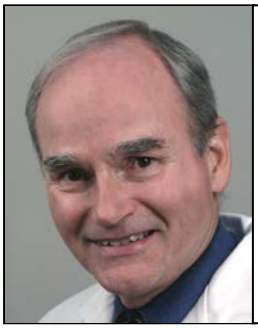
Clinical Chemistry

To P or Not to P: That Is the Question



Article:
J.C. Boyd and T.M. Annesley.
To P or Not to P: That Is the Question
Clin Chem 2014;60:909-910.
http://www.clinchem.org/content/60/7/909.extract

Guest:
Dr. James Boyd is Associate Professor of Pathology and Director of Clinical Pathology at the University of Virginia in Charlottesville.

Bob Barrett: This is a podcast from C*linical Chemistry,* sponsored by the Department of Laboratory Medicine at Boston Children's Hospital. I am Bob Barrett.

There are few scientific papers that do not include the use of the *P* value to evaluate the statistical significance of results. However, use of this statistic may be misleading, as noted by a recent paper by Regina Nuzzo in the journal *Nature*. That paper served as a basis for commentary with additional examples by Drs. Jim Boyd and Tom Annesley in the July 2014 issue of *Clinical Chemistry*.

James Boyd is Associate Professor of Pathology and Director of Clinical Pathology at the University of Virginia in Charlottesville, and he is our guest in today's podcast.

Dr. Boyd, your recent editorial considers the question of what constitutes a proper statistical analysis of data gathered in scientific experiments and refers readers to a recent article in nature by Regina Nuzzo. Dr. Nuzzo and many biostatisticians for that matter seem to be urging researchers to reconsider their use of statistical analysis, and to recognize the limitations posed by conventional statistics. This whole idea is confusing. Haven't we always used *P* values less than 0.05 to establish statistical significance? What are the problems with doing this?

James Boyd: Bob, let me begin by saying that *P* values don't tell the whole story. *P* values provide a measure of the statistical likelihood of a set of observations, but they do not provide a good indication of the overall importance of the observations.

Let me give you a little example. So in 1998, there was a scientific correspondence report that was published in the journal *Nature,* and the authors of this report demonstrated that there were small differences in the adult height of Austrian military recruits, with those recruits that had birthdays in the spring being about a quarter inch taller on average than those with birthdays in the fall. And this actual observed 0.6 cm difference in height between the spring and fall birthdays is what we call the effect size.

So this 0.6 cm or quarter inch height difference was not only hard to explain but it was unlikely to be biologically important. Nevertheless, because these data were gathered from over 500,000 recruits, the small observed difference in height was found to be very highly statistically significant. How important is such a finding? Well, I think there are very few people who would consider the quarter inch difference in height to be important. I think this serves as a good example of the dictum that just because a given finding is statistically significant, does not establish that it's biologically important.

Additional problem with *P* values is the one of falsely significant results. When we use the threshold of 0.05 for statistical significance, this allows for a 5% or 1 in 20 chance of coming to a false conclusion that there is a significant effect on the data, when in fact there is none. And so this problem of reaching false conclusion compounds when multiple statistical tests were conducted at one time, such as when looking for an optimal set of biomarkers, for distinguishing a particular disease from others.

Hypothetically, if we were to make measurements of 50 biomarkers, say in two groups, healthy individuals and people with disease X, and we knew in advance that none of these markers was informative, we would still expect to find on average two or three markers that demonstrated statistically significance between group differences in marker values when we use the *P* value of 0.05.

Now there are statistic approaches to the correct this problem and hold the overall rate of reaching a false conclusion to 5%. And researchers who have done experiments and evaluated multiple biomarkers and have conducted multiple statistical tests without making corrections for the increased rate of falsely significant findings, need to be aware of them and they need to acknowledge this concern and to point out the need for follow-up confirmatory experiments.

Bob Barrett: So if *P* values are not optimal what other approaches might be more suitable?

James Boyd: Well, one approach that would allow researchers to make a more direct assessment of the importance of the data, is to report the actual observed between group effect size with a confidence interval. This approach allows a rapid assessment of how large the true effect might be, and if the range of effect size within the confidence interval approaches or encompasses zero, that would suggest, would the data hold less overall importance.

And there is another approach that's gaining popularity that's called the use of the Bayesian methods or Bayesian framework, and for those of us in laboratory medicine, this is very similar in concept to predictive value theory. Predictive value reports the probability of having the disease after having performed a diagnostic test, and we can compare that probability with the probability of having the same disease in a given patient before we ever did the test.

And in statistical hypothesis testing; the analogy would go that if a researcher could estimate the odds that a given hypothesis was true before performing the experiment then those odds could be factored into an overall calculation of odds that the hypothesis is true after the experiment has been conducted.

Usually the odds at a given hypothesis is true are much smaller than the 50-50 chance or one on one odds commonly assumed with traditional hypothesis testing, and as Dr. Nuzzo pointed out in her article in *Nature*, if we assume in advance, the experiment of hypothesis has equal odds--50-50 chance of being true--and after performing the experiment, we find a *P* value of less than 0.05, there is still by a commonly used formula a 29% residual chance of there being no real effect.

This residual chance that there's no effect increases progressively as you lower the before-hand odds, the experimental hypothesis might be true. So, this is a very non-intuitive result and the fact that there is a high residual chance of no real effect even when the experiment yielded statistically significant findings could explain why there have been so many instances reported recently of initial significant experimental findings that could not be verified in subsequent studies.

Bob Barrett: Dr. Boyd, what is the rationale for doing follow up confirmatory experiments?

James Boyd: The most important reason is that a single experiment and statistical analysis isn't necessarily going to account for all the factors that might play a role in producing a given experimental outcome. There could be some nuances of the population in which the experiment was conducted or of the assays that were used that in themselves led to a statistically significant finding. So a repetition of the same experiment in a different population or using a different assay method could lead to entirely different results.

The other factor that we need to think about in the need for confirmatory experiments is the phenomenon I've already pointed out that even when there are initially statistically

significant findings, after an experiment has been conducted, there is still a large residual chance that no real effect existed in the first place. And this large residual chance of no effect as predicted by the Bayesian methods is a growing area of concern in the area therapeutic drug studies for instance, where there been several reports recently confirming its truth, where are a large fraction of initial external findings could not be reproduced in independent studies.

Bob Barrett: Why don't journals provide guidelines on the proper use of statistics?

James Boyd: Well, this is harder to do than you might think. The range of statistical methods that could be used for a given analysis is so broad that it's really hard to be prescriptive about what method to use. In many data sets, any two biostatisticians could analyze the data by entirely different statistical methods and yet still come up with equally valid statistical analyses. So, at *Clinical Chemistry,* we've chosen editorially to encourage readers to become more knowledgeable in statistics and to seek help when needed from professional biostatisticians. We've also decided to periodically present examples of data analysis problems for which the statistical method that was used, exerted an important effect on the conclusions reached. In the current issue of the journal for instance, Tom Annesley and I present just such a set of data where often employed traditional methods such as Student t-test or Mann-Whitney U test can fail to disclose really important information of the data that could be gotten by simply looking at graphical presentations.

Bob Barrett: Well then, before we go Dr. Boyd, any final thoughts on statistics?

James Boyd: Statistics has always been a controversial topic and one with which each new generation of scientists has had to grapple. As Mark Twain once remarked, there are three kinds of lies: lies, damned lies, and statistics. And so I can only encourage researchers to seize the day and make further efforts to understand various statistical approaches in their optimal applications in the analysis of data.

Bob Barrett: Dr. James Boyd is Associate Professor of Pathology and Director of Clinical Pathology at the University of Virginia in Charlottesville, and he is been our guest in the podcast from *Clinical Chemistry* from on “To *P* or not to *P*, that is the question.”

I'm Bob Barrett, thanks for listening.