EDITORIAL

Inappropriate use of statistical power

We are pleased to add this typescript, *Inappropriate use of statistical power* by Raphael Fraser to the *BONE MARROW TRANSPLANTATION* Statistics Series. The authour discusses how we sometimes misuse statistical analyses after a study is completed and analyzed to explain the results. The most egregious example is *post hoc* power calculations.

When the conclusion of an observational study or clinical trial is *negative*, namely, the data observed (or more extreme data) fail to reject the *null hypothesis*, people often argue for calculating the observed statistical power. This is especially true of clinical trialists believing in a new therapy who wished and hoped for a favorable outcome (rejecting the *null hypothesis*). One is reminded of the saying from Benjamin Franklin: *A man convinced against his will is of the same opinion still.*

As the authour notes, when we face a negative conclusion of a clinical trial there are two possibilities: (1) there is no treatment effect; or (2) we made a mistake. By calculating the observed power after the study, people (incorrectly) believe if the observed power is high there is strong support for the null hypothesis. However, the problem is usually the opposite: if the observed power is low, the null hypothesis was not rejected because there were too few subjects. This is usually couched in terms such as: *there was a trend towards…* or *we failed to detect a benefit because we had too few subjects* or the like. Observed power should not be used to interpret results of a negative study. Put more strongly, observed power should not be calculated after a study is completed and analyzed. The power of the study to reject or not the null hypothesis is already incorporated in the calculation of the *p* value.

The authour use interesting analogies to make important points about hypothesis testing. Testing the *null hypothesis* is like a jury trial. The jury can find the plaintiff guilty or not guilty. They cannot find him innocent. It is always important to recall failure to reject the *null hypothesis* does not mean the null hypothesis is true, simply there are insufficient evidence (data) to reject it. As the author notes: *In a sense, hypothesis testing is like world championship boxing where the null hypothesis is the champion until defeated by the challenger, the alternative hypothesis, to become the new world champion*.

The authour include a discussion of what is a *p*-value, a topic we discussed before in this series and elsewhere [1, 2]. Finally, there is a nice discussion of confidence intervals (frequentist) and credibility limits (Bayesian). A frequentist interpretation views probability as the limit of the relative frequency of an event after many trials. In contrast, a Bayesian interpretation views probability in the context of a *degree of belief* in an event. This belief could be based on prior knowledge such as the results of previous trials, biological plausibility or personal beliefs (*my drug is better than your drug*). The important point is the common mis-interpretation of confidence intervals. For example, many researchers interpret a 95 percent confidence interval to mean there is a 95 percent chance this interval contains the parameter value. This is wrong. It means, if we repeat the identical study many times 95 percent of the intervals will contain the true but unknown parameter in the population. This will seem strange to many people because we are interested only in the study we are analyzing, not in repeating the same study-design many times.

We hope readers will enjoy this well-written summary of common statistical errors, especially *post hoc* calculations of observed power. Going forth we hope to ban statements like *there was a trend* towards… or *we failed to detect a benefit because we had too few subjects* from the Journal. Reviewers have been advised. Proceed at your own risk. Robert Peter Gale MD, PhD, DSc(hc), FACP, FRCP, FRCPI(hon), FRSM, Imperial College London, Mei-Jie Zhang PhD, Medical College of Wisconsin.

Bone Marrow Transplantation (2023) 58:474–477; https://doi.org/10.1038/s41409-023-01935-3

INTRODUCTION

The concept of statistical power was developed by Jerzy Neyman and Egon S. Pearson in the 1930s [3] following initial work by Ronald Fisher. One of its main application was in planning scientific research. However, the concept remained largely unpopular until the early 1960s. One reason for this revival was an article written in 1962 by Jacob Cohen [4]. In his review he concluded that reported studies had, on average, a 48% chance of obtaining a statistically significant result. Since then power analysis has been widely used in planning scientific research. Unfortunately, some researchers incorrectly apply statistical power retrospectively. The problem is exacerbated by statisticians who advocate its use and by statistical software packages that provide "observed power" in conjunction with data analyses. Ironically,

Received: 15 June 2021 Revised: 25 April 2022 Accepted: 13 June 2022
Published online: 3 March 2023

Cohen's use of *post hoc* power analyses in his review may have contributed to the current dilemma. Additionally, many reviewers and editors of medical journals often ask investigators to provide details of the observed statistical power of their study.

Statistical power is the probability of a statistically significant result given there is a treatment effect. In general, when there is a negative study (i.e., no treatment effect or no association between a co-variate and an outcome), it is often argued that the observed statistical power can help in the interpretation or evaluation of the study—observational or prospective. The traditional and widely accepted standard of hypothesis testing is to protect the investigator from falsely concluding a treatment is effective when it is not. Within a hypothesis testing framework, a negative study offers two possibilities (1) we fail to reject the *null hypothesis* when it is true or (2) type II error, namely we failed to reject the *null hypothesis* when the alternative hypothesis is true (Table 1). Many researchers reason that if the observed power is high and the *null hypotheses* was not rejected, this is strong evidence supporting the *null hypothesis*. Unfortunately, a hypothesis testing framework does not allow us to decipher between the two possibilities of a negative study. After the data is observed you either made an error or you did not but you can never tell if you did.

Inappropriate use of statistical power for data analytic purposes is prevalent in the research community. Many statisticians have identified the problem [5–9] but the problem remains and is strongly entrenched by its users today. Here, we describe inappropriateness of retrospective power analyses and pitfalls of using observed power to interpret results of negative studies. Instead of discussing the many mis-interpretations of statistical power, although some are mentioned, we seek to give clear definitions and explanations to help readers identify them. A concept closely related to observed statistical power is *p* value (observed significance level), probably the most misunderstood concept in all of statistics is the *p* value.

WHAT IS A *P* VALUE?

Every statistical method depends on a collection of assumptions about how the data were collected and analyzed and how the analysis results were selected for presentation. The full set of assumptions is embodied in a statistical model. The *p*-value is the conditional probability of a test statistic equal to or more extreme as that observed, given the collection of assumptions are true [2]. We discussed the *p* value in a prior typescript in this series and elsewhere in the context of haematopoietic cell transplants [10, 11].

The *p*-value encompasses the complete collection of assumptions. Hence it can be viewed as a measure of how incompatible the observed data are with the model assumptions. However, it does not tell us which assumptions are incorrect. Thus, the *p* value could be small because the *null hypothesis* is unlikely to be true. It could also be small because multiplicity (or multiple comparisons) was ignored during the analyses or because the study protocol was violated. A large *p* value tells us the observed data are not unusual given the model assumptions. It should be clear from the above definition that *p* values tell us less than what most people think they do. This definition is lacking from far too many textbooks and articles attempting to define it. From this point, we assume all the model assumptions are correct and rarely mention them again. Many of the mis-understandings about statistical power stem from a lack of a clear understanding of the concepts and definitions in hypothesis testing.

HYPOTHESIS TESTING

Hypothesis testing is perhaps the most widely used statistical analysis tool in all of research yet it is not well understood by many resulting in claims beyond the scope of what is possible. One reason for this may be the lack of understanding of the logic of hypothesis testing. The hypothesis testing method is an indirect strategy for conducting research. Hypothesis testing can be likened to the US judicial system where a person is presumed innocent until proven guilty. In a court of law, based on the evidence presented, the verdict is "guilty" or "not guilty" instead of "guilty" or "innocent". The court never declares a person as innocent; only that they are not guilty. This is a subtle but important difference. Not guilty does not mean innocent. A not guilty verdict simply means there was insufficient evidence to eliminate all reasonable doubt about guilt from the minds of the jurors.

Similarly, we assume the null hypothesis is true then based on the evidence collected (i.e., the data) we reject the *null hypothesis* or fail to reject the null hypothesis. If we reject the null hypothesis, the alternative hypothesis is accepted as the "new truth" and becomes the new "king of the hill." The expression, "fail to reject the null hypothesis" implies that there is insufficient evidence to reject the *null hypothesis* but does not necessarily mean the *null hypothesis* is true.

In hypothesis testing we cannot determine the probability a hypothesis is true. We can only assess whether the observed or more extreme data are consistent with the assumption the *null hypothesis* is true. If the data observed (or more extreme data) would be unlikely when the *null hypothesis* is assumed true, we reject the *null hypothesis* in favor of the alternative hypothesis. It is important to note that rejection of the *null hypothesis* does not establishes the truth of the alternative hypothesis but rather constitutes evidence favoring the alternative hypothesis. In a sense, hypothesis testing is like world championship boxing where the *null hypothesis* is the champion until defeated by the challenger, the alternative hypothesis, to become the new world champion.

OBSERVED STATISTICAL POWER AND *P* VALUE

Observed power should not be used to interpret the results of a statistically non-significant study. The observed power is estimated using the observed treatment effect and variability of a completed study. This can be a completed clinical trial or an observational study. Advocates of observed power argue if we fail to reject the *null hypothesis* with high observed power, this is evidence supporting the *null hypothesis*. To explain why this reasoning is incorrect we first note there is a *one-to-one* relationship between the observed power and *p* value; small *p* values are associated with high observed power [7]. Figure 1

Table 1. Possible outcomes in hypothesis testing.

The Truth	**Decision and Result**		
	Null Hypothesis		
	Reject	**Fail to Reject**	**Total**
Null Hypothesis True	α (Type I Error)	1-α (Confidence Level)	1.0
Alternative Hypothesis True	1-β (Statistical Power)	β (Type II Error)	1.0

displays this *one-to-one* relationship. In the Figure both studies are not statistically significant at the 5% level. Study A has higher observed power than study B. This implies that study A will have a smaller *p* value than study B meaning study B has stronger evidence for *favoring* the null hypothesis than study A despite having the lower observed power of the two studies. And study A has stronger evidence *against* the *null hypothesis*. The lesson: high statistical power and non-significance does not imply support for the *null hypothesis* over the alternative hypothesis. [12] High power can coincide with non-significance regardless of whether the power is computed prospectively or from the data. Additionally, based on the relationship between the observed power and *p*-value, the observed power is determined completely by the *p*-value and does not contribute to the interpretation of results [13] also illustrated via a simulation study the folly of observed power. They showed a large variability in the observed power compared with the true power indicating the observed power is not informative for any practical purpose.

PROBABILITY AND STATISTICAL POWER

Statistical power should never be used to interpret the result of analyses. Statistical power, types-I and -II error are generally used to plan studies. However, these probabilities are meaningless after data collection and statistical analyses are completed. On a fundamental level, statistical power is a probability and has a relative frequency interpretation. Statistical power is the conditional probability of rejecting the *null hypothesis* given that the alternative hypothesis is true and the model assumptions are correct. Statistical power does not refer to the likelihood of success of a particular study but rather what happens, on average, by replicating the same study many times. To many researchers this interpretation is awkward and not very helpful. In fact, some would argue this contributes to the mis-interpretation of statistical power. Additionally, a probability is a way to quantify the likelihood of an event which has not yet occurred. Once an event occurs, the probability becomes irrelevant. For example, if a person with multiple myeloma died yesterday. We would never ask, “what is the probability that the same person dies today?” Similarly, once data has been collected or observed, estimating statistical power is not meaningful. What many researchers including statisticians do is to pretend the event did not occur and calculate the so-called power using the observed data or other data. Another, more subtle, idea is: “if we had a similar person, like the one before, what is the probability this person will die?” This is one justification for calculating retrospective power. However, there are good reasons we should not adopt this approach. Because of its relative frequency interpretation, any power calculation done after data collection and/or statistical analyses refers to subsequent studies and cannot refer to the current one.

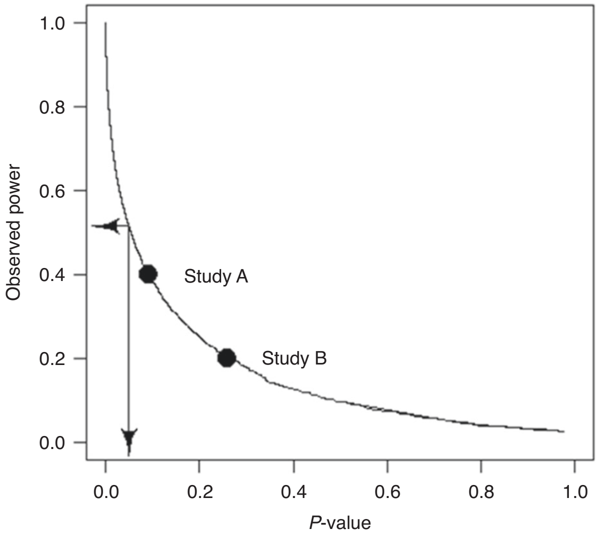

Fig. 1 Observed power as a function of the *p* value for a two-sided two-sample *t* test with significance level 5%. When the *p* value is 0.05 the estimated power is 50%.

CONFIDENCE INTERVALS AND STATISTICAL POWER

Confidence intervals are important in interpreting results of a study. They are more useful in this respect than statistical power. Confidence intervals are a way of quantifying uncertainty about an estimate (e.g., magnitude of treatment effect). We can think of a confidence interval as the set of plausible range of values for an estimate computed from sample data which is likely to include the true but unknown population value. The confidence level is a probability and indicates how likely it is that the interval estimate captures the population parameter. Thus, the confidence level has a relative frequency interpretation. If we consider all possible randomly selected samples of the same sample size from a population, the confidence level is the percentage of those samples for which the confidence interval includes the population value provided all model assumptions are correct. The confidence level expresses only how often the confidence interval procedure works on average. Importantly, it does not tell us the probability a specific interval includes the population value. Once we observe the data, that is, the data has been collected and the confidence interval constructed, the probability the interval contains the population value is either 0 or 100 percent. It either contains the population value or it does not, since any population parameter is a constant, not a random variable. We cannot properly evaluate or interpret the results (e.g., clinically important effect) of a study without knowing the confidence interval for the effect size or treatment effect.

Confidence intervals should not be used to calculate power. We gain no information from power calculations based on the plausible values outside the confidence interval because the data have already told us these are unlikely values. Power calculations based on values inside the plausible range can be very high. However, it would be mis-leading to interpret this value as representing an upper bound on the population value.

DISCUSSION

There are alternative definitions to statistical power and confidence intervals in Bayesian statistics. In frequentist statistics, these concepts do not provide the information many if not most researchers assume they do. For example, many researchers interpret a 95% confidence interval to mean there is a 95% chance this interval contains the true but unknown parameter value in the population. This is incorrect. It means, if we repeat the study many times, ninety-five percent of the intervals will contain the true but unknown parameter value. Obviously, this is an odd interpretation because we are mostly interested in the study we are conducting, not repetitions. Statistical power has the same relative frequency interpretation. For instance, a 90% power means if we repeat the same study many times, 90% of the studies would reject the *null hypothesis* when the alternative hypothesis is true. Therefore, statistical power and confidence intervals do not apply to a single study but what happens were we to repeat the same study under similar conditions many times.

In hypothesis testing we do not know the probability that the null or alternative hypothesis is true. For most researchers this idea is intuitive. Fortunately, Bayesian statistics address this issue. For example, under the Bayesian framework statistical power applies to your specific study given the data and not what happens if a study were to be repeated many times. Bayesian confidence intervals also known as credible intervals is the probability that the computed interval contains the true but unknown parameter value given the data, provided the model assumptions are correct. In Bayesian hypothesis testing we can have multiple alternative hypotheses instead of just one. We can estimate the probability the null or the alternative hypotheses are true.

The role of power calculations for data analytic purposes is conceptually flawed and analytically misleading. The idea that we can use statistical power, retrospectively or prospectively, to interpret results of a study is very common. A major reason for this mis-understanding may have to do with poor understanding of the concepts of hypothesis testing and probability. Statistical power, confidence level, type I error and type II error are all probabilities. These probabilities have a relative frequency interpretation and their use is limited to *unobserved data*. Once we have observed the data these probabilities are not helpful. There is little merit in calculating the statistical power once the results of the study are known or the data are collected and analyzed. Researchers should recognize limitations to hypothesis testing and avoid making claims beyond the realm of hypothesis testing. Many reviewers and researchers trying to interpret the results of a negative study say, "the study was underpowered" or "the sample size was too small." Hypothesizing what could have made the study successful is not helpful because performing a larger study offers no guarantee the observed treatment effect will not get smaller or introduce more variability. These quantities are unknown and remain so until after the data are collected. It is like speculating on whether we think the price of a stock will go up or down. Simply put, it is not a good research practice. We recommend statistical power only be used for planning research studies and discourage its use of interpreting results of a negative study.

Raphael A. Fraser[1]✉
[1]*Medical College of Wisconsin, Milwaukee, WI, USA.*
✉*email: raphael.fraser@gmail.com*

REFERENCES

1. Gale RP, Zhang MJ. What is the P-value anyway? Bone Marrow Transplant. 2016;51:1439–40.
2. Gale RP, Hochhaus A, Zhang MJ. What is the (P-) value of the P-value? Leukemia. 2016;30:1965–7.
3. Neyman J, Pearson ESIX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond. Series A, Contain Pap Math Phys Character. 1933;231:694–706.
4. Cohen J. The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol. 1962;65:145.
5. Cox DR. Some problems connected with statistical inference. Ann Math Statist. 1958;29:357–72.
6. Zumbo DB, Hubley AM. A note on misconceptions concerning prospective and retrospective power. J R Stat Soc: Series D. 1998;47:385–88.
7. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
8. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2002;134:663–94.
9. Senn SJ. Power is indeed irrelevant in interpreting completed studies. BMJ. 2002;325:1304.
10. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.
11. Gale RP, Zhang MJ. What's the p-value anyway? Bone Marrow Transplant. 2016;51:1439.
12. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Ann Epidemiol. 2012;22:364–68.
13. Zhang Y, Hedo R, Rivera A, Rull R, Richardson S, Tu XM. Post hoc power analysis: is it an informative and meaningful analysis? Gen Psychiatry. 2019;32:4.

ACKNOWLEDGEMENTS

Supported by National Institutes of Health/National Cancer Institute grant, 5U24 CA76518-22 at CIBMTR. I thank Profs. Robert Peter Gale (Imperial College London) and Mei-Jie Zhang (Medical College Wisconsin) for helpful suggestions.

AUTHOR CONTRIBUTIONS

RF was the sole contributor for this article. Mei-Jie Zhang and Robert Peter Gale added the Series Editors Introduction upon completion of the article.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Raphael A. Fraser.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.