Journal of Cardiothoracic and Vascular Anesthesia 000 (2023) 1–3

Contents lists available at ScienceDirect

Journal of Cardiothoracic and Vascular Anesthesia

journal homepage: www.jcvaonline.com

Editorial

The Art of the Null Hypothesis—Considerations for Study Design and Scientific Reporting

SINCE THE ADVENT of the scientific method, hypothesis testing has been a crucial tool for drawing inferences from research studies. In medical research, conventional null hypothesis testing compares a null hypothesis H_0 (typically that there is no difference between 2 or more differently exposed groups) with an alternative hypothesis H_a (usually that a difference exists).[1] Because 2 comparator groups rarely have identical outcomes, statistical methods for hypothesis testing assess the likelihood that observed differences between the groups result from random chance.[2] This assessment is critical for scientific inference; if the observed findings are unlikely to be from chance alone, then the scientist should reject the null hypothesis in favor of a feasible alternative. This editorial outlines the basics of study design to enable rigorous null hypothesis testing for scientific inference, and suggestions for manuscript language to succinctly communicate those findings in scientific reports. We also discuss the common problem of multiple-hypothesis testing in research, the appropriate considerations for these study designs and analyses, and how to describe them in manuscripts.

Defining and Testing a Null Hypothesis

A critical first step in null hypothesis testing is stating the study objectives and hypothesis clearly, typically at the end of the study introduction.[3] The outcomes must be clear, objective, specific, and self-evident to the reader, given the study background in the introduction.[1,3] Although this may seem intuitive, it is common for initial journal submissions to state only vague hypotheses (or none at all). The hypothesis statement is often framed in terms of the alternative hypothesis; the null hypothesis is typically inferred. An excellent example of a clear hypothesis statement comes from a recent study by He et al. examining total intravenous anesthesia (TIVA) or volatile anesthesia in cardiac surgery. The authors state they "tested the hypothesis that compared with propofol-based TIVA, volatile anesthesia was associated with fewer pulmonary complications in adults undergoing cardiac surgery...."[4] The hypothesis statement clearly defines the alternative hypothesis; the reader can easily infer the null—there is no significant difference between TIVA and volatile anesthesia in postoperative pulmonary complications. This establishes an easily interpretable null hypothesis test. If there are differences in the risk of postoperative pulmonary complications between patients receiving TIVA and volatile anesthesia, the authors can investigate the probability that these differences arose from chance alone and choose to either reject or not reject the null hypothesis.

In conventional hypothesis testing, a generally accepted threshold to reject the null hypothesis typically has been $\leq 5\%$; in other words, if the probability of the observed result occurring by chance alone is $<5\%$, the null hypothesis should be rejected.[1] This 5% rejection threshold commonly is referred to as "Type I error" (or α), which is the probability of incorrectly rejecting the null hypothesis (and accepting the alternative hypothesis) when the null hypothesis is true. For a null hypothesis for which there is no difference between study groups, the probability of the observed results occurring by random chance typically is referred to as the "probability value" or "p-value." Importantly, p-values suggesting rejection of the null hypothesis (classically < 0.05) do not *a priori* indicate the null hypothesis is false; instead, they indicate the observed results are unlikely to have occurred by chance alone, and it is more reasonable to accept an alternative hypothesis instead of the null.[2] Rejecting the null hypothesis in scientific manuscripts can be done with language indicating that findings are "significantly different," "significantly greater/less than," or "significantly associated" among groups. The prefix "significantly" implies that the difference was unlikely to occur by chance while still appropriately reserving a remote possibility that the null hypothesis may be correct. The more improbable study results are due to chance (eg, if p-values are less than 0.01, 0.001, 0.0001, etc), the more robustly the study supports an alternative hypothesis.[5]

Conversely, if the observed study results have a greater probability than 5% of occurring by chance alone (eg, a p-value > 0.05), the null hypothesis cannot be rejected, even though there may be a "true" difference between groups. Importantly, this does not mean the null hypothesis is true, only that the study results do not support its rejection. Failure to reject the null hypothesis when it is false is referred to as a

https://doi.org/10.1053/j.jvca.2023.02.026

"Type II error" (often quantified as β). Manuscript language must reflect this uncertainty. Because failure to reject the null hypothesis does not prove the null hypothesis is correct, authors should not claim that 2 groups are "similar," "equal," or that there is "no difference" between groups when they are unable to reject the null hypothesis. Similarly, authors should refrain from describing results as "trending toward statistical significance" when the results are close to a critical p-value threshold but ultimately do not cross it. Instead, specific language such as there is "no significant difference" or "no significant association" is more appropriate, leaving room that the study may have appropriately failed to reject the null hypothesis due to a Type II error. He et al. again excellently demonstrated this concept, stating in their discussion, "an anesthetic maintenance regimen with a volatile anesthetic was not statistically superior to propofol-based TIVA regarding the occurrence of pulmonary complications."[4] This statement clearly summarizes the study that the null hypothesis could not be rejected while leaving open the possibility that true differences between groups may exist but could not be detected.

Multiple Hypothesis Testing

Research studies frequently have numerous outcomes, and standard null hypothesis testing for multiple endpoints requires modifications. When multiple independent hypotheses are assessed simultaneously, the risk of making a Type I error increases. When performing a single hypothesis test with an α-threshold of 5% on a null hypothesis known to be correct, the probability of incorrectly rejecting it is $1\text{-}0.95^1 = 5\%$. However, if 5 independent null hypotheses known to be true are tested, and each were held to the same threshold, then the probability of incorrect rejection of at least 1 of the 5 true null hypotheses increases to $1\text{-}0.95^5 = 23\%$.[6] Failure to account for multiple comparisons produces incorrect null hypothesis rejection and Type I error, misleading the researcher into believing a significant difference exists when none is present.

Prespecifying the multiple hypotheses and an appropriate statistical approach correcting for multiple tests is crucial to prevent inadvertent bias and incorrect interpretation of study results. Once the multiple hypotheses are identified, these can be considered a "family," and the global family-wise error rate can be set with an α-threshold of 5%. The correct null hypothesis test does not examine whether any single tested hypothesis meets the α-threshold of 5%; rather, it is the probability that an individual hypothesis meets the α-threshold while accounting for the probability of Type I error with each independent hypothesis. The simplest correction for family-wise error rate is the Bonferroni correction—dividing the α-threshold by the number of hypotheses, as seen below.

$$P_{critical} = \alpha \,/\, (\text{number of independent hypotheses})$$

For a study with 5 hypotheses, a critical p-value of 0.05/5 = 0.01 may be considered significant. Alternatively, the calculated p-values for each hypothesis can be multiplied by the total number of hypotheses in the family and the resultant values compared with a standard α-threshold of 5%. This correction ensures that the overall study retains a global Type I error rate of 5%. However, it raises the threshold for each hypothesis to be rejected as not occurring from chance alone. Other variations for multiple hypothesis testing corrections, such as the Bonferroni-Holm correction or the Benjamin-Hochberg false discovery procedure, are also available.[7-9] Regardless, it is typically best practice to "maximize α" by specifying a single primary outcome (or composite outcome) while reserving exploratory endpoints as secondary outcomes.

Zhuo et al. demonstrated the superb application of multiple hypothesis correction in their study assessing 3 different risk prediction models with 2 separate outcomes (30-day and 1-year mortality) in valvular cardiac surgery. Because a total of 6 hypotheses were tested (three models x 2 outcomes each = 6 hypotheses), the authors stated, "For C-statistic analysis, a p-value < 0.008 was chosen to define statistical significance, as a Bonferroni correction was used to minimize type I error by accounting for multiple testing procedures (p-value of 0.05 divided by 6 total hypotheses…)"[10] Due to this correction, Zhuo et al. correctly failed to reject the null hypothesis for one of their hypothesis tests despite a p-value of 0.02, as it did not meet the corrected $P_{critical}$ value of 0.008. This correction improved the authors' findings' robustness and overall study quality. Similar to this study, authors must prespecify their multiple hypotheses and method for correcting family-wise error rates to enable their work to be generalized to future research and clinical care.[11]

Conclusion

Although imperfect, null hypothesis testing remains a core tenet of statistical inference in biomedical research. For successful execution, clear null and reasonable alternative hypotheses must be stated, ideally in the study introduction, with specific outcomes to be assessed. A failure to reject a null hypothesis does not prove it is correct; we recommend specific manuscript language to convey this uncertainty. When assessing multiple hypotheses, correction for family-wise error rate with a Bonferroni or other statistical correction is required and critical to draw the appropriate inference. Applied correctly, null hypothesis testing remains a powerful tool to assist researchers and clinicians in sorting scientific observations that may occur due to random chance from those more likely associated with a true finding.

Declaration of competing interest

M.W.V. receives royalties from the Dana-Farber Cancer Institute & Novartis for a patent licensing agreement regarding a novel cancer immunotherapy in preclinical development.

Christian T. O'Donnell, MD
Vikram Fielding-Singh, MD, JD
Matthew W. Vanneman, MD
Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA

References

1 Nizamuddin SL, Nizamuddin J, Mueller A, et al. Developing a hypothesis and statistical planning. J Cardiothorac Vasc Anesth 2017;31:1878–82.

2 Kacha AK, Nizamuddin SL, Nizamuddin J, et al. Clinical study designs and sources of error in medical research. J Cardiothorac Vasc Anesth 2018;32:2789–801.

3 Vetter TR, Mascha EJ. In the beginning-there is the introduction-and your study hypothesis. Anesth Analg 2017;124:1709–11.

4 He LL, Li XF, Jiang JL, et al. Effect of volatile anesthesia versus total intravenous anesthesia on postoperative pulmonary complications in patients undergoing cardiac surgery: A randomized clinical trial. J Cardiothorac Vasc Anesth 2022;36:3758–65.

5 Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: A case for a Fragility Index. J Clin Epidemiol 2014;67:622–8.

6 Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. BMJ 1995;310:170.

7 Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.

8 McLaughlin MJ, Sainani KL. Bonferroni, Holm, and Hochberg corrections: Fun names, serious changes to p values. PM R 2014;6:544–6.

9 Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol 2018;71:353–60.

10 Zhuo DX, Bilchick KC, Shah KP, et al. MAGGIC, STS, and EuroSCORE II risk score comparison after aortic and mitral valve surgery. J Cardiothorac Vasc Anesth 2021;35:1806–12.

11 McCullough JM, Kaplan B. A random walk through large data: Caveats regarding the potential for false inference. Transplantation 2016;100:18–22.