

David L. Schriger, MD, MPH

From the UCLA Emergency Medicine Center, UCLA School of Medicine, Los Angeles, CA.

Editor's note: This new feature, "Brief Commentary," is a discussion focusing on 1 or 2 key points about the study on which it is written—strengths, weaknesses, where it fits in the context of other studies, controversies, how it should or should not change our clinical practice, or even how it illustrates some important principle of science or methodology. It is not meant to be as long or complete as an "Editorial," but more of a focused comment that does not attempt to be a complete discussion of a paper.

How Do We Draw Inference From "Negative" Studies?

See related article, p. 57.

[*Ann Emerg Med.* 2003;41:69-71.]

Every research paper published in *Annals of Emergency Medicine* is peer reviewed by both content reviewers and methodology and statistical reviewers. As part of the review process for the Mion et al¹ manuscript, the paper's editor noted that "the authors did not meet their recruitment goal of 800 subjects (they ended up with 650)" and asked the methodology reviewer, "since they report no important difference in groups for most of their end points, could this be a type 2 error?" How do we answer this editor's question, a question that arises often with trials that do not reach "statistical significance"?

Unfortunately, the answer depends on the context. Most investigators regrettably continue to use a hypothesis testing model to conduct clinical research. In this model, the investigators determine the sample size needed for a clinical trial of a dichotomous out-

Copyright © 2003 by the American College of Emergency Physicians.

0196-0644/2003/\$30.00 + 0

doi:10.1067/mem.2003.25

come measure by: (a) stating the likely result in the control group, (b) stating the smallest improvement in the intervention group that would be important to detect, (c) specifying an α level (the probability of a type I error, that is, deeming an observed difference statistically significant when the difference was actually the result of chance), (d) specifying the desired power (the probability that the study will be statistically significant if the true difference is that specified in b), and (e) calculating the number of participants required to achieve this power for this difference. In the article by Mion et al,¹ the authors determined that approximately 400 patients in each study limb were needed to achieve 90% power with an α of .05 to detect an absolute decrease of 10% (from 30% to 20%) in emergency department return rates.

The hypothesis testing model leaves the investigators with only one option for interpreting their trial. They must calculate a P value and compare it to α . If this P value is less than α , they declare that the null hypothesis (no difference between limbs) is rejected. If, as was the case in the Mion et al¹ study, the P value is greater than α , they declare that the null hypothesis cannot be rejected.

Thus, within the framework of traditional hypothesis testing statistics, the answer to the editor's question is, "This particular trial did not reject the null hypothesis." That is the only statement that can be made. This feels unsatisfying, and researchers faced with this situation commonly use a variety of strategies to patch the situation. A common one is to do post-hoc power calculations based on the number of patients actually enrolled and the observed results. This faulty practice has been eloquently critiqued by Goodman and Berlin,² yet continues to be used in the medical literature.

The solution to this dilemma is to step outside the hypothesis testing model that medical researchers have misguidedly clung to for the past half century. Hypothesis testing was not developed for clinical trials in medical research and fits poorly with the way clinicians and researchers think about experimentation.³ How often do we execute a single trial and believe that that trial

alone will lead to the categoric rejection or adoption of a diagnostic test or therapy? We conduct a trial with the goal of estimating the difference between clinical strategies and how certain we are of our estimate. From this perspective, sample size calculations should not be performed with the goal of ensuring sufficient power for a given α level and clinical difference, but to ensure sufficient precision in the estimate of effect. For example, a researcher may hope to find a difference of 10% between study limbs and may want the study to be sufficiently large that the confidence interval (CI) around this value will range from 8% to 12%. Another researcher may be content with a smaller study that will (if all goes as expected) generate a CI that ranges from 6% to 14%. When we abandon hypothesis testing, we shift the purpose of sample size planning from preparing for a single hypothesis test to achieving a desired precision for the study.

An easy rule of thumb to use when planning studies is that a study performed with the number of participants needed to achieve 90% power (as calculated using widely available software) will produce a CI for the expected difference (Δ) between groups that ranges from approximately $\Delta - 0.6 \Delta$ to $\Delta + 0.6 \Delta$, and a study using the number of participants needed to produce 80% power will yield a larger CI than roughly spans $\Delta - 0.7 \Delta$ to $\Delta + 0.7 \Delta$.² Thus, in the Mion et al¹ study, if 824 participants had been enrolled (the number needed for 90% power for a 10% difference from 30% to 20%), then the expected CI would have been 4% to 16%. Although many sample size calculation software packages perform only traditional power calculations, most statisticians can use Monte Carlo simulation to calculate the expected precision of various sample sizes, and there exists off-the-shelf software that does a good job with this task (see <http://www.studysize.com> for one example).

Within the estimation framework, we are not restricted to asking the single question, "Is the hypothesis rejected?" We are permitted to make interpretations based on the CI for the observed result. For example, it can be calculated from Mion et al¹ that 30-day ED utilization in the intervention group was 3% (95% CI -6% to 12%) higher than in the control group. If we assume

no prior knowledge about the topic (a “noninformative prior” in Bayesian terminology), this interval can be used to answer the editor’s question. The interval tells us that the most likely true value is that the intervention increases the number of ED reusers by 3%, but that there is a 95% chance that the true value lies between a 9% increase and a 6% decrease. It also tells us that the true value is unlikely (<5% chance) to lie outside this range. Therefore, if the investigators truly believed that the intervention would not be worth doing unless it produced at least a 10% improvement, they could feel quite comfortable deciding that this particular intervention was not worth pursuing. On the other hand, if the investigators thought that a 4% improvement was important, then this study is not inconsistent with such a true value, and it may be premature to scrap this intervention. This example illustrates how effect estimation produces more meaningful and relevant interpretations than hypothesis testing.

The effect estimation model can be further enhanced by the consideration of likelihoods or the use of Bayesian priors, a detailed consideration of which is beyond the scope of this commentary.^{4,5} Briefly, however, in the Bayesian framework we incorporate prior information into our interpretation of the current study’s results. If, instead of having no prior information about this issue, we were aware of 10 large randomized trials, each of which produced a 7% to 15% decrease in utilization, there would be more support for the notion that the current trial, through chance or bias, has produced spurious results. We would not call this “type II error” because we are not testing a hypothesis, but by synthesizing the old and new knowledge into a new estimate of true effect, we would be incorporating the belief that the data from the current experiment may not be the best estimate of the truth.

In summary, the hypothesis testing model is the wrong framework for conducting clinical research. It is through the thorough assessment of CIs, likelihood ratios, or Bayesian posterior intervals that the most meaningful interpretations of study results can be achieved.

Reprints not available from the author.

Address for correspondence: David L. Schriger, MD, MPH, UCLA Emergency Medicine Center, 924 Westwood Boulevard, #300, Los Angeles, CA 90024; 310-794-0593, fax 310-794-0599; E-mail schriger@ucla.edu.

REFERENCES

1. Mion LC, Palmer RM, Meldon SW, et al. Case finding and referral model for emergency department elders: a randomized clinical trial. *Ann Emerg Med.* 2003;41:57-68.
2. 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.
3. Schriger DL. Problems with current methods of data analysis and reporting, and suggestions for moving beyond incorrect ritual. *Eur J Emerg Med.* 2002;9:203-207.
4. Royall RM. *Statistical Evidence: A Likelihood Paradigm. Monographs on Statistics and Applied Probability 71.* Boca Raton, FL: Chapman & Hall/CRC; 1997.
5. Matthews RAJ. Why should clinicians care about Bayesian methods? Available at: <http://ourworld.compuserve.com/homepages/rajm/jspib.htm>. Accessed September 13, 2002.