Do You Have Power? Considering Type II Error in Medical Education

Gail M. Sullivan, MD, MPH
Richard S. Feinn, PhD

Perhaps related to years of participating in clinically focused journal clubs, medical education researchers usually feel comfortable using quantitative methods—the "how much?" questions. However, in contrast to large patient trials, sample sizes in medical education, and particularly in graduate medical education (GME) projects, may be small. If there really is a difference between groups, educators may not uncover it with a small sample size. This can lead researchers to opine in their Discussion section: "Although we found no statistically significant differences between our innovation and the prior approach, this is probably because the sample size was too small." In other words, the authors believe there *is* a difference, but *the study lacked power.* This type of nonfinding or nonconclusion is reported often in medical education studies, but it is not very helpful. In fact, some experts believe that not considering the power of a study beforehand is irresponsible.[1] Yet there are times, such as in novel or exploratory work, that problems with study "power" may be inevitable.

Can we sensibly address the question of power to detect differences? In most instances, we can, even in medical education research with typically small sample sizes. This brief introduction is designed to help beginning researchers and readers wade into the murky waters of power calculations and sample sizes—and emerge unscathed.

Review: Type I vs Type II Errors

Type I errors can produce *false positive* findings; if we reject the null hypothesis (that there is no difference between groups we are comparing) when it is actually true, ie, *there is no difference*. We usually choose a small number, such as 0.05 or less, for the type I error level, or alpha, for single or a few comparisons.[2] The lower the alpha level, the less likely we will make false positive conclusions: *less likely* being the operative words.[3]

In contrast, type II errors can produce *false negative* findings: not rejecting the null hypothesis (that there is no difference between groups) when it is not true—*there is a difference*. When planning a study we usually choose 0.20 as the type II error (beta) level. However, when choosing alpha and beta, one should also consider the research question and real-world effects of overlooking an actual difference vs claiming differences that do not exist (see TABLE 1).

The *power* of a study to find differences is 1 - beta, which is 0.80, or 80%, if beta is chosen at 0.20. Power is the likelihood of correctly rejecting the null hypothesis (that there is no difference) when it is not true. Power answers the question: if an effect, of a specified magnitude, really occurs, what is the chance that a trial, of a certain sample size, will find a *statistically significant* result given the chosen alpha level? The greater the power of a test, the more confidence we have that we will be able to detect a difference between groups.[3] A study power set at 80% accepts a chance of 1 in 5 (20%) of missing a difference that really exists. Researchers may set the power at 90% to lower the chance of missing a real difference to 1 in 10.

Whether a difference exists and how big that difference is are not under our control; these are features of the intervention, setting, and subjects under study. However, we can control the sample size of the trial: how many trainees are included or how many assessments are examined.

In medical education research we are usually looking for moderate or large differences. For example, there may be a real difference in resident satisfaction for a new program vs the existing program, on a Likert-type scale of 1 to 5. But if the means are 3.16 vs 3.27, this difference is not meaningful in an educational sense, no matter if statistically significant.[4] (Note that it takes a large sample to demonstrate a real, but tiny difference as in this example.) Or a national study may find that a new wellness initiative shows a decline in burnout measures from 28% to 27%. This difference appears real, as the *P* value is lower than our chosen alpha level cutoff. Do we care? No. In contrast to clinical medicine, in medical education we generally want larger differences to justify decisions (see TABLE 2).

DOI: http://dx.doi.org/10.4300/JGME-D-21-00964.1

TABLE 1
Type II vs Type I Error, and Power

Type II	Type I
False negative error ▪ Not rejecting the null hypothesis (there is no difference in groups) when it is false: there is a difference	False positive error ▪ Rejecting the null hypothesis (there is no difference between groups) when it is true: there is no difference
Beta error: often 0.20	Alpha error: often 0.05[a]
Power ▪ Probability of rejecting the null hypothesis (no difference) when it is false ▪ Probability that a test of significance will find a difference if it exists ▪ Equals 1 - beta	If *P* value is above .05 ▪ Cannot conclude there is a difference; evidence is insufficient to conclude there is a difference ▪ With sufficient sample size and considering educational differences of moderate or large size, more likely to be a true finding If *P* value is below .05 ▪ Evidence suggests there is a real difference between groups, but it may not be an educationally meaningful difference (consider magnitude)
Power depends upon ▪ Effect size (magnitude of the difference) ▪ Standard deviation: variability or variance in the variable measured ▪ Alpha (type I) error chosen[b] ▪ Sample size tested[b]	Statistical significance depends upon ▪ Effect size (magnitude of difference) ▪ Sample size[b] ▪ Alpha level chosen[b]

[a] May require adjustment for multiple comparisons.[2]
[b] Under investigator's control.

When to Think About Power and Sample Sizes

It is best to perform a power analysis *before* doing a study to maximize the ability to detect differences that do exist. At the least, power considerations must precede any look at the data or data analysis.

Power depends upon the actual size of the difference (ie, effect size), variability or variance in the variables we are measuring (eg, standard deviation), level of significance we choose (alpha), and the sample size. Only the last 2 of these are within our control—the level of significance (alpha) and the sample size. As sample size increases, beta decreases, and thus power to find a real difference increases. Most people accept that a power of 80% is reasonable, which means selecting a beta or type II error of 0.20. Ideally the choice of power level—or the flip side, type II error—depends upon how serious the consequences are of making a type II error (false negative finding), which relates to subsequent decisions that will be based on the findings. For example, will the findings affect a high-stakes assessment? Will they lead to removing a resident rotation? Consequences also should include how much time and effort is involved in conducting the study to avoid wasting these precious resources. You want to avoid spending a lot of time and resources conducting a study only to realize, later, that you are unable to reach a definitive conclusion because the study was underpowered.

In addition, there is a tradeoff in that, as alpha increases, beta decreases, which must be factored into study plans. Consider which is more critical in informing subsequent decisions: avoiding false positive findings (type I errors) or avoiding false negative findings (type II errors).

TABLE 2
Cohen's Guide to Effect Sizes (Magnitude of Differences)[4]

Effect Size	General Interpretation
<0.1	Trivial effect
0.1–0.3	Small effect
0.3–0.5	Medium effect
0.5–0.7	Large effect
>0.7	Very large effect

Note: Effect size = difference/standard deviation.

Calculating Sample Sizes

To calculate the sample size, you need the chosen alpha and beta error levels, the expected minimum effect size (magnitude of difference), as well as expected variability in the outcome variable.[5] Researchers often wonder how to determine the effect size, when the comparisons under study have not been done before or not in the same way. In fact, medical education researchers can rarely search the literature and find likely numbers for expected differences. One strategy is to ask experts—knowledgeable medical educators: What is the *minimum difference* that would convince them that one approach is better

BOX Calculating Sample Size

1. Determine or estimate the baseline outcome average (eg, control or comparison group).
2. From literature or experts, estimate the least difference that is educationally meaningful for the context.
3. Estimate the variability in outcomes expected (from past history or experts).
4. Choose type II (beta) error (such as 0.20) or power (such as 0.80), depending upon the importance of subsequent decisions.
5. Choose type I (alpha) error, often 0.05.[2]
6. Consider potential losses of subjects (trainees, faculty) or other losses during the study.
7. Use a sample size calculator[7] or consult a friendly biostatistician.

than another? This strategy is also used in clinical research. For example, how big a difference is clinically meaningful, on a scale of 0 to 70, for a cognitive scale? In one study, clinicians chose a 4-unit difference as meaningful.[6]

Next, the expected variability will need to be determined or estimated. If pilot work has been done, this may generate an estimate of variability. A literature review might reveal the standard deviation for a scale or, if a literature search yields no information, knowledgeable experts could weigh in. Finally, you must decide on the alpha level (usually 0.05) and power (usually 0.80) and enter these pieces of information into a sample size calculator to determine the sample size[7] (see BOX).

With an extremely large sample size, the power is great to find a very small difference, which may not be meaningful educationally. Thus, there is always a balance between reality—how many subjects or tests can you afford to include—vs what is a meaningful difference, in a given educational context.

Consider an example: the Associate Dean for GME wishes to determine whether a full day interactive conference on professionalism will result in fewer residents put on probation for professionalism issues in a large institution with rising numbers of incidents. The current program uses several required online case-based videos. Fifty incidents were reported last year (5% of 1000 total trainees), with an average of 4% overall for the past 5 years (25, 30, 45, 50, and 50 annually, respectively). The GME Executive Committee decides that a difference of 15 fewer incidents (ie, 3.5%), compared to last year's report of 5% of incidents, would be meaningful, given the cost and faculty effort of the new strategy. In this case, a comparison of proportions will be performed. The proportions, along with the chosen alpha and beta levels, are entered into a sample size calculator for a test of proportions. It turns out, even with this very large "minimum" difference, the sample size needed, using an alpha of 0.05 and beta of 0.20 (with power of 0.80), is 1505[7]; the study would require 2 years. This is because the incidence of the outcome measure—here professionalism reports—is not common in this population. The 30% relative reduction in reports (5% down to 3.5%) corresponds to just a 1.5% absolute difference.

Another example concerns a program director (PD) who is interested in determining if a new required subspecialty rotation will increase the average In-Training Examination subscore, for the 100 residents in the program, by at least 5 points (to 55); the current average subscore is 50. Will 1 year—with 100 subjects in the new rotation—be sufficient to compare to the prior average? Using an estimated standard deviation of 10, obtained from the previous year's data, the PD will need 60 to 70 residents to determine if the rotation improved the average score by at least this much; thus, 1 year is feasible.[7]

Making more than one inference from your data requires different power calculations, which will not be discussed here.

What About Confidence Intervals?

The confidence interval (CI), with a level often set at 95%, estimates that the true mean difference is within this interval range. In testing differences between groups, when the 95% CI excludes 0, one is more confident that the finding is not due to chance (a type I error). CIs depend upon power calculations, in that statistical power affects the width (or precision) of the CI: as power (sample size) increases, the width of the CI narrows around the effect size difference.

Reporting Power Calculations

Although accurate reporting of sample size calculations has improved in the clinical literature—up to 34% in a 2009 review of high impact general medical journals[8]—in our experience it remains low in medical education studies. For this reason, some journals, including the *Journal of Graduate Medical Education*, are hesitant to consider studies with small numbers of participants or outcomes, such as less than 40. The precise method and numbers used for calculating the sample size should be described in the Methods section. If this will require too many words, authors may add this information in the supplementary data. We recognize the difficulty of making assumptions about effect size and sample variability (standard deviation), but the reasoning surrounding your assumptions should be outlined for readers.

Transparency will enhance credibility, as well as aid in potential future replication of your work.

If a power calculation was not done prior to the study, which might occur with new educational experiences that must get underway before study considerations, authors should state the lack of a power calculation in the Methods section: a single sentence will suffice. It is not possible to calculate a sample size after the study data has been collected and analyzed, due to risk of bias. A post hoc calculation may be useful for your next project, but it will not serve as evidence when discussing a lack of difference observed in the current study.

Conclusions

We welcome quantitative studies and expect that power calculations will usually be performed before you assemble or analyze data, preferably even earlier, at the very beginning of projects. Often there are no existing studies that can provide an estimate of what size difference you may expect for use in calculating a sample size. Not a problem! Assemble a few wise experts, local or national, and ask them how large a difference would be needed to determine that a difference, in this context, is educationally meaningful. What is the minimum difference, with the outcome measure(s) you plan to use, that would be convincing regarding the value of the intervention under study? Use this number to calculate the likely needed sample size. Provide specific details of your reasoning and process in the Methods section. Given the small size of many education programs, you may need to repeat interventions or assessments, or add sites, to garner sufficient numbers; this will also increase the generalizability of your results to other settings and subjects. In the Discussion section, be transparent about error and never, ever, present a post hoc power calculation as the justification for why you did not find a difference.

Please let us know if this article is helpful, and how you have tackled the thorny problems of sample size and power in your medical education projects (Twitter @JournalofGME).

References

1. Shreffler J, Huecker MR. Type I and type II errors and statistical power. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing; 2021.
2. Sullivan GM, Feinn RS. Facts and fictions about multiple comparisons. *J Grad Med Educ*. 2021;13(4):457–460. doi:10.4300/JGME-D-21-00599.1
3. Greenland S, Senn SJ, Rothman, KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350. doi:10.1007/s10654-016-0149-3
4. Sullivan GM, Feinn RS. Using effect size—or why the P value is not enough. *J Grad Med Educ*. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1
5. Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. *Med J Aust*. 2003;177(7):256–257. doi:10.5694/j.1326-5377.2003.tb05240.x
6. Raina P, Santaguida P, Ismaila A, et al. Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. *Ann Intern Med*. 2008;148(5):379–397. doi:10.7326/0003-4819-148-5-200803040-00009
7. ClinCalc. Sample size calculator. https://clincalc.com/Stats/SampleSize.aspx. Accessed September 30, 2021.
8. Charles P, Giraudeau B, Dechartres A, et al. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*. 2009;338:b1732. doi:10.1136/bmj.b1732

Gail M. Sullivan, MD, MPH, is Associate Director for Education, Center on Aging, and Professor of Medicine, University of Connecticut Health Center, and Editor-in-Chief, *Journal of Graduate Medical Education* (*JGME*); and **Richard S. Feinn, PhD,** is Associate Professor of Medical Sciences, Quinnipiac University, and Biostatistics Editor, *JGME*.

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut Health Center, gsullivan@uchc.edu, Twitter @DrMedEd_itor