Balancing statistical significance and clinical relevance

Charity J. Morgan, PhD^a

^a Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

Received Mar 22, 2018; accepted Mar 22, 2018

doi:10.1007/s12350-018-1267-y

Charity Morgan,
Statistical Consultant,
Journal of Nuclear Cardiology.

Recently, I was reviewing the results of a statistical analysis with a clinical collaborator. We had performed a number of statistical tests and the P value from one of these tests was $P = .07$.

As is common, we had pre-specified that we would use a P -value less than .05 as our cut-off value for determining statistical significance. My collaborator asked me how we should report the results of this test, wondering “Would it be fair to say that there is a trend towards a significant finding?”

This situation is one that I have found myself in many times over. Reporting a statistically significant result is relatively easy. (And it’s always great to be able to point to a significant finding!) But what, if anything, should we say about our non-significant results?

This is far from a new question. And for many statisticians, it’s not even a difficult one. According to the fundamentals of statistical theory, a P -value either is or isn’t statistically significant. So language such as “trending towards significant,” “almost statistically significant,” or “verging on significance” is misleading and inaccurate. However, it’s not hard to find examples

of authors discussing and interpreting non-statistically significant results across several disciplines.¹ If this question (and its straightforward answer) is such well-worn territory, why does this issue persist?

For starters, one could make a case that negative findings may still be informative. Reporting only our positive, or statistically significant, findings is a form of publication bias.² Some journals have taken steps to address this situation by indicating their willingness to publish null findings.³ There have even been journals established specifically for publishing non-statistically significant results (for example, the *Journal of Pharmaceutical Negative Results* or the now-defunct *Journal of Negative Results in Biomedicine*). So including the non-significant findings in our reports of our results is entirely justified.

But having reported a non-statistically significant result, it seems only natural that we should say *something* about it—especially when the P -value is in the .05 to .1 range—even if we know better than to describe that result as “approaching significance.”

Some authors may try to provide explanations of why a hypothesized result was not statistically significant. However, Hewitt et al. note that this practice can result in another type of bias, known as *interpretative bias*, which consists of downplaying undesirable results or equivocating when a desired result is not found.⁴ Often this equivocation takes the form of pointing to the sample size as the reason why a result was not statistically significant. Some will go even further and try to bolster this claim by calculating what is known as the “observed” (or “post hoc”) power and present the results of those calculations as evidence that the study was underpowered. However, Hoenig and Heisey have demonstrated that observed power calculations are based on circular logic and are therefore statistically meaningless.⁵ Yes, underpowered studies can miss true findings. But the converse is not necessarily true: a failure to demonstrate statistical significance is not always due to sample size.⁶

Reprint requests: Charity J. Morgan, PhD, Department of Biostatistics, University of Alabama at Birmingham, 1720 Second Avenue South, Birmingham, AL 35294-0022; cjmorgan@uab.edu

J Nucl Cardiol 2018;25:707–8.

1071-3581/\$34.00

Copyright © 2018 American Society of Nuclear Cardiology.

So if we shouldn't say a non-significant result shows a trend towards significance and we can't argue non-significant results are due to insufficient sample size, what *can* we say in the face of non-statistically significant findings? Typically, I lean towards "Report but don't interpret." For example, wording such as "We observed a lower complication rate for Treatment A (4%) than for Treatment B (5%); however, this difference was not statistically significant ($P = .07$)" could be used. And when discussing a non-statistically significant result, I may point to sample size as a *potential* contributing factor, but make a point of considering other possibilities. After all, sometimes we don't find a statistically significant result because there isn't one to be found.

And of course, statistical significance is not everything. With a large enough sample, even a tiny effect size can be statistically significant. What good is a statistically significant result without clinical relevance? While determining if a result is meaningful is usually best left to those with the appropriate clinical background, statistics can still be useful here as well.

Hypothesis testing and P -values can tell us if a result is statistically significant. But calculating estimates and confidence intervals for effect sizes such as hazard ratio, odds ratio, relative risks, or mean differences will help clinicians determine the relevance of those results. Compare, for example, the following reported results: "Treatment A had a statistically lower risk of complications than Treatment B ($P < .001$)." versus "The rate of complications for Treatment A (4.2%) was significantly lower than for Treatment B (4.21%) ($P < .001$)." The former may make our results sound more exciting than they actually are, but the latter gives the full picture: statistical significance was demonstrated but clinical relevance may be minimal.

When presenting results, we must keep this balance between significance and relevance in mind. When either statistical significance or clinical relevance is absent, we must tread carefully. I don't believe that results with so-called "borderline" P -values are necessarily meaningless. For one, such results can serve as a starting point for future research. But we must resist the temptation to over-interpret non-statistically significant findings. And if the clinical relevance of our results is questionable, we must ask ourselves whether these results are worth reporting at all.

So what did I say to my collaborator when she asked me how we should discuss our non-significant result? I advised her not to call the result a "trend" and suggested she report the estimated effect size and its P -value and refrain from interpreting the result any further.

Disclosures

The author has no conflict of interest to disclose.

References

1. Gibbs NM, Gibbs SV. Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research. *Br J Anaesth* 2015;115:337-9.
2. Granqvist E. Why science needs to publish negative results. Amsterdam: Elsevier; 2015.
3. Spiegel B, Lacy BE. Negative is positive. *Am J Gastroenterol* 2016. <https://doi.org/10.1038/ajg.2016.476>.
4. Hewitt CE, Mitchell N, Torgerson DJ. Listen to the data when results are not significant. *BMJ Br Med J* 2008;336:23-5.
5. Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am Stat* 2001;55:19-24.
6. Wood J, et al. Trap of trends to statistical significance: Likelihood of near significant P value becoming more significant with extra data. *BMJ* 2014;348:g2215.