The self-fulfilling prophecy of post-hoc power calculations

Christos Christogiannis, Stavros Nikolakopoulos, Nikolaos Pandis, and Dimitris Mavridis
Ioannina, Greece, Utrecht, the Netherlands, Bern, Switzerland, and Paris, France

Researchers often resort to post-hoc power calculations, and unfortunately, this practice is often encouraged by reviewers, editors, and journals. In this short article, we attempt to clarify the shortcomings of post-hoc power calculations by demonstrating its irrelevance at the analysis stage, as opposed to power calculations at the design stage of a study. We discuss the false but sometimes convenient and promising interpretation that an effect went undetected because of low power. We advise against using post-hoc power as it contains no more information than that carried by the observed *P* value.

Randomized controlled trials (RCTs) are thought to provide the highest level of evidence with respect to the evaluation of medical interventions. One of the key elements that makes evidence generated from RCTs robust is their prospective nature. Researchers have to declare in advance (before data collection) what the research question is, how such a question is formulated into a decision problem (usually by employing the null hypothesis significance testing [NHST] framework), which includes true states of nature (parameter values) and actions or decisions (reject or not reject a null hypothesis [H_0] comparison with an alternative hypothesis [H_1]). The NHST framework is governed by what is usually termed operational characteristics, that is, probabilities of correct and wrong decisions, under assumed parameter values. There are 2 possible errors in the NHST framework. The type I error (also known as significance level) is the probability to reject the null hypothesis when it is true (false positive [We use the terms positive and negative to refer to studies with statistically significant and not statistically significant results respectively]) and the type II error is the probability of not rejecting the null hypothesis when the H_1 is true (false negative). By specifying such quantities, a priori, researchers can make probabilistic statements with respect to their decision.

Am J Orthod Dentofacial Orthop 2022;161:315-7
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2021.10.008

The power of a statistical test is the probability of rejecting the H_0 when the H_1 is correct (true positive). It is the complement probability of type II error (power = 1 − type II error). Generally, we use the term power, also known as statistical power, to denote power calculations done before data analyses, which constitute standard practice in prospective RCTs on scientific and ethical grounds.[1] The statistical power is subsequently used to calculate sufficient sample sizes for an a priori determined effect to be detected with an a priori defined type I error. Statistical power is paramount for answering the research question with adequate certainty and not exposing more patients than necessary to potentially ineffective or even harmful treatments.

The power calculation is very useful for the design of a study but is not related to the analysis of the data and the treatment effect estimation.[2] Note also that for a given type I error and magnitude of effect, the statistical power is a mere function of the sample size. The larger the sample, the higher the power and vice versa. Many studies are by design underpowered; a common problem in many trials in which clinically important effects are not statistically significant.[3] This emphasizes why we should not emphasize statistical significance alone but focus more on point and interval estimation. Low statistical power is not a problem only when nonsignificant results are observed (eg, $P > 0.05$). Even if we get statistically significant results, a low-powered study can be misleading, given a large amount of uncertainty of small studies.

To illustrate the concepts, let us consider the following example from the field of orthodontics: researchers wish to evaluate the comparative efficacy of self-ligating (SLB) vs conventional brackets (CB) in the treatment of mandibular crowding, and they consider the irregularity index (II) as a primary outcome.[4] They assume that the II in untreated patients eligible for the trial has a mean of 5 mm and a standard deviation (SD) of 3 mm. They expect that SLB will, on average, reduce the II 1.5 mm more than CB, 5 months after treatment initiation, and this effect size is considered clinically relevant. They want their study to have 80% power to detect such a difference if such a difference exists and 2.5% type I

error probability (1-sided). Standard power calculations show that they need 64 patients per group (allocation ratio, 1:1) for these design criteria to be met. What do these numbers actually mean? In other words, if 64 patients per arm are included:

1. Assuming that the mean II is 5 mm, SD is 3 mm, and the H_0 of no difference is true (ie, the true difference in mean reduction of II between SLB and CB is zero, treatments are equally efficacious), the probability of falsely rejecting H_0 and thus concluding efficacy of SLB is 2.5%.
2. Assuming that the mean II is 5 mm, SD is 3 mm, and the difference at which the power is set (ie, the true difference in mean reduction of II between SLB and CB is 1.5 mm), the probability of rejecting H_0 and thus concluding efficacy of SLB is 80%.

The true difference in efficacy is unknown and will remain unknown, even after observing and analyzing data from a sample. True difference refers to a population parameter that is impossible to observe. If the true difference were known, there would be no need to conduct a study. At best, one can get an estimate with some uncertainty (usually quantified by confidence intervals). This uncertainty can be reduced by increasing the sample size. The only thing researchers can control in this decision framework is the operational characteristics (probabilities of correct and wrong decisions) under assumed parameter values (see example above).

POST-HOC POWER

Post-hoc power analysis, also known as observational or retrospective power, is conducted after the study has been completed, and it uses the observed sample size and estimated effect, along with the assumed type I error to calculate observed power.

Most times, the need or demand for post-hoc power calculations arises from the difficulty of researchers to handle the nonrejection of the null hypothesis. Post-hoc power is usually employed to calculate whether the sample size was sufficient to detect a statistically significant effect and researchers falsely believe that a low post-hoc power would justify a nonstatistically significant result.[5] Although power is often not considered an issue when a study has rejected the null hypothesis, we still see post-hoc power calculations in an effort by the researchers to appease any credibility concerns because of small sample sizes. However, an underpowered study can still be falsely positive and therefore misleading.

Although awareness was raised in the statistical and medical literature on several occasions,[2,5-8] post-hoc power calculations are persisting, occasionally motivated by journals and reviewers' requests[9] and misguided suggestions in the literature.[10]

Post-Hoc Power and P Value

We typically reject the null hypothesis by computing a *P* value less than the assumed type I error (significance level). The *P* value is the probability of observing an effect equal or larger than the one observed in the sample when the null hypothesis is true. A small *P* value (eg, <0.05) means that the estimated effect is not very likely when the null hypothesis is true. Hence, we infer that the null hypothesis is probably not true, and we reject it. It is straightforward to show that post-hoc power is a one-to-one function of the *P* value.[5] This essentially means that post-hoc power is merely a transformation of the observed *P* value and provides no further insight with respect to the observed data. The Figure shows the relationship between observed power and *P* value for a z-test. The larger the *P* value, the smaller the observed power and vice versa.

MISREPRESENTATION OF POST-HOC POWER

An important characteristic of the observed power and *P* value relationship is that for a *P* value equal to the significance level (eg, 0.05), the observed power is 50% (Fig). That said, any study with a nonstatistically significant result at a given significance level (ie, *P* value >0.05) would have at most 50% post-hoc power, or, in other words, it will be misinterpreted as underpowered. Employing such reasoning, anytime we do not find a positive effect, we can attribute it to an underpowered study and incorrectly imply that had a larger sample size been used, the treatment of interest would have been proven efficacious.[11] In contrast, any study with a statistically significant result would have an observed power of at least 50%, and for *P* values close to zero, we will have 100% observed power! Hence, post-hoc power ends up being a self-fulfilling prophecy making results more attractive as post-hoc power will corroborate a positive result and attribute a negative one to small sample size. One can think of the paradox that an adequately powered study (eg, designed to have a power of 80% for a treatment effect that was considered relevant prospectively) may have a very small post-hoc power if the trial is negative, especially if it demonstrates an effect size close to zero. This is explained because you need a very large sample to detect a true effect close to zero but detecting such an effect was not within the aims of the study. Clinical trials aim to confirm whether there is an effect that is of practical or clinical significance. In our example, we are interested in the sample size needed to detect the minimum clinically relevant effect that would lead us to suggest the use of SLB.

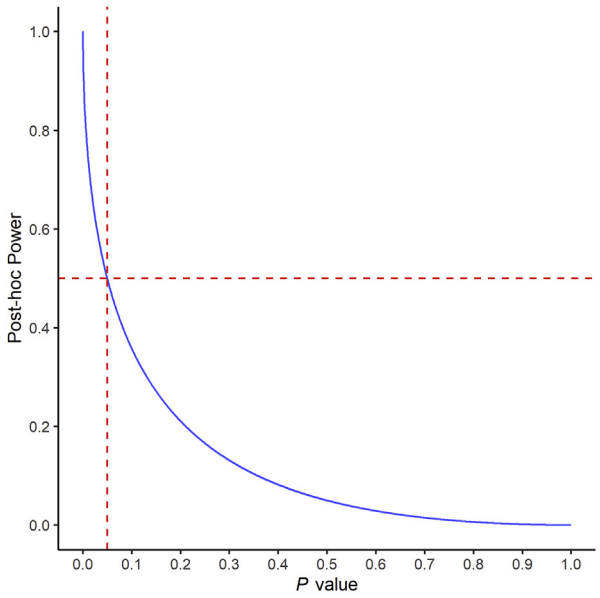

Fig. *P* value is presented in the horizontal axis and post-hoc power on the vertical axis for a z-test with a 5% significance level. By drawing a vertical dashed line in 0.05 and a horizontal dashed line in 0.5, the 2 lines are crossed. Hence, whenever a hypothesis is not rejected ($P > 0.05$), post-hoc power will always be <50%.

To support our statement, let us go back to our example and assume that the null hypothesis has not been rejected ($P > 0.05$). We calculate the post-hoc power and find a reduced power compared with a power of 80%, which was assumed when designing the trial. Note that the trial still has, and always will, a power of 80% for the effect size assumed at the design stage (1.5 mm II reduction, assuming known variance). We hypothetically postulate that an increased sample size would lead to statistically significant findings. However, any effect will become asymptotically (for very large sample sizes) statistically significant. The actual magnitude of the effect is crucial. Back to our example, we designed the study to detect a 1.5 mm II reduction, and this is the effect we believe is clinically important. Note that the larger the effect we assume, the smaller the required sample size. Post-hoc power uses the observed effect, which may be far from the effect we would want to detect to infer the treatment is efficacious.

The true question that researchers should ask themselves before conducting a study is whether they did or did not take into consideration an effect size that is close to the true effect size and whether that effect size is of clinical relevance. Ideally, one should adequately power a study (collect enough sample size) for a minimum clinically relevant effect size. In that case, any calculation of post-hoc power becomes irrelevant.

In conclusion, post-hoc power is misinterpreted as inadequate power for trials with nonstatistically significant results, and it does not provide any extra information in the analysis. On the contrary, the power of a test is a prestudy design characteristic of paramount importance. Post-hoc power is a self-fulfilling prophecy that falsely justifies any negative result as a product of a small sample size. Power is defined a priori to determine the sample size needed to estimate a certain effect with a certain type I error. We recognize that in practice, many researchers decide on the sample size based solely on feasibility and select the effect needed to have a certain level of power. This strategy, just like the use of post-hoc power, is ill-advised. Matters get worse as most software provides tools that easily estimate post-hoc power. Journals and reviewers also typically ask for post-hoc power calculations. We urge researchers to resist such demands and resort to the vast amount of literature and regulatory guidelines explaining the reasons for avoiding such practice.[5-8,12,13]

REFERENCES

1. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358-62.
2. 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.
3. Ioannidis JPA, Stanley TD, Doucouliagos H. The power of bias in economics research. Econ J 2017;127:F236-65.
4. Pandis N, Polychronopoulou A, Eliades T. Self-ligating vs conventional brackets in the treatment of mandibular crowding: a prospective clinical trial of treatment duration and dental effects. Am J Orthod Dentofacial Orthop 2007;132:208-15.
5. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001;55:19-24.
6. Zumbo BD, Hubley AM. A note on misconceptions concerning prospective and retrospective power. J Royal Statistical Soc D 1998;47:385-8.
7. Thomas L. Retrospective power analysis. Conserv Biol 1997;11:276-80.
8. Jiroutek MR, Turner JR. Why it is nonsensical to use retrospective power analyses to conduct a postmortem on your study. J Clin Hypertens (Greenwich) 2018;20:408-10.
9. Bacchetti P. Peer review of statistics in medical research: the other problem. BMJ 2002;324:1271-3.
10. Bababekov YJ, Stapleton SM, Mueller JL, Fong ZV, Chang DC. A proposal to mitigate the consequences of type 2 error in surgical science. Ann Surg 2018;267:621-2.
11. Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. BMJ 2014;348:g2215.
12. Committee for Proprietary Medicinal Products. Points to Consider on Switching Between Superiority and Non-inferiority. London, United Kingdom: European Medicines Agency; 2000.
13. Lenth RV. Two sample-size practices that I don't recommend. In: Proceedings of the Section on Physical and Engineering Sciences. Indianapolis: American Statistical Association; 2000, p. 8-11.