American Journal of Gastroenterology
© 2004 by Am. Coll. of Gastroenterology
Published by Blackwell Publishing

ISSN 0002-9270
doi: 10.1111/j.1572-0241.2004.40592.x

METHODOLOGY

The Value of a *p*-Valueless Paper

Jason T. Connor, M.S.
Department of Statistics and H.J. Heinz III School of Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania

As is common in current biomedical research, about 85% of original contributions in *The American Journal of Gastroenterology* in 2004 have reported *p*-values. However, none are reported in this issue's article by Abraham *et al.* who, instead, rely exclusively on effect size estimates and associated confidence intervals to summarize their findings. Authors using confidence intervals communicate much more information in a clear and efficient manner than those using *p*-values. This strategy also prevents readers from drawing erroneous conclusions caused by common misunderstandings about *p*-values. I outline how standard, two-sided confidence intervals can be used to measure whether two treatments differ or test whether they are clinically equivalent.

(Am J Gastroenterol 2004;99:1638–1640)

INTRODUCTION

Within this issue of *The American Journal of Gastroenterology* is an unusual article. Abraham *et al.*'s article, "Sedation *vs* No Sedation in the Performance of Diagnostic Upper Gastrointestinal Endoscopy," (1) is a well-run double-blinded randomized controlled trial. But there is a conspicuous absence in this article, a complete lack of *p*-values and statistical hypothesis tests. There is, however, absolutely no lack of clarity in the article's conclusions.

Abraham *et al.*'s article provides an excellent opportunity to highlight the similarities and differences between statistical hypothesis tests and statistical effect size estimation. When performing hypothesis tests we as researchers are making a dichotomous decision about an unknown reality—performing a test and drawing one of two broad conclusions; for example, choosing between "The treatment is more (less) effective than the placebo," or, if we are careful, "There is not sufficient evidence to indicate the treatment and the placebo have different effects," an oftentimes ambiguous conclusion.

Instead, a more enlightening inference is to estimate the difference in effectiveness (which may be zero, may be a small, clinically insignificant difference, or a large clinically relevant difference) between two or more treatments or groups and then to indicate our confidence in our estimate. By providing estimates and statements about their certainty, we allow our readers to draw their own conclusions regarding the important decision of whether the effect is real and clinically relevant rather than making decisions for them.

First, let us compare the different conclusions that arise when using interval estimation *versus* using hypothesis tests and their associated *p*-values. The primary result from the Abraham aricle is, "Overall, 61% of EGDE were successful, 76% active *vs* 46% placebo; (unadjusted odds ratio 3.77; 95% CI: 2.5–5.7)." The results are reproduced in Table 1.

This study estimates the odds of successful endoscopy to be 3.8 times higher in the sedated than nonsedated group; the confidence interval implies that if our estimate is true and the study were run many more times with the same sample size, 95% of the time the estimated odds ratio will be between 2.5 and 5.7. The width of this interval is inversely proportional to the precision of our estimate—the narrower the interval, the greater the precision, or confidence, of the stated estimate. Be careful—this confidence interval does not mean that there is a 95% chance that the true, albeit unknown, odds ratio is between 2.5 and 5.7.

Similarly, the investigators could have chosen to use the difference of proportions and reported, "Sedation leads to a 30% higher success rate (95% CI 21%–39%) than nonsedation," indicating that if this estimate is true, then in repeated trials, 95% of trials would see differences in rates between 21% and 39%. Such a conclusion allows the reader to judge whether he believes the possible effect sizes indicate a clinically relevant difference.

Many other authors would have chosen to state, "Overall, 61% of EGDE were successful; 76% active *vs* 46% placebo ($p < 0.0001$)." This conclusion is based upon the same statistical methodology, but the former statements using an estimate of the effect size with a confidence interval rather than a *p*-value are more informative to the practicing physician. A *p*-value fails to indicate the magnitude of the treatment effect because it incorporates two factors—estimated effect size and precision of the estimate—into one number in order to make a decision rather than an inference about effect size. Combining these two factors results in the loss of information. Therefore, it is often ambiguous whether a small *p*-value (typically indicating a difference in treatment effects) indicates a large effect size estimated with high uncertainty or a small effect size estimated with low uncertainty or whether a large *p*-value (typically indicating lack of evidence to

Table 1. Successful Endoscopies from Abraham *et al.*

	Successful	Percentage	Odds
Active	160/210	76	3.20:1
Placebo	96/209	46	0.85:1

Odds ratio: 3.77; 95% 2.5 – 5.7.
Difference of proportions: 30%, 95% CI 21 – 39%.

conclude that a difference exists between treatments) indicates no or a small effect estimated with high precision, a moderate effect estimated with low precision, or even no or a small effect estimated with low precision.

THE MISINTERPRETATION AND MISUSE OF *p*-VALUES

p-values not only provide less information and offer less decision-making power to the reader, but they are also frequently misinterpreted by both researchers and readers alike. *p*-values are also haphazardly used in the literature as they are frequently tossed into articles parenthetically while the hypotheses they are meant to test are never explicitly stated leaving the reader to guess the null hypothesis. Their misuse and misinterpretation is so common that the famous statistician, R.A. Fisher, who popularized hypothesis tests and *p*-values in the early 20th century, spent much of the rest of his life denouncing their misuse and misinterpretation applied researchers and statisticians(2).

While *p*-values have a role in statistical inference, particularly in regulatory settings, there is a growing trend in some medical journals to almost completely replace *p*-values with estimates of effect size and their confidence intervals. For instance, *Annals of Emergency Medicine* discourages *p*-values for all but the primary *a priori* test of interest (3).

We will take this opportunity to remind precisely what *p*-values mean, what they do not mean, and what their limitations are, particularly compared to using estimates of effect size and confidence intervals in their place.

For the above test, the null hypothesis is that the true, though unknown, percentage of successes in sedated patients equals the true, though unknown, percentage of successes in nonsedated patients (the placebo group), or formally, Ho: $p_{\text{active}} = p_{\text{placebo}}$. A common misconception is to interpret the *p*-value, in this example, $p < 0.0001$, as "The probability the null hypothesis is true is less than 0.0001." This all-too-common (albeit straightforward and understandably desirable) interpretation is simply incorrect. Rather, the *p*-value means that "Assuming the null hypothesis is true, that the treatments truly are equally effective, then the probability of observing a difference this large or larger is less than 0.0001." The logic for why a small *p*-value leads to rejecting the null hypothesis is that if the data are unlikely assuming the null hypothesis is true, then the null hypothesis must not be true. In this case, the observed data are extremely unlikely if the two treatments are equally effective in promoting successful endoscopy, and therefore we may conclude that the treatments are probably not equally effective. This awkward logic is difficult to fully comprehend and provides a roundabout inference, but it is the strongest inference we can make without assuming some formal prior knowledge about the treatments' efficacies, which entails Bayesian inference, a topic for another day (4).

Another common misinterpretation of *p*-values in the medical literature is a conclusion that two treatments are similar based upon a large *p*-value, *e.g.*, for a *p*-value of 0.19, authors in *AJG* recently stated "Kaplan-Meier survival of patients who had [live donor liver transplant] was similar to those who had cadaver transplantation" (5). The use of a hypothesis test and *p*-value to indicate a similarity in treatments is inappropriate unless an equivalence test is being performed (6, 7). The large *p*-value may indicate a certainty that the two treatments are nearly equally effective but it may also indicate uncertainty about the magnitude of the effect due an imprecise estimate usually due to an insufficient sample size. A confidence interval is again far more insightful. Let us imagine two hypothetical studies that both result in a test with *p*-value = 0.4 (Table 2).

A small trial of 52 patients results in 50% successes in the treatment group *versus* 38.5% successes in the placebo group. A χ^2 test shows *p*-value = 0.4. Another larger study of 4,000 patients results in 50% successes in the treatment group, and 48.65% in the placebo group but the same χ^2 *p*-value of 0.4. The two trials and their hypothesis tests result in the same *p*-values but different effect sizes between treatments (11.5% *vs* 1.35%) due to different sample sizes (52 *vs* 4000). So *p*-values alone would lead to the same inference. In the worst case, authors may incorrectly write or readers might improperly interpret both *p*-values of 0.4 to mean "The treatment and the placebo are equally effective."

AND HOW INTERVAL ESTIMATION MAKES IT EASIER

Using confidence intervals to make inferences provides greater insight and formally incorporates the difference in sample sizes into the inference. The smaller study indicates that while the observed difference was large, 11.5%, repeating similar trials would result in estimates ranging anywhere from –15% to 38%. Depending upon the disease, few clinicians would claim that the two treatments have similar success rates because the treatment could be 15% worse to 38% better

Table 2. Two Hypothetical Clinical Trials Producing the Same *p*-Value but Different Inferences

	Small Study		Large Study	
	Successful	Percentage	Successful	Percentage
Treatment	13/26	50	1000/2000	50.0
Placebo	10/26	38.5	973/2000	48.65
χ^2 *p*-value	*p*-value = 0.40		*p*-value = 0.40	
Difference (95% CI)	11.5% −15% to 38%		1.35% −1.8% to 4.4%	

than the placebo. Essentially, this hypothetical trial provided very little usable information. For the larger study, a difference of just 1.35% was observed and the confidence interval indicates that similar repeated trials would result in estimates ranging from only –1.8 to 4.4%. Many clinicians may view every possible value in this range as a clinically insignificant difference. Therefore, the p-value = 0.4 in the large study results from two treatments that are truly similar, whereas the p-value = 0.4 in the small study results from an imprecise estimate due to small trial. Another ambiguity never encountered is when we rely on effect size estimates and confidence intervals when reporting our results.

Another advantage of interval estimation is avoiding awkward equivalence trials when it is desired to prove that the two treatments are equally effective. As just demonstrated, narrow intervals around 0 (for a difference) or 1 (for a ratio) may be interpreted as resulting from clinically similar treatments. In a hypothesis testing framework, this would be the equivalent of accepting the null hypothesis—which is not permitted. Standard hypothesis tests result in the rejection of the null hypothesis, finding a difference in treatments or groups, or "failure to reject" the null hypothesis. Never as researchers do we accept the null hypothesis, though sometimes in order to prove that the two treatments are equally effective, this is exactly what we would like to do. While a variety of specialized methods called equivalence trials exist to prove that the null hypothesis is true (within clinically insignificant bounds) (6, 7), standard interval estimation most simply allows for this conclusion without these methods.

NEGATIVE TRIALS?

Another drawback that results from the permeating sanctity of $\alpha = 0.05$ in scientific literature is that trials that fail to reach statistical significance are said to be "negative trials." As shown above, however, interpreting the results of the large trial with narrow confidence intervals provides a very precise, meaningful, and relevant estimate of the treatment effect. Similarly, in the small trial the trial should not be interpreted as negative, but rather as not very informative. The treatment may be very effective. The confidence interval indicates that more data are necessary to reach a reliable conclusion.

Similarly, in large trials, a p-value may be deceivingly small due to a small, clinically insignificant effect size that happens to be estimated with high precision. Making inferences based upon the p-value alone may result in changing practices unnecessarily to a marginally better but potentially more expensive or invasive treatment. The resulting confidence interval from the same data would show a very tight interval around a point close to 0 (for a difference in treatment effects) or 1 (for a ratio of effects). In this case, reporting a confidence interval would show the reader a more informative inference, that we are very confident that there is a very small benefit.

SUMMARY

Any time a statistical hypothesis test is performed we are making a decision, and when we report the results of the test *via* a p-value, we are telling our readers the decision we made.

An increasingly popular practice, however, is to use estimation over testing so that readers can ponder the observed effect size, the range of effect sizes seen if similar trials were to be repeated (the 95% confidence interval), and then to let them draw their own decisions based upon your data and your discussion of the problem. Furthermore, when large p-values would lead to ambiguous decisions, interval estimation illuminates whether the trial fails to provide sufficient information or whether treatments truly are similar.

Reprint request and correspondence: Jason T. Connor, M.S., Department of Statistics and H.J. Heinz III School of Public Policy Management, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213.

Received April 9, 2004; accepted May 29, 2004.

REFERENCES

1. Abraham NS, Fallone CA, Mayrand S, et al. Sedation versus no sedation in the performance of diagnostic upper gastrointestinal endoscopy: A Canadian randomised controlled cost-outcome study. Am J Gastroenterol 2004;99.
2. Fisher R, Statistical methods and scientific inference. 3rd ed. New York: MacMillan, 1973.
3. Cooper RJ, Wears RL, Schriger DL. Reporting research results: Recommendations for improving communication. Ann Emerg Med 2003;41:561–64.
4. Spielgelhalter DJ, Myles JP, Jones DR, et al. An introduction to Bayesian methods in health technology. Br Med J 1999;319:508–12.
5. Maheshwari A, Yoo HY, Thuluvath PJ. Long-term outcome of liver transplantation in patients with PSC: A comparative analysis with PBC. Am J Gastroenterol 2004;99:538–42.
6. Blackwelder WC. Proving the null hypothesis in clinical trials. Controlled Clin Trials 1982;3:345–53.
7. Blackwelder WC. Equivalence trials. Encyclopedia of biostatistics. London: Wiley, 1998.