Knee Surgery, Sports Traumatology, Arthroscopy (2022) 30:3924–3928
https://doi.org/10.1007/s00167-022-07188-9

EDITORIAL

Fragility Part I: a guide to understanding statistical power

Sophia J. Madjarova[1] · Ayoosh Pareek[1] · Christina M. Eckhardt[2] · Arjun Khorana[1] · Kyle N. Kunze[1] · Mattheu Ollivier[3] · Jón Karlsson[4] · Riley J. Williams III[1] · Benedict U. Nwachukwu[1]

Received: 20 September 2022 / Accepted: 27 September 2022 / Published online: 7 October 2022

Abstract

The aim of this paper is to close the knowledge-to-practice gap around statistical power. We demonstrate how four factors affect power: *p* value, effect size, sample size, and variance. This article further delves into the advantages and disadvantages of a priori versus post hoc power analyses, though we believe only understanding of the former is essential to addressing the present-day issue of reproducibility in research. Upon reading this paper, physician–scientists should have expanded their arsenal of statistical tools and have the necessary context to understand statistical fragility.

There is a clear knowledge-to-practice gap in statistical literacy that separates basic science and implementation in clinical practice with the latter trailing behind. Reasonably, closing this gap has become in part a matter of increasing research participation throughout the medical community and particularly necessitates early exposure for physicians in-training. Participation in research is currently emphasized, if not required, by most residency programs in the United States regardless of specialty as per the Accreditation Council for Graduate Medical Education's guidelines, with similar requirements in Europe as well [19]. But, unfortunately, mandated readings and journal clubs do not fully ensure the development of competent physician-scientists.

Statistical literacy among medical trainees remains low, while the statistical complexity of research continues to increase [2, 9, 17]. This disparity poses a serious threat to evidence-based medicine as the next generation of physicians struggles to engage with the work of their peers and determine what to implement into clinical practice; however, this is not an insular issue, and many fully trained medical professionals demonstrate poor statistical understanding. However, straightforward interventions have been shown to improve statistical literacy. In a study assessing the statistical literacy of 169 medical students and 16 senior instructors, Jenny et al. demonstrated that as short as a 90-min informational session on statistics could increase the median percentage correct from 50 to 80% on the 10-question Quick Risk Test [10]. This study suggests, as perhaps is already inferable, that a lack of statistical literacy among medical professionals is not due to inability to understand statistical concepts, but rather is due to a lack of instruction and accessibility. There is, therefore, a need to provide formal dedicated training and resources to increase the accessibility of statistics for existing as well as future generations of doctors.

The term *fragility* has recently gained traction in the clinical research community, and in particular, within the field of orthopaedic research [6, 13–15]. However, before defining fragility, a topic to be discussed later in this series, a more traditional and perhaps familiar term, i.e., (statistical) *power* is defined. In clinical research studies, a null hypothesis, which assumes that two groups are the same, is usually tested. In other words, that there are no significant differences between these groups. The *p* value measures how well study results fit with the normal distribution for the null hypothesis. The value corresponding to the significance level alpha (α), and it can be manipulated by the researcher. Arbitraril,y alpha is chosen to be 0.05, and a *p* value less than alpha ($p \leq 0.05$) indicates the presence of a difference

✉ Ayoosh Pareek
ayooshp@gmail.com

[1] Sports Medicine and Shoulder Service, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA

[2] Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA

[3] Institut du Movement Et de l'appareil Locomoteur, Aix-Marseille Université, Marseille, France

[4] Department of Orthopaedics, Sahlgrenska University Hospital, Sahlgrenska Academy, Gothenburg University, Gothenburg, Sweden

between the experimental groups due to an effect beyond chance [11].

When researchers mistakenly reject the null hypothesis and wrongly conclude there is an observable effect (false positive), this results in a Type I error (Fig. 1). The probability of a Type I error is equivalent to alpha. When a two-tailed α of 0.05 is selected as the significance threshold by the researcher, 95 of 100 trials are expected to detect the true effect, and the remaining 5 are false positives. Conversely, Type II errors occur when we mistakenly fail to reject the null hypothesis (false negative) when a true effect is present (Fig. 1). The probability of a Type II error with the Greek letter beta (β).

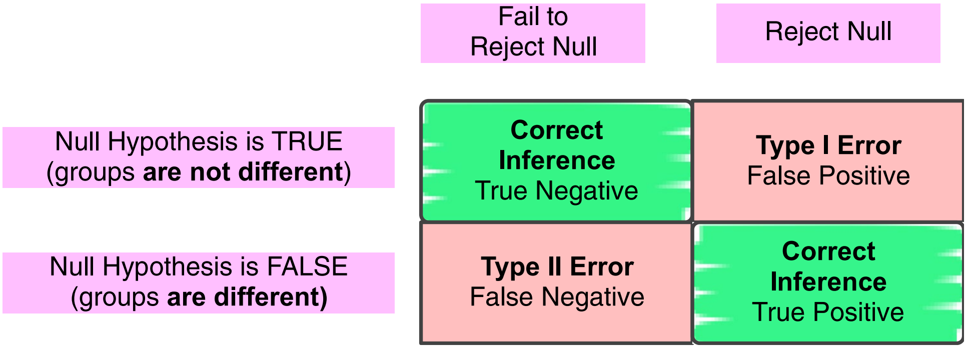
Power is the probability of correctly rejecting the null hypothesis, or alternatively the probability of correctly detecting a true effect when an effect is present. In other words, power is the probability avoiding a Type II error ($1 - \beta$) [12]. If power were 0.6 (60%) for a study, it suggests that if we were to run an experiment 100 times, 60 of those iterations would demonstrate an effect. Traditionally, a threshold of 0.8 (80%) is used as an acceptable level of power, though different thresholds may be reasonable depending on the study design [16]. However, some critics believe this threshold is unattainable even for high quality studies, particularly when considering studies with inherently small samples, such as many of those related to surgical science or rare diseases [4]. For example, studies with inherently small sample sizes including some surgical studies may still be able to yield important clinical findings. However, particularly with respect to null findings, it is always critical to bear in mind whether a study was adequately powered to detect an effect.

Ideally, power is calculated a priori, meaning before the investigation begins. This means that the conditions of a study can be set to ensure the power threshold (0.8) is achieved. There are four main variables that can be manipulated to increase power: *alpha* (α), *sample size, effect size,* and *variance*. As previously discussed, the α is regarded as a threshold for rejecting or failing to reject the null hypothesis; therefore, raising the p value threshold will make observing a difference more likely since there is a wider range of acceptable p values. Since power is a measure of likelihood to detect a difference when a difference is truly present, manipulating to a p value to increase likelihood of rejecting the null will increase power.

Power can also be modified by manipulating the sample size. Since power is ideally calculated before a trial is run, it can be used to determine how many data points, patients, or samples are needed to detect the underlying effect. In this way, increasing sample size makes obtaining adequate power more likely; however, increasing sample size also increases the likelihood of Type I errors (false positives) [16]. Moreover, it is not always economically or practically feasible to increase a sample size (for example, monetary reasons), and it may delay project completion [16]. Thus, there is a delicate balance when determining sample size. How we choose to calculate sample size relies on what kinds of variables the study is analyzing (continuous, binomial, categorical) and whether the study design is within-subjects or between-subjects [18]. Calculating sample size or preforming a power analysis by hand can be quite complex, but automated calculation is prevalent using available resources (free and paid) as shown in Table 1.

The final component needed for a power analysis or sample size calculation is the effect size, which is the strength of the relationship between the variables. Alternatively, effect size can be defined as the magnitude of difference between the two groups. For example, when comparing the odds of pain relief among a group of patients who received surgical intervention to a group who were managed conservatively, an effect size or odds ratio of 1.5 would suggest the group who received surgery had 50% higher odds of achieving pain relief. The effect size differs between experiments even when parameters are completely identical; therefore, the true effect size is often estimated by creating a normal distribution from the other experimental effect sizes [18]. A larger, more detectable effect size correlates with increased power and hence a lower sample size will be needed. However, when performing a power calculation, the estimated effect size should be derived from existing literature and prudent clinical judgement. The importance of designing experiments with eventual statistical analysis in mind is clearly

Fig. 1 Relationship between actual truth of the null hypothesis and statistically-based acceptance or rejection



	Fail to Reject Null	Reject Null
Null Hypothesis is TRUE (groups **are not different**)	**Correct Inference** True Negative	**Type I Error** False Positive
Null Hypothesis is FALSE (groups **are different)**	**Type II Error** False Negative	**Correct Inference** True Positive

Table 1 Suggested resources for sample size calculation

Program	Cost of use	Link
G*Power	Free	http://www.gpower.hhu.de
PASS	Paid	https://www.ncss.com/software/pass
nQuery	Paid	https://www.statsols.com/nquery-sample-size-and-power-calculation-for-successful-clinical-trials
*R Packages**		
Pwr	Free	https://cran.r-project.org/web/packages/pwr
Trial Size	Free	https://cran.r-project.org/web/packages/TrialSize/
PowerUpR	Free	https://cran.r-project.org/web/packages/PowerUpR/
powerSurvEpi	Free	https://cran.r-project.org/web/packages/powerSurvEpi/index.html
Programs (Paid)		
Systat	Paid	https://systatsoftware.com/
Statistica	Paid	https://www.tibco.com/products/data-science
SAS (PROC POWER)	Paid	https://support.sas.com/documentation/onlinedoc/stat/141/power.pdf
SPSS (SamplePower)	Paid	https://www.ibm.com/docs/en/spss-statistics/saas?topic=features-power-analysis
STATA (power)	Paid	https://www.stata.com/features/power-and-sample-size/
PASS	Paid	https://www.ncss.com/software/pass
Microsoft excel		
XLSTAT	Paid	https://www.xlstat.com/en/
PowerUp	Free	https://www.causalevaluation.org/power-analysis.html
Online resources		
UCSF sample size calculator	Free	Sample Size Calculators (sample-size.net)
Genetic power calculator	Free	https://zzz.bwh.harvard.edu/gpc/
Power and sample size	Free	http://powerandsamplesize.com/Calculators/

*R is a commonly used statistical programming language. These R packages include code that defines various useful functions. Many more R packages are available for download on the Comprehensive R Archive Network: https://cran.r-project.org/

evident when considering this aspect of power. The effect size can also be set depending on known values of clinical significance, such as minimal clinically important difference (MCID) [16, 18]. It should be noted that well-powered studies that may find statistically significant results do not necessarily yield clinically meaningful results. Thus, defining the effect size with values such as MCID in mind may help reduce confusion when discussing statistical significance in clinical research.

Power is also affected by the variance. Translated into layman's terms, variance is a measure of spread. Increased variance can generate heterogeneity of the treatment effect, which can decrease the likelihood of observing an underlying effect. Increased spread may, therefore, result in lower power. However, it is important to balance the homogeneity of the population against the need for externally valid and generalizable results.

To demonstrate the sample size calculation process, we pose an a priori power analysis using the Power and Sample size free online calculator (Table 2). We are interested in comparing the time to ACL reinjury after reconstruction for female soccer players versus non-soccer athletes [1]. To complete this survival analysis, we need to determine the sample size needed for the included cohorts matched according to a 1:1 ratio. We will analyze this using a two-tailed hypothesis test, splitting alpha in half at both extremes of the normal distribution (0.025). We have chosen a two-tailed test, because we are interested in seeing if there is a *difference* between soccer players and non-soccer athletes. If we were interested in seeing if non-soccer athletes do *better than* soccer players, a one-tailed test would be appropriate. Though tempting, especially because significance is more easily achieved when alpha is applied to only one tail, we need to analyze both the possibility that non-soccer athletes do better than or worse than soccer players, meaning we must analyze using a two-tailed *p* value.

Moving on to the Power and Sample size online calculator, we will select the option for analyzing two-tailed hypothesis testing for time-to-event data, since this is what a survival curve reveals. According to convention, alpha is defined as 0.05 and the desired power as 0.8. A hazard ratio, or an effect size, must then be estimated as well. This is the ratio of how often ACL reinjury occurs in soccer players versus non-soccer athletes, over time. In this case we will estimate a hazard ratio of 2, indicating an assumption that soccer players will have an ACL tear twice as often

Table 2 Definitions and key terms

Type I error

False positive

Probability of Type I error is represented by alpha (α), which is manipulated by researchers but conventionally defined as 0.05

Type II error

False negative

Probability of Type II error is represented by beta (β)

Power

The likelihood of avoiding a Type II error ($1 - \beta$)

Dependent on alpha, sample size, effect size, and variance (spread)

Conventionally defined as 0.8

Effect Size

Strength of relationship between variables

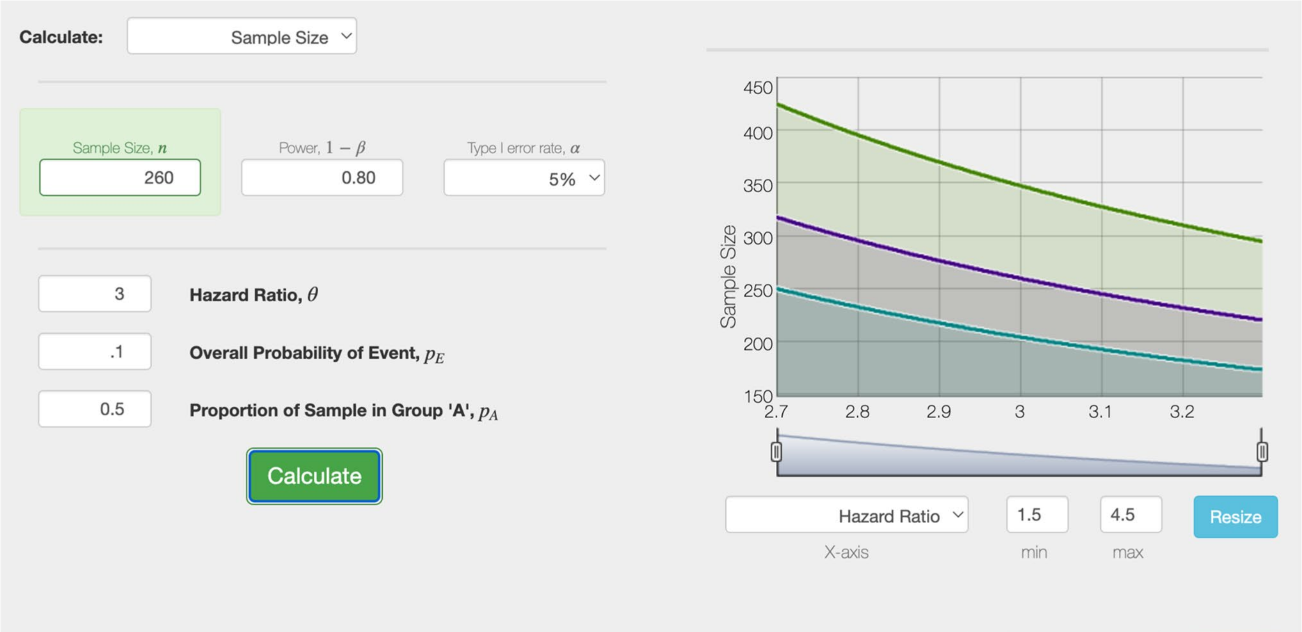
Differs between experiments even when conditions are identical

True effect size is often estimated from normal distribution of other experimental effect sizes

Can also be defined with respect to MCID

as non-soccer players. We also assume an overall ACL tear rate of 10%, and an equal number of patients in each group (equal proportion of sample in soccer and non-soccer groups). Under these assumptions, we would need a sample size of 653 patients. This decreases to a sample size of 260 patients with a hazard ratio of 3 (Fig. 2).

In the case above, an a priori power analysis was conducted, which can be very helpful when designing a new study. However, a post hoc power analysis can be performed after the study has been completed if necessary. Supporters of post hoc power analyses argue that in studies evaluating surgical interventions, which often have smaller sample sizes, traditional power analyses can be difficult to perform, since the effect size may not be known or may be difficult to estimate for never-performed surgical procedures [3, 4]. Furthermore, retrospective power analyses can be useful when a statistically non-significant result is obtained in a small study. For example, a null finding may stem from low power or a truly small effect. A post hoc power analysis can help distinguish these findings. However, post hoc power analyses can be somewhat redundant, because they depend directly on the p value. The less significant the p value from



Calculate: Sample Size

Sample Size, n: 260

Power, $1 - \beta$: 0.80

Type I error rate, α: 5%

3 Hazard Ratio, θ

.1 Overall Probability of Event, p_E

0.5 Proportion of Sample in Group 'A', p_A

X-axis: Hazard Ratio

min: 1.5

max: 4.5

Fig. 2 Example of sample size calculation for time-to-event data. (Figure adapted from http://powerandsamplesize.com/Calculators/Test-Time-To-Event-Data/Cox-PH-2-Sided-Equality.)

the study, the larger the suggested sample size derived by the post hoc power calculation [5, 7, 8]. There can also be great variance or "noise" for an observed effect size, since even identically run experiments may vary in the magnitude of difference between groups, which undermines any direct calculation of power from the results. However, Bababekov et al. claim that redundancy can be an effective way to both ensure that results are being clearly communicated and to indicate whether further investigation on the topic may be useful [3, 4]. Moreover, if a *p* value indicates insignificance, but the power is extremely low, a researcher may benefit from changing their experimental design to ensure they are not missing a real world effect. Thus, while post hoc power calculations need to be interpreted with caution, there are some situations in which determining whether a study was adequately powered may be important for interpreting study results.

Amidst the current discussion about improving reproducibility in clinical research, understanding power is necessary to maintain the integrity of published information. Moreover, an understanding of the importance of power mandates a more robust statistical education at large due to a need for better understanding of experimental design and subsequent statistical analysis. In the case of reporting power, we claim it is essential to teach researchers how to perform a priori power analyses before getting caught up in post hoc power, which can often lead to more confusion than clarity. Increasing research participation is a worthwhile and necessary goal for the future of evidence-based medicine. Thus, it is important to remember that the quality of the resulting research depends on an effort to improve our understanding of the statistical underpinnings of clinical research.

Funding The authors received no funding for the study.

Declaration

Conflict of interest The authors declare no conflicts of interest with regards to the manuscript.

Ethical approval Ethical approval was not required for this manuscript.

References

1. Allen MM, Pareek A, Krych AJ, Hewett TE, Levy BA, Stuart MJ, Dahm DL (2016) Are female soccer players at an increased risk of second anterior cruciate ligament injury compared with their athletic peers? Am J Sports Med 44:2492–2498
2. Anderson BL, Williams S, Schulkin J (2013) Statistical literacy of obstetrics-gynecology residents. J Grad Med Educ 5:272–275
3. Bababekov YJ, Chang DC (2019) Post hoc power: a surgeon's first assistant in interpreting "negative" studies. Ann Surg 269:e11–e12
4. Bababekov YJ, Hung Y-C, Hsu Y-T, Udelsman BV, Mueller JL, Lin H-Y, Stapleton SM, Chang DC (2019) Is the power threshold of 0.8 applicable to surgical science?—empowering the underpowered study. J Surg Res 241:235–239
5. Dziak JJ, Dierker LC, Abar B (2020) The interpretation of statistical power after the data have been gathered. Curr Psychol 39:870–877
6. Ehlers CB, Curley AJ, Fackler NP, Minhas A, Chang ES (2021) The statistical fragility of hamstring versus patellar tendon autografts for anterior cruciate ligament reconstruction: a systematic review of comparative studies. Am J Sports Med 49:2827–2833
7. Gelman A (2019) Don't calculate post-hoc power using observed estimate of effect size. Ann Surg 269:e9–e10
8. Gelman A (2019) Comment on "post-hoc power using observed estimate of effect size is too noisy to be useful." Ann Surg 270:e64
9. Hellems MA, Gurka MJ, Hayden GF (2007) Statistical literacy for readers of *pediatrics* : a moving target. Pediatrics 119:1083–1088
10. Jenny MA, Keller N, Gigerenzer G (2018) Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany. BMJ Open 8:e020847

11 Madjarova SJ, Williams RJ, Nwachukwu BU, Martin RK, Karlsson J, Ollivier M, Pareek A (2022) Picking apart *p* values: common problems and points of confusion. Knee Surg Sports Traumatol Arthrosc. https://doi.org/10.1007/s00167-022-07083-3

12. Norton BJ, Strube MJ (2001) Understanding statistical power. J Orthop Sports Phys Ther 31:307–315
13. Parisien RL, Constant M, Saltzman BM, Popkin CA, Ahmad CS, Li X, Trofa DP (2021) The fragility of statistical significance in cartilage restoration of the knee: a systematic review of randomized controlled trials. Cartilage 13:147S-155S
14. Parisien RL, Trofa DP, Cronin PK, Dashe J, Curry EJ, Eichinger JK, Levine WN, Tornetta P, Li X (2021) Comparative studies in the shoulder literature lack statistical robustness: a fragility analysis. Arthrosc Sports Med Rehabil 3:e1899–e1904

15 Parisien RL, Trofa DP, O'Connor M, Knapp B, Curry EJ, Tornetta P, Lynch TS, Li X (2021) The fragility of significance in the hip arthroscopy literature: a systematic review. JBJS Open Access. https://doi.org/10.2106/JBJS.OA.21.00035

16. Serdar CC, Cihan M, Yücel D, Serdar MA (2021) Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Med (Online) 31:27–53
17. Shreffler J, Thomas A, Huecker M (2022) An analysis of statistical terminology applied in emergency medicine literature methods. Am J Emerg Med 58:251–254
18. Uttley J (2019) Power analysis, sample size, and assessment of statistical assumptions—improving the evidential value of lighting research. Leukos 15:143–162
19. Accreditation Council for Graduate Medical Education (2022,) Common program requirements (Residency). https://www.acgme.org/what-we-do/accreditation/common-program-requirements/