Moving beyond significance testing: Confidence intervals in clinical research

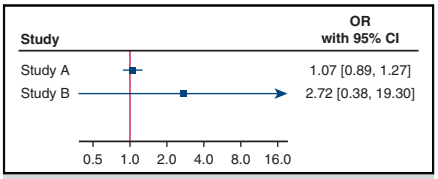
Paul F. Visintainer, PhD

The New England Journal of Medicine recently refined its guidelines for statistical reporting of research studies.[1] The new guidance is in part due to the recent position articles published in *The American Statistician* critical of the use of *P* value thresholds and statistical reporting of results.[2,3] In fact, the entire issue (*Am Stat*. 2019;73[Suppl 1]) has more than 40 articles, editorials, and commentaries on the current state of statistical testing in research and why we need to change. The title of the lead editorial captures the essence of the issue: "Moving to a World Beyond $p < 0.05$."[3] It is long, long overdue.

The limitations and misuse of *P* values to summarize results in clinical research have been discussed in great detail for decades.[4-7] Contrary to its intent, significance testing has been elevated to the status of the criterion for scientific and clinical relevance. Wasserstein and colleagues[3] refer to it as a "tyrant," stating that "no *P* value can reveal the plausibility, presence, truth, or importance of an association or effect."

Yet, the application of *P* values as evidence for precisely these concerns persists. Further, focusing solely on whether an effect is or is not statistically significant provides no information regarding the magnitude of the clinical effect under study or the degree of confidence we have in the result. Indeed, it is difficult to see what useful information a clinician might derive from knowing the *P* value.

Why do we continue to misinterpret significance testing in this manner? The practice persists because we fall into a pattern of writing in a way that mimics what we read.[2] Perhaps we believe that somehow the phrase "statistically significant" indicates that we have met some high standard of scientific integrity or scientific proof. Furthermore, many investigators know no other approach to reporting or interpreting results. If journals begin to impose restrictions on reporting *P* values or otherwise relegate them to a less important role, how should we describe the results of our study? If we are to rise to the challenge of Wasserstein and colleagues[3] to stop writing "statistically significant," what do we write instead?



Two "nonsignificant" studies providing substantially different degrees of precision regarding "nonsignificance."

CENTRAL MESSAGE

Greater use of confidence intervals can lead to stronger critiques and richer, more relevant discussions of clinical research.

This Invited Expert Opinion provides a perspective on the following papers: N Engl J Med. 2019;381(3):285–286. https://doi.org/10.1056/NEJMe1906559 Am Stat. 2016;70(2):129-133. https://doi.org/10.1080/00031305.2016.1154108 Am Stat. 2019;73(suppl 1):1-19. https://doi.org/10.1080/00031305.2019.1583913.

The purpose of this commentary is to show how we might begin to move beyond a "*P* value less than .05" standard and make small changes in the approach to reporting study results. One change that can be implemented immediately is to expand the use of confidence intervals (CIs) in reporting study effects.[8] The suggestion is not new nor is it without its own limitations.[4,9] However, elevating point estimates and CIs to a prominent role in presenting and interpreting study results will not only lead to a richer discussion of study results but also make transparent the variability inherent in all research. CIs not only capture the information of the *P* value but also allow the author to communicate a much richer narrative of study results, and this can be accomplished in clinical terms.

In this piece, I assume that the reader has a familiarity with CI and *P* values. As such, I do not spend time deriving CIs from a technical perspective, nor discuss their limitations, only insofar as these relate to the examples.

From the Department of Medicine, Office of Research, University of Massachusetts–Baystate, Springfield, Mass.

Received for publication Oct 29, 2019; revisions received Dec 28, 2019; accepted for publication Jan 16, 2020; available ahead of print April 30, 2020.

Address for reprints: Paul Visintainer, PhD, Epidemiology and Biostatistics Research Core Office of Research, 3rd Floor, 3601 Main St, Springfield, MA 01199 (E-mail: paul.visintainer@baystatehealth.org).

J Thorac Cardiovasc Surg 2021;161:1373-6
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.01.120

Textbooks on biostatistics and epidemiology provide the technical aspects of CIs and how they relate to *P* values, as well as how they may be misused.[9-12] Throughout the text, I use only 95% CIs in the examples. Although the examples and interpretations apply to other levels, the 95% CI is the level most commonly reported because it corresponds to the 5% critical test level.

In the following examples, I suggest an approach to help clinicians interpret study results on the basis of point estimates and CIs. We can ask 3 questions:

1. Is the point estimate (eg, the odds ratio [OR], the relative risk, the difference in means) clinically realistic?
2. Are both limits of the CI clinically realistic?
3. Would a narrower width substantially improve my clinical interpretation of the study results?

These questions arise in part from the work of Matthews,[13-15] who developed a Bayesian method for assessing credibility of outcomes for clinical trials. The questions engage the investigator and reader in a debate on clinical relevance and reasonableness of the estimates rather than suppress discussion because results were deemed "significant" or not. In a way, the questions lead readers and researchers to conduct an informal meta-analysis by helping them imagine how the results might compare relative to other studies, as in a forest plot.

CONFIDENCE INTERVALS CAPTURE THE INFORMATION OF *P* VALUES

I hesitate to describe this characteristic lest we simply supplant the *P* value with another term that propagates the old ways. I suspect, though, that CIs are likely being used in this manner. That is, if we estimate an effect like the OR, relative risk, or hazard ratio (HR) and the 95% CI does not include 1.0, then the risk estimate is statistically significant at *P* equals .05. This is because most often the value of the null hypothesis is 1.0. If we estimate a difference between 2 effects, for example, the difference between 2 means using a *t* test, then a 95% CI that does not contain zero would be significant at *P* equals .05. Because, most often, in these comparisons the value of the null hypothesis is 0. However, to stop here and state that the effect is statistically significant would be an aberration in the use of CIs.

CONFIDENCE INTERVALS SUMMARIZE STUDY RESULTS IN CLINICALLY MEANINGFUL TERMS

Consider the statement regarding the risk of atrial fibrillation (AF) and blood transfusion: "No statistical significant relationship was found between the number of red blood cell units transfusion during surgery ($P = .7$) and during hospital stay ($P = .2$) with the occurrence of postoperative AF …"[16] This is a statistical statement indicating that the null hypothesis was not rejected. However, the statement provides the reader no useful clinical information. What is the clinical effect?

Liu and colleagues,[17] in their meta-analysis, report an OR and 95% CI for the study by Vlahou and colleagues[16] as 0.63 (95% CI, 0.14-2.84). The OR is an estimate of the relative magnitude of the clinical effect, and the CI shows the range of possible effects that are consistent with the study data. Assuming that the model is valid and incorporated important confounders, the best estimate of the clinical effect expected from this study is a 37% reduction in the odds of AF from transfusion. However, we cannot rule out that other effects that are less than or greater than a 37% reduction are also possible. On the basis of the confidence limits, the study cannot rule out that transfusions may be associated with an 86% reduction in the odds of AF (OR, 0.14) or possibly a large 184% increase in the AF odds (ie, OR, 2.84). The confidence bounds indicate that this wide range of clinical effects are consistent with the study data.

CONFIDENCE INTERVALS PROVIDE INSIGHT INTO THE PRECISION OF THE ESTIMATE

The CI allows the reader to judge the precision of point estimate in clinical terms (ie, 37% reduction in AF risk). Is an 86% reduction in readmission risk a reasonable clinical effect from transfusion? Does a CI that ranges from a profound 86% reduction in risk to a trivial 184% increase in risk provide sufficient precision on the point estimate?

The width of the CI reflects the amount of variability inherent in the data. In estimating risk ratios, in particular, wide CIs usually arise from small sample sizes overall or a small number of events, even if the sample size appears large. Although we do not know the reason for a degree of imprecision in an estimate, the CI, not the *P* value, allows us to judge the precision.

In this example, the confidence bounds indicate that a wide range of possible AF risk estimates associated with transfusion (from substantially reducing risk to substantially increasing risk) are consistent with the data, and at least one confidence bound seems unrealistic (ie, the lower bound of 86% reduction). Thus, we may conclude that the risk estimate of a 37% reduction in the odds of AF is so imprecise as to provide little information regarding the association of AF risk and blood transfusion. It is important to note that the estimated OR of 0.63 may not be incorrect (the validity of the estimate is a consequence of the study design and sampling methods). Rather, the estimate is just very imprecise.

Contrast these results with those reported by Paone and colleagues.[18] In their study of blood transfusion and clinical outcomes after coronary artery bypass grafting, they found the adjusted risk of AF was 1.21 (95% CI, 1.10-1.33). Thus, blood transfusions (1 or 2 units vs none) increased the odds of AF by 21%. The 95% CI indicates that risk estimates of a 10% increase up to a 33% increase are consistent with the data. Further, the narrow width of the CI suggests a high

degree of precision in the estimate and the confidence bounds (ie, 1.10 and 1.33) appear clinically reasonable. In this case, estimates were based on a sample size of 16,835 patients. Notice that in discussing the estimate by Paone and colleagues,[18] there was no need to invoke "statistical significance" to interpret the result.

NOT ALL SIGNIFICANT RESULTS ARE EQUALLY IMPORTANT

The study by Choi and colleagues[19] found a significant association between AF and blood transfusion, as did Paone and colleagues.[18] For both studies, the association was highly significant at P less than .001. Given these similar significant P values, one might conclude that the studies are similar in the type and precision of information provided about the association. Yet, examining the CIs leads to a different conclusion.

As noted earlier, the study by Paone and colleagues[18] estimated the OR of AF due to blood transfusion to be 1.21 with a 95% CI ranging from 1.10 to 1.33. The study by Choi and colleagues[19] reported an adjusted OR of 5.32 with a 95% CI of 2.80 to 10.11. The OR of 5.32 appears unreasonably large, as does the upper confidence bound of 10.11. Both estimates suggest substantial instability in the estimated effect. Again, large risk ratios and unrealistic confidence bounds frequently arise when there are too few events to provide stable estimates of risk. Although the study by Paone and colleagues[18] reports a risk estimate with a high degree of precision, the study by Choi and colleagues[19] reports an OR that appears exaggerated, an upper confidence bound that is unrealistic, and a CI that is wide. Thus, regardless of their similar levels of statistical significance, we may have far more confidence in the estimates from Paone and colleagues'[18] study than those of the study by Choi and colleagues.[19]

CAN A NONSIGNIFICANT RESULT HAVE A HIGH DEGREE OF PRECISION?

Just as significant results may be reported with different degrees of precision, nonsignificant results may be reported similarly. In comparing video-assisted thoracotomy surgery with open surgery for lobectomy, Licht and colleagues[20] and Murakawa and colleagues[21] found no "significant" difference in survival between the 2 approaches. In multivariable analyses, Licht and colleagues[20] reported an HR of 0.98 (95% CI, 0.80-1.22). Murakawa and colleagues[21] reported an HR for video-assisted thoracotomy surgery of 0.56, (95% CI, 0.19-1.66), using propensity score analysis. Although both studies are not significant, the clinical effect reported by Licht and colleagues[20] is estimated with more precision than the estimate reported by Murakawa and colleagues[21] because the sample size is substantially larger. Although the HR reported by Murakawa and colleagues[21] may be considered reasonable (HR, 0.56), the lower bound of the CI is exaggerated (HR, 0.19) and unrealistic. The wide CI indicates substantial instability in the estimated HR. Thus, regardless of the P values, we may conclude that the study by Licht and colleagues[20] provides stronger support of no difference in survival between the 2 surgical approaches than the study by Murakawa and colleagues.[21]

SAMPLE SIZE AND CONFIDENCE WIDTH

As we encourage their use in reporting study results, we have to be cognizant that intentional use of CIs needs to be incorporated into the design phase of studies. Investigators are aware that a priori sample size calculations are a critical component of study designs. Frequently, these calculations are based on testing an alternative hypothesis against a null hypothesis with sufficient power to reject the null at a critical alpha level (ie, $P = .05$). In other words, the approach is based on statistical hypothesis testing without consideration for the degree of precision for the estimated effect.

For example, suppose we want to test whether an analgesic administered postoperatively will reduce opioid use among elder patients undergoing cardiac surgery. Suppose our estimate of the average morphine milligram equivalents (MME) among our target population is 100 (standard deviation, 40). Suppose further, we consider a reduction of 25 MMEs to an average of 75 MMEs an important clinical effect. With this information, a sample size of 41 patients per group would provide 80% power to detect this effect size, assuming a 2-sided test at a critical value of 5%.

Now suppose we want a precise estimate on the effect by specifying that the 95% confidence width should be no larger than 20 MMEs. Under this scenario, a sample size of 134 patients per group would be required to provide 80% probability that our 95% confidence width would be no larger than 20 MMEs.

Therefore, although we encourage greater use of CIs in presenting study results, we need to attend to their importance in sample size calculations.

CONCLUSIONS

The goal of research is not to find statistical significance. The goal of research is to derive valid estimates of phenomena under study with as much precision as reasonably possible. Research is a complex process, resulting from the successful interplay of intricate activities as study design, sampling, variable measurement, bias and confounding assessment, and data analysis, to name a few. We must recognize that underlying the research process is randomness, variability, and uncertainty. Statistical methods quantify this variability so it can be evaluated. Statistics, as Gelman[22] reflects, is not a method that transforms randomness into certainty. Rather than ignoring uncertainty or otherwise implying that it has been made trivial by finding "statistical significance," we should move it front and center and embrace it.[23] To this end, expanding the

use of CIs will promote a richer and more relevant discussion, not only statistically but also clinically, among clinical researchers and their audience.

Conflict of Interest Statement

The author reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, et al. New guidelines for statistical reporting in the journal. *N Engl J Med*. 2019;381:285-6.
2. Wasserstein RL, Lazar NA. The ASA statement on P-values: context, process, and purpose. *Am Stat*. 2016;70:129-33.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$" *Am Stat*. 2019;73(Suppl 1):1-19.
4. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-50.
5. Goodman S. A dirty dozen: twelve P-value misconceptions. *Semin Hematol*. 2008;45:135-40.
6. Berger JO, Sellke T. Testing a point null hypothesis: the irreconsilability of p-values and evidence. *J Am Stat Assoc*. 1987;82:112-22.
7. Bakan D. The test of significance in psychological research. *Psychol Bull*. 1966;66:423-37.
8. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature*. 2019;567:305-7.
9. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
10. Szklo M, Nieto FJ. Epidemiology: Beyond the Basics. 4th ed. Burlington, MA: Jones & Barlett Learning; 2019.
11. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research. 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.
12. Woodward M. Epidemiology: Study Design and Data Analysis. 3rd ed. Boca Raton, FL: CRC Press; 2014.
13. Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. *Drug Inf J*. 2001;35:1469-78.
14. Matthews RAJ. Beyond 'significance': principles and practice of the analysis of credibility. *R Soc Open Sci*. 2018;5:171047.
15. Matthews RAJ. Moving towards the post $p < 0.05$ era via the analysis of credibility. *Am Stat*. 2019;73(Suppl):202-12.
16. Vlahou A, Diplaris K, Ampatzidou F, Karagounnis L, Drossos G. The role of blood transfusion in the development of atrial fibrillation after coronary artery bypass grafting. *J Thorac Cardiovasc Surg*. 2016;64:688-92.
17. Liu S, Li Z, Liu Z, Hu Z, Zheng G. Blood transfusion and risk of atrial fibrillation after coronary artery bypass graft surgery: a meta-analysis of cohort studies. *Medicine (Baltimore)*. 2018;97:e9700.
18. Paone G, Likosky DS, Brewer R, Theurer PF, Bell GF, Cogan CM, et al. Transfusion of 1 and 2 units of red blood cells is associated with increased morbidity and mortality. *Ann Thorac Surg*. 2014;97:87-94.
19. Choi YS, Shim JK, Hong SW, Kim DH, Kim JC, Kwak YL. Risk factors of atrial fibrillation following off-pump coronary artery bypass graft surgery: predictive value of C-reactive protein and transfusion requirement. *Eur J Cardiothorac Surg*. 2009;36:838-43.
20. Licht PB, Jørgensen OD, Ladegaard L, Jakobsen E. A national study of nodal upstaging after thoracoscopic versus open lobectomy for clinical stage I lung cancer. *Ann Thorac Surg*. 2013;96:943-50.
21. Murakawa T, Ichinose J, Hino H, Kitano K, Konoeda C, Nakajima J. Long-term outcomes of open and video-assisted thoracoscopic lung lobectomy for the treatment of early stage non-small cell lung cancer are similar: a propensity-matched study. *World J Surg*. 2015;39:1084-91.
22. Gelman A. The problems with P-values are not just with P-values [discussion]. *Am Statistician*. 2016;70:1-2.
23. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat*. 2019;73(Suppl 1):235-45.