Post Hoc Power Analysis: An Idea Whose Time Has Passed?

Marc Levine, Ph.D., and Mary H. H. Ensom, Pharm.D., FASHP, FCCP

Using a hypothetical scenario typifying the experience that authors have when submitting manuscripts that report results of negative clinical trials, the pitfalls of a post hoc analysis are illustrated. We used the same scenario to explain how confidence intervals are used in interpreting results of clinical trials. We showed that confidence intervals better inform readers about the possibility of an inadequate sample size than do post hoc power calculations.
(Pharmacotherapy 2001;21(4):405–409)

Over the past 30 years, randomized, controlled, clinical trials have become the standard for generating the best evidence for the efficacy of drug therapy. By the late 1970s, however, it became obvious that investigators frequently were reporting results of clinical trials in which too few subjects were included to determine whether treatments under investigation were superior to placebo. Such studies lacked sufficient statistical power to detect potentially clinically important effects. One survey[1] noted this problem and sampled 71 clinical trials from the literature that had negative findings. The authors calculated that 67 of the 71 trials had a greater than 10% risk of missing a true therapeutic improvement of 25%, and that 34 of these negative trials had confidence intervals consistent with a 50% improvement. These and similar observations[2–4] led to a general recognition that clinical trials needed to be designed with great attention paid to statistical power.

Before a clinical trial is started, its power is defined as the probability of detecting an effect as large or larger than that used in the design of the trial when such an effect exists. Put another way, the power of a statistical test used to analyze the data of a trial is "the probability that the test will lead to rejection of the null hypothesis in favor of the alternative when the null hypothesis is indeed false."[5] In recent years, many research grant committees and professional journals began to expect investigators to provide details of their sample size estimate when outlining the methods of a clinical trial. With this change in practice, however, it became necessary to develop a rational approach to the interpretation of the results of negative trials (i.e., those in which the null hypothesis is not rejected). With the increasing use of power analysis in the a priori estimation of sample size, a logical extension of this concept was to recommend that negative trials be evaluated with a post hoc approach to power analysis. Authors and readers were encouraged to address the question, "What was the power of the study to detect the observed effect?"[6] It then became common for manuscript reviewers and journal editors to insist on a post hoc power analysis when authors submitted manuscripts on studies with negative results. This is still largely the case, despite that post hoc power analysis was shown by a few authors to be incorrect and misleading.[6, 7] (As anecdotal evidence, recent reviewers' comments on a research manuscript requested, "What was the power of this study to detect a real difference?" and "the power of the tests to detect differences should be given and discussed.")

We believe that post hoc power analysis should not be applied to the results of negative trials, and we encourage the rational use of confidence

From the Faculty of Pharmaceutical Sciences, University of British Columbia, and the Department of Pharmacy, Children's and Women's Health Centre of British Columbia, Vancouver, British Columbia, Canada (both authors).

Address reprint requests to Marc Levine, Ph.D., Faculty of Pharmaceutical Sciences, University of British Columbia 2146 East Mall,Vancouver, British Columbia, Canada V6T 1Z3.

intervals in both the design and interpretation of results of clinical trials.

Use of Power in Planning a Study

To address the use of power analysis after a study has been completed, we present a hypothetical study from its planning stage through interpretation of results. Throughout the scenario, we refer to the Z distribution for statistical analyses, as the literature on sample size and power is often based on the standard normal distribution. Most investigators would analyze the data based on the *t* distribution, being familiar with *t* tests. With a sample size of 25/group or more, however, the *t* distribution closely approximates the Z distribution.

An investigator is interested in the effect of a polymorphism in the cytochrome P450 (CYP) 2C19 gene[8] on the clearance of a new drug. Based on in vitro and preliminary in vivo data, she hypothesizes that the clearance of the new agent will be lower among individuals who are heterozygous for one of the CYP2C19 polymorphisms than among those who are homozygous for the wild-type allele. The clearance of the drug in those who are homozygous wild-type is expected to be approximately 6 ml/minute/kg, with a standard deviation of 1.8 ml/minute/kg. The investigator believes that if heterozygous individuals have a reduction in clearance of 25% or more, this would be clinically important; prior evidence suggests that an effect of this magnitude is possible.

The usual approach at this stage is to estimate the sample size required to detect the effect thought to be both important and possible. The concept of detection of an effect is based on the statistical notion of rejection of the null hypothesis when it is false (i.e., when there is a nonchance difference between groups). The investigator can use the concept of power to estimate a sample size for the study.[2] If she wishes to detect a difference of 25% between groups with a power of 0.9 (i.e., a 90% probability of rejecting the null hypothesis when it is false), then the β error (type II error) is 0.1 (i.e., 1 - power). The number of subjects required in each group (n) can then be calculated as follows[5]:

$$n = \frac{2\sigma^2 (Z_\alpha + Z_\beta)^2}{(\mu_1 - \mu_2)^2}$$

where n is the sample size required in each group (assuming equal-size groups), σ^2 is the expected variance (usually of the control group), Z_α is the Z value corresponding to a one-tailed α of 0.05, Z_β is the Z value corresponding to a one-tailed β of 0.1, and $\mu_1 - \mu_2$ is the difference in means to be detected (in this case, a difference of 25% relative to homozygous wild-type individuals). For the proposed study, n can be calculated to be approximately 25 from:

$$n = \frac{2(1.8)^2(1.64 + 1.28)^2}{(6 - 4.5)^2}$$

The a priori power of this study should be interpreted as follows: if it is true that there is a real difference in clearance of 25% or more between groups, then if this study is repeated a very large number of times, in 90% of those repetitions the difference that is observed will be statistically significant at a p value less than or equal to 0.05. This is important because the notion of power includes all the possible cases (10% of all repetitions) in which the result will not be statistically significant, despite the real difference between groups. By convention, investigators typically estimate a study sample size to achieve a power of greater than or equal to 80%.

Let us assume that this investigator decided to enroll 25 patients in each group, with an a priori p value for statistical significance of 0.05. At the conclusion of the study, she found that mean drug clearance ± SD was 6.3 ± 1.6 ml/minute/kg and 5.7 ± 1.4 ml/minute/kg in the homozygous and heterozygous groups, respectively (i.e., clearance was 9.5% lower in the heterozygous group; one-tailed Z test, p=0.08). This result was not statistically significant, and the author concluded that there was insufficient evidence to rule out chance as an explanation for the observed difference between groups. This conclusion is appropriate given the evidence, but authors, reviewers, and editors frequently attempt, erroneously, to obtain additional information from the data by calculating the post hoc power of the study.

Post Hoc Power of the Study

The following scenario typifies the experience that authors have when submitting manuscripts reporting results of negative trials. Based on reviewers' comments, journal editors require the investigator to perform a post hoc power analysis to determine whether the study had adequate power (usually meaning adequate power to

detect the observed effect). Although the Methods section of the manuscript included the a priori sample size estimate, the investigator complies, in order to secure publication of the manuscript, and she determines the following power analysis.

Given the a priori criterion for significance, the observed difference between groups, the sample size obtained, and the variance in the control group (homozygotes), Z_β was calculated to be 0.38. This is equivalent to a β error of 0.41, and the post hoc power of the study (1 - β) was 0.59. The author therefore concluded that the study had only a 59% probability of detecting the observed difference (9.5%), if an effect of this magnitude truly exists; thus, the sample size may have been too small to detect the effect of the polymorphic allele on the clearance of the new drug. The editors are satisfied, and they approve the manuscript for publication because they believe it is important to publish results of well-conceived and -designed trials regardless of whether the results are positive.

Inappropriateness of a Post Hoc Power Analysis

The application of power analysis to the results of a study is inappropriate for a number of reasons. First, it can lead the reader to the incorrect conclusion that the β probability (41% in the present case), based on 1 - the calculated power, is the probability that the observed result was a false-negative one. Thus, the stated or implied conclusion is that the effect may be real but that there were too few subjects in the study. This interpretation of the β value is incorrect because, like the p value, it must be interpreted as both a conditional and a frequency probability.[9]

A correct interpretation of the calculated β value from this study would be as follows: if the true difference between the groups is equal to or greater than the observed effect (9.5% in this case) and the variance is as observed, then the null hypothesis will be erroneously rejected in 41% of the cases, were this study to be repeated a large number of times. Of course, this study was performed once, and no probability can be assigned to a singular, observed result. Thus, we have no method for deciding whether this one case was a false-negative or a true-negative finding.

An analogous interpretation should be applied to the observed p value (0.08 in this case). The p value is not, as many investigators believe, the probability that the observed difference between groups occurred by chance. No probability can be applied to an observed result.[6] Correctly interpreted, the p value is as follows: if there truly is no difference between the groups, then a result as large or larger than the one observed would be expected to occur by chance alone in 8% of the cases, in our hypothetical study, were this study to be repeated a large number of times. When such is the case, an investigator has a basis for making a decision and would adopt a criterion probability (typically, 0.05) before the study, which would serve as the basis for making a decision about rejection of the null hypothesis. In our hypothetical study, the decision was not to reject the null hypothesis, as the p value exceeded the criterion. Because there is no meaningful decision criterion for the β value, no real decision can be made.

This leads to the second problem with the use of post hoc power analyses. When a clinical trial leads to a negative result, the calculation of power based on the observed results will always lead to a low value.[6] Thus, even if it were reasonable to do a power analysis of the observed result of a negative trial, it will always lead to a low value for power. As there is no decision criterion, the investigator and reader are therefore no further ahead in interpreting the negative result. There are other, more fundamental reasons for the inappropriateness of post hoc power analyses[9]; however, the noted problems should be sufficient to warrant consideration of alternative approaches to interpreting results of negative trials.

Use of Confidence Intervals for Interpretation of Results of Negative Trials

Because post hoc power calculations do not provide a meaningful method for evaluating the results of a negative trial, several authors have suggested that confidence intervals should be used instead.[6, 7] To illustrate this approach, the observed difference between the means in our hypothetical study was 0.61 ml/minute/kg. The 95% confidence interval (CI) for the difference was estimated to be -0.67 to 1.89 ml/minute/kg, based on the following:

$$95\%CI = (\mu_1 - \mu_2) \pm 1.96\left[\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

where $(\mu_1 - \mu_2)$ is the difference between the means, σ^2_1 and σ^2_2 are the variances; n_1 and n_2 are the sample sizes of the two groups,

respectively; and 1.96 is the value of Z that includes 95% of the area under the normal distribution curve. The confidence interval should be interpreted as follows: if this experiment were repeated a large number of times, each repetition would have an associated confidence interval for the difference between groups, and 95% of those intervals will contain the true difference. This study was done only once, and the confidence interval may or may not contain the true difference between the homozygous wild-type and the heterozygous subjects' clearances. However, it does provide useful information. The fact that zero is contained in the interval indicates that, given the data, we should not rule out the possibility that there is no difference between the groups. The upper and lower bounds of the confidence interval are affected by both the sample size of the study and the variance of the data and thus provide further information about the possible magnitude of the difference in clearance.

Based on data from this study, it is possible that the clearance of the drug in heterozygotes is 0.67 ml/minutes/kg greater than that in homozygotes. Conversely, the data are consistent with the possibility that the clearance of the drug in heterozygotes is 1.89 ml/minutes/kg lower than that in homozygotes. This represents a 30% reduction in mean clearance in the heterozygous group. Whereas the result does not provide sufficient evidence to reject the null hypothesis, the data are statistically consistent with as much as a 30% difference between groups.

Many authors and reviewers mistakenly take the confidence interval to be that range of values within which the true difference between groups can be said to lie with 95% probability. Although this interpretation is incorrect, it is close to what investigators actually wish to know. Another approach to evaluating results of clinical studies, the Bayesian method, is fundamentally different and leads to interpretations of results that are more intuitive for many investigators than the standard interpretations. Although there are fundamental questions about the appropriateness of Bayesian methods, the Bayesian approach increasingly is being recommended for the analysis of clinical research data and for evidence-based decision making.[10]

Editors and readers of a paper reporting a negative result should know whether the study was planned with adequate power to detect the hypothesized effect. In our hypothetical example, the study was well planned and the confidence interval around the observed difference between groups indicates that a 25% difference cannot be excluded by the data. The observed difference and variance can be useful in planning a future study to address this question. It is essential for investigators to be aware that a negative result such as this can arise for a number of reasons: the real difference between groups is less than the hypothesized amount; there is no difference between groups; the variance of the observed data was greater than anticipated; there were confounding factors in the conduct of the study or analysis of the data that led to a smaller difference than actually exists; and the real difference between groups is as great or greater than hypothesized, but this result is a case of a type II error, that is, one of the 10% of cases (when power = 0.9) expected to produce a false-negative result.

Conclusion

Post hoc power analysis based on observed results from a negative study is inappropriate and should not be recommended by reviewers or editors. Instead, reviewers should focus on whether the a priori sample size estimate was reasonable and whether the study met the target enrollment. After results are obtained, confidence intervals should be used to estimate the magnitude of effects that are statistically consistent with the data.[6, 7] Thereafter, it is up to the authors to interpret the potential clinical importance or application of the findings with due regard for the observations of other investigators.

Unfortunately, by 1994, despite the increased attention paid to sample size estimation in the literature, many negative trials apparently still had inadequate power to detect clinically important effects, and in many cases sample size estimates were either not reported or were poorly described.[11] This is a matter of concern, as studies with inadequate sample sizes can lead to discrepancies in the literature and confusion about the benefits of a particular therapy.[12] The use of confidence intervals in reporting the results of negative studies will better inform readers about the possibility of an inadequate sample size than will post hoc power calculations.

Acknowledgement

The authors thank Dr. James McCormack for his reading of an early draft of this manuscript and for his thoughtful suggestions.

References

1. Freiman JA, Chalmers TC, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. N Engl J Med 1978;299:690–4.
2. Donner A. Approaches to sample size estimation in the design of clinical trials—a review. Stat Med 1984;3:199–214.
3. Detsky AS, Sackett DL. When was a negative clinical trial big enough?: how many patients you needed depends on what you found. Arch Int Med 1985;145:709–12.
4. Makuch RW, Johnson MF. Some issues in the design and interpretation of "negative" clinical studies. Arch Int Med 1986;145:986–9.
5. Everitt BS. Statistical inference. In: Statistical methods for medical investigators, 2nd ed. London, UK: Edward Arnold, 1994:37–52.
6. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Int Med 1994;121:200–6.
7. Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiol 1992;3:449–52.
8. Xie HG, Stein CM, Kim RB, et al. Allelic, genotypic and phenotypic distributions of S-mephenytoin 4′-hydroxylase (CYP2C19) in healthy Caucasian populations of European descent throughout the world. Pharmacogenetics 1999;5:539–49.
9. Chow SL. Statistical significance: rationale, validity and utility, London, UK: SAGE Publications Ltd., 1996.
10. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Int Med 1999;130:1005–13.
11. Moher D, Duhlberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122–4.
12. Ottenbacher KJ, Maas F. How to detect effects: statistical power and evidence-based practice in occupational therapy research. Am J Occup Ther 1999;53:181–8.