

Richelle J. Cooper, MD, MSHS
Robert L. Wears, MD, MS
David L. Schriger, MD, MPH

From the University of California–Los Angeles Emergency Medicine Center, University of California–Los Angeles School of Medicine, Los Angeles, CA (Cooper, Schriger); and the Department of Emergency Medicine, University of Florida Health Science Center Jacksonville, Jacksonville, FL (Wears).

Copyright © 2003 by the American College of Emergency Physicians.

0196-0644/2003/\$30.00 + 0

doi:10.1067/mem.2003.135

Reporting Research Results: Recommendations for Improving Communication

See editorial, p. 565.

[*Ann Emerg Med.* 2003;41:561-564.]

Of the many new recommendations contained in the revised “Instructions for Authors” (see page 18A), those regarding hypothesis testing and *P* values provoke the most controversy. A number of readers and authors have commented on them, and some disagree with the change. The purpose of this editorial is to explain why we prefer authors to emphasize estimation over hypothesis testing, that is, to use point and interval estimates rather than *P* values.

Although our recommendations differ from the practice at many journals, they can hardly be considered new or novel. A large body of literature provides ample evidence of the problems with hypothesis testing and *P* values (dating back to R. A. Fisher, who invented the *P* value and spent a good part of his subsequent career criticizing its misinterpretation and misuse).¹⁻¹⁵ Although statisticians, theoreticians, and methodologists may continue to differ about the exact nature of the advantages and disadvantages of hypothesis testing and its alternatives, there is little disagreement with the proposition that *P* values are commonly misunderstood and misinterpreted.¹⁶ Despite this, the medical research community, like an old dog unable to learn new tricks, is caught in a vicious cycle. Authors continue to report *P* values liberally, in part because they believe journals expect and may demand them.

Journals print them because authors (whom they wish not to offend) report them, and they believe readers expect them. Readers, seeing them everywhere, may expect them, but we venture to say, hardly desire them. We hope with our new emphasis on estimation, to break or at least attenuate this cycle, and replace it with a style of reporting that we hope will be more informative and enlightening. Emergency physicians have always been willing to break from dogma, tradition, and received wisdom to improve patient care, and we hope this change will be met with the same attitude.

A *P* value estimates the probability of getting a result as extreme or more extreme than that observed, assuming that the null hypothesis is true and the study was free of bias or measurement error.⁹ In other words, it tells us about the probability of data given a hypothesis.¹² For example, a one-sided *P* value of .04 means that, if the null hypothesis is true (eg, that Drug A and Drug B are equivalent), there is a 4% chance of seeing the observed data or data even more extreme. It does not mean that the probability of the drugs being equivalent is 4%, nor does it mean that the probability that the drugs differ is 96%, although both these misinterpretations are common.¹⁶ The problem is that clinicians and researchers are far more interested in the probability of a hypothesis, given the data (eg, there is a 73% chance that Drug A is at least 10% better than Drug B) than the other way around.¹⁷ However, this probability can only be calculated using Bayesian methods.¹⁸ In addition, *P* values confound effect size with sample size and provide virtually no information about the strength of the evidence supporting or refuting the null hypothesis.¹⁹

If *P* values do not provide useful information for clinicians, and Bayesian methods are not yet generally available, what should we use instead? Confidence intervals (CIs) are better than hypothesis tests because they emphasize estimation by providing a range of values that are plausibly compatible with the data.^{20,21} A *P* value of .06 (“not statistically significant” using an α of .05) could mean that the absolute difference was 20% with a 95% CI of -1% to 44%. It could also mean that the absolute difference was 0.02% with a 95% CI of -0.001% to 0.041%. This example illustrates how reporting a CI

shifts the focus from “is this result statistically significant?” to “are there clinically important values in this range, and is the range narrow enough for comfort?” The importance of this shift cannot be overstated and is why we encourage authors to present difference measures with their CIs instead of *P* values.

Yet CIs are not without problems.^{22,23} They are based on the same classical statistical model as hypothesis tests and share the same set of assumptions regarding randomization and the absence of bias and measurement error. They are also subject to misinterpretation. Just as the *P* value is not the probability that the null hypothesis is true, the 95% CI is not the interval that has a 95% chance of containing the true value. It is the interval that, if calculated in 100 repetitions of a study, will contain the true value in 95. This meaning is awkward, not directly relevant (who cares about trials that did not happen), and nonintuitive. The interval that has a 95% chance of containing the true value is called a posterior interval (also called the credible interval or credible set) and can be obtained only from a Bayesian analysis.^{18,22-25}

Because CIs share many of the problems of *P* values, we ask that authors use them with appropriate caution. First, CIs should not be used as a back door to hypothesis testing. By determining whether the CI includes or excludes the null value (0 for differences, 1 for ratios), one can use the CI for hypothesis testing. We discourage this because it ignores the other useful information contained in the CI. Second, authors should recognize that the CI only reflects sampling uncertainty. It does not capture other forms of uncertainty and bias. There are classical and Bayesian techniques that can be used to adjust CIs for common forms of bias.²⁶ Finally, as noted previously, CIs are not posterior intervals and should not be interpreted as such. We encourage CIs because they are a step in the right direction but urge authors to be forthcoming about their limitations and consider using Bayesian methods when appropriate.

This change in emphasis should not be interpreted as a ban on all *P* values. A rule banning all *P* values would be no more sensible than a rule demanding their inclusion.^{15,27} Authors should understand that no manu-

script will be rejected or accepted on the basis of a P value being included; that is not how our peer review process works. What we do require is that authors carefully consider their analytic choices and provide a reasonable rationale for their selections. During the peer review of submitted manuscripts, we as methodology and statistical editors will engage in a dialogue with the authors about their choices, if necessary, and during this dialogue, special cases and exceptions can be addressed.

There are some circumstances where reporting P values are clearly inappropriate, some where they are clearly appropriate, and inevitably, some where they are in the middle. P values should not be used when describing baseline characteristics of randomized groups,²⁸ when comparing results among subgroups,²⁹⁻³³ or when reporting the results of observational studies.³⁴

Conversely, there are some situations where summarizing an analysis by a P value is the only reasonably compact option available. For example, an analysis of variance designed to test the null hypothesis that the outcome in several groups is identical can only be reasonably summarized by giving the respective means (with CIs) and a P value. Most other multiple degree of freedom tests would also qualify here. Types of survival analysis, analysis heterogeneity, receiver operating characteristic curves, or time series may also be best reported with a P value.

Intermediate circumstances, where hypothesis tests do not add much (in our opinion) would include testing a specific a priori hypothesis regarding the main outcome in a 2-group randomized controlled trial. Here, a P value (eg, $P=.03$, not $P<.05$) could be reported if accompanied by a point estimate of the difference and its CI. We discourage this practice, however, because it shifts the emphasis from estimation to hypothesis testing without adding any appreciable benefit.

Our request to emphasize estimation over hypothesis testing and CIs over P values is one part of a larger mission. That mission is to provide readers with articles that candidly and clearly describe what was done and bring them as close to the data as possible.³⁵⁻³⁷ The goal

is to empower readers with sufficient information that they can justifiably reach their own conclusions about the meaning of the study. In most circumstances, P values do not serve this mission.

Reprints not available from the authors.

Address for correspondence: Richelle J. Cooper, MD, MSHS, University of California—Los Angeles Emergency Medicine Center, 924 Westwood Boulevard, Suite 300, Los Angeles, CA 90024; 310-794-0583, fax 310-794-0599; E-mail richelle@ucla.edu.

REFERENCES

1. Rozeboom W. The fallacy of the null hypothesis significance test. *Psychol Bull.* 1960;57:416-428.
2. Cohen J. The earth is round ($P<0.05$). *Am Psychol.* 1994;47:997-1003.
3. Krueger J. Null hypothesis significance testing. On the survival of a flawed method. *Am Psychol.* 2001;56:16-26.
4. Walker AM. Reporting the results of epidemiologic studies (Different Views). *Am J Public Health.* 1986;76:556-558.
5. Gallagher EJ. $P<0.05$: threshold for decerebrate genuflection. *Acad Emerg Med.* 1999;6:1084-1087.
6. Salsburg DS. The religion of statistics as practiced in medical journals. *Am Stat.* 1985;39:220-223.
7. Krantz DH. The null hypothesis testing controversy in psychology. *J Am Stat Assoc.* 1999;44:1372-1381.
8. Sterne JAC, Smith GD. Sifting the evidence—what's wrong with significance tests? *BMJ.* 2001;322:226-231.
9. Rothman KJ, Greenland S. Approaches to statistical analysis. In: Rothman KJ, Greenland S, eds. *Modern Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998:181-200.
10. Altman DG. Statistics in medical journals. *Stat Med.* 1983;1:59-71.
11. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.
12. Schriger D. Problems with current methods of data analysis and reporting, and suggestions for moving beyond incorrect ritual. *Eur J Emerg Med.* 2002;9:203-207.
13. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol.* 1993;137:485-496.
14. Fisher R. *Statistical Methods and Scientific Inference*. 3rd ed. New York, NY: Macmillan; 1973.
15. Goodman SN. Of P -values and Bayes: a modest proposal. *Epidemiology.* 2001;12:295-297.
16. Wulff HR, Andersen B, Brandenhoff P, et al. What do doctors know about statistics? *Stat Med.* 1987;6:3-10.
17. Royall RM. *Statistical evidence: a likelihood paradigm*. Monographs on statistics and applied probability 71. 1st ed. Boca Raton, FL: Chapman & Hall/CRC; 2000:xvi, 191.
18. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. *Ann Emerg Med.* 1993;22:1328-1336.
19. Rothman K. Writing for epidemiology. *Epidemiology.* 1998;9:333-337.
20. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med.* 1994;121:200-206.
21. Young KD, Lewis RJ. What is confidence? Part 1: the use and interpretation of confidence intervals. *Ann Emerg Med.* 1997;30:307-310.

REPORTING RESEARCH RESULTS

Cooper, Wears & Schriger

-
22. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987;77:195-198.
 23. Matthews R. Bayesian statistical methods: what, why—and when. *J Altern Complement Med*. 1998;4:361-363.
 24. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130:1005-1013.
 25. Spiegelhalter DJ, Myles JP, Jones DR, et al. An introduction to Bayesian methods in health technology. *BMJ*. 1999;319:508-512.
 26. Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal*. 2001;21:579-583.
 27. Poole C. Low *P*-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291-294.
 28. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990;335:149-153.
 29. Mills JL. Data torturing. *N Engl J Med*. 1993;329:1196-1199.
 30. Smith GD, Ebrahim S. Data dredging, bias or confounding. *BMJ*. 2002;325:1437-1438.
 31. Yusuf S, Wittes J, Probstfield J, et al. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266:93-98.
 32. Greenland S, Rothman KJ. Multiple comparisons in fundamentals of epidemiologic data analysis. In: *Modern Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 1998:225-229.
 33. Matthews JNS, Altman DG. Statistics Notes: interaction 2: compare effect sizes not *P* values. *BMJ*. 1996;313:808.
 34. Colorado State University, College of Natural Resources Web site. 326 Articles/Books Questioning the Indiscriminate Use of Statistical Hypothesis Tests in Observational Studies. Available at: <http://www.cnr.colostate.edu/~anderson/thompson1.html>. Accessed June 10, 2002.
 35. Tufte ER. *Visual Display of Quantitative Information*. Cheshire, CT: Graphic Press; 1983.
 36. Schriger DL, Cooper RJ. Achieving graphical excellence: suggestions and methods for creating high-quality visual displays of experimental data. *Ann Emerg Med*. 2001;37:75-87.
 37. Cleveland WS. *Visualizing Data*. Summit, NJ: Hobart Press; 1993.