NEJM Evidence

Published July 22, 2025
NEJM Evid 2025;4(8)
DOI: 10.1056/EVIDctw2400393

CLINICAL TRIALS WORKSHOP

# Multiplicity Control in Clinical Trials

Amy LaLonde, Ph.D.,[1] and Steven E. Nissen, M.D.[2]

## Abstract

Statistical testing of more than one hypothesis has the potential to increase the risk of wrongly concluding that the result for a given end point is statistically significant (false discovery). This review is designed to acquaint nonstatisticians with traditional approaches for controlling type I error and with the seemingly complex procedure known as graphical testing.

## Introduction

A simple clinical trial tests a single hypothesis (primary end point), with success or failure determined by rejection of a null hypothesis ($H_0$) at a significance level (alpha) that typically requires the type I error rate (P value) to be 0.05 or lower. However, many contemporary clinical trials are designed to test more than one hypothesis, termed $H_1$, $H_2$, $H_3$, etc., to establish the efficacy relative to control for more than one primary end point or multiple treatment groups (e.g., dose levels, devices, or procedures). There may also be interest in testing several additional hypotheses (secondary end points) if the result for the primary end point is statistically significant.

Unless carefully constructed, statistical testing of more than one hypothesis has the potential to increase the risk of wrongly concluding that the result for a given end point is statistically significant (false discovery), termed familywise error rate (FWER).[1] This review is designed to acquaint nonstatisticians with traditional approaches for controlling type I error and with the seemingly complex procedure known as graphical testing.

## Traditional Approaches

The simplest approach to control the FWER for multiple hypotheses in superiority, noninferiority, and equivalence trials is the Bonferroni procedure, which splits alpha equally among the hypotheses to be tested.[2] For two hypotheses ($H_1$, $H_2$), alpha is split by two ($P=0.025$); with four hypotheses, the required P value is 0.0125. Different fractions of an alpha of 0.05 can be assigned to each hypothesis (weighted Bonferroni), as long as the total alpha is 0.05. The Bonferroni procedure is very conservative — the least likely of the common approaches to introduce type I error (false positives). This comes at the expense of increased type II error (false negatives). Bonferroni corrections are typically employed when evaluating the dual primary end points for a trial assessing more than one drug dose or intervention. The trial will be deemed successful if either comparator is statistically better than the control, at this corrected threshold.[3]

*The author affiliations are listed at the end of the article.*

*Dr. Nissen can be contacted at nissens@ccf.org or at Cleveland Clinic JB-20, 9500 Euclid Ave., Cleveland, OH 44195.*

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

**Table 1. Overview of Common Multiple Testing Procedures.**

| Class of Procedure | Description | Considerations for Use |
|---|---|---|
| Bonferonni | No predefined hierarchy; test all $m$ hypotheses at $\alpha/m$. | Most conservative; may be appropriate when testing a dual primary end point or multiple doses (trial is successful if either end point is achieved). |
| Holm, Hochberg, and Hommel | No predefined hierarchy; these procedures rank P values and test hypotheses based on the ranks and/or the number of hypotheses. | Less conservative among the nonhierarchical tests; commonly used when assessing post hoc or hypothesis-generating analyses, for example, heterogeneity of treatment effect. |
| Benjamini–Hochberg | A step-up method for controlling false discovery rate. | A less conservative procedure that controls the false discovery rate when there are multiple hypotheses, rejecting the null hypotheses that exceed a critical value. |
| Dunnett | Specifically tests whether any of the treatment groups are significantly different from the control group, without comparing treatments with each other. | Appropriate for testing multiple treatment groups to the same control. |
| Hierarchical or graphical testing | Hypotheses are ordered based on clinical importance and/or likelihood of success with or without alpha splitting across multiple hypotheses. | Appropriate for analysis of primary and key secondary end points in trials with many hypotheses with varying degrees of uncertainty and clinical relevance. |

To create more powerful approaches without inflating the type I error, several alternative procedures have been defined and mathematically validated. The Holm procedure, commonly referred to as a step-down procedure, ranks P values from multiple hypotheses from smallest to largest, then tests the hypothesis with the smallest P value at an alpha of $0.05/m$, where $m$ is the number of hypotheses.[2] If the smallest P value is greater than $0.05/m$, testing stops. However, if the first test is significant, the next lowest P value is tested at an alpha of $0.05/m–1$. In the simplest example, for two end points, the hypothesis with the lowest P value is tested at an alpha of 0.05/2, or 0.025, and, if significant, the second end point is tested at an alpha of 0.05/1, or 0.05.[4] The Hochberg procedure, conversely, begins with the largest P value and steps up.[5]

For FWER control, several alternatives exist, including the Hommel procedure, a variant of the step-up procedure, and Dunnett's test, which leverages the dependence across the comparisons to control FWER and improve testing efficiency for multiple comparisons of various active groups to a single control group with the same end point.[6,7] Table 1 provides an overview of these traditional closed testing procedures and some scenarios to consider when choosing one approach over another.

Procedures to limit the false discovery rate (FDR) represent an alternative to FWER control as an approach to multiple comparisons. Rather than control the probability of any type I error, FDR procedures represent multiplicity adjustments that limit the proportion of type I errors, incorrect rejections of the null hypothesis (false positives). In other words, FDR procedures limit the ratio of false positives to the total number of rejections of the null hypothesis. FDR-controlling procedures are particularly important when there are many statistical tests performed on a data set, a problem commonly arising in genomics research. However, these procedures provide less stringent control of type 1 errors (the likelihood of a single false positive) than FWER procedures such as the Bonferroni correction. The Benjamini–Hochberg procedure balances the risk of type I and type II error (false negatives) by ordering P values and setting a desired FDR (often 5%) to reject the null hypothesis only for those P values that exceed a critical value.[8]

## Secondary End Points and the Need for Hierarchical Testing Strategies

Contemporary clinical trials typically have multiple secondary end points. Those that are considered critical to assessing clinical benefits of a therapy are termed key secondary end points, which prespecifies that they will be controlled for type I error. Sequential testing is a traditional approach to control FWER that prespecifies a fixed order of testing. If the primary end point has a P value of 0.05 or less, the first-ranked secondary end point is tested at an alpha of 0.05, and, if below this threshold, the next end point in sequence is also tested at an alpha of 0.05. The process continues until an end point in the sequence is not below this threshold, after which the remaining end points are deemed not statistically significant and may not be individually tested (Fig. 1A).[9] In most cases, any end points that are not significant are not included by regulatory authorities in the product label.

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

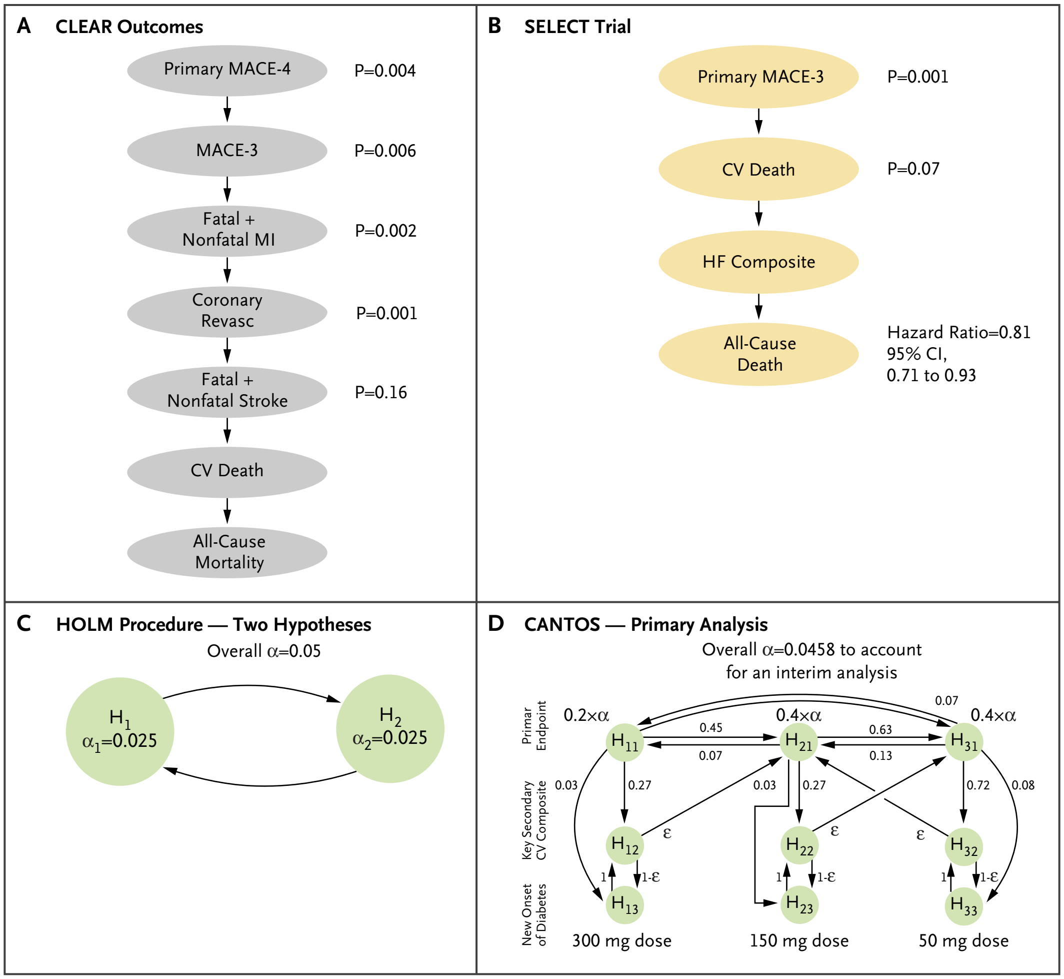

Figure 1. Examples of Conventional and Graphical Testing Approaches.

Panel A shows a hierarchical testing procedure in which the first four end points are statistically significant and the last three end points are not. Panel B shows a hierarchical testing scheme in which only the first end point is significant and the next three are not significant. Panel C shows the Holm procedure, depicted using a graphical testing framework. Panel D shows a complex graphical testing procedure with nine individual end points. CANTOS denotes Canakinumab Antiinflammatory Thrombosis Outcome Study; CI, confidence interval; CLEAR, cholesterol lowering via bempedoic acid, an ACL inhibiting regimen; CV, cardiovascular; epsilon, weight assigned to an edge (less than or equal to 1); HF, heart failure; HR, heart rate; MACE-3, three-point major adverse cardiac events; MACE-4, four-point major adverse cardiac events; MI, myocardial infarction; Revasc, revascularization; and SELECT, semaglutide effects on cardiovascular outcomes in people with overweight or obesity.

To understand the implications of failing to demonstrate statistical significance, consider the recent example from an outcome trial in which cardiovascular death was the first secondary end point in the fixed sequence and was associated with a P value of 0.07[10] (Fig. 1B). The remaining end points failed, and P values for all subsequent end points in the fixed sequence were not reported, including all-cause mortality with a hazard ratio of 0.81 and an upper bound of the 95% confidence interval of 0.93.

## Graphical Testing

To overcome the limitations of these traditional testing procedures, Bretz et al. proposed graphical testing as an approach for preserving FWER.[11] Figure 1C provides an example depicting the Holm procedure for two hypotheses as a graph, showing a traditional approach to multiplicity control within the graphical testing framework. Graphical

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

testing can be extremely complex (Fig. 1D).[11,12] The graph represents each hypothesis as a node and connects these nodes with arrows assigned numerical weights that dictate the proportion of alpha that will be propagated when that hypothesis is evaluated successfully, i.e., the P value is less than the allocated alpha. The graph's arrows dictate the sequence of weighted Bonferroni-like adjustments to enable testing of several hypotheses until enough end points fail to achieve significance, halting further testing.

Figure 2A represents a hypothetical trial with a single primary hypothesis, $H_1$, and four key secondary hypotheses, $H_2$–$H_5$. These key secondary end points are prespecified because of their clinical relevance and importance in conveying a more complete picture of the benefits and risks for patients. Figures 2B–2F demonstrate how testing would proceed at the conclusion of such a hypothetical trial. In Figure 2B, $H_1$ is tested at a two-sided significance level alpha of 0.05, so the observed P value of 0.02 achieves statistical significance, and alpha is subsequently split 0.7 and 0.3 between $H_2$ and $H_3$, respectively. Figure 2C demonstrates the alpha propagation from the first test ($\alpha$=0.05*0.7=0.035) and shows that, with a P value of 0.01, $H_2$ achieves significance; an alpha of $0.035\times1$ is then passed to $H_4$, which achieves significance with a P value of 0.02. Figure 2D demonstrates significance for $H_5$. Finally, in Figure 2E, $H_3$ fails to achieve statistical significance. If these key secondaries were tested hierarchically, as in Figure 2F, testing would cease after $H_3$ fails to achieve statistical significance.

The example above illustrates how graphical testing is applied at the completion of the trial, but fails to convey the scientific rigor required to build the testing scheme in a statistical analysis plan. Much like a hierarchical testing scheme, end points placed higher in the scheme will be tested earlier. The placement of the nodes and the weights should reflect both the clinical importance of the end point and the anticipated likelihood of achieving statistical significance. Therefore, creation of the graph requires close collaboration between clinicians who provide the assessment of the value of each end point in translating the benefit of a therapy to patients and providers and statisticians who can provide input on the marginal power for end points of interest. Furthermore, since achieving statistical significance is typically required for inclusion of end points in the product label, input from regulators is valuable. The ordering of end points can be viewed as an optimization problem; maximizing the likelihood of success and the value of the data at the end of the trial. The graph, as well as the resulting P values, should be included in the reporting of trial results.

In the hypothetical example in Figure 2, suppose $H_1$ to $H_5$ represent hypotheses for time to first occurrence of the following end points: $H_1$ is four-point major adverse cardiovascular events (MACE-4) (nonfatal myocardial infarction, nonfatal stroke, death from cardiovascular causes, or coronary revascularization), $H_2$ is three-point major adverse cardiovascular events (MACE-3) (nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes), $H_3$ represents fatal or nonfatal stroke, $H_4$ is fatal or nonfatal myocardial infarction, and $H_5$ is coronary revascularization. If these key secondary end points were tested hierarchically, as in Figure 2F, only MACE-4 and MACE-3 would be demonstrated as statistically significant. Figure 2A, compared with Figure 2F, exemplifies the value of prespecifying a graphical testing scheme with careful consideration of clinical importance and statistical power to achieve significance on all but the coronary revascularization end point.

Because outcome trials are typically long in duration, external data from other trials and the evolving health care landscape can impact the initial assumed efficacy and/or clinical relevance of each of the key secondary end points. The graphical testing scheme is usually finalized in the statistical analysis plan, an extension of the protocol. For example, in Figure 2A, researchers may have assessed that $H_2$ and $H_3$ had similar clinical relevance and placed them in the first level of testing following the primary end point. The researchers may have allocated more weight, 0.7, to $H_2$ if there was more statistical power for this end point, and thus greater likelihood that the alpha would propagate further to $H_4$. One could also argue that if the statistical power for $H_2$ is higher than for $H_3$, then the $H_3$ end point should receive more weight than $H_2$, increasing the likelihood of success. This simple thought exercise demonstrates that the construction of the graphical testing scheme is both an art and a science.

There is no limit to the number of tests that can be considered within a graphical testing scheme; however, investigators should limit testing to those end points of valid scientific interest and reasonable power. As with any multiplicity adjustment procedure, any end points included in a graphical testing scheme that are not tested after previously tested end points have exhausted alpha are not deemed statistically significant. However, there may also be end points not included in the scheme, where the conclusions should explicitly state that these end points were not tested. Owing to the cost of a type I error, graphical testing is most appropriate for the confirmatory trials, where data interpretation informs clinicians and patients, and is less valuable for exploratory trials, driving future research. Researchers may choose to introduce multiplicity control into these

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

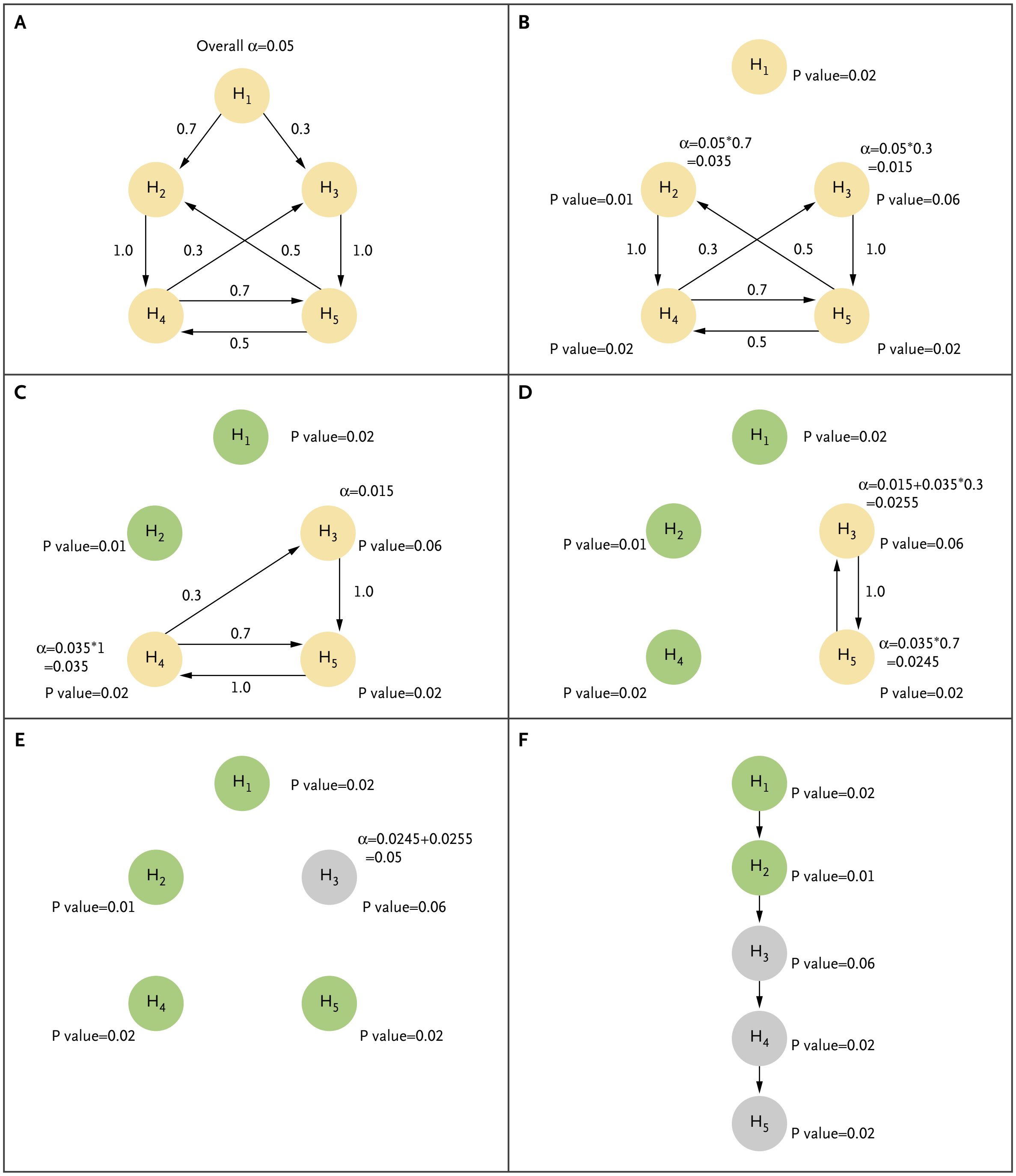

Figure 2. Application of Graphical Testing Framework Compared with a Traditional Approach to Multiplicity Adjustments.

Panel A shows a graphical testing procedure with five end points and the weights assigned to each in the prespecified analysis plan. Panels B, C, D, and E show how these weights were applied and resulted in statistical significance for all except hypothesis $H_3$. Panel F shows how a traditional approach would yield a nonsignificant result for hypotheses $H_4$ and $H_5$. The yellow circles, or nodes, denote the untested hypotheses; green circles denote hypotheses deemed statistically significant; gray circles denote hypotheses not deemed statistically significant.

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

early-phase trials, but given the more exploratory nature of such trials, the focus should be on a single pivotal end point sufficiently powered to support initial approval.

Safety end points that are part of the confirmatory strategy and desired for labeling claims could be considered, but it is not advisable to adjust for multiplicity for safety end points to minimize false-negative findings that multiplicity adjustments could obscure. This approach often leads to false-positive safety findings, which is scientifically acceptable because of the importance of safety in trials of drugs or medical devices.

Further complexity of determining the number and ordering of the graphical testing scheme arises in the case of group sequential clinical trials, which is out of scope for this review. Careful examination of the graphical testing scheme used at each interim and final analysis is necessary to control the overall type I error, and researchers should refer to more technical discussions.[11]

Researchers aiming to utilize a graphical testing approach to multiplicity control can utilize the gMCP package available in R.[13,14]

## Conclusion

There are also some disadvantages to multiplicity control, including complexity in trial design and interpretation and an increased likelihood of type II error (i.e., reduced statistical power), which may require an increased sample size, particularly in the case of multiple primary end points. However, large outcome trials are massive endeavors; the ability to test multiple hypotheses can provide considerable value to the medical community and help the trial sponsor justify the high cost. Trialists should always design trials to provide an answer to the primary objective, but where there is considerable value in allowing several scientific hypotheses to be tested in a single trial, increasing the power to support such testing may be weighed against the potential need for future trials to examine these hypotheses. Researchers should consider the clinical significance (i.e., the value to patients, prescribers, and payers) as well as the power (i.e., the anticipated effect size) to select end points for consideration. Given the size and duration of the trials, the study team (investigators and sponsor) has an ethical obligation to patients to rigorously design the final graphical testing plan.

### Disclosures

Author disclosures are available at evidence.nejm.org.

### Author Affiliations

[1] Eli Lilly and Company, Indianapolis

[2] Cleveland Clinic Coordinating Center for Clinical Research, Cleveland Clinic, Cleveland

### References

1. Tukey JW. The problem of multiple comparisons. Princeton, NJ: Princeton University, 1953.
2. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:1-62.
3. U.S. Food and Drug Administration. Multiple endpoints in clinical trials. Silver Spring, MD: FDA, 2022 (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials).
4. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65-70 (http://www.jstor.org/stable/4615733).
5. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800-802. DOI: 10.1093/biomet/75.4.800.
6. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75:383-386. DOI: 10.1093/biomet/75.2.383.
7. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096-1121. DOI: 10.1080/01621459.1955.10501294.
8. 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. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
9. Nissen SE, Lincoff AM, Brennan D, et al. Bempedoic acid and cardiovascular outcomes in statin-intolerant patients. N Engl J Med 2023;388:1353-1364. DOI: 10.1056/NEJMoa2215024.
10. Lincoff AM, Brown-Frandsen K, Colhoun HM, et al. Semaglutide and cardiovascular outcomes in obesity without diabetes. N Engl J Med 2023;389:2221-2232. DOI: 10.1056/NEJMoa2307563.
11. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med 2009;28:586-604. DOI: 10.1002/sim.3495.
12. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease. N Engl J Med 2017;377:1119-1131. DOI: 10.1056/NEJMoa1707914,
13. Rohmeyer K, Klinglmueller F. gMCP: a graphical approach to sequentially rejective multiple test procedures. R package version 0.6-5. March 25, 2024 (http://cran.r-project.org/package=gMCP).
14. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011 (http://www.R-project.org).

NEJM Evidence is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from evidence.nejm.org at University of Manchester (man) / England on August 9, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.