ELSEVIER

Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative

SANDER GREENLAND, MA, MS, DrPH

This article summarizes arguments against the use of power to analyze data, and illustrates a key pitfall: Lack of statistical significance (e.g., $p > .05$) combined with high power (e.g., 90%) can occur even if the data support the alternative more than the null. This problem arises via selective choice of parameters at which power is calculated, but can also arise if one computes power at a prespecified alternative. As noted by earlier authors, power computed using sample estimates ("observed power") replaces this problem with even more counterintuitive behavior, because observed power effectively double counts the data and increases as the *P* value declines. Use of power to analyze and interpret data thus needs more extensive discouragement.
Ann Epidemiol 2012;22:364–368.

KEY WORDS: Counternull, Power, Significance, Statistical Methods, Statistical Testing.

INTRODUCTION

Use of power for data analysis (post hoc power) has a long history in epidemiology (1). Over the decades, however, many authors have criticized such use, noting that power provides no valid information beyond that seen in *P* values and confidence limits (2–9). Despite these criticisms, recommendations favoring post hoc power have appeared in many textbooks, articles, and journal instructions, especially as a purported aid for interpreting a "nonsignificant" test of the null. Although such recommendations have dwindled in mainstream journals, as Hoenig and Heisey note (6), a search on "power" through journal archives reveals that the practice and its encouragement survives (10). Furthermore, it is still common in internal reports, especially for litigation, where it may be used to buttress claims of study adequacy when in fact the study has inadequate numbers to reach any conclusion.

Statistical power is the probability of rejection ("significance") when a given non-null value (the alternative) is correct. That is, power is the probability that $p < \alpha$ under the alternative, where α is a given maximum allowable type I error (false positive) rate. Among the problems with power computed from completed studies are these:

1. Irrelevance: Power refers only to future studies done on populations that look exactly like our sample with respect to the estimates from the sample used in the power calculation; for a study as completed (observed), it is analogous to giving odds on a horse race after seeing the outcome.
2. Arbitrariness: There is no convention governing the free parameters (parameters that must be specified by the analyst) in power calculations beyond the α-level.
3. Opacity: Power is more counterintuitive to interpret correctly than *P* values and confidence limits. In particular, high power plus "nonsignificance" does not imply that the data or evidence favors the null (6).

The charge of irrelevance can be made against all frequentist statistics (which refer to frequencies in hypothetical repetitions), but can be deflected somewhat by noting that confidence intervals and one-sided *p* values have straightforward single-sample likelihood and Bayesian posterior interpretations (11, 12). I therefore review the arbitrariness and opacity issues with the goal of illustrating them in simple numerical terms. I then review how "observed power" (power computed using sample estimates), which is supposed to address the arbitrariness issue, aggravates the opacity issue. Like many predecessors (2–9), I conclude that post hoc power is unsalvageable as an analytic tool, despite any value it has for study planning.

THE ARBITRARINESS OF POWER

A *P* value has no free parameter and a confidence interval has only one, α, which is inevitably taken to be 0.05. In contrast, in addition to α, power also depends on the alternative and at least one background parameter (e.g., baseline incidence); because there is no convention regarding their choice, power can be manipulated far more easily than a *p* value or a confidence interval. The reason for lack of

From the Department of Epidemiology and Department of Statistics, University of California, Los Angeles, Los Angeles, CA.
Address correspondence to: Sander Greenland, MA, MS, DrPH, University of California, Department of Epidemiology and Department of Statistics, Campus 177220, Los Angeles, CA 90095-1772. Tel.: +1 310 455 1197; Fax: +1 310 455 1428. E-mail: lesdomes@ucla.edu.
Received October 28, 2011. Accepted February 3, 2012. Published online March 3, 2012.

1047-2797/$ - see front matter
doi:10.1016/j.annepidem.2012.02.007

Selected Abbreviations and Acronyms

FDA = U.S. Food and Drug Administration
RR = relative risk

convention is not hard to understand: The alternative and any background parameter are too context specific (even more context specific than an α-level).

The following example, although extreme, is real and illustrates the plasticity of power calculations compared with *P* values and confidence intervals. While serving as a plaintiff statistical expert concerning data on the relation of gabapentin to suicidality, I was asked to review pooled data from randomized trials as used in a U.S. Food and Drug Administration (FDA) alert and report (13) regarding suicidality risk from anti-epileptics (the class of drugs to which gabapentin belongs) and defense expert calculations. The defense expert statistician (a full professor of biostatistics at a major university and ASA Fellow) wrote:

> Assuming that the base-rate of suicidality among placebo controlled subjects is 0.22% as stated in the FDA alert, we would have power of 80% to detect a statistically significant effect of gabapentin relative to placebo for gabapentin alone in the 4932 subjects (2903 on drug and 2029 on placebo) used by FDA in their analysis, once the rate for gabapentin reached 0.70%, or a relative risk of 3.18. This computation reveals that even for the subset of gabapentin data used by FDA in their analysis, a significant difference between gabapentin and placebo would have been consistently detected for gabapentin alone, once the incidence was approximately three times higher in gabapentin treated subjects relative to placebo (14, p. 7).

The computation and conclusion do not withstand scrutiny. With regard to problem 2 above, note that

(a) There were only 3 cases observed in the 28 placebo-controlled gabapentin trials contributing to these numbers, and only one case among the placebo groups; thus, actual observed baseline rate in the gabapentin trials was 1/2029 = 0.05%. The figure of 0.22% used in the expert's calculation was more than four times this rate; it is not from placebo-controlled trials of gabapentin, but is instead from all 16,029 placebo controls in 199 randomized trials of all types of anti-epileptics. The gabapentin trial controls are only 2029 of 16,029 or 13% of these controls; furthermore, only 7% of the gabapentin trial patients were psychiatric (high suicide risk), compared with 29% of patients in other trials (13, Table 8), so the lower rate in gabapentin controls is unsurprising.

(b) The value of the relative risk (RR) as 3.18 in the power calculation is back-calculated to produce 80% power, rather than determined from context; for example, there was no plaintiff claim that an effect this large was present. In many legal contexts, a guideline used for tort decisions is instead RR = 2, based on the common notion that this represents a (2 − 1)/2 = 50% individual probability of causation. This notion is incorrect in general, but tends to err on the low side of the actual probability of causation at RR = 2 (15–17); thus, RR = 2 is still useful as a pragmatic upper bound on the RR needed to yield 50% probability of causation.

If one uses the baseline rate of 0.22% cited by the expert, the power for detecting RR = 2 is under 25%; if one uses instead the 0.05% seen in the gabapentin trials, the power for detecting RR = 2 is under 10%. Thus the power reported by the defense expert was maximized by first taking the higher risk population as the source of the baseline rate, and then finding an RR that would yield the desired power.

Regardless of one's preference, the figures illustrate the dramatic sensitivity of the power calculations to debatable choices. Of course, all the powers are arguably irrelevant to inference (problem 1) (4–9): The mid-*P* 95% odds-ratio confidence limits (8, Ch. 14) from the same combined data are 0.11, 41, whereas the approximate risk-ratio limits (8, Ch. 14) after adding ½ to each cell are 0.15 and 8.8, both showing that there is almost no information in the gabapentin trials about the side effect at issue.

POWER IN A PERFECT RANDOMIZED TRIAL

In the previous example, the low adverse event rate in controls severely limited the actual (before trial) power and after trial precision. However, genuinely high power can coincide with nonsignificance, regardless of whether the power is computed before the study or from the data under analysis. This phenomenon seems to especially challenge intuitions. Hence, I provide a simple, hypothetical example (with reasonable rates for common safety evaluation settings) in which there is high power for RR = 2 and the *P* value for testing RR = 1 (the null *P* value) exceeds the usual significance cutoff α of 0.05, yet standard statistical measures of evidence favor the alternative (RR = 2) over the null (RR = 1). The example is designed to exclude other issues such as bias, with a rare outcome and large case numbers to keep the computations simple (although the figures resemble those seen in large postmarketing evaluations).

Suppose a series of balanced trials randomize 1000 patients to a new treatment, 1000 to placebo treatment,

AEP Vol. 22, No. 5
May 2012: 364–368

TABLE 1. Hypothetical randomized trial data exhibiting "nonsignificance" and high power, yet evidential measures favor RR = 2 over RR = 1

	New treatment	Placebo
Adverse events	48	32
Total	1000	1000

with no protocol violations, losses, unmasking, and so on, leading to the combined data in Table 1.

From conventional 2 × 2 table formulas treating the log RR estimate as an approximately normal variate (see Appendix) we would then find

- $P = .07$ (and thus "not significant at the .05 level") for the null hypothesis that the RR is 1.
- Assuming the 32 events observed arm were as expected in the placebo group, the power for RR = 2 at $\alpha = 0.05$ computed from these data is over 85%.

Based on these results, do the data favor RR = 1 over RR = 2?

Here are some relevant statistics to answer the question:

a) The RR estimate is 1.50; in proportional terms, 1.50 is closer to 2 than to 1.
b) The 95% confidence limits are 0.97 and 2.33; in proportional terms, 1 is closer the lower limit than 2 is to the upper limit.
c) The likelihood ratio comparing RR = 2 vs. RR = 1 is about 2.3.
d) The P value for RR = 2 is 0.20, 3 times the p value for RR = 1.
e) The value of RR having the same p value and likelihood as the null (the "counternull" (18)) is about $1.5^2 = 2.25$, which is further from the RR estimate than is 2.

Thus, despite "nonsignificance" ($p > .05$ for RR = 1) and power approaching 90% for RR = 2 at $\alpha = 0.05$, the results favor RR = 2 over RR = 1 whether one compares them using the point estimate, the confidence interval, their likelihoods, their p values, or the counternull value.

OBSERVED POWER

To avoid the arbitrariness problem, post hoc power analyses often focus on "observed power," that is, the power computed using the point estimates of the parameters in the calculation (the baseline rate and effect size). One problem with observed power is that it will make most any study look underpowered (5): In approximately normal situations with $\alpha = 0.05$, such as those common in epidemiologic studies and clinical trials, the observed power will usually be less than 50% when $p > \alpha$ (although moderate exceptions can occur (6)). In the hypothetical example, the observed power is only about 45%.

Observed power is plagued by nonintuitive behavior, traceable to the fact that the alternative used in an observed power calculation varies randomly and may be contextually irrelevant; hence, the observed power is also random like a p value, rather than fixed in advance as in ordinary power calculations (6). One consequence is that, just as a p value can be far from the false-positive (type I error) rate of the test (19), so observed power can be far from the true-positive rate (sensitivity) of the test. Even more startling is the "power approach paradox" detailed by Hoenig and Heisey (6): Among nonsignificant results, those with higher observed power are commonly interpreted as stronger evidence for the null, when in fact just the opposite is the case. Observed power is merely a fixed transform of the p value, which grows as the p value shrinks; thus, higher observed power corresponds with a lower P value and lower relative likelihood for the null (6). In other words, higher observed power implies more evidence against the null by common evidence measures, even if the evidence is "nonsignificant" by ordinary testing conventions.

Observed power also involves and encourages a double counting of data. To illustrate, consider the following statement: "We observed no significant difference ($p = .10$) despite high power." Introducing observed power alongside p gives the impression that one has two pieces of information relevant to the null. But because observed power is merely a fixed transform of the null p value, it adds no new statistical information; it just an awkward rescaling of the null p value that is even harder to interpret correctly than that p value (which is notorious for its misinterpretation (8, 20, 21) even though one-sided p values do have simple Bayesian interpretations (11)). In contrast, confidence limits cannot be constructed from a single p value, and thus do supply additional and more easily interpreted information beyond a single p value.

DISCUSSION

There are elements of arbitrariness in all analyses. For all their problems, conventions are an obstacle to manipulation of results. Thus, although a p value can vary tremendously depending what value of a measure (such as RR) is being tested, convention has decreed the null p value (e.g., for RR = 1) as one that must be included if testing is done. Of course, such conventions have side effects, and arguably many of the objections to statistical testing and p values stem from the focus on the null testing. But, as with power, these objections would be partially addressed if a conventional alternative value was always tested as well (e.g.,

RR = ½ or RR = 2 depending on the directions observed and expected for the association).

Likewise, the convention of fixing the test criterion α at 0.05 is arbitrary, but has likely prevented its manipulation. This convention has carried over into interval estimation as the nearly universal 95% level seen in both confidence intervals and posterior intervals, and remained in place despite attempts to unseat it by using a 90% level (22). From a precision perspective, however, shifting to 90% has modest implications, as it narrows approximate normal intervals by only $1 - 1.645/1.960 = 16\%$; furthermore, the reader is warned of this narrowing by the statement of 90% accompanying the interval. In contrast, power changes arising from shifts in the baseline rate or alternative can have far more spectacular impact, and yet come with no reference point, simple calculation, or even intuition to warn of this impact.

The latter arbitrariness problem has led to use of observed power, which brings a host of its own problems. Nonetheless, one might ask if observed power or the like remains useful for speculating how much power a future study would have. I would question even that much utility: The observed data are almost never the only source of information on which to base such a forecast. The alternative of interest should be at least partly determined by what effect size is considered important or worth detecting, rather than the noisy and possibly biased estimate observed from existing data.

Calculating power from data using a fixed alternative of genuine interest is a partial answer to the problems of observed power, but brings back the arbitrariness issue. And it still depends on study-peculiar features (such as the observed baseline rate and exposure allocation ratio or prevalence) that would unlikely apply to a different study population. In fact, it could be advantageous to alter these features for future studies, as power can be sensitive to design choices like allocation ratios (or case-control ratios in case-control studies), which can be improved relative to past studies.

In sum, use of power in data analysis and interpretation (as opposed to research proposals) is more prone to grave misinterpretation than are other statistics. Chief among them is the mistake that "high power" in the face of nonsignificance means the null is better supported than the alternative, a mistake still exploited in unpublished reports even if no longer common in epidemiologic articles. Thus, contrary to some articles (10) but in agreement with many others (2–9) I argue that power analysis is only useful in discussing sample size requirements of further studies; if there are specific alternatives of interest in an analysis, the *P* value for those alternatives should be given in place of power. This means, in particular, that we need to accustom ourselves and students to concepts (such as power and smallest detectable effect) that can be detrimental to inference from existing data even if they are useful for study planning.

The problem of "underpowered studies" (10, 23) that post hoc power is supposed to address is an artifact of focusing on whether $p < \alpha$ (fixed-level testing) in individual studies. A study can contribute useful data no matter how small and underpowered it is, as long as it is interpreted with proper accounting for its final imprecision. Once its data are in, "underpowered" needs to be replaced by its post-trial analog, imprecision—a problem immediately evident and addressed when using confidence intervals (4–9, 24). Unlike *p* values and power, those intervals also supply the minimum information needed to combine individual study results in a meta-analysis, which is the most direct way of addressing imprecision.

REFERENCES

1. Beaumont JJ, Breslow NE. Power considerations in epidemiologic studies of vinyl chloride workers. Am J Epidemiol. 1981;114:725–734.
2. Cox DR. The planning of experiments. New York: Wiley; 1958.
3. Greenland S. On sample-size and power calculations for studies using confidence intervals. Am J Epidemiol. 1988;128:231–237.
4. Smith AH, Bates M. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology. 1992;3:449–452.
5. Goodman SN, Berlin J. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200–206.
6. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
7. Senn S. Power is indeed irrelevant in interpreting completed studies (letter). BMJ. 2002;325:1304.
8. Rothman KJ, Greenland S, Lash TL, eds. Modern epidemiology. 3rd ed. Philadelphia: Lippincott-Wolters-Kluwer; 2008.
9. Hooper R. The Bayesian interpretation of a P-value depends only weakly on statistical power in realistic situations. J Clin Epidemiol. 2009;62:1242–1247.
10. Halpern SD, Barton TD, Gross R, Hennessy S, Berlin JA, Strom BL. Epidemiologic studies of adverse effects of anti-retroviral drugs: how well is statistical power reported? Pharmacoepidemiol Drug Safety. 2005;14:155–161.
11. Cox DR, Hinkley DV. Theoretical statistics. New York: Chapman and Hall; 1974.
12. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J Am Stat Assoc.. 1987;82:106–111.
13. Office of Biostatistics. Statistical review and evaluation: antiepileptic drugs and suicidality. Bethesda, MD: U.S. Food and Drug Administration; 2008.
14. Gibbons RD. Supplemental expert report of March 19, 2009 in re: Neurontin Marketing, Sales and Liability Litigation, U.S. District Court of Massachusetts (Case 1:04-cv-10981-PBS).
15. Robins JM, Greenland S. The probability of causation under a stochastic model for individual risks. Biometrics. 1989;46:1125–1138 [Erratum: 1991;48:824.].
16. Greenland S. The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem. Am J Public Health. 1999;89:1166–1169.
17. Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. Jurimetrics. 2000;40:321–340.

18. Rosenthal R, Rubin DB. The counternull value of an effect size: a new statistic. Psychol Sci.. 1994;5:329–334.
19. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. Am Stat. 2001;55:62–71.
20. Goodman SJ. A dirty dozen: twelve P-value misconceptions. Semin Hematol. 2008;45:135–140.
21. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. Jurimetrics. 2011;51: 113–129.
22. Rothman KJ. Modern epidemiology. Boston: Little Brown; 1986.
23. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285:1987–1991.
24. Poole C. Low *P* values or narrow confidence intervals: which are more durable? Epidemiology. 2001;12:291–294.

APPENDIX.

Statistics for Table 1 were computed from the usual normal approximation to the log risk-ratio estimator $\hat{\beta}$ (the Wald method), where β is the log risk-ratio parameter ln(RR) (8,Ch. 14). Suppose the sample (observed) log risk ratio is b and the estimated asymptotic standard deviation of $\hat{\beta}$ is s. Let $\Phi(z)$ is the standard cumulative normal distribution (area below z). Then $\Phi(-z) = 1-\Phi(z)$ is its complement and the following approximations are useful for tables in which all counts exceed 4:

1) The 95% confidence limits for RR are $\exp(b \mp 1.96s)$.
2) The one-sided P values for $RR \leqslant e^{\beta}$ and $RR \geqslant e^{\beta}$ are $\Phi(-(b-\beta)/s)$ and $\Phi((b-\beta)/s)$.
3) The two-sided P value for $RR = e^{\beta}$ is $2\Phi(-|b-\beta|/s)$, twice the minimum of the 1-sided P values.
4) The rejection rates of the one-sided 0.025-level tests of $RR \leqslant 1$ and $RR \geqslant 1$ given $RR = e^{\beta}$ are $\Phi(\beta/s-1.96)$ and $\Phi(-\beta/s - 1.96)$.
5) The power of the two-sided 0.05-level test of $RR = 1$ given $RR = e^{\beta}$ is the sum of the one-sided 0.025-level rejection rates, $\Phi(\beta/s - 1.96) + \Phi(-\beta/s - 1.96)$.
6) The likelihood ratio for $RR_2 = \exp(\beta_2)$ relative to $RR_1 = \exp(\beta_1)$ is $\exp(-[(\beta_2 - b)^2 - (\beta_1 - b)^2]/2s^2)$.

Statistics for Table 1 were computed using $b = \ln(1.5)$ and $s = (1/48 + 1/32 - 2/1000)^{\cdot}$ in these formulas. Because of the large case numbers, using the two-binomial likelihood for the table instead of the normal approximation changes the answers only slightly, for example, the approximate ratio of likelihoods for RR = 2 versus RR = 1 is 2.3, whereas the exact ratio is 2.4.