© 2018. Published by The Company of Biologists Ltd | Journal of Cell Science (2018) 131, jcs215707. doi:10.1242/jcs.215707

STICKY WICKET

Mea culpa, mea powerful culpa

Mole

Okay, it's pouring rain here and everything, including me, is soggy. And when I'm soggy, I get moody, and when I'm moody there's not much else for it but to overindulge in 'tea,' which isn't actually tea. But that isn't why I'm writing today (while trying to dry out a bit, in every way): recently I posted a column (doi: 10.1242/jcs.206375) in which I made a statement that was simply dead wrong. I admit it, indeed, I've learned quite a bit from the correction.

Thanks to alert reader, R. J. Calin-Jageman, my error was laid bare. The following is an excerpt from his letter (*italics* are my comments).

> In an otherwise pithy overview of the absurdities of grant review (*thank you!*), Mole delivers a dangerously misleading broadside on reforms aimed at enjoining sample-size planning: Somehow, journals have taken to asking how our animal studies were powered – and here's the point: If the results are statistically significant then, indeed, our study is appropriately powered. (*Yes, I said that, and yes, it was absolutely wrong – read on!*)
>
> This is a common but pernicious misconception. An underpowered study is one that is uninformative relative to the research question being asked. The *p*-value obtained does not alter the fact that the results are uninformative and not to be relied upon. On the contrary, underpowered studies only achieve statistical significance by leveraging sampling error; publishing these types of studies leads to a distorted and polluted research literature. The easiest way to appreciate the pitfalls of low power is to examine the confidence intervals obtained. For example, imagine conducting a correlational study but collecting only 12 data points. This sample size is adequate only for very strong relationships, providing >80% power only if $r>0.71$ in the population (assuming a two-tailed test with $\alpha=0.05$). Suppose you obtain $r=0.00$, 95% CI [−0.57, 0.57], $p<1$. Clearly, this is not statistically significant. The long confidence interval shows that the result is consistent with a very wide range of possibilities, meaning that the sample is not very informative and we cannot conclude much about the real correlation. Now suppose you obtain $r=0.58$, 95% CI [0.01, 0.87], $p<0.05$. This finding is statistically significant, but the CI shows that it is not much more informative. A wide range of correlation values are plausible, including many that are vanishingly small and scientifically intractable for further study. The only safe conclusion from such an underpowered study is that a negative correlation is fairly unlikely. The lesson here is unsettling but clear: underpowered samples are uninformative even when statistical significance has been achieved. More comforting, though, is that adequate sample sizes are informative regardless of statistical significance.

Well said, R.J.C.-J. My lack of education in statistics is woeful, but I think the message is clear: if a study is not properly powered, statistical significance is simply not meaningful. But what should we do about it? In the area of discovery research (which some people like to call 'basic' research), we must perform experiments to know if our ideas have any merit; often this requires fairly large numbers of animals. Technically, we should examine the preliminary results, however 'significant', and then conduct another, properly powered, experiment to see if our conclusions are indeed correct. But often, if we are using complex *in vivo* analyses involving multiple controls (as is often the case with genetically engineered animals and techniques such as RNA-seq) the costs of such experiments are prohibitive. I know, that isn't a valid excuse. Alternatively, we can determine, *post hoc*, if our sample sizes sufficiently powered our experiment to support the conclusions we wish to make. Statisticians note (I think) that such determination *after* the fact is not valid, although the reasons for this escape my very limited understanding of statistics (as I said, I am woefully uninformed).

But I do have a suggestion (of course I do). If we strive for transparency in our publications, determining after the fact if our experiments are properly powered, and noting that this is what we did, readers will have at least a bit more information on which to base their own conclusions regarding the robustness of our reported results. This can only help our endeavor: As I've said before, the value of a reported finding is not whether it is *true*, but whether it is *useful* in gaining an understanding of a process that leads to more reproducible results. That is what we all do. We can do it better.

Usually, when I read a paper that represents discovery research, it is the weight of evidence, and how it holds together to create a picture of how things might actually work, that determines whether or not I will incorporate it into my own views. Even if every experiment is not appropriately powered (probably we should power our immunoblots and the differences they show, but I doubt this will happen beyond a few repeats for most experiments), how the results fit the other conclusions of the paper help us to reach our own conclusions. We should *never* believe everything that we read, and this insight into appropriate powering of key experiments is another valuable tool in our arsenal.

Look at that, I'm dried out. Must have been my heated embarrassment for having been so wrong.