Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

SPECIAL FEATURE: THE POST HOC POWER CONTROVERSY

Response: The Proliferation and Misinterpretation of "As Safe As" Statements in Surgical Science: A Call for Professional Discourse to Search for a Solution

We want to thank Mr Griffith and his co-author for their professionalism in voicing their opinions, which, unfortunately, is becoming increasingly rare in the days of social media, an issue we will come back to at the end of this letter. Their professionalism should be congratulated and even celebrated.

We would like to begin our response to their thoughtful letter by noting that most misunderstandings occur when the context is lost. Although we appreciate all the excellent comments that have been raised regarding our article, we would like to take a moment to redescribe the exact context of the problem that we are concerned about, describe why we are not concerned about the other scenarios that have been raised by some readers, and then respond to a few specific technical comments.

The context

In a nutshell, in the surgical literature, we have an epidemic of "as safe as" statements that are trying to promote a new surgical therapy without sufficient data.

We are concerned about studies that, for example, report that a novel treatment has a complication rate of 20%, whereas the traditional treatment has a complication rate of 10%. In addition, the authors nevertheless assert that the novel treatment is "as safe as" the traditional treatment and advocate for its adoption, based on nothing other than these two percentages and a P value of $P > 0.05$. Those of us in the surgical community are very familiar with this problem. In addition, it has been reported repeatedly over the past 2 decades.[1,2] Similar problems are seen outside of surgery, but the problem may be particularly worse in surgery because surgical innovations—at least when they do not involve new drugs or devices—do not fall under any regulatory body. And so while innovations in nonsurgical fields are almost always examined by a competent regulatory body, they are not in surgery, and so surgical innovations are almost always promoted simply with these types of studies. In fact, nearly all of the reversals in surgical innovations can be traced back to these studies with type II errors. It is a huge problem.

To state it another way, we are concerned about (1) small studies that are already published, and continue to be published, that (2) reported nonsignificant P values, but (3) with nontrivial effect sizes, and yet (4) the authors claim that two treatments are "as safe as" each other. All four of these conditions must exist, for our proposal to be applicable. These articles often do not report sample size calculations, either reflecting an omission in the article writing, or perhaps reflecting a true oversight in the study design stage. Unfortunately, they occur quite often in surgery, perhaps to the surprise of our statistical colleagues.

What we are not concerned about, and are not proposing

The concern that Mr Griffith and others have raised about our article is valid, but they are really about different contexts that are different from the ones we are concerned about. For example,

(1) We are not concerned about large studies that reported a nonsignificant trivial effect size. The average readers would not be motivated to dissect the statistics further or consider adopting a new therapy, if the reported event rates were only 10.1% in one group *versus* 10.0% in another group. Moreover, these large studies are rare in surgery, and they often occur after surgical innovations have been adopted by the majority already, and as such, they are outside of the context of the "unsafe surgical innovation" problem that we are concerned about. This, again, is a major difference between surgery and nonsurgical science: Although most nonsurgical innovations, like pharmaceuticals, are introduced after a large trial, surgical innovations are nearly always introduced with small studies.

Both authors contributed to conception of the manuscript, drafting of the article, and final approval of the submitted version.

(2) We are also not concerned about small studies that happen to have detected a large difference that was statistically significant. This issue is also rare in surgery because in reality, these small studies in surgery almost never detect a statistical difference. Although we recognize the flaws of *P* values, the reality is the average readers of surgical journals are still impressed with findings of $P < 0.05$ and would not question the statistics further. However, what they might do is turn their attention toward nonstatistical qualitative issues; for example, they might question the study design, or question whether the study population was representative and generalizable. And so there are safe guards in place when these rare small but significant studies are reported.
(3) We are also not proposing to replace thoughtful study designs and appropriate pre hoc power analysis with our approach of "hacking" the studies and reverse engineering their powers after a study is performed. We agree strongly that pre hoc power analysis is the most appropriate thing to do. However, our study, like several others in the literature, was a literature review research study meant simply to raise alarm for the readers because so many of these studies are already out there in the surgical literature. We do not advocate that the investigators perform post hoc analysis on a regular basis in the future; we merely want to propose a possible approach that the curious readers may use to dissect studies that they question.
(4) We are also not questioning the value of Bayesian statistics and other advanced statistical approaches that may limit the impact of these small studies. However, for the readers of these articles that are already in the literature, it would be impossible for them to be able to do this themselves based on the typical data reported in the studies.

Observed effect size?

Hoping that we have focused the reader's attention on the exact scenario we are concerned about, we now turn to the comment regarding minimum clinically important difference (MCID).

We agree that it would be more ideal to attempt power calculation with an MCID. What prior authors have performed is to use a single, "prespecified" effect size. However, we do not feel that the "single effect size" approach is reasonable because a reasonable argument can be made that the MCID for, say, urinary tract infection should be different from the MCID for death. Therefore, we decided to use different effect sizes for different studies. But then we ran into several practical barriers that led to our approach of using observed effect sizes:

(1) Unfortunately, there is simply no standard and objective definition of MCID for most clinical outcomes. What is clinically meaningful difference in wound complication rates? What is it for readmission rates? And in the rare cases that they do have standards, someone can easily argue that they should change based on the clinical scenarios, and/or for different patient populations.
(2) The reality is the MCID is almost always "made up" in the surgical literature—if the investigators bother to make them up at all, and/or report them in their articles. In fact, if it was not subjectively made up by some "experts", it was often obtained from some prior studies—in essence, the observed effect size of a study often becomes the "prespecified" effect size for future studies. Therefore, the rationale for using observed difference is that they provide a convenient starting point, given the absence of prespecified effect size in the surgical literature (either from the authors, and/or from the literature). It is not ideal, but it is a convenient starting point for the readers.
(3) Moreover, the studies we are concerned about all have nontrivial effect sizes, much larger than any reasonable MCID. The average observed effect size for the studies we reviewed was 0.86! Therefore, by using the observed effect sizes, we were essentially calculating a more conservative estimate of their power and giving them the benefit of the doubt.

Nevertheless, even if we had used the same single, "prespecified" effect size as in previous studies, our results would be identical. As one would expect, among studies whose average effect size is 0.86, switching to a predefined effect size of 0.25 would result in larger sample size requirement and lower power calculation. In other words, the power values that we presented could be construed as the upper bounds of the power that the authors might have reasonably calculated.

Post hoc power with observed effect sizes is simply a transformation of the *P* value?

As Mr Griffith and other statistical colleagues have pointed out, a high *P* value is redundant to low power. We completely agree. Then why do it?

Our thinking is rooted in what health literacy experts have taught us: To communicate not in what we think is best, but in a way that the readers understand. And to the extent that the average readers of surgical literature only understand alpha and beta—and because a high *P* value could be misused to justify an unwarranted claim of "as safe as"—we feel that reporting power offers the only other chance to prevent misinterpretations.

Being redundant is not the same as being wrong.

We do not argue against the value of other advanced statistical approaches to prevent misinterpretations. Unfortunately, the typical readers of surgical journals probably would not understand Bayesian statistics. Even confidence intervals are often misinterpreted. They would still point to the fact that it overlaps the value of no difference and still try to justify their claim of "as safe as". They would miss the fact that the magnitude of the confidence intervals is uncomfortably large.

Encourage more underpowered studies?

We also want to acknowledge Mr Griffith's concern that relaxing the power threshold could "encourage more

underpowered studies to proceed" and "simply increase the number of studies with nonsignificant findings". However, the reality is that these studies are already being published——again, accounting for as much as two-third of the surgical literature by some studies.[1,2] The problem is these studies are being published with a misleading conclusion, attempting to promote a novel therapy without valid justification. We agree with Mr Griffith that these studies should still be considered for publication because there could be some "diamonds in the rough". What we want to make sure is that these authors will frame their conclusions appropriately, with some words of caution, so as to limit the impact of these small studies. We disagree that these studies should be excluded altogether—that would be throwing the baby out with the bathwater. Many surgical innovations do, in fact, come out of these "underpowered" studies. Therefore, we are essentially proposing a middle ground, to continue to allow surgeon scientists to publish these small and limited studies, but to avoid the harms of inappropriate conclusions.

Professionalism on social media

Finally, I want to again thank Mr Griffith and his co-author for setting an example of professionalism in sharing their opinions. Unfortunately, the rise of social media has sometimes allowed a circumvention of due process, and this has enabled the influence of false experts and unprofessional behaviors to contaminate a meaningful discussion.

We believe there are three key steps to professional discourse: (1) submission of concerns to an objective third party, who (2) scrutinizes the concern to see if there is a prima facie case to be made, and who then (3) respects the concept of due process, and invites a response within a reasonable time frame. However, what we are seeing on social media is circumvention of these steps.

Instead of submitting their concerns to an independent third party, what we often see on social media are cases of "self-validation", that is, someone who professes to be both judge and jury. In our legal system, we protect against this possibility with the concept of "probable cause" which, in most cases, requires law enforcement officers to make a case to a judge before being permitted to conduct invasive searches. However, on social media, we came across individuals who made claims that our work was "wrong", but their conclusion was completely self-validated. A telltale sign is that these people would focus on describing their "conclusions", but often omit a detailed description of their methods or data. For example, it turns out many of the so-called "errors" involved differences that existed in the thousandth digit, which, if revealed to a neutral judge, would have been dismissed as meaningless. In addition, they often resort to emotional labels that are unprofessional.

The importance of the neutral third party—that is, the traditional journal editor—cannot be overstated. Although this may not be fail-safe, it can at least screen out some false experts. However, on social media, we now see people who exist behind pseudonyms, not allowing their qualifications and background to be verified. We also see people lacking any organizational affiliations that can vouch for their credentials or integrity. Or worse, we see people who have no relevant formal training—background searches reveal that many of these "experts" are "self-taught". We came across individuals who claim that they are unable to reproduce our work, implying that the work is nefarious and irreproducible—but in all likelihood, in the absence of formal training, it is more likely that they do not have the skills to do the work and/or to interpret the results correctly. Nevertheless, these individuals would push the former narrative and conveniently leave out the latter possibility. These people may be able to produce legitimate-looking documents, often with legitimate-looking citations. However, these citations are often from noncredible, unverifiable sources, or are simply self-citations (another reflection of their tendency to want to be both judge and jury).

Last but not least, a democratic society cherishes the concept of due process and giving the other side reasonable time to respond. Unfortunately, this is often ignored on social media. The operative issue is "reasonable time". For example, we came across an individual who emailed us on a Sunday at midnight with a request for clarification, but then posted on a website by 8 AM Monday morning that he had not received information from us, suggesting that we had something to hide, and continuing his attack. He neglected to mention that the reason he had not received any response was probably because he had only given us 8 h to respond—from Sunday midnight to Monday morning at 8! This is another reason why a neutral third party is imperative, to serve as a repository for challenges and responses, so that a reasoned and balanced discussion can be presented.

Unfortunately, this is not the first time this has happened in the recent history of social media.[3,4] Our case is not the first, and it will certainly not be the last.

It is our hope that we would abstain from using a forum where unprofessional discourse is allowed to flourish, with "expert" comments from unverifiable persons, and where judgment is often pronounced without due process. Otherwise, we would be allowing a hostile takeover of the academic scientific process. As social media evolves, let us call for professionalism. Let us set an example for the rest of our society on how to behave on social media. Let us use the Internet for what it was intended: To collaborate and brainstorm for solutions together.

REFERENCES

1. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature - equivalency or error? *Arch Surg*. 2001;136:796–800.
2. Ahmed Ali U, Ten Hove JR, Reiber BM, van der Sluis PC, Besselink MG. Sample size of surgical randomized controlled trials: a lack of improvement over time. *J Surg Res*. 2018;228:1–7.
3. Longo DL, Drazen JM. Data sharing. *N Engl J Med*. 2016;374:276–277.
4. *Trolling on Social Media is Never a Good Look – That Applies to Academics Too*; 2018. The Guardian; 2018. Available at: https://

www.theguardian.com/higher-education-network/2018/jan/12/trolling-on-social-media-is-never-a-good-look-that-applies-to-academics-too.

David C. Chang, PhD, MPH, MBA*
Sahael M. Stapleton, MD, MBA
Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

**Corresponding author*. Massachusetts General Hospital, 165 Cambridge Street, Suite 403, Boston, MA 02114. Tel.: +1 617 643 6730; fax: 617-724-9811.
E-mail address: dchang8@mgh.harvard.edu (D.C. Chang)

0022-4804/$ – see front matter

https://doi.org/10.1016/j.jss.2020.03.074