JAMA Guide to Statistics and Methods

Gatekeeping Strategies for Avoiding False-Positive Results in Clinical Trials With Many Comparisons

Kabir Yadav, MDCM, MS, MSHS; Roger J. Lewis, MD, PhD

Clinical trials characterizing the effects of an experimental therapy rarely have only a single outcome of interest. In a previous report in *JAMA*,[1] the CLEAN-TAVI investigators evaluated the benefits of a cerebral embolic protection device for stroke prevention during transcatheter aortic valve implantation. The primary end point was the reduction in the number of ischemic lesions observed 2 days after the procedure. The investigators were also interested in 16 secondary end points involving measurement of the number, volume, and timing of cerebral lesions in various brain regions. Statistically comparing a large number of outcomes using the usual significance threshold of .05 is likely to be misleading because there is a high risk of falsely concluding that a significant effect is present when none exists.[2] If 17 comparisons are made when there is no true treatment effect, each comparison has a 5% chance of falsely concluding that an observed difference exists, leading to a 58% chance of falsely concluding at least 1 difference exists. The formula $1-[1-\alpha]^N$ can be used to calculate the chance of obtaining at least 1 falsely significant result, when there is no true underlying difference between the groups (in this case α is .05 and N is 17 for the number of tests).

To avoid a false-positive result, while still comparing the multiple clinically relevant end points used in the CLEAN-TAVI study, the investigators used a serial gatekeeping approach for statistical testing. This method tests an outcome, and if that outcome is statistically significant, then the next outcome is tested. This minimizes the chance of falsely concluding a difference exists when it does not.

Use of the Method

Why Is Serial Gatekeeping Used?

Many methods exist for conducting multiple comparisons while keeping the overall trial-level risk of a false-positive error at an acceptable level. The Bonferroni approach[3] requires a more stringent criterion for statistical significance (a smaller *P* value) for each statistical test, but each is interpreted independently of the other comparisons. This approach is often considered to be too conservative, reducing the ability of the trial to detect true benefits when they exist.[4] Other methods leverage additional knowledge about the trial design to allow only the comparisons of interest. In the Dunnett method for comparing multiple experimental drug doses against a single control, the number of comparisons is reduced by never comparing experimental drug doses against each other.[5] Multiple comparison procedures, including the Hochberg procedure, have been discussed in a prior JAMA Guide to Statistics and Methods.[2]

Description of the Method

A serial gatekeeping procedure controls the false-positive risk by requiring the multiple end points to be compared in a predefined sequence and stopping all further testing once a nonsignificant result is obtained. A given comparison might be considered positive if it were placed early in the sequence, but the same analysis would be considered negative if it were positioned in the sequence after a negative result. By restricting the pathways for obtaining a positive result, gatekeeping controls the risk of false-positive results but preserves greater power for the earlier, higher-priority end points. This approach works well to test a sequence of secondary end points as in the CLEAN-TAVI study or to test a series of branching secondary end points (**Figure**).

Steps in serial gatekeeping are as follows: (1) determine the order for testing multiple end points, considering their relative importance and the likelihood that there is a difference in each; (2) test the first end point against the desired global false-positive rate (ie, .05) and, if the finding does not reach statistical significance, then stop all further testing and declare this and all downstream end points nonsignificant. If testing the first end point is significant, then declare this difference significant and proceed with the testing of the next end point; (3) test the next end point using a significance threshold of .05; if not significant, stop all further testing and declare this and all downstream end points nonsignificant. If significant, then declare this difference significant and proceed with the testing of the next end point; and (4) repeat the prior step until obtaining a first nonsignificant result, or until all end points have been tested.

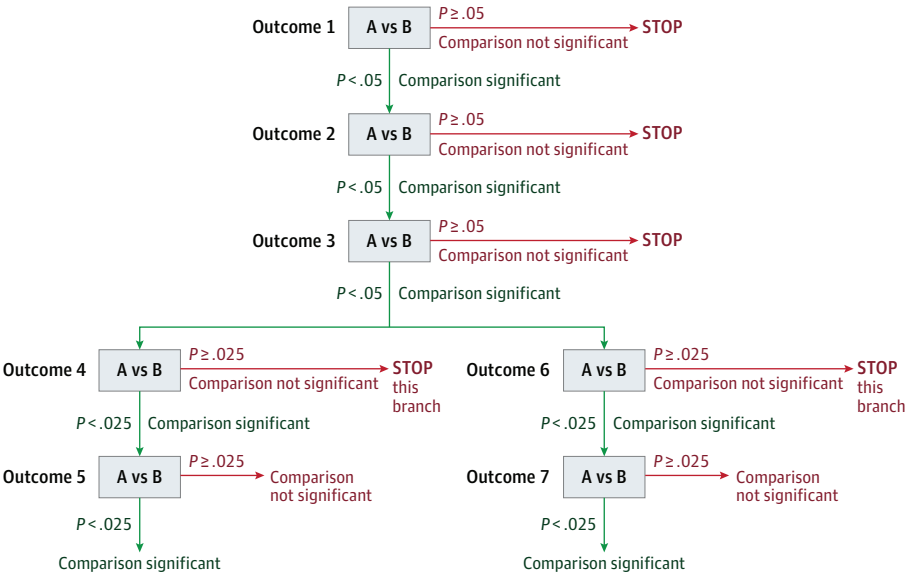
As shown in the Figure, this approach can be extended to test 2 or more end points at the same step by using a Bonferroni adjustment to evenly split the false-positive error rate within the step. In that case, testing is continued until either all branches have obtained a first nonsignificant result or all end points have been tested. For example, a neuroimaging end point could be used as a single end point for the first level, reflecting the assumption that if an improvement in an imaging outcome is not achieved then an improvement in a patient-centered functional outcome is highly unlikely, followed by a split to allow the testing of motor functions on one branch and verbal functions on the other. This avoids the need to prioritize either motor or verbal function over the other and may increase the ability to demonstrate an improvement in either domain.

Serial gatekeeping provides strict control of the false-positive error rate because it restricts multiple comparisons by sequentially testing hypotheses until the first nonsignificant test is found, and, *no matter how significant later end points appear to be*, they are never tested. The advantage is increased power for detecting effects on the end points that appear early in the sequence because they are tested against .05 rather than, eg, .05 divided by the total number of outcomes tested using a traditional Bonferroni adjustment. By accounting for the importance of certain hypotheses over others and by grouping hypotheses into primary and secondary groups, gatekeeping allocates the trial's power to be consistent with the investigators' priorities.[6]

What Are the Limitations of Gatekeeping Strategies?

Gatekeeping strategies are a powerful way to incorporate trial-specific clinical information to create prespecified ordering of hypotheses and mitigate the need to adjust for multiple comparisons

Figure. Criteria for Statistical Significance That Would Be Used in a Hypothetical Gatekeeping Strategy



This Figure shows the criteria for statistical significance that would be used in a hypothetical gatekeeping strategy in which there are 3 levels each with a single end point, followed by 2 levels with 2 end points each. The 3 end points are each tested in order against a criterion of .05. All testing stops as soon as 1 result is nonsignificant. If all are significant then a pair of fourth-level end points are tested, and to preserve the required significance of .05 at that level across 2 end points, the criterion for statistical significance is adjusted with a Bonferroni correction value of .025 for each. If 1 or both of these end points is significant at .025, then the next end point in the branch is tested, against a criterion of .025. If 1 or both are nonsignificant, no further testing occurs. If any outcome tested along a given pathway is not statistically significant, no further outcomes along that branch are tested because they are assumed to be nonsignificant.

at each stage of testing. The primary challenge in using gatekeeping is the need to prespecify and truly commit to the order of testing. The resulting limitation is that if, in retrospect, the order of outcome testing appears ill chosen (eg, if an early end point is negative and important end points later in the sequence appear to suggest large treatment effects), then there is no rigorous, post hoc method for statistically evaluating the later end points. This highlights the importance of having a clear data analysis strategy determined before the trial is started, and maintaining transparency (eg, publishing the study design and analysis plan on public websites or in journals).

How Was Gatekeeping Used in This Case?

The CLEAN-TAVI investigators used a gatekeeping strategy to compare several magnetic resonance imaging end points along with neurological and neurocognitive performance.[1] The first was the primary study end point, the number of brain lesions 2 days after TAVI. Secondary end points were only tested if the primary one was positive. Then, up to 16 secondary end points were tested in a defined sequence. The study was markedly positive, with the primary and many secondary end points demonstrating benefit. The first 8 comparisons were reported in detail in the publication—in their prespecified order—retaining the structure of the gatekeeping strategy.[1]

How Should the Results Be Interpreted?

The CLEAN-TAVI clinical trial demonstrated the efficacy of a cerebral protection strategy with respect to multiple imaging measures of ischemic damage. The use of the prespecified gatekeeping strategy should provide assurance that the large number of imaging end points that were compared was unlikely to have led to false-positive results.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Yadav, Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Yadav); Berry Consultants, LLC, Austin, Texas (Lewis).

Corresponding Author: Kabir Yadav, MDCM, MS, MSHS, Department of Emergency Medicine, 1000 W Carson St, Box 21, Torrance, CA 90509 (kabir@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

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. Haussig S, Mangner N, Dwyer MG, et al. Effect of a cerebral protection device on brain lesions following transcatheter aortic valve implantation in patients with severe aortic stenosis. *JAMA*. 2016;316(6):592-601.

2. Cao J, Zhang S. Multiple comparison procedures. *JAMA*. 2014;312(5):543-544.

3. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310(6973):170-170.

4. Hommel G, Bretz F, Maurer W. Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat Med*. 2007;26(22):4063-4073.

5. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.

6. Dmitrienko A, Millen BA, Brechenmacher T, Paux G. Development of gatekeeping strategies in confirmatory clinical trials. *Biom J*. 2011;53(6):875-893.