Editorial

On What *P* Means to Me: An Encouragement to Investigators

Jeffrey Pomerance, MD, MPH

Statistical significance seems to be the hallmark of success in today's scientific investigation. However, statistical significance and clinical meaning are two different issues. For example: Groups A and B each are composed entirely of infants weighing 2485 gm and 2500 gm, respectively. We then sample 100 babies from each of these groups. We find a statistically significant difference in the weights of the infants in each group. The average and only weight in group A is 2485 gm, and the average and only weight in group B is 2500 gm. The *p* value is $<$0.001 by the Student's *t*-test, a highly significant finding. More recently, statistical comparisons are being made with "odds ratios" to estimate "relative risks." Also provided is a range of relative risks representing a confidence interval of 95%. If this confidence interval does not cross 1.0, then the result has a *p* value of $<$0.05. Usually, an exact *p* value is provided. Same implications, new format. No matter which format is used, the point is, does anyone care? In the example above, likely only those placing infants into categories of $<$2500 gm and $\geq$2500 gm would care. So point one is that statistical significance and clinical significance are two different matters. If the study groups are large enough and the variations within the groups small enough, any difference, no matter how small, can be shown to be statistically significant.

On the flip side of the coin, we frequently find a difference between groups that does interest us (i.e., is of clinical importance), but because of "insufficient numbers," our *p* value is not $<$0.05, the holy grail of statistical evaluation. What does a *p* value mean anyway? When the *p* value is $<$0.05, this means that the odds of finding the difference in question is likely to have occurred by chance alone $<$5% of the time. In other words, the difference in question likely has a cause other than random chance $>$95% of the time. However, no matter how small the *p* value, the difference uncovered is only an association; cause and effect is never demonstrated. This is why randomized, double-blinded studies are so important. It is hoped that the process of randomization and double blinding will make it unlikely that known and unknown factors are the cause of the differences demonstrated. This may or may not occur, but that is the hope. Also, it is important to remember that if we compare many outcome variables between two groups, we should expect, by chance alone, that 1 of 20 comparisons would have a *p* value of $<$0.05.

Greater Baltimore Medical Center, Baltimore, MD.

Address correspondence and reprint requests to Jeffrey Pomerance, MD, MPH, Greater Baltimore Medical Center, 6701 North Charles Street, Room 2358, Baltimore, MD 21204.

What if our *p* value turns out to be 0.049 vs 0.051? In the former case, our results will likely get the attention they justly deserve. In the latter case, our results will likely die an ignominious death. In most tables, a *p* value of 0.051 gets the big NS—not significant. Frequently, we are not even told how close to significance the value was. And yet these two *p* values are as alike as, well, two *p*s in a pod. To put it another way, these two *p* values are not likely to be statistically significantly different from each other. It probably isn't fair to turn statistical evaluation loose on itself!

And what if, God forbid, our *p* value should turn out to be 0.22? What does this mean? The first thing this means is that the study will probably not get published. The second, and more important meaning is that the difference encountered is likely to be the result of chance alone 22% of the time. In other words, the difference uncovered is likely to be due to a cause other than chance alone 78% of the time. The difference found is much more likely than not to have a cause other than just chance occurrence. Nonetheless, many of us would put this idea aside and take on a more promising project. This may or may not be a good decision.

Sometimes we choose inappropriately to entirely ignore *p* values. For example: one study involving 50 subjects each in the experimental and control groups has a *p* value of 0.042, but a much larger identically designed study that has 500 subjects each in the experimental and control groups also has a *p* value of 0.042. Which one is more meaningful? To have obtained identical *p* values, the smaller study must have found a much larger difference between the two groups than the larger study. However, the meaning of the two studies is no different. That is, the differences observed in each of the two studies are likely to have been caused by chance alone 4.2% of the time. It is true that the larger study is more likely to have uncovered rarer adverse effects of treatment.

How did we get to choose a *p* value of $<$0.05 as our measure of significance? The fact is, it is only a convention, not a rule of nature. There are probably more clinically important, but statistically "nonsignificant" results lying around in long-lost files than all of the published results combined. The point is for the investigator who has an idea that she or he thinks is of merit, and the differences encountered are of clinical interest, but the *p* value does not meet traditional significance, to persevere. Do not give up on your idea! Repeat the study with sufficient numbers (power) so that the question is likely to be answered to the satisfaction of the scientific community.

Enough said. I have to go *p*.