American Journal of Epidemiology
Vol. 00, No. 00
https://doi.org/10.1093/aje/kwac019
Advance Access publication:

Practice of Epidemiology

Casting New Light on Statistical Power: An Illuminating Analogy and Strategies to Avoid Underpowered Trials

Michaela Kiernan* **and Michael T. Baiocchi**

* Correspondence to Dr. Michaela Kiernan, Stanford Prevention Research Center, Stanford University School of Medicine, 3180 Porter Drive, MC 5702, Palo Alto, CA 94304-1212 (e-mail: mkiernan@stanford.edu).

Initially submitted December 8, 2020; accepted for publication January 28, 2022.

Current standards for methodological rigor and trial reporting underscore the critical issue of statistical power. Still, the chance of detecting most effects reported in randomized controlled trials in medicine and other disciplines is currently lower than winning a toss of a fair coin. Here we propose that investigators who retain a practical understanding of how statistical power works can proactively avoid the potentially devastating consequences of underpowered trials. We first offer a vivid, carefully constructed analogy that illuminates the underlying relationships among 3 of the 5 essential parameters—namely, statistical power, effect size, and sample size—while holding the remaining 2 parameters constant (type of statistical test and significance level). Second, we extend the analogy to a set of critical scenarios in which investigators commonly miss detecting intervention effects due to insufficient statistical power. Third, we highlight effective pragmatic strategies for the design and conduct of sufficiently powered trials, without increasing sample size.

analogy; data visualization; effect size; randomized controlled trial; retention; rigor; sample size; statistical power

Abbreviation: NIH, National Institutes of Health.

Across disciplines, randomized controlled trials are the most rigorous study design for establishing whether and how much an intervention has an effect on an outcome—specifically, whether the intervention itself, rather than other (nuisance) variables, causes an outcome measure to change. The rigor of randomized controlled trials is due in part to intentionally integrating key methodological principles into the trial design to anticipate and reduce potential biases, the need for which has been strongly underscored by the recent National Institutes of Health (NIH) guidance to improve the "rigor and reproducibility" of NIH clinical trials and preclinical experiments (1, 2). Two key principles for designing randomized trials include randomizing participants to either an intervention or control group, rather than allowing research staff or participants to select the group, and ensuring that research staff assessing the trial's outcome measures are "masked" (i.e., research staff are unaware which group participants are randomized to) (3).

Statistical power is also a key methodological principle for designing randomized controlled trials, but it is less well understood and integrated into randomized trials by investigators. Statistical power is the probability that a trial's intervention effect will be detected, if the effect is there. The goal is to have a reasonable chance of detecting the intervention effect size (i.e., detecting a target difference between the intervention and control groups on a primary outcome) (4). The definition of a reasonable chance has relied on a convention of $\geq$80% statistical power (5), recently raised to $\geq$90% by some funders (5). Despite the pitfalls of arbitrary and dichotomous cutoffs, and keeping in mind that greater statistical power is preferred, a trial is typically designed to attain sufficient statistical power of $\geq$80%, or said another way, that there is an $\geq$80% chance (or probability of $\geq$0.8) that the trial's intervention effect will be detected, if the effect is there (5).

The reality for actual trials may be quite different, as concern exists about the ability of current trials to detect clinically important effect sizes. In a comprehensive analysis of 11,852 meta-analyses comprising 136,212 randomized clinical trials in medicine from 1975–2014, only 50% of the

meta-analyses ($n = 5{,}903$) observed a statistically significant intervention effect size for their reported outcomes (6), never mind exhibited clinical significance.

As a rule, assessing statistical power per se is reserved for designing trials, not completed trials (7). Yet, prudently extracting particular data from completed trials and meta-analyses could inform the design of future trials (6, 8), while still recognizing that important distinctions exist among individual trials. The authors of the comprehensive analysis estimated the statistical power for a similar set of hypothetical future trials—using sample sizes from the completed individual trials and using observed intervention effect sizes from an individual trial's respective meta-analysis (not effect sizes of the individual trials themselves). Remarkably, even for the subset of significant meta-analyses, only 15% of recent individual trials (2000–2014) were estimated to have sufficient power to detect an effect size in a future trial as large as the observed effect size from the individual trial's respective meta-analysis. Indeed, median statistical power was only 23% to detect similar effect sizes in future trials (6). Stepping back, and caveats aside, a mere 23% chance of detecting an effect, if it is there, is not only far below the convention of ≥80% power but also disconcertingly lower than winning a coin toss, assuming a fair coin.

Although the reported outcomes above may have included primary outcomes with sufficient power and secondary outcomes without, recent evidence in subfields of medicine suggests that reporting of primary and secondary outcomes remains disturbingly fluid: 21% of trials published primary outcomes that were not prespecified, and 8% published prespecified secondary outcomes subsequently converted to primary outcomes (9). Nevertheless, whether reported outcomes are primary, secondary, or exploratory, their results have entered the literature and currently shape clinical conclusions.

To rectify the pernicious problem of underpowered trials, investigators in medicine and other areas of science are now asked to follow recommended guidelines to comprehensively state a trial's key statistical parameters a priori, including planned sample size and statistical power (4); register a trial and its planned sample size on https://clinicaltrials.gov/ or other trial registration websites before initiating recruitment (10); and follow detailed requirements for reporting a trial's results in academic journals, including the trial's effect size, statistical power, and planned and actual sample sizes (3). Unfortunately, these guidelines have not been sufficient to substantially increase statistical power in recent trials.

WHY THE DISCONNECT?

We speculate that investigators may be aware of statistical power but not thoroughly understand it. Investigators may only vaguely recall a statistically accurate but not terribly practical statement that "statistical power is the probability of rejecting the null hypothesis given that the alternative hypothesis is true" or may have a general nagging sense that "false negatives are bad." Additionally, investigators may focus on statistical power during grant submission but then prioritize other issues during subsequent research stages or may feel pressured to finish up trial recruitment within a certain time period to meet research grant milestones, so they "get what they get." A recent review suggests that only 50% of trials reached target sample sizes (11).

Here, we propose that investigators who retain a practical understanding of how statistical power works can proactively avoid the potentially devastating consequences of underpowered trials. To cast new light on the pernicious problem of underpowered trials, we first offer a vivid, carefully constructed analogy illuminating the underlying relationships among 3 of the 5 essential parameters—namely, statistical power, effect size, and sample size—while holding the remaining 2 parameters constant (type of statistical test and significance level). Second, we extend the analogy to a set of critical scenarios in which investigators commonly miss detecting intervention effects due to insufficient statistical power. Third, we highlight effective pragmatic strategies for the design and conduct of sufficiently powered trials, without increasing sample size.

An illuminating analogy for statistical power

Imagine you are lying in bed one night and you think you hear the deep, distinctive hoot of a great horned owl. You've never seen this large, nocturnal owl before (12), so you get up to look for it. You grab the little penlight you use for reading in bed and shine it out the back door. Unsurprisingly, with such a little amount of light, you cannot see very far. Indeed, you would miss seeing the owl, if it was there.

Next, you grab the hefty outdoor flashlight you use for walking outside at night. You shine this bigger light out the back door, and—wow, high in the tree is a large, 2-foot-tall, adult great horned owl staring down at you with large, round, yellow eyes.

In these scenarios, the amount of light you use to look for the owl is analogous to a trial's sample size. The probability of seeing the owl, if it is there, is the trial's degree of statistical power. The size of the owl is the trial's effect size. If a trial has a larger sample size, other statistical parameters being equal, you will have greater statistical power and be more likely to see the owl, if it is there. However, the amount of light you use does not affect whether the owl is there. The owl is either there—or not.

Unfortunately, there is no such thing as a perfect flashlight. The amount of light from the hefty outdoor flashlight you use for finding a large adult owl is woefully underpowered to find a baby owl. Using your hefty outdoor flashlight to find a baby owl is as futile as using your little penlight to find a large adult owl. In both cases, you would miss seeing the owl, if it is there.

Indeed, given that the relationship between sample size and effect size is negative (that is, inverse) and nonlinear, substantially larger sample sizes are required to detect small effect sizes with sufficient statistical power than for either medium or large effect sizes. That is perhaps counterintuitive—sometimes, for something small, there is a correspondingly small price. But to detect small effect sizes, there is a very high price, including having to recruit substantially larger sample sizes. In contrast, considerably smaller sample sizes—which are more feasible to recruit—can easily detect medium effect sizes with sufficient statistical power. Finding

a medium-sized, albeit ungainly, juvenile owl is substantially easier than finding a baby owl.

But those charming baby owls are still pretty tempting. Why not just flip on your backdoor light that is equipped with stadium-level wattage? If an owl is not there, this excessively large blast of light may reveal only a pile of teensy, regurgitated owl pellets, not a baby owl. Faced with ambiguous evidence of desiccated pellets rather than a live owl, you cannot imagine waking up your household to boast you "saw" an owl. Yet, studies with excessively large sample sizes often report findings with teensy, inconsequential effect sizes (i.e., extremely small effect sizes) that may indeed be statistically significant given the excessively large sample sizes but that are clinically meaningless. This practice distracts the field by squandering scarce research funds, wasting researchers' and participants' time on a subsequent rash of unneeded, uninformative future studies.

Studies with excessively large sample sizes (and indeed studies of all sample sizes) are also often subjected to unruly hordes of examinations by legions of research teams, leading to a statistically fraught phenomenon known as multiple hypothesis testing (aka "*P*-hacking" or data dredging) (13–15). If done without adjusting the significance level for individual hypotheses (16), multiple hypothesis testing boosts the chance of finding something (or anything) that appears to be statistically significant but is actually a transitory illusion. The owl you think you see is instead a windblown plastic bag caught high in a tree.

Suppose you have successfully recruited a sample size that is able to detect a large effect size with sufficient statistical power. You are not out of the woods yet. A pervasive, neglected issue can silently sabotage ongoing trials. Trials that successfully recruit their planned sample sizes, but later experience poor retention as participants continue to drop out over the course of a trial, lose substantial statistical power. This loss of statistical power can substantially reduce a trial's ability to detect even large effect sizes. Using your hefty, outdoor flashlight with dying batteries to find an adult owl is as futile as using your little penlight to find an adult owl. Once again, you would miss seeing the owl, if it is there.

Underpowered trials have potentially devastating consequences

One clear consequence of an underpowered trial is that investigators can completely miss detecting the effect of the very intervention they are testing. As described above and in Figure 1, there are a set of critical scenarios where intervention effects are commonly missed due to insufficient statistical power—recruiting initial sample sizes that are too small for expected effect sizes; fruitlessly chasing after teensy, clinically meaningless effects; and failing to retain initially sufficient sample sizes.

Another common consequence of underpowered trials is that investigators and, perhaps more importantly, consumers of trial results—including other researchers, field practitioners, journalists, and the general public—may misinterpret the lack of results from an underpowered trial as evidence that the trial intervention itself "didn't work." This is an incorrect conclusion, given that the "absence of evidence is not evidence of absence" (17), which can prematurely close off potentially worthwhile lines of inquiry. A more appropriate conclusion is that the trial's intervention effect may not have had a good enough shot (or probability) of being detected because of the lack of statistical power, even if the trial's intervention indeed worked. An indication that a trial was underpowered, despite a sizable intervention effect size, is the presence of wide confidence intervals, suggesting low precision and a large degree of uncertainty regarding the size of the actual effect (7).

A final consequence of underpowered trials is ethical (18). Underpowered trials provide "low informational value" (6), as they suffer from a lack of rigor and reproducibility. A recent NIH guidance underscores the importance of including sample size calculations in grant applications to ensure sufficient statistical power in both preclinical experiments and clinical trials (1, 2). Underpowered trials lead to lack of reproducibility (i.e., tests of the same intervention producing different and inconsistent conclusions), which needlessly wastes immense government, foundation, and industry funding of thousands of trials annually; undermines efforts of research teams to retain trial participants; and compromises the invaluable contributions of trial participants.

Strategies to avoid underpowered trials

Although median sample sizes per randomized trial group in medicine have doubled to nearly 100 since 1975, median effect sizes themselves have remained small, leaving median power increasing only to an unacceptable 23% (6). Yet, trials across disciplines could be explicitly designed to routinely detect larger effect sizes as a minimum, requiring considerably smaller (and more feasible) sample sizes, thus enhancing trial feasibility, rigor, and reproducibility (1, 2).

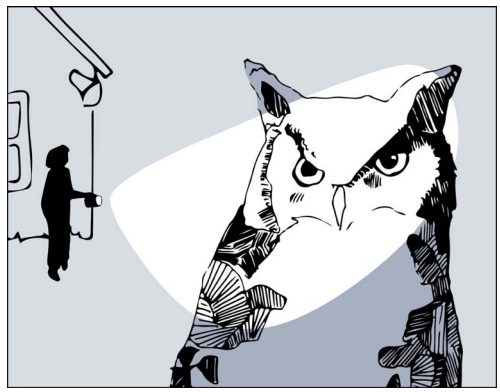
HOW TO DO THIS

Chose to detect an adequately large effect size

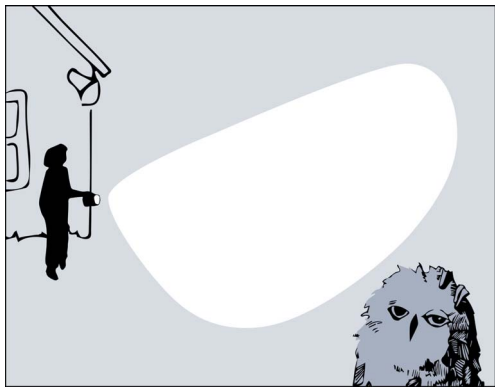
One important strategy is to avoid chasing clinically meaningless effects from the start. When testing whether an intervention group affects an outcome more than a control group (or more generally, a comparator group) in a randomized trial, the intervention effect size is defined as the target difference between the intervention and comparator groups on a primary outcome and should be specified using the original scale (4), such as the mean difference between the 2 groups on a continuous outcome.

Ideally, when designing a trial, the choice of a target difference for a proposed primary outcome needs to be large enough to be clinically relevant, realistic, and important to stakeholders (4, 19). To evaluate clinical relevance, examine the empirical relationship of the proposed target difference relative to a separate, clinically meaningful outcome in prior trials (4), such as proposing a particular reduction in systolic blood pressure between the intervention and comparator group as a target difference empirically shown to reduce cardiovascular events and mortality in

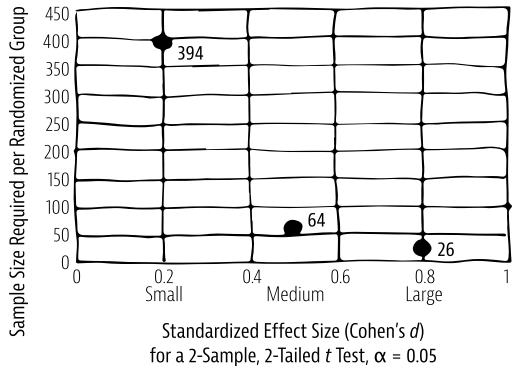
Scenario 1. Studies with small sample sizes will *not* have sufficient statistical power to detect even large effect sizes.



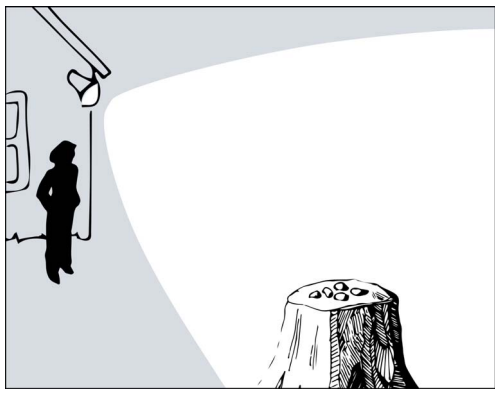
Scenario 2. Studies with large sample sizes have sufficient statistical power to detect large effect sizes.



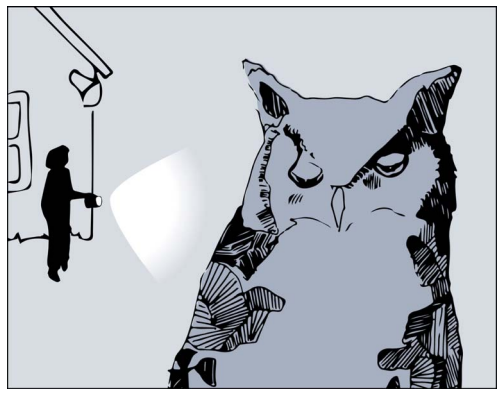
Scenario 3. Studies with large sample sizes will *not* have sufficient statistical power to detect small effect sizes.



Scenario 4. Studies require substantially larger sample sizes to have sufficient statistical power to detect small effect sizes. Considerably smaller (and more feasible) sample sizes can detect medium effect sizes.



Scenario 5. Studies with excessively large sample sizes have sufficient statistical power to detect teensy but clinically meaningless effect sizes.



Scenario 6. Studies that successfully recruit planned sample sizes but later experience poor retention will *not* have sufficient statistical power to detect even large effect sizes.

Figure 1. An illuminating analogy for statistical power. (A wee bit of artistic license was taken with details of ornithology and optical physics to illustrate relationships among statistical parameters.)

prior large-scale, sufficiently powered trials with similar sample characteristics (20, 21). If a target difference is not adequately large to be clinically relevant, viable options exist, such as wisely modifying the research question to assess a primary outcome earlier in the disease process, given the current state of the science.

Cohen's *d*, a standardized effect size (22), can be a valuable heuristic technique for systematically thinking about

how to design a trial with an adequately large effect size—by 2 distinct routes. Cohen's *d* neatly separates 2 important quantities: the mean target difference between the 2 groups (the numerator) divided by the variability within the 2 groups (the denominator), specifically the pooled standard deviation (given standard assumptions such as independent samples and equal sample sizes) (5). Thus, think about increasing the standardized effect size by not only increasing the target difference between groups (the numerator) but also decreasing the variability within groups (the denominator) (23).

To increase the numerator, increase the impact of the intervention dose. To increase dose, investigators typically increase the difference between the intervention and comparator groups by designing or selecting a stronger intervention to test against the comparator group. However, selecting a stronger intervention is not feasible in all trials, and selecting a deliberately weak comparator group has ethical ramifications, especially for high-risk subgroups.

Ironically, the small, observed median effect sizes in medicine cited above may be due in part to investigators underestimating how well the planned comparator group will do rather than overestimating how well the intervention group will do (24). Recent NIH Expert Panel recommendations suggest that selecting optimal comparator groups should be strategically informed by the type of research question for a given research stage rather than relying on a single, accepted, familiar, artificial, or "gold" standard within a discipline (25). For example, the selected comparator group may differ when exploring early on "whether (the intervention) works at all" versus determining later on "how well (the intervention) works relative to a clinically relevant alternative" (25).

As mentioned above, Cohen's *d* can be used as a heuristic technique for systematically thinking about not only how to increase the target difference (the numerator), but also, and distinctly, how to decrease the variability (the denominator). Specifically, reduce measurement variability of trial outcomes, such as intentionally taking reliable, duplicate, and paired measurements (5, 26). For example, for body weight, use a reliable, accurate scale to take 2 measurements for each individual, and take these 2 measurements both at the beginning and end of the trial. Often overlooked, the strategy of reducing variability to increase the standardized effect size (and thus reduce sample size) is compelling. Given a continuous outcome and standard assumptions (Cohen's $d = 0.5$, 80% statistical power, 2-sample *t* test, 2-tailed $\alpha = 0.05$, Student's *t* distribution), reducing the standard deviation of 1 by 25% reduces the estimated sample size by 42%. Indeed, merely by reducing the variability, the same target difference (on the original scale) can be more easily detected.

Ensure excellent intervention fidelity

A second and underutilized strategy that greatly increases intervention impact and statistical power at any (and every) dose is ensuring excellent intervention fidelity for the entire trial (23). For instance, when evaluating intervention receipt, set explicit benchmarks to evaluate not only whether participants received any of the intervention at all (e.g., percentage of participants who attended $\geq$1 intervention sessions) but, importantly, whether most participants received high levels of the intervention (e.g., whether $\geq$80% of participants attended $\geq$80% of intervention sessions). Assess such benchmarks early and often (in both intervention and comparator groups) so suboptimal fidelity is quickly addressed (27). In 3 recent weight-management trials, the percentage of participants meeting a high benchmark early on (attended $\geq$80% of intervention sessions) ranged considerably across trials (13.9%, 52.2%, and 79.0%) (27).

Best practices for intervention fidelity exist (28, 29) and continue to be refined (30, 31). The NIH Behavioral Change Consortium recommended comprehensively assessing 5 dimensions: intervention design (e.g., intended dose), provider training (e.g., standardized format), treatment delivery (e.g., delivered dose by providers), treatment receipt (e.g., received dose such as participant intervention attendance), and treatment enactment (e.g., participant skills performance) (28). However, implementation remains underutilized. In 2005, only 30% of health behavior trials assessed provider adherence with an explicit mechanism (e.g., audio recording or provider self-report) (28). By 2017, assessing provider adherence with an explicit mechanism had doubled to 67%, but 70% of trials did not specify the assessment's reliability and validity (30). Intervention fidelity strategies can be implemented more consistently, while being careful not to fundamentally change the intervention, including leveraging unobtrusive, time-stamped digital approaches.

Maintain excellent retention

A third strategy for avoiding underpowered trials is to maintain excellent retention of trial participants for the entire trial. To do this, integrate not only conventional extrinsic strategies such as financial incentives and appointment reminders (32) but also innovative retention strategies intentionally implemented prior to randomization and included in trial protocols from the start (27, 33). With the Methods-Motivational Interviewing approach (27, 33), interactive orientation sessions are held for potential participants well before randomization, giving potential participants time and space. Sessions are composed of: 1) easy-to-understand research methods modules that foster participant research literacy and nurture participant buy-in during recruitment for essential trial procedures such as randomization and retention at future follow-up visits; and 2) facilitated use of motivational interviewing techniques to explore and diffuse possible ambivalence about participating in a research trial or changing trial-related behaviors (27, 33). Offered via in-person, online, or telephone formats for small groups of individuals, these interactive orientation sessions are associated with 8.8% higher rate of intervention session attendance and 11.4%–17.3% higher rates of trial retention at 12- and 18-month clinic visits (27, 34). Methods-Motivational Interviewing strategies can work individually. In online experiments, an easy-to-understand, visually powerful 1-page infographic letter—illustrating the detrimental impact of dropouts on trial conclusions and that a "true picture" of trial outcomes was preferred scientifically by trial investigators regardless of individual participant success or failure—substantially

improved participant research literacy and participant trust in the research team relative to a control letter (35).

Design innovative early-phase studies

What if sufficient power is not possible? Do not initiate miniature versions of larger, proposed full-scale randomized trials just to "collect a little data." Often, such data are inappropriately used to estimate effect sizes for larger full-scale trials (19). These effect sizes, including variance estimates, are more likely to be extreme (high or low) and unstable, exactly because they are based on small sample sizes (19). Indeed, if there was sufficient statistical power for the miniature-version trials, larger full-scale trials would not be needed (19).

Rather, a fourth strategy for avoiding underpowered trials is to leverage pilot resources to design a strategic set of innovative early-phase studies rather than a miniature trial. The NIH-funded Obesity-Related Behavioral Intervention Trials (ORBIT) Consortium proposed a systematic framework for behavioral treatment development clearly delineating several early phases, useful for defining, refining, and optimizing interventions prior to conducting phase III efficacy trials (36). Building on ORBIT, a subsequent NIH-sponsored workshop compiled a rich range of early-phase study designs (37), spanning mixed methods, *N*-of-1 trials, adaptive treatments, fractional factorial trials that identify independent and interactive intervention components, and others. Integrate these designs with valuable, prespecified, numeric milestones, such as feasibility and acceptability metrics (38), to fully inform future large, full-scale, sufficiently powered trials (36). Indeed, innovative early-phrase designs conducted by different research teams with diverse methodological expertise can avoid myriad small individual trials and instead collectively inform a few, definitive, full-scale, multisite, collaborative trials amply powered to compellingly answer a discipline's central, pressing priorities (39).

ENHANCING BEHAVIOR CHANGE FOR INVESTIGATORS

Behavior change of any kind, whether encouraging adults to become more physically active or urging investigators to conduct sufficiently powered trials, requires a comprehensive array of behavioral change techniques (40) beyond simply advising individuals to implement a behavior for their own good, regulating their behavior, or outing them as punishment if they fail. Indeed, how individuals learn can facilitate not only what they retain but, even more importantly, how well they "transfer" or apply what they have retained to novel situations and challenges (41).

Here, to provide a practical understanding of statistical power, we incorporated 4 evidence-based practices from the education field that improve how individuals learn, retain, and transfer new, complex concepts (41). First, we carefully constructed an analogy that built upon on individuals' prior knowledge from a familiar context (flashlights in the dark) to explicitly introduce the less familiar, complex, underlying relationships among 3 essential parameters of statistical power. Second, we integrated "just-in-time telling" by intentionally waiting to explain the underlying relationships among statistical parameters until after the analogy was introduced, because individuals transfer "deep structure" among concepts more successfully if they first work it out on their own (41). Third, we leveraged the use of contrasting cases (the set of critical scenarios when intervention effects are commonly missed) that systematically isolated each statistical parameter in turn, essential for developing "expert mastery" of nuanced important distinctions among concepts, not just bringing general awareness to novices (41). Fourth, we integrated worked examples (the explicit explanations and specific strategies to increase statistical power without increasing sample size) to facilitate future efficient transfer (41). To complement these evidence-based practices, we harnessed the data visualization technique of "small multiples" for visual depiction of the analogy scenarios, which repeated basic elements across images to accentuate noteworthy elements that changed across images (parameters of statistical power) (42).

LOOKING AHEAD

Investigators who retain a practical understanding about the underlying relationships among statistical power, sample size, and effect size will be better equipped to prioritize statistical power in future trial decision making, avoid relying on sample size estimates erroneously driven by budget or staff constraints, and ensure that adequate resources are allocated for trial retention. Discerning that penlights cannot detect owls of any size exposes the tough reality of false negatives and the ultimate futility of underpowered trials. No matter how hard research teams work (43), if trials are underpowered, intervention effects are likely to be missed, even when the effects are there.

ACKNOWLEDGMENTS

Author affiliations: Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California, United States (Michaela Kiernan); and Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, United States (Michael T. Baiocchi).

This work was supported by the National Heart, Lung, and Blood Institute at the National Institutes of Health (grant R01 HL128666 to M.K.) and a Stanford Cancer Institute Innovation Award (grant to M.K.). The Stanford Cancer Institute is a National Cancer Institute-designated Comprehensive Cancer Center. M.K. used personal funds to pay the artist who created the illustrations in Figure 1.

We thank Lee Cline for her marvelous owls, integration of information design, and attention to detail (https://www.leecline.com/). We also thank the following colleagues for their insightful feedback: Dr. Nancy Ellen Kiernan, Dr. Michael T. Kiernan, Dr. Marily A. Oppezzo, Dr. Erin A.

Vogel, Dr. Disep I. Ojukwu, Dr. J. Bradley Zuchero, Dr. Jamie M. Doyle, and Ashley Gray.

The views expressed in this article are those of the authors and do not reflect those of the National Institutes of Health or the Stanford Cancer Institute.

Conflict of interest: none declared.

REFERENCES

1. National Institutes of Health. Guidance: rigor and reproducibility in grant applications. https://grants.nih.gov/policy/reproducibility/guidance.htm. Accessed: February 7, 2021.
2. Ramirez FD, Motazedian P, Jung RG, et al. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res.* 2017;120(12):1916–1926.
3. Moher D, Hopewell S, Schultz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.
4. Cook JA, Julious SA, Sones W, et al. DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ.* 2018;363:k3750.
5. Browner WS, Newman TB, Hulley SB. Estimating sample size and power: applications and examples. In: Hulley SB, Cummings SR, Browner WS, et al., eds. *Designing Clinical Research*. 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.
6. Lamberink HJ, Otte WM, Sinke MRT, et al. Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *J Clin Epidemiol.* 2018;102:123–128.
7. 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.
8. Turner RM, Bird SM, Higgins JPT. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLoS One.* 2013;8(3):e59202.
9. Lancee M, Lemmens CMC, Kahn RS, et al. Outcome reporting bias in randomized-controlled trials investigating antipsychotic drugs. *Transl Psychiatry.* 2017;7(9):e1232.
10. Zarin DA, Crown WH, Bierer BE. Issues in the registration of database studies. *J Clin Epidemiol.* 2020;121:29–31.
11. Walters SJ, dos Anjos B, Henriques-Cadby I, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. *BMJ Open.* 2017;7(3):e015276.
12. Cornell Lab of Ornithology. *All About Birds: Great Horned Owl Overview*. Ithaca, NY: Cornell Lab of Ornithology. https://www.allaboutbirds.org/guide/Great_Horned_Owl/overview. Accessed February 7, 2021.
13. Smith GD, Ebrahim S. Data dredging, bias, or confounding: they can all get you into the *BMJ* and the Friday papers. *BMJ.* 2002;325(7378):1437–1438.
14. Benjamini Y. Simultaneous and selective inference: current successes and future challenges. *Biom J.* 2010;52(6):708–721.
15. Young SS, Karr A. Deming, data and observational studies: a process out of control and needing fixing. *Significance.* 2011;8(3):116–120.
16. Shaffer JP. Multiple hypothesis testing. *Annu Rev Psychol.* 1995;46:561–584.
17. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. *BMJ.* 1995;311:485.
18. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA.* 2002;288(3):358–362.
19. 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(5):484–489.
20. Beddhu S, Chertow GM, Greene T, et al. Effects of intensive systolic blood pressure lowering on cardiovascular events and mortality in patients with type 2 diabetes mellitus on standard glycemic control and in those without diabetes mellitus: reconciling results from ACCORD BP and SPRINT. *J Am Heart Assoc.* 2018;7(18):e009326.
21. SPRINT Research Group, Lewis CE, Fine LJ, et al. Final report of a trial of intensive versus standard blood-pressure control. *N Eng J Med.* 2021;384(20):1921–1930.
22. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
23. Lipsey MW. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage Publications; 1990.
24. Moroz V, Wilson JS, Kearns P, et al. Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. *Trials.* 2014;15:481.
25. Freedland KE, King AC, Ambrosius WT, et al. The selection of comparators for randomized controlled trials of health-related behavioral interventions: recommendations of an NIH expert panel. *J Clin Epidemiol.* 2019;110:74–81.
26. Kraemer HC. To increase power in randomized clinical trials without increasing sample size. *Psychopharmacol Bull.* 1991;27(3):217–224.
27. Jake-Schoffman DE, Brown SD, Baiocchi M, et al. Methods-Motivational Interviewing approach for enhanced retention and attendance. *Am J Prev Med.* 2021;61(4):606–617.
28. Borrelli B, Sepinwall D, Ernst D, et al. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *J Consult Clin Psychol.* 2005;73(5):852–860.
29. Bellg AJ, Borrelli B, Resnick B, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychol.* 2004;23(5):443–451.
30. Walton H, Spector A, Tombor I, et al. Measures of fidelity of delivery of, and engagement with, complex, face-to-face health behaviour change interventions: a systematic review of measure quality. *Br J Health Psychol.* 2017;22(4):872–903.
31. Walton H, Spector A, Williamson M, et al. Developing quality fidelity and engagement measures for complex health interventions. *Br J Health Psychol.* 2020;25(1):39–60.
32. Brueton VC, Tierney JF, Stenning S, et al. Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. *BMJ Open.* 2014;4(2):e003821.
33. Goldberg JH, Kiernan M. Innovative techniques to address retention in a behavioral weight-loss trial. *Health Educ Res.* 2005;20(4):439–447.
34. Mayhew M, Leo MC, Vollmer WM, et al. Interactive group-based orientation sessions: a method to improve adherence and retention in pragmatic clinical trials. *Contemp Clin Trials Comm.* 2020;17:100527.

35. Kiernan M, Oppezzo MA, Resnicow K, et al. Effects of a methodological infographic on research participants' knowledge, transparency, and trust. *Health Psychol.* 2018; 37(8):782–786.
36. Czajkowski SM, Powell LH, Adler N, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. *Health Psychol.* 2015;34(10): 971–982.
37. Naar S, Czajkowski SM, Spring B. Innovative study designs and methods for optimizing and implementing behavioral interventions to improve health. *Health Psychol.* 2018; 37(12):1081–1091.
38. Voils CI, Adler R, Strawbridge E, et al. Early-phase study of a telephone-based intervention to reduce weight regain among bariatric surgery patients. *Health Psychol.* 2020;39(5): 391–402.
39. Freedland KE. The Behavioral Medicine Research Council: its origins, mission, and methods. *Health Psychol.* 2019; 38(4):277–289.
40. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med.* 2013;46(1):81–95.
41. Schwartz DL, Chase CC, Oppezzo MA, et al. Practicing versus inventing with contrasting cases: the effects of telling first on learning and transfer. *J Educ Psychol.* 2011;103(4): 759–775.
42. Tufte ER. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press; 1983.
43. Fisher RA. *The Design of Experiments*. 9th ed. New York, NY: Hafner Press; 1971.