

Post-hoc power analysis: a conceptually valid approach for power based on observed study data

Natalie E Quach ,¹ Kun Yang,¹ Ruohui Chen,¹ Justin Tu,² Manfei Xu,³ Xin M Tu,¹ Xinlian Zhang¹

To cite: Quach NE, Yang K, Chen R, *et al.* Post-hoc power analysis: a conceptually valid approach for power based on observed study data. *General Psychiatry* 2022;**35**:e100764. doi:10.1136/gpsych-2022-100764

Received 01 July 2022
Accepted 23 August 2022

ABSTRACT

Power analysis is a key component of planning prospective studies such as clinical trials. However, some journals in biomedical and psychosocial sciences request power analysis for data already collected and analysed before accepting manuscripts for publication. Many have raised concerns about the conceptual basis for such post-hoc power analyses. More recently, Zhang *et al* showed by using simulation studies that such power analyses do not indicate true power for detecting statistical significance since post-hoc power estimates vary in the range of practical interests and can be very different from the true power. On the other hand, journals' request for information about the reliability of statistical findings in a manuscript due to small sample sizes is justified since the sample size plays an important role in the reproducibility of statistical findings. The problem is the wording of the journals' request, as the current power analysis paradigm is not designed to address journals' concerns about the reliability of the statistical findings. In this paper, we propose an alternate formulation of power analysis to provide a conceptually valid approach to the journals' wrongly worded but practically significant concern.

INTRODUCTION

Power analysis is critical to designing and planning prospective studies in biomedical and psychosocial research. It provides critically important sample sizes needed to detect statistically significant and clinically meaningful treatment differences and evaluate cost–benefit ratios so that studies can be conducted with minimal resources without compromising scientific integrity and rigour. Thus, power analysis is informative for prospective studies, that is, studies that are yet to be conducted. However, the last author of this paper has been receiving numerous requests from domain experts to perform power analysis for data already analysed and reported in submitted manuscripts. Although the reasons for such 'post-hoc' power analysis are never provided by the journals considering publications of the manuscripts, our understanding is that they are likely due to the sample sizes of the data analysed, that is,

whether the limited sample sizes are sufficient to reliably detect significant treatment differences reported in the manuscripts.

As statistical power describes the probability, or likelihood, of an event to occur in the future, such as a statistically significant treatment or exposure effect in a study, post-hoc power analysis is clearly flawed since power analysis is being performed for an event that has already occurred (ie, the treatment or exposure difference already exists in the study data) regardless of whether the difference is statistically significant. Many have raised concerns on such conceptual grounds.^{1–5} Despite these efforts, some journals continue to request post-hoc power analysis as part of their decision-making process in publishing manuscripts. On the other hand, even if an approach or method is conceptually flawed, it may still provide useful information.

For example, for addressing missing follow-up data in longitudinal data, the last observation carried forward (LOCF) is conceptually flawed when used as a general statistical strategy to deal with missing data during follow-up assessments. However, in some cases, LOCF is still used to provide information about treatment differences. Consider a longitudinal study on a disease of interest in which the subjects' health conditions will deteriorate over time. Estimates of changes over time under LOCF provide information for the mean change of health conditions in the best scenario since follow-up missing data are likely due to deteriorated health conditions.

Unfortunately, this is not the case with post-hoc power analysis. Zhang *et al*¹ examined the utility of post-hoc power analysis in comparing two groups using simulation and found that post-hoc power estimates are generally not informative about the true treatment difference unless used for large effect size and/or large sample size. For medium



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, UC San Diego, La Jolla, California, USA

²Department of Orthopedics, Emory Healthcare, Emory University, Atlanta, Georgia, USA

³Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Correspondence to

Dr Xinlian Zhang;
xizhang@health.ucsd.edu

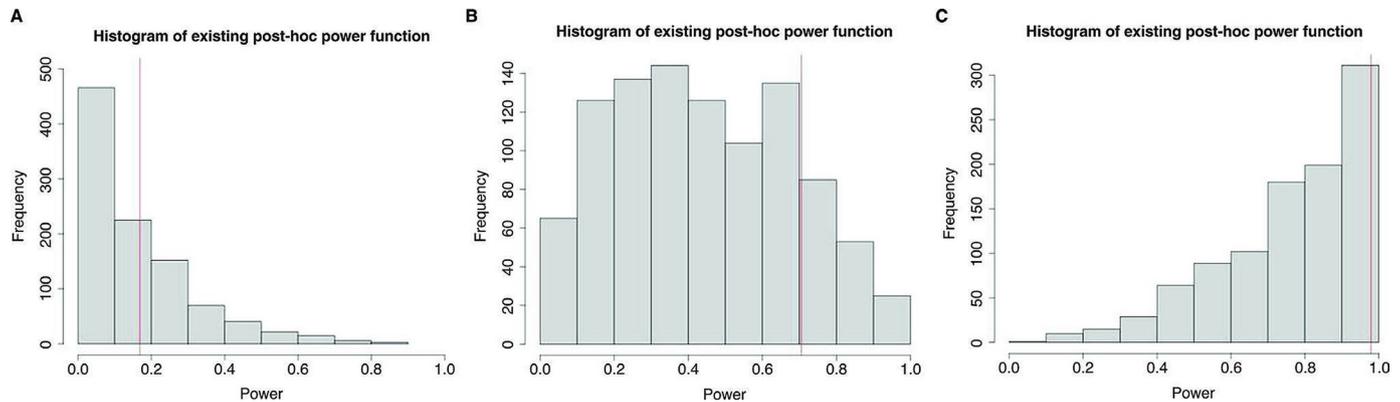


Figure 1 Histogram of power from existing post-hoc power function, along with power from the prospective power function for sample size $n=50$, based on 1000 Monte Carlo runs for effect size (A) $\Delta=0.2$, (B) $\Delta=0.5$ and (C) $\Delta=0.8$.

effect size, Cohen’s $d=0.5$, post-hoc power estimates will vary uniformly between 0 and 1 even for a sample size of $n=100$ per group.

Although requests for post-hoc power analysis present a conceptual conflict with the power analysis paradigm, the rationale for wanting to know if the sample size is sufficient or insufficient to detect statistically significant treatment difference is a meaningful one, especially when the sample size is relatively small. In this paper, we discuss conceptually valid approaches to help capture journals’ concerns about the reliability of statistical findings.

POST-HOC POWER ANALYSIS FOR COMPARING TWO POPULATION MEANS

Within the power analysis paradigm, the reason why post-hoc power analysis is conceptually flawed is the misinterpretation of parameters for power analysis in prospective studies. In fact, standard power analysis is ill-posed for assessing the reliability of significant statistical findings from data of a completed study. We discuss one more appropriate formulation that extends the current power analysis paradigm to address the fundamental flaw in the existing approach.

For convenience, consider two independent samples and let Y_{ik} denote a continuous outcome of interest from subject i and group k ($1 \leq i \leq n_k, k = 1, 2$). For simplicity and without loss of generality, we assume that for both groups Y_{ik} follows a normal distribution population mean μ_k and common population variance σ^2 , denoted $N(\mu_k, \sigma^2)$ ($1 \leq i \leq n_k, k = 1, 2$). The most popular hypothesis for comparing two groups is whether the population means are the same between the two groups, that is:

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs \quad H_a : \mu_1 - \mu_2 = \delta \neq 0, \quad (1)$$

where δ is a known constant and H_0 (H_a) is known as the null (alternative) hypothesis. The hypothesis in Equation 1 is known as a two-sided hypothesis as no direction of effect is specified in the alternative hypothesis H_a . If a directional effect is also indicated, such as in Equation 2, the hypothesis is called a one-sided hypothesis.

$$\begin{aligned} H_0 : \mu_1 - \mu_2 \leq 0 \quad vs \quad H_a : \mu_1 - \mu_2 = \delta > 0 \\ or \quad H_0 : \mu_1 - \mu_2 \geq 0 \quad vs \quad H_a : \mu_1 - \mu_2 = \delta < 0 \end{aligned} \quad (2)$$

In addition to stating that the two population means are different, it also indicates that the population mean for group 1 is larger (or smaller) than that for group 2 under the alternative. Since two-sided hypotheses are much more popular in practice, we focus on the two-sided hypothesis throughout the rest of the discussion unless stated otherwise. However, all results and conclusions derived apply to the one-sided hypothesis.

Note that when testing the hypothesis in Equation 1 as in data analysis, δ is an unknown constant and p values are calculated based on the null H_0 without any knowledge about δ in the alternative H_a . For power analysis, however, this mean difference must be specified, which actually is the most important parameter for power analysis.

In practice, the normalised difference, or Cohen’s $d = \frac{|\mu_1 - \mu_2|}{\sigma}$, is often used since it is invariant under linear transformation.⁶ This invariance property plays a significant role in statistical analysis. For example, consider comparing gas mileage between two types of vehicles, such as sport utility vehicles (SUVs) and sedans. If this study is conducted in the USA, miles per gallon of gas will be used to assess gas mileage for each vehicle. If the study is conducted in Canada, kilometres per gallon of gas will be used to record gas mileage for each car. Although the means μ_k for the two classes of vehicles are different, the effect size is the same regardless of whether kilometres or miles are used to measure distance travelled per gallon of gas. With the unitless effect size, the hypothesis in Equation 1 can be expressed as:

$$H_0 : d = 0 \quad vs \quad H_a : d = \Delta. \quad (3)$$

We will use effect size d and the hypothesis in Equation 3 in what follows unless stated otherwise.

Note that all popular statistical models such as t-tests and linear regression have the invariance property under linear transformation so that the same level of statistical significance is reached regardless of measurement units used. For example, in the above gas mileage example, if we model the outcome of miles travelled per gallon of gas

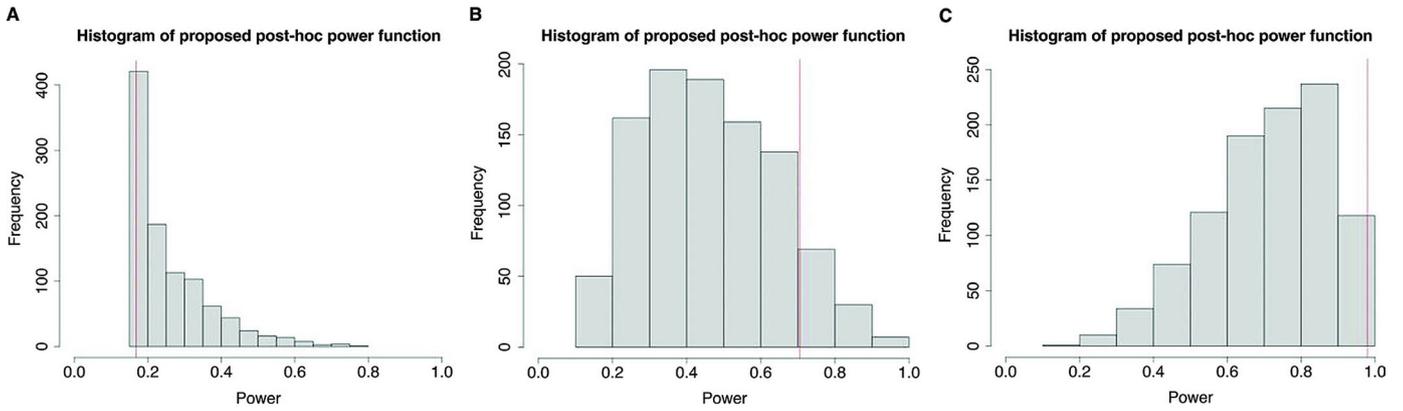


Figure 2 Histogram of power from a proposed post-hoc power function, along with power from the prospective power function for sample size $n=50$, based on 1000 Monte Carlo runs for effect size (A) $\Delta=0.2$, (B) $\Delta=0.5$ and (C) $\Delta=0.8$.

as a function of manufacturers in addition to differences between SUVs and sedans using linear regression, we will get different estimates of regression parameters (coefficients) and standard errors, but same test statistics (F and t statistics) and p values.

In clinical research, the magnitude of d is used to indicate meaningful treatment difference or exposure effects. This is because statistical significance is a function of sample size and any small treatment difference can become statistically significant with a sufficiently large sample size. Thus, statistical significance cannot be used to define the magnitude of treatment or exposure effects. Defined only by the population parameters, effect size is a meaningful measure of the magnitude of treatment or exposure effects. Equation 3 indicates that both the null and alternative hypotheses only involve population parameters. This characterisation of the statistical hypothesis is critically important since this fundamental assumption is violated when performing post-hoc power analysis.

For power analysis, we want to determine the probability to reject the null H_0 in favour of the alternative H_a under H_a for the hypothesis in Equation 3. To compute power, we need to specify the H_0 , H_a , type I error α , and sample size n_1 and n_2 , with n_k denoting the sample size for group k ($k=1,2$). Given these parameters, power is the probability of rejecting the null H_0 when the alternative H_a is true:

$$\begin{aligned} \psi(n_1, n_2, \alpha, H_a(d = \Delta)) &= \Pr(\text{Reject } H_0 \mid H_a) \\ &= \Pr\left(\left|Z + \frac{\Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \geq Z_{\alpha/2}\right) \\ &= \begin{cases} 1 - \phi\left(Z_1 - \frac{\alpha}{2} - \frac{\Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) & \text{if } \Delta > 0 \\ \phi\left(-Z_1 - \frac{\alpha}{2} - \frac{\Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) & \text{if } \Delta < 0 \end{cases} \quad (4) \end{aligned}$$

where $\Pr(A|B)$ denotes the conditional probability of the occurrence of event A given event B , Z denotes a random variable following the standard normal distribution $N(0,1)$, $Z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of $N(0,1)$, and $\phi(z)$ denotes the cumulative distribution

function of the standard normal distribution $N(0,1)$. If we condition on the null H_0 instead of H_a in Equation 4, we obtain a type I error, as in Equation 5:

$$\begin{aligned} \text{type I error}(n_1, n_2, \alpha, H_0) &= \Pr(\text{Reject } H_0 \mid H_0) \\ &= \Pr(|Z| \geq Z_{\alpha/2}) \\ &= 2[1 - \phi(Z_{\alpha/2})] \\ &= \alpha \end{aligned} \quad (5)$$

which is the probability of rejecting the null H_0 when the null is true. For power analysis, we generally set type I errors at $\alpha=0.05$, so that $Z_{\alpha/2}=Z_{0.025}=1.96$.

In practice, we often set power at some prespecified levels and then perform power analysis to determine the minimum sample size to detect a prespecified effect size Δ with the desired level of power. For example, if we want to determine sample size n per group to achieve, say, 0.8 power with equal sample between two groups, we can obtain such minimum n by solving for n in the following equation⁷:

$$\psi(n, n, \alpha, H_a(d = \Delta)) \geq 0.8 \quad (6)$$

When applying Equation 3 for a post-hoc power analysis in a study, we substitute the observed effect size $\hat{\Delta}_n$ in place of the true effect size Δ . This observed $\hat{\Delta}_n$ is calculated based on the observed study data with sample size n_1 and n_2 :

$$\hat{\Delta}_n = \frac{|\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}|}{s_n}$$

where $\bar{Y}_{k\bullet}$ is the sample mean of group k and s_n is the pooled sample standard deviation (SD):

$$\bar{Y}_{k\bullet} = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}, \quad s_n = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\bullet})^2 + \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_{2\bullet})^2 \right]}$$

Unlike Δ , the observed effect size $\hat{\Delta}_n$ is computed based on a particular sample in the study and thus subject to sampling variability. Unless for an extremely large sample size, $\hat{\Delta}_n$ may deviate substantially from Δ . Thus, power is calculated based on $\hat{\Delta}_n$:

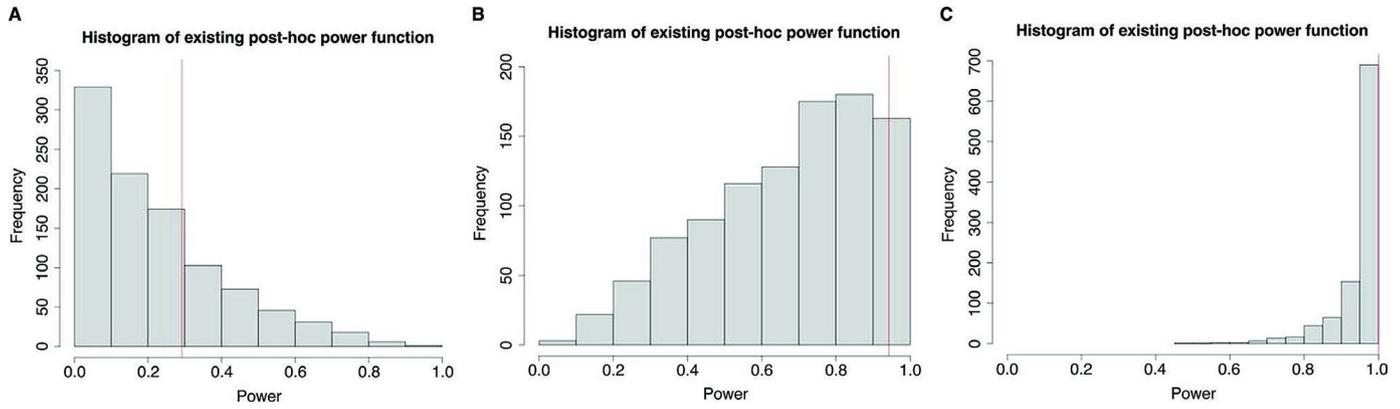


Figure 3 Histogram of power from existing post-hoc power function, along with power from the prospective power function for sample size $n=100$, based on 1000 Monte Carlo runs for effect size (A) $\Delta=0.2$, (B) $\Delta=0.5$ and (C) $\Delta=0.8$.

$$\psi(n_1, n_2, \alpha, H_a(d = \hat{\Delta}_n)) = \begin{cases} 1 - \phi \left(Z_{1-\frac{\alpha}{2}} - \frac{\hat{\Delta}_n}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) & \text{if } \hat{\Delta}_n > 0 \\ \phi \left(-Z_{1-\frac{\alpha}{2}} - \frac{\hat{\Delta}_n}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) & \text{if } \hat{\Delta}_n < 0 \end{cases} \quad (7)$$

Equation 7 only indicates the probability to detect the sample effect size $\hat{\Delta}_n$, which can be quite different from power estimates computed based on the true population effect size Δ . Except for large sample sizes n , power estimates based on $\hat{\Delta}_n$ can be highly variable, rendering them uninformative about the true effect size Δ .

On the other hand, post-hoc power analysis based on Equation 7 also presents a conceptual challenge. Under the current power analysis paradigm, power is the probability to detect a population-level effect size Δ . This effect size is specified with complete certainty. For example, if we set $\Delta=0.5$, it means that we know that the difference between two population means of interest is 0.5. For post-hoc power analysis, we compute power using $\hat{\Delta}_n$ as if this was the difference between the two population means. Due to sampling variability, $\hat{\Delta}_n$ varies around Δ and the two can be substantially different. Indeed, as illustrated in Zhang *et al.*¹ post-hoc power based on Equation 7 can be misleading when used to indicate power for Δ based on Equation 4.

Thus, for post-hoc power analysis to be conceptually consistent with power analysis for prospective studies and informative about the population effect size Δ , we must account for the sampling variability in $\hat{\Delta}_n$. Although $\hat{\Delta}_n \neq \Delta$, $\hat{\Delta}_n$ is generally informative about Δ , with diminishing uncertainty as sample size increases, a phenomenon known as the central limit theorem (CLT) in the theory of statistics.⁸ By quantifying the variability in $\hat{\Delta}_n$ and incorporating such variability in specifying the alternative hypothesis, we can develop new post-hoc power analysis to inform our ability to detect Δ .

By the CLT, the variability of $\hat{\Delta}_n$ is described by a normal distribution $N(\Delta, \frac{1}{n_1} + \frac{1}{n_2})$. Thus, in the absence of knowledge about Δ , values closer to $\hat{\Delta}_n$ are better candidates for Δ , while values more distant from $\hat{\Delta}_n$ are less likely to be good candidates for Δ . By giving more weights to values closer to $\hat{\Delta}_n$ and less weights to values more distant from $\hat{\Delta}_n$, the normal distribution centred at $\hat{\Delta}_n$, $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$, quantifies our uncertainty about Δ . Thus, for post-hoc power analysis, we replace the alternative hypothesis in Equation 3 involving a known population effect size Δ with a set of candidate values for Δ with their candidacy described by the distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$:

$$H_0 : d = 0 \quad \text{vs} \quad H_a : d \sim N\left(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2}\right) \quad (8)$$

The hypothesis in Equation 8 is fundamentally different from the hypothesis in Equation 3 for regular power analysis for prospective studies. Unlike Equation 3, there are more than one candidate value for Δ and post-hoc power analysis must consider all such candidate d s with their relative informativeness for Δ described by the distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. Thus, a sensible way to achieve this is to average power estimates over all such candidates according to their relative informativeness for Δ described by the distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. However, since there are infinitely many such Δ s, we need to use integrals in calculus to perform this averaging.

Let $f_{\hat{\Delta}_n}(\Delta)$ denote the density function of the normal distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. Then by averaging the power function $\psi(n_1, n_2, \alpha, H_a(d = \Delta))$ in Equation 7 over all plausible values of Δ weighted by the distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$, we obtain power for the hypothesis in Equation 8:

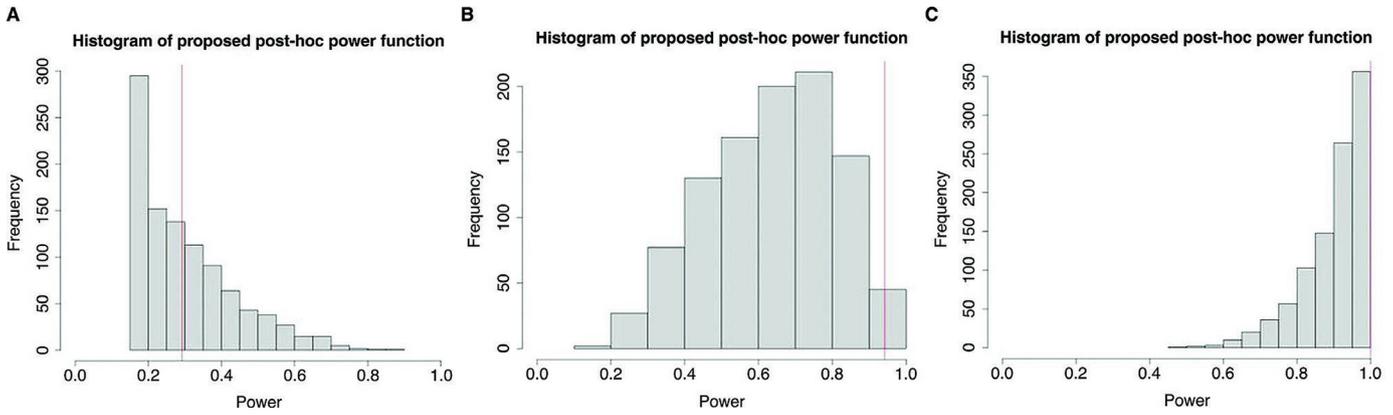


Figure 4 Histogram of power from proposed post-hoc power function, along with power from the prospective power function for sample size $n=100$, based on 1000 Monte Carlo runs for effect size (A) $\Delta=0.2$, (B) $\Delta=0.5$ and (C) $\Delta=0.8$.

$$\tau \left(n_1, n_2, \alpha, H_a(d : N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})) \right) = \int_{-\infty}^{\infty} \psi(n_1, n_2, \alpha, H_a(d = x)) f_{\hat{\Delta}_n}(x) dx \tag{9}$$

where $\int_{-\infty}^{\infty} h(x) dx$ denotes the value of the integral of the function $h(x)$ over the interval $(-\infty, \infty)$.

In the new formulation of hypothesis in Equation 8, we essentially treat true effect size d as a random variable, rather than a (known) constant as under the current power analysis paradigm. This perspective by viewing unknown population parameters as random variables is not new and in fact a well-established statistical paradigm known as Bayesian inference. Under this alternative paradigm, the choice about an unknown population parameter such as d within the current context does not have to be an unknown constant, but can vary over a range of possibilities following a distribution that reflects our knowledge about the true d . For example, in the hypothesis in Equation 8, our knowledge about d is informed by an observed $\hat{\Delta}_n$, with its variability described by the normal distribution $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. In contrast, the traditional hypothesis for post-hoc power analysis in Equation 7 treats $\hat{\Delta}_n$ as the absolute truth, which completely violates the fundamental assumption of the power analysis paradigm.

The new formulation also allows one to build up knowledge about d . For example, if we use a distribution to describe our knowledge about d prior to an observed $\hat{\Delta}_n$ based on the study data, we can then integrate our a priori knowledge with $\hat{\Delta}_n$ to obtain a distribution to describe our improved knowledge about d and use it in Equation 8. This improved knowledge about d obtained by combining our initial knowledge with observed $\hat{\Delta}_n$ from a real study is known as a posterior distribution. The distribution that describes our initial knowledge is called a prior distribution. By using a posterior distribution as a new prior with data from an additional study, we can derive a new posterior distribution. We can keep updating our knowledge by repeating this process.

For example, within the current study context, we may start without any knowledge about d , in which case we can use a non-informative prior distribution, or a constant. After obtaining $\hat{\Delta}_n$ from a real study with sample size $n=n_1+n_2$, our posterior is $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. If there is a new $\hat{\Delta}_m$ from an additional study about d with sample size $m=m_1+m_2$, we then obtain a new posterior distribution that integrates information from both observed $\hat{\Delta}_n$ and $\hat{\Delta}_m$, which is still a normal but with a different mean and variance $N(\hat{\Delta}_{nm}, \sigma_{nm}^2)$, where $\hat{\Delta}_{nm}$ and σ_{nm}^2 are given by:

$$\hat{\Delta}_{nm} = \frac{\frac{m_1 m_2 \hat{\Delta}_m}{m_1 + m_2} + \frac{n_1 n_2 \hat{\Delta}_n}{n_1 + n_2}}{\frac{m_1 m_2}{m_1 + m_2} + \frac{n_1 n_2}{n_1 + n_2}}, \quad \sigma_{nm}^2 = \frac{1}{\frac{m_1 m_2}{m_1 + m_2} + \frac{n_1 n_2}{n_1 + n_2}} \tag{10}$$

By setting $m_1=m_2=0$, the above normal reduces to $N(\hat{\Delta}_n, \frac{1}{n_1} + \frac{1}{n_2})$. To see this, first set $m_1=m_2=m$, then simplify $\hat{\Delta}_{nm}$ and σ_{nm}^2 in Equation 10 to:

$$\hat{\Delta}_{nm} = \frac{\frac{m^2 \hat{\Delta}_m}{2m} + \frac{n_1 n_2 \hat{\Delta}_n}{n_1 + n_2}}{\frac{m^2}{2m} + \frac{n_1 n_2}{n_1 + n_2}}, \quad \sigma_{nm}^2 = \frac{1}{\frac{m^2}{2m} + \frac{n_1 n_2}{n_1 + n_2}} \tag{11}$$

Setting $m=0$, the $\hat{\Delta}_{nm}$ and σ_{nm}^2 in Equation 11 reduce to $\hat{\Delta}_{nm} = \hat{\Delta}_n$ and $\sigma_{nm}^2 = \frac{1}{n_1} + \frac{1}{n_2}$. Thus, we may view a non-informative prior as an observed $\hat{\Delta}_m$ from a study with zero sample size.

ILLUSTRATIONS

In this section, we use Monte Carlo simulation to compare the three types of power analysis, that is, the regular power analysis for a prospective study in Equation 4 and the two post-hoc power analyses with one based on observed effect size in Equation 7 and the other on the new paradigm in Equation 8. In all cases, we set a two-sided alpha at $\alpha=0.05$ and Monte Carlo sample size at 1000.

We again assume a normal distribution $N(\mu_k, \sigma^2)$, with μ_k denoting the (population) mean of group k and σ^2 the

common (population) variance. We set the population parameters as follows:

$$\mu_1 = 0, \quad \mu_2 = \mu_1 + \Delta, \quad \sigma^2 = 1. \quad (12)$$

Since $\sigma^2=1$, the difference between the means, Δ , is Cohen's d , which is interpreted as $\Delta=0.2$, $\Delta=0.5$ and $\Delta=0.8$ for small, medium and large effect size, respectively.⁶ For convenience, we assume a common sample size for both groups, that is, $n_1=n_2=n$. We set Δ and n to different values so we can see how power estimates from the three different approaches change as a function of the two parameters.

Given all these parameters, we can readily evaluate the prospective power function in Equation 6. For post-hoc power analysis, power is not a constant and varies according to observed effect size $\hat{\Delta}_n$. We use Monte Carlo simulation to capture such variability. Given the parameters in Equation 12 and sample size n , we simulate sample, Y_{ik} , from $N(\mu, \sigma^2)$, compute the sample effect size $\hat{\Delta}_n$ and evaluate the post-hoc power functions in Equation 7 and Equation 9. Thus, unlike prospective power, both post-hoc power functions depend on the observed effect size $\hat{\Delta}_n$. The difference is that Equation 7 treats $\hat{\Delta}_n$ as the true effect size, while Equation 9 acknowledges sampling variability in $\hat{\Delta}_n$ and uses its sampling distribution to inform the underlying effect size Δ .

The variance of $\hat{\Delta}_n$ is $Var(\hat{\Delta}_n) = \frac{2}{n}$, which will be close to 0 and $\hat{\Delta}_n$ will be close to Δ for a very large sample size n . In this case, both post-hoc power will be close to the prospective power. For small and moderate sample sizes, all three power values will differ from each other. Given Δ , the prospective power only has one value for each sample size, while the two post-hoc power approaches will have different values for different observed $\hat{\Delta}_n$. For small and moderate sample sizes, post-hoc power values will have large variabilities and may not be informative about the power based on the true Δ .

Shown in figure 1 are the histograms of power from the existing post-hoc power approach based on 1000 Monte Carlo sample sizes for effect size $\Delta=0.2$ (figure 1A), $\Delta=0.5$ (figure 1B) and $\Delta=0.8$ (figure 1C), along with power from the prospective power function (vertical line) for sample size $n=50$. As expected, power increased in both cases as the true effect size Δ became larger; the sample mean for the post-hoc power was 0.167, 0.443 and 0.761, respectively. For all three effect sizes, there was a large amount of variability in the post-hoc power, covering the entire range of power function between 0 and 1.

Shown in figure 2 are the histograms of power from the proposed post-hoc power approach based on 1000 Monte Carlo sample sizes with the mean difference $\Delta=0.2$ (figure 2A), $\Delta=0.5$ (figure 2B) and $\Delta=0.8$ (figure 2C), along with power from the prospective power function (the vertical line) for sample size $n=50$. As in figure 1, the mean power increased from

0.259 to 0.463 to 0.711 when Δ changed from 0.2 to 0.5 to 0.8. Unlike in figure 1, power was larger than 0.1 in all three cases. Moreover, the sample SDs for the three effect sizes were 0.111, 0.177 and 0.162 for the proposed, compared with 0.151, 0.239 and 0.194 for the traditional power method. The smaller variability of the proposed post-hoc power was the result of being less sensitive to variability in $\hat{\Delta}_n$, compared with the existing post-hoc power.

Shown in figure 3 are the histograms of power from the existing post-hoc power approach based on 1000 Monte Carlo sample sizes for effect size $\Delta=0.2$ (figure 3A), $\Delta=0.5$ (figure 3B) and $\Delta=0.8$ (figure 3C), along with power from the prospective power function (the vertical line) for sample size $n=100$. As in figure 1, as Δ changed from 0.2 to 0.5 to 0.8, the mean power increased from 0.226 to 0.670 to 0.951. Again, post-hoc power was quite variable, covering the entire range in all cases, except for $\Delta=0.8$, where the minimum was 0.48. With $\Delta=0.8$ and $n=100$, the smaller sampling variability $\hat{\Delta}_n$ along with the large effect size led to power values close to 1, greatly reducing the variability of post-hoc power, as compared with $n=50$.

Shown in figure 4 are the histograms of power from the proposed post-hoc power approach based on 1000 Monte Carlo sample sizes with the mean difference $\Delta=0.2$ (figure 4A), $\Delta=0.5$ (figure 4B) and $\Delta=0.8$ (figure 4C), along with power from the prospective power function (the vertical line) for $n=100$. As in figure 3, the mean power grew from 0.303 to 0.636 to 0.901 as Δ increased from 0.2 to 0.5 to 0.8. The sample SDs for $\Delta=0.2$, $\Delta=0.5$ and $\Delta=0.8$ were 0.135, 0.174 and 0.086, compared with 0.184, 0.221 and 0.071 for the existing power in figure 3. Except for $\Delta=0.8$, the proposed post-hoc power was less variable than its existing counterpart. For $\Delta=0.8$, the trend was reversed and post-hoc power was more variable for the proposed than the traditional approach. This is made clear by comparing the two histograms in figure 3C and figure 4C. For $\Delta=0.8$, there was much less variability in power from the traditional than the proposed approach; about 70% of power values were between 0.95 and 1 for the traditional approach, compared with about 30% for the proposed approach. Thus, this smaller variability of power from the traditional approach again reflects its higher sensitivity to the observed $\hat{\Delta}_n$ than the proposed alternative.

DISCUSSION

In this paper, we proposed a new approach for post-hoc power analysis. Unlike the existing approach, the proposed alternative is conceptually sound and can be applied to gauge variability of statistical findings and thus inform reliability of such findings. The approach is illustrated using simulated data by comparing with the traditional approach. The simulation study results show that the proposed approach is less sensitive to observed effect sizes and is more

informative about power estimates based on the underlying true and observed effect size. The simulation results also show that even when performed correctly post-hoc power analysis may yield power values that are different from the power based on the underlying effect size under the current paradigm. On the other hand, we believe that the current paradigm may not provide useful power estimates for real prospective studies. As true effect sizes are rarely known for certain in most studies, treating our guessed effect sizes as such ground truth in computing power is flawed practice. Thus, power analysis for prospective studies also needs to account for uncertainty about true effect sizes to provide practical and useful power estimates. Work is underway to extend the proposed post-hoc power approach to power analysis for prospective studies.

Twitter Natalie E Quach @NatalieEQuach

Contributors All authors participated in the discussion of the statistical issues and worked together to develop this paper. JT, XMT, XZ and MX reviewed the literature and discussed the conceptual issues with the conventional post-hoc power analysis approach. All authors contributed to the development of the new post-hoc power analysis approach proposed in this paper. NEQ, KY and RC developed the simulation algorithms and associated R codes and performed the simulation study. JT, XMT, XZ, MX and NEQ worked together to draft and finalise the manuscript.

Funding The project described was partially supported by the National Institutes of Health (grant UL1TR001442) of CTSA funding.

Disclaimer The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.



Natalie E Quach has been a Biostatistics PhD student in the Division of Biostatistics and Bioinformatics at the UC San Diego Herbert Wertheim School of Public Health and Human Longevity Science since 2022. She graduated with Latin honors from UC San Diego, USA in 2022, receiving a Bachelor of Science in Applied Mathematics. She is currently a researcher and teaching assistant at UC San Diego. Her main research interests include causal inference and clinical trials.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Natalie E Quach <http://orcid.org/0000-0001-9382-5826>

REFERENCES

- Zhang Y, Hedo R, Rivera A, *et al*. Post hoc power analysis: is it an informative and meaningful analysis? *Gen Psychiatr* 2019;32:e100069.
- Heonig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* 2001;55:19–24.
- Korn EL. Projection from previous studies. A caution. *Control Clin Trials* 1990;11:67–9.
- Kraemer HC, Mintz J, Noda A, *et al*. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;63:484–9.
- Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* 2001;21:405–9.
- Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Mahwah (NJ): LEA, 1988.
- Tu XM, Kowalski J, Zhang J, *et al*. Power analyses for longitudinal trials and other clustered designs. *Stat Med* 2004;23:2799–815.
- Tang W, He H, XM T. *Applied categorical and count data analysis*. Boca Raton (FL): Chapman & Hall/CRC, 2012.