AJNR

No Significant Difference … Says Who?

Andrew T. Trout, Timothy J. Kaufmann and David F. Kallmes

AJNR Am J Neuroradiol 2007, 28 (2) 195-197
http://www.ajnr.org/content/28/2/195

No Significant Difference . . . Says Who?

In the August 2006 issue of *AJNR*, Dr. Cloft presented a discussion of the meaning and importance of *P* values in the medical literature. As readers of the literature, many of us too often merely look at the results of a trial and take the authors at their word regarding the statistical significance of their findings. Dr. Cloft's main point was that, as readers, we need to critically evaluate the findings that are being presented to us. Not only do we need to consider the questions that Dr. Cloft posed for *positive* statistical findings (Why should I care? Is the result consistent with my experience? Were the right tests and the right numbers used?) but also we need to equally critically evaluate negative findings. If an author tells us that there was no significant difference between groups, should we believe it? This depends on many issues and raises the question of *statistical power*.

P values describe the risk of making a type I error (α) (that is, the risk of concluding that there is a significant difference between groups when in fact there is no such difference) (Table 1). Equally important, however, is statistical power, which is intimately related to the risk of making a type II error (β), which is the risk of concluding that there is no significant difference between groups when in fact such a difference exists (Table 1). More specifically, statistical power is the ability to detect a difference between 2 groups or 2 results and is defined as 1-β. Power is determined by 1) sample size (larger studies are inherently more powerful), 2) effect size (larger effects are easier to detect), 3) result variability (large standard errors/deviations blur the data), 4) the accepted α (being willing to accept lower levels of significance makes a difference more likely to be detected), and 5) the type of statistical test being used (nonparametric tests, appropriate whenever data are not normally distributed [Table 2], are by definition less powerful than parametric tests).

A review of the literature reveals that a discussion of the importance of statistical power has been taking place across many subspecialties of medicine. It is disturbing, however, to note that authors in various fields (Interventional Radiology, Cardiology, OB/GYN, Orthopedics, Family Practice, and others) have found that large numbers of manuscripts in the literature are underpowered to detect a difference between the patient groups.[1-6] We suspected the same situation existed in the *AJNR*. However, it is problematic to determine power in a post hoc fashion. Thus, in an effort to explore whether articles published in the *AJNR* had adequate power, we undertook to determine whether, among the manuscripts describing no significant difference, the authors provided power information or confidence intervals to the reader. In other words, was there adequate information provided to the reader to allow them to interpret statistically nonsignificant results?

On August 8, 2006, we searched all past issues of *AJNR* available on-line (abstracts from January 1980 to November 1994; full text from January 1995 to the present) by using the search term "no significant difference." This query returned 372 articles. Review articles, editorials, commentaries, and the few articles for which only abstracts were available were excluded ($n = 12$). We briefly reviewed the abstract of each of the remaining articles to identify those in which the primary conclusion was a finding of no difference ($n = 43$). Two additional articles were subsequently excluded after review indicated that the negative finding was not a substantial focus of the study, and another article was excluded because it was clearly identified as a pilot study. Each of the 40 articles were then reviewed in detail to assess for study design (prospective versus retrospective), study type (clinical versus laboratory), statistics type (parametric versus nonparametric), discussion of sample size calculations or statistical power, and presence of confidence intervals for the described results. The results of this search are given in Table 3, but the important finding is that only 3 (7.5%) of the manuscripts reported the level of power that the study had to detect a significant difference.

Table 1: Error types in statistical analysis

	Reality	
Result	Difference Exists between Groups in Population	No Difference Exists between Groups in Population
Accept null hypothesis	Type II error (β)	True
Reject null hypothesis	True	Type I error (α)

Table 2: Examples of nonparametric statistical tests

χ^2
Fisher exact
Wilcoxon signed-rank
Kruskal-Wallis
Mann-Whitney U
Spearman rank correlation

The issue of statistical power is clearly not on our radar screen as authors, editors, peer reviewers, and readers. None of us would accept a claim of statistical significance in the absence of numeric support and yet apparently we are willing to accept a claim of nonsignificance without the necessary support. We believe that this stems from a lack of understanding of the concept and importance of statistical power in the literature. The purpose of this article is to educate ourselves, and *AJNR* readers and authors, and to push for advancement in the statistical quality of the literature being published in this journal.

Power for Authors

Ideal study design incorporates both hypothesis generation and sample size calculations. Sample size calculations incorporate a predefined level of statistical power. The author defines the effect size of interest, the α value (0.05 by convention), and the degree of power he or she would like to have to detect a significant effect. She or he then uses these values and either known or predicted standard deviations to calculate the required sample size (equations for calculation of sample size are readily available in the statistical literature). Ideally, we should determine sample size through a prospective calculation; all too often, however, we arbitrarily select a sample size for our study. Given that sample size is defined by a fixed equation incorporating the variables de-

Table 3: Results of search of *AJNR* articles with the term "no significant difference" on August 8, 2006

	n	%
Study design		
Prospective	28	70
Retrospective	10	25
Indeterminate*	2	5
Study type		
Lab/Animal	7	17.5
Statistics		
Parametric	22	55
Nonparametric	11	27.5
Both	7	17.5
Clearly indicated power	3	7.5
Raised the issue of power	6	15
Provided confidence intervals for results	2†	5

* Study design could not be determined from the *Methods* section.
† Neither of these studies indicated the power of their analysis

scribed above, it is easy to see that by arbitrarily selecting a sample size, we force the other elements of the equation (including statistical power). This is problematic both practically (a sample size that is too small forces either a low-powered study or an adequately powered study that can only detect a large effect size) and ethically (excessive or useless patient exposure in inappropriately powered studies). More importantly, an underpowered study may find no significant difference even in the presence of a real difference in the population. It is in everyone's best interest that prospective sample-size calculations are performed. Although describing these calculations is beyond the scope of this manuscript, it is important to note that the convention is to set statistical power at 80%–90% (just as α [*P*] is typically set at 5%).

As authors it is not enough to simply perform sample size calculations. These calculations should be adequately described in the methods section of published articles. That is, we should tell our readers the size of the effect we were looking for and the power of our statistical analysis. At the very least, if we are not performing sample size calculations, we owe it to our readers to describe the strength of our analyses. We should at least provide for the reader a prospective assessment of the power of our analysis so they may accurately interpret negative findings. This can be calculated using the sample size equations, defining the effect size of interest (a calculation only meaningful if it incorporates an effect size of interest that was defined before the analysis was performed), and back-calculating the power of the analysis for our given sample size.

Power for Readers

From the readers' perspective, power is important in interpreting the results of a study. When an author states that "no significant difference exists," the meaning of this finding depends on whether the study had the power to detect a difference in the first place. Just as we look for a *P* value to support a significant finding, we should look to power when considering a nonsignificant finding. Few readers would change clinical practice based on a conclusion of "no significant difference" from a study with minimal power to demonstrate that difference. Thus, as we are critically reading the literature, we should always ask ourselves, "Did the authors have sufficient power to detect a clinically relevant difference?" Unfortunately, unless the author provides a sample size or power calculation, this question cannot be definitively answered. One might think that it would be possible to back-calculate the power of the analysis knowing sample size and the observed effect size. However, this is problematic because power calculations are uniquely a prospective concept and are based on assumptions and pretrial data. Once the trial has been completed, if the author has not provided information regarding their prospective power calculations, or the assumptions inherent in their analysis (eg, the effect size of interest, assumed measurement variability), post hoc power calculations are largely unhelpful.[7] For example, imagine an experiment in which the researcher was interested in the extent of occlusion of cerebral aneurysms with 2 different coil types. The study results demonstrate that between the 2 groups of patients, a 14% greater rate of occlusion was observed with coil A than coil B but this result is nonsignificant on statistical analysis. Post hoc power analyses using the data provided in the manuscript may conclude that the authors did not have sufficient power to detect a significant difference between the groups given the observed 14% difference. This conclusion is faulty, however, in that we know nothing about the pretrial assumptions made by the authors. If they had designed their study to detect a 20% difference, any observed difference less than 20% falls within the group of nonsignificant effects. Post hoc power calculations using observed effect sizes less than that for which the study was designed will always underestimate the prospective power of the study.[7,8] Thus it is misleading to perform post hoc power calculations using the observed data. That being said, we as readers should still be alert to the issue of statistical power and critically analyze nonsignificant findings. Specifically, for studies that do not calculate power, it is better to assume that no differences found means that the conclusions were merely unproved.[1]

Confidence Intervals

Because of the confusion surrounding post hoc power analysis, some statisticians have suggested that the use of 95% confidence intervals should be encouraged.[7,8] Confidence intervals are useful in that they incorporate the element of power and give a more accurate representation of the findings of an analysis. Confidence intervals tell the reader exactly the range of values with which the data are statistically compatible.[7] That is, they define all of the potential results that are supported by the data. Even in the absence of a statistically significant result, some results are not supported by the data and can be ruled out. Take the example described above in which there was a 14% difference in occlusion between groups. If the 95% confidence interval for these data were −2% to 16%, it would be clear to the reader that even though the study was not able to detect a 14% difference as significant, the data were not consistent with a 30% difference.

Not only do studies published in the specialty journals have difficulty with statistical power, but randomized controlled trials in the major journals (*JAMA*, *Lancet*, *The New England Journal of Medicine*) have also been shown to be underpowered. (It is important to note that this finding, and those in the specialty journals, is based on post hoc power calculations. As described previously, these are problematic.) As readers and authors, it should be our goal to strengthen the evidence in support of our therapies. One way to do this is to regularly

incorporate statements of statistical power in our literature. We respectfully submit that, in an effort to further strengthen the value of the science presented in this journal, power figures and sample size calculations should be required elements of all published manuscripts except for descriptive studies, pilot studies, and editorials.

References

1. Bernstein J, McGuire K, Freedman KB. **Statistical sampling and hypothesis testing in orthopaedic research.** *Clin Orthop Relat Res* 2003;(413):55–62
2. Fox N, Mathers N. **Empowering research: statistical power in general practice research.** *Fam Pract* 1997;14:324–29
3. Huang W, LaBerge JM, Lu Y, et al. **Research publications in vascular and interventional radiology: research topics, study designs, and statistical methods.** *J Vasc Interv Radiol* 2002;13:247–55
4. Mittendorf R, Arun V, Sapugay AM. **The problem of the type II statistical error.** *Obstet Gynecol* 1995;86:857–59
5. Williams JL, Hathaway CA, Kloster KL, et al. **Low power, type II errors, and other statistical problems in recent cardiovascular research.** *Am J Physiol* 1997;273:H487–93
6. Moher D, Dulberg CS, Wells GA. **Statistical power, sample size, and their reporting in randomized controlled trials.** *JAMA* 1994;272:122–24
7. 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;121:200–06
8. Smith AH, Bates MN. **Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies.** *Epidemiology* 1992;3:449–52
9. Cleophas RC, Cleophas TJ. **Is selective reporting of clinical research unethical as well as unscientific?** *Int J Clin Pharmacol Ther* 1999;37:1–7

Andrew T. Trout
Virginia Mason Medical Center
Seattle, Wash
Timothy J. Kaufmann and David F. Kallmes
Department of Radiology
Mayo Clinic
Rochester, Minn