Is There a Role for Spin Doctors in Med Ed Research?

Gail M. Sullivan, MD, MPH

As an optimistic person, I understand the desire on the part of others to view the world through rose-colored glasses, see the cup as half filled, and generally look on the bright side of things. The world of medical education is replete with those who honestly view events more positively than skeptically. One might even say that optimism is an essential trait for those who toil in the medical education fields today. In the public arena, spin is associated with reframing the current story into one that is favorable to the speaker with the goal of manipulating public opinion. In fact entire careers are made through functioning as a "spin doctor." However, medical education research suffers greatly from the influence of this spin in selling the research results in order to publish. Spin blurs the distinction between what may work and what likely will not.

In this article I will define spin in education research, discuss why the practice is harmful, and propose safeguards to avoid this type of manipulation in crafting and writing education research. I can be considered an expert in this area because I violated most of these precepts early in my career. My comments pertain primarily to quantitative research where these concerns are most prevalent.

The Fishing Expedition

My all-time favorite in the "spin" technique toolbox is the presentation of the sole statistically significant result in the conclusions—and title—of a research article. This is common in clinical research, even in illustrious journals, and also seen in medical education publications. Often these articles have not preselected a finite, planned number of associations for testing. Instead, the authors have looked at every possible association between known factors and the outcome of interest. The analysis does not correct for multiple associations and—voilà!—1 association is found, $P = .04$, between a medical student's score on 1 subscale out of many and the choice of pediatrics as a specialty. Only this single positive finding is presented in the title and conclusions. Although the article's limitations discuss the fact that the association may be spurious, this is buried in the Discussion section, which may not be read by all. One of my mentors, the late Dr Alvan Feinstein, termed this activity "going on a fishing expedition." That is acceptable for pilot work, but not acceptable for a study worthy of publication. If the work has value, for example, as it reports on topics that are understudied, a finite number of associations should be planned, all results should be reported, and the study should be presented as a negative, preliminary study.

I have to pause here to reflect on the term *negative study*. This is common parlance but not an accurate use of the word *negative*. These studies fail to show a difference or association that is unlikely to be due to chance (to a prespecified likelihood): The evidence supports the null hypothesis. Thus, these findings are not "negative" except in the sense that the authors may be disappointed by them. We probably need a better term.

P Values Going Solo

My second favorite spin technique is the inclusion of low *P* values in the abstract results without the accompanying effect size, or impact. If a difference does exist between the groups studied, no matter how small, this difference can be found with a sufficiently large sample size. All other things being equal, the larger the sample size, the lower the *P* value. However, the *P* value does not reflect the size of the difference between groups. For example, in a large study of 10 000 subjects the new (and astronomically expensive) cancer drug shows a difference versus the older, generic drug with a P value $< .001$, thus not likely due to chance. However, the average difference in life expectancy was 5 days.

Authors can avoid this issue fairly easily: always include impact (absolute difference, effect size, or odds ratio with confidence interval) in the abstract results and always consider impact in analyzing the meaning of your findings. For example, an intervention added a workshop, 5 online web modules, and team rounds to your baseline rotation and compared outcomes before and after the enhancement, with a large number of residents in both groups. The average performances on the outcome measure were significantly different with a very low *P* value. However, the absolute improvement was an average of 3 out of 100 points on the outcome measure. There may be very good reasons for this finding, such as the outcome measure not actually assessing what you are trying to teach. However, the results are still the results and should be reported as showing little impact.

Gail M. Sullivan, MD, MPH, is Editor-in-Chief, *Journal of Graduate Medical Education*, and Professor of Medicine, University of Connecticut.

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut, 253 Farmington Avenue, Farmington, CT 06030-5215, gsullivan@uchc.edu

DOI: http://dx.doi.org/10.4300/JGME-D-14-00338.1

BOX **Definition of Spin in Research**

Implying or stating more positive conclusions than the methods and results demonstrate

Amazing Precision for Likert Scale Means

This spin technique is actually a subset of the one described above. The authors find differences in outcomes assessed by Likert-type scales, and report the means to the 100th decimal point. For example, the means are reported as 3.73 versus 3.51, with a scale range of 1 to 5. Reporting 3 numbers may suggest that the overall difference is more important than it actually is. Reporting Likert scales to the 100th decimal point implies a precision that does not exist: that survey responders understand the difference not just between 3.5 and 3.7—which many of us would find problematic—but also between 0.03 and 0.01, which is likely impossible. To avoid this trap, do not report Likert-type scale means to the 100th decimal point and always include the impact or effect size with the *P* value.

Dramatic Graphs

Beloved by pharmaceutical reps, this technique uses a graphic that shows a large difference between the favored drug versus the placebo or another drug. This can be accomplished by using a very fine scale for the y-axis—which is also labeled in small print. This is another way to obscure a trivial impact. The differences appear dramatic because the scale used is fine. Often there are great colors and pictures, too, which may distract you. Another variation of this occurs when the y-axis does not start at zero and the units include a reduced range, such as from 3 to 5 for a scale that ranges from 0 to 7.

Power Calculation: Yes—Trends: Usually Not

It is sometimes difficult to comprehend how results that show no difference between educational strategies are transformed into a success by the time the reader reaches the authors' conclusions. In the Limitations section these articles may include a discussion that the article was likely underpowered to find a difference (that the authors believe must exist) between the different groups. That is, β error is the reason that a difference was not observed. The authors omitted doing a power calculation, using a best estimate of a meaningful effect size, before the study. One should always do a power calculation in the planning stage, as post hoc power/sample size calculations violate the necessary assumptions. Authors sometimes justify the absence of a power calculation by saying that they had a fixed number of residents with no ability to increase the sample size. However, one can delay publication to run the intervention with a second cohort, such as in the following year, which may double the sample size. Having a fixed number of subjects does not excuse the lack of a power calculation. In articles of this type, the authors will argue in the Discussion section that their small sample size likely resulted in no differences seen. By the Conclusion section the lack of findings are described as "promising."

Another example is the reporting of "trends" in the results as notable findings. As the *P* value did not reach the predetermined level, the differences observed between groups may be due to chance: Evidence supports the null hypothesis. Yet, one group did better than the other and this is reported as a key finding: a trend. Unless the research is in a novel area, in which we have no good idea of the likely effect size, trends should not be reported as findings of note. In contrast, in cutting-edge research, a power calculation must use a "best guess" effect size. Here one may have assumed a medium effect size and calculated a sample size on the basis of this estimate. If a difference does exist but the effect size is actually small, one would have needed a larger sample size to reach the predetermined *P* value. However, usually effect sizes of interest to medical educators are medium or large—versus the small effect sizes commonly found in clinical research studies. Thus, in education research, "trends" usually should not be reported unless the study is investigating a completely new area.

Remember that potential reasons for the study findings—including lack of findings—should be thoroughly explored in the Discussion section. In some instances the lack of difference is false: A difference does really exist and a new, carefully planned study will find this difference next time. For this article and at this point in time if your work shows no difference, most often you should conclude with this finding rather than report a promising trend.

The Research Findings—Conclusions Disconnect

Our enthusiasm and passion regarding medical education research sometimes lead us to leap to conclusions that are unrelated to the questions we have studied. This disconnect is seen when authors attempt to jump far past their actual findings. For example, a study that finds declining resident test scores over time, for 1 topic at 1 program, concludes that a required rotation in this area should be added to the residency requirements in this discipline. Here the spin aspect is that since the conclusions—changing national requirements—are of enormous importance to the field, the research is similarly essential. In reality the study showed a decline in scores, but determined neither the causes nor the solutions. This overreaching can badly mislead readers.

TABLE **Spin Techniques and Solutions**

Techniques	Solutions
The fishing expedition	Preplan all comparisons Choose a limited number of comparisons Adjust *P* value for multiple associations Report all findings, not just those that are statistically significant: avoid cherry-picking
Reporting significant *P* values without effect size	Report impact, ie, absolute difference, effect size, or odds ratio with confidence interval, in conjunction with *P* values—no solo *P* values
Likert-type scale means reported to 100th decimal point	Report Likert-type scale means only to the 10th decimal point
Misleading graphic display of results	Use a fair scale to represent data
Attributing lack of differences between groups to a too small sample size	Perform power/sample size calculation before study Do not report trends; exception (rare) for "cutting-edge" research
Overreaching in conclusions	Relate conclusions strictly to results
Overconfidence in the intervention or association	Be open to the possibility that the evidence will support the null hypothesis Analyze possible causes for lack of findings, including that the intervention may be ineffective (or the association does not exist) Explore reasons why the intervention may be ineffective, or no more effective than the comparison intervention

The Intervention Always Works

Finally, for some authors, the intervention always works or the association always exists. In the Limitations section every possible reason for a negative study is explored thoroughly and quite well, with 1 glaring omission: the possibility that the intervention is ineffective. In many articles that I read, with absence of differences or associations, this possibility is not even mentioned. Sometimes authors will mention this possibility at the end of the Limitations section. "And finally we have to consider that perhaps the intervention is not as effective as we think it is." They know that this cannot be true, but include this statement for the sake of completeness. Although I am quite willing to believe that every intervention works—if you teach them, they will learn[1]—the new intervention may not be any better than the comparison intervention. It is important to explore why the intervention might not be effective and how this affected the results. This type of discussion may nudge the field forward.

Final Thoughts

Medical education research emphatically does not need spin doctors: We are drowning in what we do not know about how learners learn and do not need further obfuscation. Studies that fail to find differences or associations and that thoroughly explore possible cause are important to publish. Don't cherry-pick which findings you include in your article conclusions and title. Match your findings—all of your findings—to your conclusions to avoid playing spin doctor. I believe it is primarily our underlying enthusiasm and passion about our subject that leads us to be overly positive and to spin results, and not that we wish to deceive or confuse. To benefit learners and their patients, we must temper our enthusiasm with thoughtful, reasonable interpretations, and above all, common sense.

References

1 Cook DA. If you teach them, they will learn: why medical education needs comparative effectiveness research. *Adv Health Sci Educ Theory Pract*. 2012;17(3):305–310.