JAMA Guide to Statistics and Methods

Multiple Comparison Procedures

Jing Cao, PhD; Song Zhang, PhD

Problems can arise when researchers try to assess the statistical significance of more than 1 test in a study. In a single test, statistical significance is often determined based on an observed effect or finding that is unlikely (<5%) to occur due to chance alone. When more than 1 comparison is made, the chance of falsely detecting a nonexistent effect increases. This is known as the problem of multiple comparisons (MCs), and adjustments can be made in statistical testing to account for this.[1]

In this issue of *JAMA*, Saitz et al[2] report results of a randomized trial evaluating the efficacy of 2 brief counseling interventions (ie, a brief negotiated interview and an adaptation of a motivational interview, referred to as MOTIV) in reducing drug use in primary care patients when compared with not having an intervention. Because MCs were made, the authors adjusted how they determined statistical significance. In this article, we explain why adjustment for MCs is appropriate in this study and point out the limitations, interpretations, and cautions when using these adjustments.

Related article page 502

Use of Method

Why Are Multiple Comparison Procedures Used?

When a single statistical test is performed at the 5% significance level, there is a 5% chance of falsely concluding that a supposed effect exists when in fact there is none. This is known as making a false discovery or having a false-positive inference. The significance level represents the risk of making a false discovery in an individual test, denoted as the individual error rate (IER). If 20 such tests are conducted, there is a 5% chance of making a false-positive inference with each test so that, on average, there will be 1 false discovery in the 20 tests.

Another way to view this is in terms of probabilities. If the probability of making a false conclusion (ie, false discovery) is 5% for a single test in which the effect does not exist, then 95% of the time, the test will arrive at the correct conclusion (ie, insignificant effect). With 2 such tests, the probability of finding an insignificant effect with the first test is 95%, as it is for the second. However, the probability of finding insignificant effects in the first and the second test is 0.95 × 0.95, or 90%. With 20 such tests, the probability that all of the 20 tests correctly show insignificance is $(0.95)^{20}$ or 36%. So there is a 100% − 36%, or 64%, chance of at least 1 false-positive test occurring among the 20 tests. Because this probability quantifies the risk of making any false-positive inference by a group, or family, of tests, it is referred to as the family-wise error rate (FWER). The FWER generally increases as the number of tests performed increases. For example, assuming IER = 5% and denoting the number of multiple tests performed as K, then for $K = 2$ independent tests, FWER = $1 - (0.95)^2 = 10\%$; for $K = 3$, FWER = $1 - (0.95)^3 = 14\%$; and for $K = 20$, FWER = $1 - (0.95)^{20} = 64\%$. This shows that the risk of making at least 1 false discovery in MCs can be greatly inflated even if the error rate is well controlled in each individual test.

When MCs are made, to control FWER at a certain level, the threshold for determining statistical significance in individual tests must be adjusted.[1] The simplest approach is known as the Bonferroni correction. It adjusts the statistical significance threshold by the number of tests. For example, for a FWER fixed at 5%, the IER in a group of 20 tests is set at 0.05/20 = 0.0025; ie, an individual test would have to have a *P* value less than .0025 to be considered statistically significant. The Bonferroni correction is easy to implement, but it sets the significance threshold too rigidly, reducing the statistical procedure's power to detect true effects.

The Hochberg sequential procedure, which was used in the study by Saitz et al,[2] takes a different approach.[3] All of the tests (the multiple comparisons) are performed and the resultant *P* values are ordered from largest to smallest on a list. If the FWER is fixed at 5% and the largest observed *P* value is less than .05, then all the tests are considered significant. Otherwise, if the next largest *P* value is less than 0.05/2 (.025), then all the tests except the one with the largest *P* value are considered significant. If not, and the third *P* value in the list is less than 0.05/3 (.017), then all the tests except those with the largest 2 *P* values are considered significant. This is continued until all the comparisons are made. This approach uses progressively more stringent statistical thresholds with the most stringent one being the Bonferroni threshold, and thus the approach can achieve a greater power to detect true effect than the Bonferroni procedure under appropriate conditions. An example in the **Table** consists of 6 tests in MCs; given a FWER of 5%, none of the tests

Table. An Example to Compare the Bonferroni Procedure and the Hochberg Sequential Procedure

		Bonferroni		Hochberg	
Test	*P* Value	Threshold	Result	Threshold	Result
1	.40	0.05/6 = 0.008	Not significant	0.05	Not significant
2	.027	0.05/6 = 0.008	Not significant	0.05/2 = 0.025	Not significant
3	.020	0.05/6 = 0.008	Not significant	0.05/3 = 0.017	Not significant
4	.012	0.05/6 = 0.008	Not significant	0.05/4 = 0.0125	Significant
5	.011	0.05/6 = 0.008	Not significant	NA	Significant
6	.010	0.05/6 = 0.008	Not significant	NA	Significant

Abbreviation: NA, not applicable.

are significant with the Bonferroni procedure. By comparison, 3 tests are significant with the Hochberg sequential procedure.

What Are the Limitations of Multiple Comparison Procedures?

Statistical procedures to control FWER in MCs were developed to reduce the risk of making any false-positive discovery. This is offset by having a lower test power to detect true effects. For example, when K = 10, the Bonferroni-corrected IER is 0.05/10 = 0.005 to control FWER at 0.05. Under the conventional 2-sided t test, for a single test in the group to be considered significant, the observed effect needs to be 43% larger than that with an IER = 0.05. When K = 20, the Bonferroni-corrected IER is 0.05/20 = 0.0025, and the observed effect needs to be 54% larger than that with an IER = 0.05. This limitation of reduced test power by controlling FWER becomes more apparent as the number of tests in MCs increases.

Why Did the Authors Use Multiple Comparison Procedures in This Particular Study?

In the study by Saitz et al, 2 tests were performed (brief negotiated interview vs no brief interview and MOTIV vs no brief interview) to determine if interventions with brief counseling were more effective in reducing drug use than interventions without counseling. With 2 tests and the IER set at 5%, the risk of falsely concluding at least 1 treatment is effective because of chance alone is 10%. To avoid the inflated FWER, the authors used the Hochberg sequential procedure.[3]

How Should This Method's Findings Be Interpreted in This Particular Study?

Saitz et al found that the adjusted P value[4] based on the Hochberg procedure was .81 for both the brief negotiated interview and MOTIV vs no brief interview. The study did not provide sufficient evidence to claim that interventions with brief counseling were more effective than the one without brief counseling in reducing drug use among primary care patients. However, the absence of evidence does not mean there is an absence of an effect. The interventions may be effective, but this study did not have the statistical power to detect the effect.

Caveats to Consider When Looking at Multiple Comparison Procedures

To Adjust or Not

If researchers conduct multiple tests, each addressing an unrelated research question, then adjusting for MCs is unnecessary. Suppose in a different study, brief negotiated interview was intended to treat alcohol use and MOTIV was intended to treat drug use. Then there is no need to adjust for MCs. This is in contrast to performing a family of tests from which the results as a whole address a single research question; then adjusting for MCs is necessary. As in the report by Saitz et al,[2] both the brief negotiated interview and MOTIV were compared with the control to draw a single conclusion about the efficacy of brief counseling interventions for drug use.

Confirmatory vs Exploratory

Bender and Lange[5] suggested that MC procedures are only required for confirmatory studies for which the goal is the definitive proof of a predefined hypothesis to support final decision making. For exploratory studies seeking to generate hypotheses that will be tested in future confirmatory studies, the number of tests is usually large and the choice of hypotheses is likely data dependent (ie, selecting hypotheses after reviewing data), making MC adjustments unnecessary or even impossible at this stage of research. "Significant" results based on exploratory studies, however, should be clearly labeled so readers can correctly assess their scientific strength.

FWER vs FDR

The main approaches to MC adjustment include controlling FWER, which is the probability of making at least 1 false discovery in MCs, or controlling the false discovery rate (FDR), which is the expected proportion of false positives among all discoveries. When using the FDR approaches, a small proportion of false positives are tolerated to improve the chance of detecting true effects.[6] In contrast, the FWER approaches avoid any false positives even at the cost of increased false negatives. The FDR and FWER represent 2 extremes of the relative importance of controlling for false positive or false negatives. The decision whether to control FWER or FDR should be made by carefully weighing the relative benefits between false-positive and false-negative discoveries in a specific study.

Definition of Family

Both FWER and FDR are defined for a particular family of tests. This "family" should be prespecified at the design stage of a study. Test bias can occur in MCs when selecting hypothesis to be tested after reviewing the data.

ARTICLE INFORMATION

Author Affiliations: Department of Statistical Science, Southern Methodist University, Dallas, Texas (Cao); Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, Texas (Zhang).

Corresponding Author: Jing Cao, PhD, Southern Methodist University, Statistical Science, Dallas, TX 75205 (jcao@smu.edu).

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Hsu JC. *Multiple Comparisons: Theory and Methods*. London, UK: Chapman & Hall; 1996.
2. Saitz R, Palfai TPA, Cheng DM, et al. Screening and brief intervention for drug use in primary care: the ASPIRE randomized clinical trial. *JAMA*. doi:10.1001/jama.2014.7862.
3. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.
4. Wright SP. Adjusted P value for simultaneous inference. *Biometrics*. 1992;48(4):1005-1013.
5. Bender R, Lange S. Adjusting for multiple testing: when and how? *J Clin Epidemiol*. 2001;54(4):343-349.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289-300.