Anesthesiology 2007; 106:415–7

Importance of Effect Sizes for the Accumulation of Knowledge

Editor's Note: Starting in January 2007, all manuscripts accepted for publication in ANESTHESIOLOGY are undergoing review by Timothy Houle, Ph.D., for statistical testing and reporting. There are two reasons for this additional review. As indicated in the following editorial, authors and readers often confuse *P* values with magnitude of effects, when it is oftentimes the latter that matters most. One goal for a universal statistical review is to remind authors to report and emphasize the magnitude of the effects they observe, in both the Results and Discussion sections, rather than restricting their comments to *P* values only. In addition, medical literature, including numerous reviews and original articles in ANESTHESIOLOGY, rely on combining results from separately published reports to reach consensus conclusions. A second goal for statistical review is to assure that statistical reporting is provided in a manner that facilitates this subsequent research. For the past several decades, there has been an escalating debate regarding the appropriate techniques to evaluate scientific hypotheses. In fact, the traditional method of significance testing is now being called into question. The discussion that follows is the first of several editorials over the coming months that will examine several possible methods used to report scientific findings. These examinations will do so by looking at the strengths and weaknesses of these methods and common reporting errors, and will begin with a comparison of interpreting *P* values *versus* effect size measures. My hope for these discussions is that they will lead to clear guidelines to statistical reporting that will eventually be incorporated into the Instructions for Authors in this journal and others in our specialty.

James C. Eisenach, M.D., Editor-in-Chief

SINCE Sir Ronald Fisher, who is widely considered the father of inferential statistics,[1] there has been ongoing debate about the proper way to evaluate scientific hypotheses. In particular, the value of significance testing, or the use of a *P* value–based criterion, has come under scrutiny, with some authors asserting that the use of *P* values should be discontinued altogether.[2] Indeed, there are several convincing arguments that the reliance on *P* values to evaluate hypotheses is problematic, specifically that it is a practice plagued by misunderstanding,[3] a perversion of the entire philosophy of hypothesis testing,[3] and even a barrier to the accumulation of knowledge.[2] Despite these acrid criticisms, the use of *P* values as the sole arbiter of evaluating hypotheses remains in widespread use. This debate over the continued use of *P* values has persisted for decades and is not likely to be settled in the near future, because the need for which the use of *P* values arose remains the same: to provide a standard for evaluating the reliability of the effects under study. Considering the litany of problems accompanying reliance on *P* values, what is the alternative?

A substantial improvement over traditional hypotheses testing is to report an effect size when publishing research. This conclusion certainly is not new in the general scientific sense[3,4] and already has been suggested in anesthesiology research.[5] Reporting effect sizes provides all of the information available from "traditional" statistical reporting but also adds an element of the magnitude of the effect that simply cannot be ascertained from *P* values alone. Despite the advantages of effect size measures over *P* values, many scientists remain largely unaware of effect size reporting and frequently neglect to report a measure of treatment (or experimental) effect. With the goal of increasing awareness of effect size reporting, what follows is an introduction to effect sizes, a description of several commonly used effect size measures, and recommendations for the inclusion of effect size measures as a standard in research reports.

An effect size simply is an index of the magnitude of the observed relation under study.[6] Whereas a *P* value provides a dichotomous indication of the presence or absence of an effect, and at most a direction of the effect (in the case of one-tailed tests), an effect size characterizes the degree of the observed effect. The simplicity of this definition conceals the complexity of the concept, especially when one considers the wide array of possible metrics, inferences, and hypotheses that could be tested. Although there are many possible choices to measure effect size, *P* values as a measure of effect are always a poor choice and should not be used.

P values are *not* measures of effect size. Quite often, statistical significance (as indicated by the *P* value) is mistaken as indicating large effects and/or meaningful effects. But a statistically significant *P* value, regardless of its size, indicates neither magnitude nor clinical meaningfulness. This is so because statistical significance confounds the size of the effect with the size of the sample

Accepted for publication November 28, 2006. The author is not supported by, nor maintains any financial interest in, any commercial activity that may be associated with the topic of this subject.

to the extent that even the most miniscule effect can be made statistically significant with a large enough sample size; or, as Thompson[7] remarks,

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a significance test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired.

Another misconception about *P* values and one that is often translated into inappropriate use as an effect size is that they are mistakenly interpreted to reflect the probability that the null hypothesis is true.[3] This reasoning, if it were valid, would logically lead to the use of a *P* value as a measure of effect, but a *P* value reflects the probability of observing the data given the null hypothesis and not the probability of the null hypothesis given the data.[3] For example, an investigator conducts a comparison between a drug and placebo condition testing a null hypothesis that drug = placebo, and finds that for the comparison $P = 0.032$. A valid interpretation would be that there is less than a 5% probability that differences of this magnitude (or greater) would be observed if in fact there were no "true" differences between the two conditions (the observed differences caused by sampling error). An appealing, but invalid, interpretation would be that there is a less than 5% chance that the drug condition is the same as placebo. Most researchers are far more interested in discovering this second probability, but this information cannot be gleaned from a single *P* value. Understanding that *P* values are not valid measures of effect size, there are several pragmatic methods by which to communicate the magnitude of an effect.

One widely used method of analyzing and reporting research findings is the comparison of mean differences. In anesthesiology research, there are several commonly used metrics (*e.g.*, procedure time, dose information, number of treatment responders, costs) that naturally lead to meaningful effect size indices when considered in the form of mean group differences. For example, if the costs of two treatment algorithms are being studied, the raw mean difference between the two algorithms could be presented, with the results described as the difference in cost. This method is greatly confounded by several factors.

The use of raw mean differences requires that the reader be familiar with the distribution of scores around the mean. For example, consider a comparison involving two postoperative treatments of pain measured using a 0–10 numerical rating scale. Is a mean treatment difference between the two groups of 1.5 points statistically meaningful? If the SD within the groups is large (*e.g.*, 5 points), this difference is less meaningful than if the variance within the groups is small (*e.g.*, 0.25 points); the interpretation of the statistical meaningfulness of this difference depends on the variability within the groups. Therefore, when using raw mean differences as a measure of effect size, it is necessary to provide estimates of the mean difference as well as estimates of confidence around those estimates (which are created using the variability of the effect as well as sample size). The use of 95% confidence intervals is recommended for this very purpose.

The practice of reporting 95% confidence intervals surrounding a point estimate not only reflects the degree of uncertainty that is encountered anytime a sample is used as a population estimate, but also provides information akin to significance testing.[2] Evaluation of the magnitude of the raw score difference and 95% confidence band can be used to evaluate both the size and direction of the effect (as in the example of the 1.5-point difference above) as well as traditional statistical significance (if the bounds of the band do not include the value specified by the null hypothesis [usually zero], the observed effect is statistically significant by conventional hypothesis testing standards). The practice of providing confidence intervals around these observed differences is made practical by widely available software applications that routinely provide the information needed for such estimates when conducting regressions, analysis of variance, and more.

There is one additional problem, however, with using a raw score metrics as a measure of effect size. When an effect is calculated using a specific metric, it is difficult to compare the findings of a given study to other studies that use a different metric. Often, the same construct is measured using substantially different strategies, thus making the comparison of effects much more complex. Returning to the example about postoperative pain, one study could have measured pain using the 0–10 numerical rating scale, whereas another could have defined the effect as the percentage of patients who requested additional pain medication. A direct comparison between studies that present effect sizes using such different metrics or that use conceptually different outcome measures presents a major hurdle in accumulating knowledge on a topic. This is precisely the reason that standardized measures of effect size were developed.

If the metric used in a study is not inherently meaningful (*e.g.*, composite scores based on factor analysis), or when the issue under study has been examined using a wide variety of outcome measures using very different metrics, standardized effect size measures are optimal for indexing the degree of effect across multiple studies. A number of standardized effect size indices have been proposed and used.[6] A complete review of these indices and their relation to each other is beyond the scope of this introduction. Regardless of their unique aspects, all standardized effect size measures share several common features. First, each of the measures quantifies the observed differences in a metric that are not unit specific

(*i.e.*, standardized). This standardization allows for much easier comparison across studies and also encourages the use of meta-analysis. Second, the strength of each index (*i.e.*, the size of the effect) is not related to sample size (unlike traditional significance testing). Although the size of the confidence interval is affected by the size of the sample on which the effect size index is based, the point estimate of standardized effect is not.

Conclusion and Recommendation

Following the lead of a growing number of scientific organizations and journals,[8] it is highly recommended that point estimates of effect size, presented with confidence intervals around these estimates, be presented for all empirical studies. By focusing on the magnitude of the effects and not just a dichotomous statistical significance judgment, we will enhance the impact of our research. This will encourage the comparison and accumulation of findings across studies and, in so doing, improve our efforts at accumulating knowledge.[2] There are numerous effect size indices available to assist with this endeavor (these will be further introduced in later editorials), and many are described in the Instructions for Authors. Based on the sage advice of Jacob Cohen, it is recommended that raw mean score differences be used when the metrics under study are meaningful and consistent across studies.[9] When this is not the case, the use of a standardized effect size measure is optimal. At the time of this writing, standardized effect size indices are widely available on most statistical software packages for many types of analyses. However, confidence intervals for these standardized indices are not in widespread use. Therefore, a point estimate of these standardized effects will have to suffice, for now, until the demand for such calculations catches up to the need.

Timothy T. Houle, Ph.D., Department of Anesthesiology, Wake Forest University School of Medicine, Winston-Salem, North Carolina. thoule@wfubmc.edu

References

1. Fisher RA: Statistical Methods for Research Workers, 14th edition. Edinburgh, Oliver and Boyd, 1970
2. Schmidt FL: Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. Psychol Methods 1996; 1:115–29
3. Cohen J: The earth is round ($p < .05$). Am Psychol 1994; 49:997–1003
4. Bakan D: The test of significance in psychological research. Psych Bull 1966; 66:1–29
5. Kain ZN: The legend of the *P* value. Anesth Analg 2005; 101:1454–6
6. Cohen J: Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Mahwah, New Jersey, Lawrence Erlbaum, 1988
7. Thompson B: Two and one-half decade of leadership in measurement and evaluation. J Counseling Dev 1992; 70:434–8. Cited from Cohen, 1994
8. Wilkinson L, APA: Task Force on Statistical Inference: Statistical methods in psychology journals. Am Psychol 1999;54:594–604
9. Cohen J: Things I have learned (so far). Am Psychol 1990; 45:1304–12