As mentioned above, in null hypothesis testing, we set out to reject the null in support of an alternative hypothesis. However, if a test fails to reach statistical significance (i.e., a researcher fails to reject the null), it cannot be said that there is no effect or difference (i.e., the difference or effect equals zero); it only means that there was a greater probability that the difference that was observed would be observed by chance. In other words, a lack of statistical significance does not necessarily mean a lack of clinical or practical significance. Furthermore, significance testing, thus the ability or power to reject the null, is dependent on sample size and does not give any indication of the relevance of a finding.[31,32] Due to these limitations, there is increasing use of effect size metrics that can better quantify the magnitude of difference, independent of significance testing.

APPROACHES TO CALCULATING EFFECT SIZES

Effect sizes provide information regarding the magnitude and direction of an observed effect. They are also vital to meta-analyses, providing a standardized way to compare results across studies. One of the most commonly used approaches when comparing 2 groups is calculating standard mean difference, and one popular standard mean difference approach is the Cohen's d. The Cohen's d mathematically translates group differences in terms of standard deviations. For example, a Cohen's d = 0.5 means 2 groups differed by a half of a standard deviation. While Cohen outlined heuristic cutoffs for interpreting Cohen's d, with d = 0.2 (small), d = 0.5 (medium), and d = 0.8 (large), Cohen[33] cautioned that this interpretation may not be applicable for all contexts and studies. Although Cohen's d is useful in estimating the effect size of differences between 2 group means, there are other metrics that can be used for other types of comparisons (e.g., odds ratio, r, numbers needed to treat). Appendix 3 lists online resources where these and other effect size metrics can be easily calculated, as well as be converted to other metrics. Understanding these effects sizes is also important concerning the overall study design. Many programs that calculate power for studies rely on effect size metrics for their computations.

As in the aforementioned considerations with multiple comparison corrections, it is important to account for dependency in calculating effect sizes; this is especially important because of their use in meta-analyses. Many common effect size calculations can be modified to account for the correlational structure of the data. A failure to account

for correlations can inflate estimates, thus leading to overestimation of a treatment's clinical importance.[34–36]

Simulation Studies

Table 5 depicts the bias induced by correlations to the Cohen's d effect size calculation.[34] As correlations increase, the effect sizes become increasing overinflated. For very high correlations (0.8), calculated effect sizes were nearly double the actual effect. These strong correlations are not uncommon in clinical research, especially in repeated-measures studies.

Table 6 summarizes a simulation study (n = 10000) comparing the effect sizes from the standard Cohen's d formula and Cohen's d corrected for correlation.[34] In this simulation, the actual effect size is 1.0. As the correlation increases in magnitude, the uncorrected effect size becomes more overinflated, while the corrected effect size becomes even more accurate.

CONCLUSIONS

Understanding the accuracy and clinical importance of results is an important issue in clinical research. However, as we note in this review, researchers, reviewers, and readers should be mindful of the limitations of significance testing and how these limitations influence the way results are reported and interpreted. The most important advice would be to make thoughtful study design choices a priori. This includes determining the number of planned comparisons, what the primary (versus secondary) end points are and finding the clinically relevant, minimally important differences in your outcomes. These a priori decisions will guide the data analysis and interpretation, as well as limit the potential problems that come with significance testing. Furthermore, it is vital that clinical researchers understand the dependency structure of their data due to the bias induced by correlation on both corrections for multiple comparisons and effect sizes. Overall, and most importantly, it is essential that researchers use the most appropriate and sound statistical tools possible to extract meaningful and accurate information from available data so that each manuscript has its maximal clinical impact on care for our patients.

DISCLOSURES

Name: Terrie Vasilopoulos, PhD.
Contribution: This author conducted the data analysis and contributed to manuscript preparation.
Attestation: Terrie Vasilopoulos approved the final manuscript.
Name: Timothy E. Morey, MD.
Contribution: This author contributed to the design of the review and manuscript preparation.
Attestation: Timothy E. Morey approved the final manuscript.
Name: Ketan Dhatariya, MD, FRCP.
Contribution: This author contributed to the design of the review and manuscript preparation.
Attestation: Ketan Dhatariya approved the final manuscript.
Name: Mark J. Rice, MD.
Contribution: This author contributed to the design of the review and manuscript preparation.
Attestation: Mark J. Rice approved the final manuscript and is the archival author.
This manuscript was handled by: Franklin Dexter, MD, PhD.

APPENDIX 1

Formula for Common Multiple Comparison Corrections

Bonferroni correction:

$$\text{Corrected threshold} = \alpha / n$$

n = number of tests performed, α = significance threshold (typically 0.05).

Sequential Bonferroni-Holm correction:

First-order P values, with first (1st) value being the smallest P value

$$\text{Corrected threshold (1st } P \text{ value)} = \alpha / n$$

$$\text{Corrected threshold (2nd } P \text{ value)} = \alpha / (n-1)$$

$$\text{Corrected threshold (3rd } P \text{ value)} = \alpha / (n-2)$$

Continue until P values are greater than calculated threshold. n = number of tests performed, α = significance threshold (typically 0.05)

Benjamin and Hochberg False Discover Rate correction:

First-order P values, with first (1st) value being the smallest P value

$$\text{Corrected threshold (1st } P \text{ value)} = \text{maximum FDR} \times (1 / n)$$

$$\text{Corrected threshold (2nd } P \text{ value)} = \text{maximum FDR} \times (2 / n)$$

$$\text{Corrected threshold (3rd } P \text{ value)} = \text{maximum FDR} \times (3 / n)$$

Continue until P values are greater than calculated threshold. n = number of tests performed, maximum FDR is analogous to α (typically 0.05).

APPENDIX 2

SAS code for the adjustment of P values (using P values from Table 2). These adjustments include approaches that assume independence and ones that account for dependence. See SAS® for detailed explanations.

```
data mc;
input Test$ Raw_P @@;
datalines;
test01 0.0001 test02 0.0002 test03 0.01
test04 0.013 test05 0.03 test06 0.04
test07 0.07 test08 0.15 test09 0.26
test10 0.52;
proc multtest inpvalues=mc Bonferroni holm fdr dependentfdr;
run;
```

Documentation: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#multtest_toc.htm

APPENDIX 3

Some Resources for Effect Size Calculations and Interpretations

Articles:

Cumming, Geoff. "The New Statistics Why and How." *Psychological Science* 25, no. 1 (2014): 7–29.
Kraemer, Helena Chmura, and David J. Kupfer. "Size of treatment effects and their importance to clinical research and practice." *Biological psychiatry* 59, no. 11 (2006): 990–996.

Durlak, Joseph A. "How to select, calculate, and interpret effect sizes." *Journal of pediatric psychology* (2009): jsp004.
Websites:
Website of Dr. Lee Beckers, from University of Colorado, Colorado Springs: http://www.uccs.edu/~lbecker/
R Psychologist website by Kristoffer Magnusson: http://rpsychologist.com/d3/cohend/

REFERENCES

1. Cao J, Zhang S. Multiple comparison procedures. JAMA 2014;312:543–4
2. Sterne JA, Smith GD. Sifting the evidence—what's wrong with significance tests? Phys Ther 2001;81.8:1464–9
3. Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–9
4. Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002;2:8
5. Wason JM, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? Trials 2014;15:364
6. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Effect 2012;5:189–211
7. Stacey AW, Pouly S, Czyz CN. An analysis of the use of multiple comparison corrections in ophthalmology research. Invest Ophthalmol Vis Sci 2012;53:1830–4
8. Holland BS, Copenhaver MD. Improved Bonferroni-type multiple testing procedures. Psychol Bull 1988;104:145
9. Eichstaedt KE, Kovatch K, Maroof DA. A less conservative method to adjust for familywise error rate in neuropsychological research: the Holm's sequential Bonferroni procedure. Neuro Rehabilitation 2013;32:693–6
10. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 2014;67:850–7
11. Holm S. A simple sequentially rejective multiple test procedure. Scan J Stat 1979;6:65–70
12. Perneger TV. Adjusting for multiple testing in studies is less important than other concerns. BMJ 1999;318:1288
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300
14. Verhoeven KJ, Simonsen KL, McIntyre LM. Implementing false discovery rate control: increasing your power. Oikos 2005;108:643–7
15. Hutson AD. A semiparametric bootstrap approach to correlated data analysis problems. Comput Methods Programs Biomed 2004;73:129–34
16. Hung HM, Wang SJ. Multiple comparisons in complex clinical trial designs. Biom J 2013;55:420–9
17. Neuhäuser M. How to deal with multiple endpoints in clinical trials. Fundam Clin Pharmacol 2006;20:515–23
18. Romano JP, Shaikh AM, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 2008;17:417–42
19. Westfall PH, Tobias R, Wolfringer R. Multiple Comparisons and Multiple Tests Using SAS. 2nd ed. Cary, NC: SAS® Press, 2011
20. Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. Vol 279. New York, NY: John Wiley & Sons, 1993
21. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Planning Inference 1999;82:171–96
22. Bland JM, Altman DG. Statistics Notes: Bootstrap resampling methods. BMJ 2015;350:h2622
23. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141–64
24. Chernick MR. Bootstrap Methods: A Guide for Practitioners and Researchers. Vol 619. New York, NY: John Wiley & Sons, 2011
25. Efird JT, Nielsen SS. A method to compute multiplicity corrected confidence intervals for odds ratios and other relative effect estimates. Int J Environ Res Public Health 2008;5:394–8
26. Ludbrook J. Multiple inferences using confidence intervals. Clin Exp Pharmacol Physiol 2000;27:212–5
27. Murphy GS, Szokol JW, Avram MJ, Greenberg SB, Shear T, Vender JS, Gray J, Landry E. The effect of single low-dose dexamethasone on blood glucose concentrations in the perioperative period: a randomized, placebo-controlled investigation in gynecologic surgical patients. Anesth Analg 2014;118:1204–12
28. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 2008;3:286–300
29. Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875–9
30. Shafer SL, Dexter F. Publication bias, retrospective bias, and reproducibility of significant results in observational studies. Anesth Analg 2012;114:931–2
31. Cumming G. The new statistics: why and how. Psychol Sci 2014;25:7–29
32. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol 2009;34:917–28
33. Cohen J. Statistical Power for the Social Sciences. Hillsdale, NJ: Laurence Erlbaum and Associates, 1988
34. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychol Methods 1996;1:170
35. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 2002;7:105–25
36. Olejnik S, Algina J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 2003;8:434–47