

Expert Review

# Post Hoc Power Calculations: An Inappropriate Method for Interpreting the Findings of a Research Study

Michael G. Heckman<sup>1</sup> , John M. Davis III<sup>2</sup> , and Cynthia S. Crowson<sup>3</sup> 

**ABSTRACT.** Power calculations are a key study design step in research studies. However, such power analysis is often inappropriately performed in the medical literature by attempting to help interpret the findings of a completed study, instead of attempting to aid in choosing an optimal sample size for a future study. The aim of this article is to provide a brief discussion of the drawbacks of performing these post hoc power calculations, and to correspondingly suggest best practices regarding the use of statistical power and the interpretation of study results. Specifically, power analysis should always be considered before any research study in order to choose an ideal sample size and/or to examine the feasibility of properly evaluating study aims, but it should never be used in order to help interpret the results of an already completed study. Alternatively, 95% confidence intervals for effect sizes (eg, odds ratio, hazard ratio, mean difference) or other relevant parameter estimates should be used when attempting to draw conclusions from results, such as the likelihood of a type II error (ie, a false negative finding).

*Key Indexing Terms:* power; sample size; type I error; type II error; effect size

Using statistical power analysis to guide sample size decisions for a future study is an important step in the design of research studies. However, there are many instances in the medical literature in which power analysis is used incorrectly in an attempt to aid in the interpretation of the results of an already completed study. The inappropriate nature of these “post hoc power calculations” has been well documented.<sup>1–9</sup> Despite this, post hoc power calculations are still provided in the medical literature relatively frequently, and it is not uncommon for journal reviewers or researchers to request that such calculations be provided. Therefore, in a continuing attempt to address this lingering issue, the aim of this article is to provide a simple discussion of the drawbacks of utilizing post hoc power calculations, and to correspondingly suggest easily implemented best practices regarding the use of statistical power and the interpretation of study results.

## Appropriate use of power calculations

Power can be defined as the probability that a statistically significant difference or association will be observed for a future

study under a set of assumptions for a given sample size. One of these assumptions is a specified true magnitude of difference or association, which is ideally chosen to be the weakest clinically meaningful difference/association.<sup>10,11</sup> As such, the general goal of performing a power analysis when designing a clinical study (assuming that the aim is to test whether a difference between patient groups or an association among variables exists) is to choose a sample size that controls the 2 types of statistical error given a specified true effect size (eg, odds ratio [OR], hazard ratio, mean difference) in the overall patient population. Specifically, these types of statistical error are type I error (ie, a false positive finding, most often chosen to be 5%) and type II error (ie, a false negative finding, most often chosen to be 20% corresponding to 80% power; Table 1). In more general terms, power analysis aids in choosing a sample size that is large enough to allow for a reasonable probability of generating meaningful conclusions from the study data, while at the same time avoiding an excessive sample size that could result in unnecessary burdens and costs to patients and investigators.

Perhaps most obviously, using power analyses to determine the sample size of a randomized controlled trial ensures that the sample size will allow for a reasonable probability of detecting a specified clinically meaningful difference between treatment groups. For example, in a recent study by Messier et al<sup>12</sup> assessing whether high-intensity strength training reduces knee pain or knee joint compressive forces in adults with knee osteoarthritis, a sample size of 372 patients (124 in each of 3 treatment groups) was targeted. This sample size resulted in 80% power at the  $P < 0.0083$  significance level (ie, after adjusting for multiple testing) to detect a mean difference of 1.1 in Western Ontario

<sup>1</sup>M.G. Heckman, MS, Division of Clinical Trials and Biostatistics, Mayo Clinic, Jacksonville, Florida; <sup>2</sup>J.M. Davis III, MD, MS, Division of Rheumatology, Mayo Clinic, Rochester, Minnesota; <sup>3</sup>C.S. Crowson, PhD, Division of Rheumatology, and Division of Clinical Trials and Biostatistics, Mayo Clinic, Rochester, Minnesota, USA.

The authors declare no conflicts of interest relevant to this article.

Address correspondence to Dr. C.S. Crowson, Division of Clinical Trials and Biostatistics, Mayo Clinic Rochester, 200 First St SW, Rochester, MN 55905, USA. Email: crowson@mayo.edu.

Accepted for publication January 7, 2022.

Table 1. Illustration of the 2 types of statistical errors.

|                               |   | Truth  |   |
|-------------------------------|---|--|---|
|                               |   | No Difference <sup>a</sup> Between Groups or Association Between Variables | Difference <sup>a</sup> Between Groups or Association Between Variables |
| Results of the research study | No statistically significant difference between groups or association between variables | True negative finding  | False negative finding (ie, type II error) <sup>b</sup>                 |
|                               | Statistically significant difference between groups or association between variables    | False positive finding (ie, type I error)                                  | True positive finding   |

<sup>a</sup> The difference is a prespecified value that is the alternative hypothesis of the statistical test. <sup>b</sup> Note that statistical power is equal to 1 minus the probability of a type II error.

and McMaster Universities Osteoarthritis Index between treatment groups.

Power analyses can also be useful for observational studies, either in the form of a prospective study of new patients or a retrospective study on an already existing patient group where data have not yet been collected. For example, such analyses can aid in evaluating the feasibility of a rigorous analysis of aims given the study population (eg, if we wish to examine risk factors for a certain outcome, how well can we do this given the data we will generate if the outcome is rare?), and if feasible, can determine ideal sample size. Additionally, power calculations can be helpful when the sample size is already fixed but there is a need to collect extra data of interest that come with financial costs. In these situations, power analysis can help decide whether these extra costs are likely to be worthwhile, and if so, whether all samples, or a smaller subset, should be included. In short, whether or not power analyses are actually conducted, they should always be considered before any research study.

### Inappropriate use of power calculations

On the other hand, performing power analysis following completion of a study in order to aid in the interpretation of its results is inadvisable for 2 reasons: (1) such a power analysis is theoretically incorrect, and (2) there is a much better and readily available alternative. Both of these issues will be discussed herein; however, we will first address the theoretically incorrect nature of performing post hoc power calculations. To illustrate this, it is first necessary to formally define probability, since power is a specific type of probability. Probability is a numerical quantity ranging from 0 to 1 that expresses the likelihood of a future event. Notably, probability in general, and correspondingly statistical power, refers only to something that may or may not happen in the future; neither of these concepts is relevant when the event of interest has already occurred. For example, the probability that a certain team will win the Super Bowl in a given year ceases to be a meaningful concept once the Super Bowl has finished. Therefore, for this reason alone, power calculations should only be performed when planning a future study that has not yet taken place. Post hoc power calculations that are performed in reference to a previous study are never appropriate, as we already know with certainty whether or not a statistically significant finding has occurred.

In our experience, a request for post hoc power calculations is the most common statistics-related comment that is made by journal reviewers in the medical literature. Additionally, post hoc power calculations are often requested by researchers prior to manuscript submission. This may be because they believe these calculations will be helpful, because they have seen these calculations presented in the literature previously and believe they are expected, or because they are aiming to preemptively address a comment by a journal reviewer.

Why is an incorrect statistical technique requested (and presented) so often? There are several likely reasons. The first and probably most common scenario occurs when a statistically significant difference or association has not been identified, resulting in the following question: “Is the lack of a statistically significant result in this study a false negative finding that is caused by an inappropriately small sample size?” This is an important question; however, performing power calculations is not an appropriate way to address it. The request for a power calculation in this scenario generally comes in 1 of 2 forms. First, there is often a desire to estimate “observed power,” or the power that the study had to detect the observed effect size assuming the observed levels of variability. For example, if a nonsignificant OR of 1.5 was reported in the study, one might wonder what power the study had to detect that OR with the sample size that was utilized. However, if a nonsignificant finding was obtained, power will always be low to detect the observed effect size,<sup>7</sup> as observed power is directly related to the obtained *P* value, with the former providing no additional information than the latter.<sup>6</sup> Therefore, calculating observed power is completely noninformative. Second, it may be of interest to estimate the power that the study had to detect a clinically meaningful effect size (eg, an OR of 2.0 might be clinically relevant in a given study). The thought process behind both these approaches is likely that a low estimate of power could signify a false negative finding. However, ignoring the fact that power in this scenario of an already completed study is undefined as previously mentioned, such an approach would be an indirect way to address the likelihood of a false negative finding.

### A sound alternative to post hoc power calculations

Fortunately, there is an alternative calculation to post hoc power calculations that is theoretically correct and is also very often

already provided in the results that we are hoping to interpret: a 95% confidence interval (CI). A 95% CI can reasonably be thought of as a range of effect sizes that are consistent with the observed data and that the true effect size is likely to lie within, and therefore directly informs us regarding whether or not a false negative finding may have occurred. Of note, the technical and somewhat long-winded interpretation of a 95% CI is that if samples of the same size as that of the current study were repeatedly taken from the same patient population and a 95% CI for the effect size calculated for each sample, 95% of these 95% CIs would contain the true population effect size. Of course, this interpretation assumes that there is no systematic error in the estimation of the effect size, such as bias or confounding.

In general, once the all the data for a given study have been collected, power analysis no longer has a part to play, and it is best to perform the analysis and interpret the results accordingly based on 95% CIs for effect sizes (along with the effect sizes themselves). For example, if the weakest clinically meaningful effect size for a given study is an OR of 1.5, a 95% CI that ranges from 0.8 to 2.2 would indicate that a clinically meaningful association is possible, whereas a 95% CI that ranges from 0.8 to 1.3 would indicate that a clinically meaningful association is unlikely. A graphical illustration regarding how to assess the likelihood of a clinically meaningful difference based on 95% confidence limits and presence or absence of a statistically significant difference is shown in Figure 1. Notably, the width of a 95% CI for an effect size is dependent on several factors related to sample

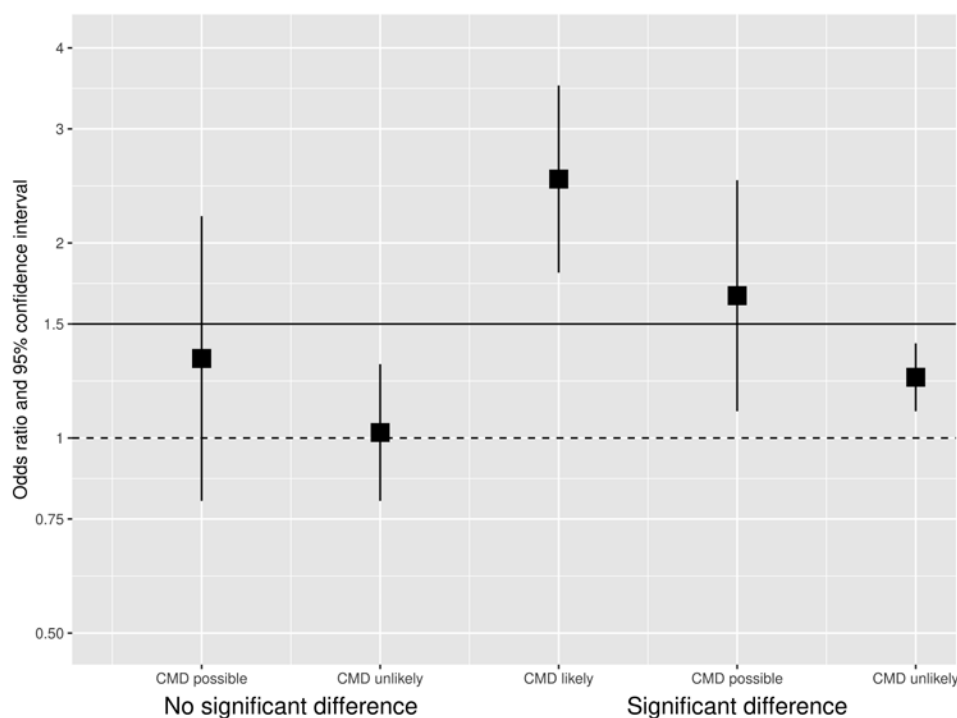
size and variability that differ depending on the hypothesis being evaluated and the types of variables being examined (ie, continuous, binary, time-to-event, ordinal). A shorter 95% CI width indicates a smaller range of likely effect sizes and therefore a more precise estimate of the true effect size.

### Comments on other power analysis scenarios

There are several other less common situations when a power analysis following the completion of a study might be requested, and we focus on 2 instances here. First, it could be of interest to estimate the power that a future study with the same sample size as the current study would have to detect a certain effect size, in order to better inform other researchers for such future studies. This is completely acceptable as long as the emphasis is solely on that future study and not on the current study that has been completed, where 95% CIs are best used to interpret the results, as previously mentioned. Second, one might have the opinion that all studies should have a power calculation and therefore any manuscript that does not contain a power statement should include one. While this is certainly a valid viewpoint at the study design stage, if a given study is already completed and a power calculation was not used to choose the sample size, performing a power calculation at that point will be of no use, as power is solely to be used to decide on the sample size of a future study.

### Suggested best practices for performing power analysis

Taking all of the above into account, 3 simple suggested best



*Figure 1.* Illustration of how to assess the likelihood of a clinically meaningful difference (CMD) based on 95% confidence limits for a scenario of comparing a binary outcome between 2 groups. Examples are provided with and without the occurrence of a statistically significant difference (ie,  $P < 0.05$ ), and all assume that an OR of 1.5 indicates a CMD in this example, which is shown with a solid horizontal line. ORs are represented by solid black square points, and 95% CIs are represented by vertical lines. A dashed horizontal line is provided for an OR of 1, indicating no difference between groups. CI: confidence interval; OR: odds ratio.

practices for performing statistical power analysis and interpreting the results of a research study are as follows (Table 2). First, power analysis should always be considered before any research study in order to choose an ideal sample size and/or to examine the feasibility of properly evaluating study aims. It should be noted that sample size decisions can also be informed by considering the precision of estimates (eg, width of 95% CIs for effect sizes) instead of, or in conjunction with, power analyses.<sup>13</sup> Second, power analysis should never be used to help interpret the results of an already completed study, or indeed for any reason other than to help inform sample size decisions for a future study. Third, 95% CIs for effect sizes (or other parameter estimates such as means, proportions, etc.) should be used along with effect sizes and *P* values when attempting to draw conclusions from results; for example, the likelihood of a false negative finding. In other words, when interpreting the results of a given research study, the best practice is to use the actual results of that study.

Table 2. Suggested best practices regarding statistical power analysis.

| Best Practice   | Should Be Avoided   |
|---|---|
| <ul style="list-style-type: none"> <li>Always consider performing a power analysis before any research study. Ideally, a statistical expert should be consulted to aid in such analysis.</li> <li>Conclusions should be made based on examination of effect sizes (eg, mean differences, odds ratios, hazard ratios) and 95% confidence intervals for those effect sizes, in conjunction with <i>P</i> values.</li> </ul> | <ul style="list-style-type: none"> <li>Performing power analysis for any reason other than to help guide sample size decisions for a future study.</li> <li>Using power analysis in order to help interpret the results of an already completed study.</li> </ul> |

## REFERENCES

- Zhang Y, Hedo R, Rivera A, Rull R, Richardson S, Tu XM. Post hoc power analysis: is it an informative and meaningful analysis? *Gen Psychiatr* 2019;32:e100069.
- Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* 1992;3:449-52.
- Nuzzo RL. Post hoc power. *PM R* 2021;13:422-4.
- Matcham J, McDermott MP, Lang AE. GDNF in Parkinson's disease: the perils of post-hoc power. *J Neurosci Methods* 2007;163:193-6.
- Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* 2001;21:405-9.
- Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* 2001;55:19-24.
- 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;121:200-6.
- Dziak JJ, Dierker LC, Abar B. The interpretation of statistical power after the data have been gathered. *Curr Psychol* 2020;39:870-7.
- Althouse AD. Post hoc power: not empowering, just misleading. *J Surg Res* 2021;259:A3-6.
- Kallogjeri D, Spitznagel EL, Jr, Piccirillo JF. Importance of defining and interpreting a clinically meaningful difference in clinical research. *JAMA Otolaryngol Head Neck Surg* 2020;146:101-2.
- Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;7:541-6.
- Messier SP, Mihalko SL, Beavers DP, et al. Effect of high-intensity strength training on knee pain and knee joint compressive forces among adults with knee osteoarthritis: the START randomized clinical trial. *JAMA* 2021;325:646-57.
- Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology* 2018;29:599-603.