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,