Some Common Misperceptions About $P$ Values

Yuko Y. Palesch, PhD

A $P$ value $<0.05$ is perceived by many as the Holy Grail of clinical trials (as with most research in the natural and social sciences). It is greatly sought after because of its (undeserved) power to persuade the clinical community to accept or not accept a new treatment into practice. Yet few, if any, of us know why 0.05 is so sacred. Literature abounds with answers to the question, What is a $P$ value?, and how the value 0.05 was adopted, more or less arbitrarily or subjectively, by R.A. Fisher, in the 1920s. He selected 0.05 partly because of the convenient fact that in a normal distribution, the 5% cutoff falls around the second standard deviation away from the mean.$^1$

However, little is written on how 0.05 became the standard by which many clinical trial results have been judged. A commentary$^2$ ponders whether this phenomenon is similar to the results from the monkeys in the stairs experiment, whereby a group of monkeys were placed in a cage with a set of stairs with some fruit at the top. When a monkey went on the steps, blasts of air descended on it as a deterrent. After a while, any monkey that attempted to get on the steps was dissuaded by the group. Eventually, the monkeys were gradually replaced by new monkeys, but the practice of dissuasion continued, even when the deterrent was no longer rendered. In other words, the new monkeys were unaware of the reason why they were not supposed to go up the steps, yet the practice continued.

In the following, I first review what a $P$ value is. Then, I address 2 of the many issues regarding $P$ values in clinical trials. The first challenges the conventional need to show $P<0.05$ to conclude statistical significance of a treatment effect; and the second addresses the misuse of $P$ values in the context of testing group differences in baseline characteristics in randomized trials. Many excellent articles and books have been published that address these topics; nevertheless, the intention of this article was to revive and renew them (using a less statistical language) to aid the clinical investigators in planning and reporting study results.

What Is a $P$ Value Anyway?

We equate $P<0.05$ with statistical significance. Statistical significance is about hypothesis testing, specifically of the null hypothesis ($H_0$) that means the treatment has no effect. For example, if the outcome measure is continuous, the $H_0$ may be that the group difference in mean response ($\Delta$) is equal to zero. Statistical significance is the rejection of the $H_0$ based on the level of evidence in the study data. Note that failure to reject the $H_0$ does not imply that $\Delta=0$ is necessarily true; just that the data from the study provide insufficient evidence to show that $\Delta \neq 0$.

To declare statistical significance, we need a criterion. The $\alpha$ (also known as the type I error probability or the significance level) is that criterion. The $\alpha$ does not change with the data. In contrast, the $P$ value depends on the data. A $P$ value is defined as the probability of observing treatment effect (eg, group difference in mean response) as extreme or more extreme (away from the $H_0$) if the $H_0$ is true. Hence, the smaller the $P$ value, the more extreme or rare the observed data are, given the $H_0$ to be true. The $P$ value obtained from the data is judged against the $\alpha$. If $\alpha=0.05$ and $P=0.03$, then statistical significance is achieved. If $\alpha=0.01$ and $P=0.03$, statistical significance is not achieved. Intuitively, if the $P$ value is less than the prespecified $\alpha$, then the data suggest that the study result is so rare that it does not seem to be consistent with $H_0$, leading to rejection of the $H_0$. For example, if the $P$ value is 0.001, it indicates that if the null hypothesis were indeed true, then there would be only a 1 in 1000 chance of observing data this extreme. So either unusual data have been observed or else the supposition regarding the veracity of the $H_0$ is incorrect. Therefore, small $P$ values ($<\alpha$) lead to rejection of the $H_0$.

In the Interventional Management of Stroke (IMS) III Trial that compared the efficacy of intravenous tissue-type plasminogen activator (n=222) and intravenous tissue-type plasminogen activator plus endovascular (n=434) treatment of acute ischemic stroke, the $\alpha$ was specified as 0.05. The unadjusted absolute group difference in the proportion of the good outcome, defined as the modified Rankin Scale score of 0 to 2, was 2.1% (40.8% in endovascular and 38.7% in intravenous tissue-type plasminogen activator).$^3$ Under the normal theory test for binomial proportions, this yields a $P$ value of 0.30, meaning that if the $H_0$ were true (ie, the treatment did not work), there would be a 30% chance of observing a difference between the treatment groups at least as large as 2.1%. Because this is not so unusual, we fail to reject $H_0$: $\Delta=0$ and conclude that the difference of 2.1% is not statistically significant.
Thinking Outside the $P<0.05$ Box

Another interpretation of the $\alpha$ is that it is the probability of rejecting the $H_0$ when in fact it is true. In other words, $\alpha$ is the false-positive probability. Typically, we choose $\alpha$ of 0.05, and hence, our desire to obtain $P<0.05$. There is nothing magical about 0.05. Why not consider the risk (or cost) to benefit ratio in the choice of the false-positive probability, the research community is willing to tolerate for a particular study? For some studies, should one consider a more conservative (like 0.01) or more liberal (like 0.10) $\alpha$? In the case of comparative effectiveness trial, where $\geq2$ treatments, similar in cost and safety profile, that have been adopted in clinical practice are tested to identify the best treatment, one might be willing to risk a higher likelihood of a false-positive finding with $\alpha$ of, say 0.10. In contrast, if a new intervention to be tested is associated with high safety risks or is expensive, one would want to be sure that the treatment is effective by minimizing the false-positive probability to, say 0.01. For a certain phase II clinical trial, where the safety and efficacy of a new treatment is still being explored, one can argue for a more liberal $\alpha$ to give the treatment a higher level of the benefit of doubt, especially when the disease or condition has only a few, if any, effective treatment options. If it should pass, it would be weeded out in a phase III trial with a more stringent significance level. Also, if the $H_0$ is widely accepted as true (perhaps, eg, in the case of hyperbaric oxygen treatment for stroke), then one might wish to be more sure that rejecting the $H_0$ implies that the treatment is effective by using $\alpha$ of 0.01 or even lower. Of course, this means a study with larger sample size has to be conducted.

Although proposing to use anything greater than an $\alpha$ of 0.05 may be challenging, especially for studies to be submitted to the US Food and Drug Administration for New Drug Application approval, scientifically sound rationale and experienced clinical judgment should encourage one to think outside the box about the choice of the $\alpha$. In doing so, one should ensure that scientific and ethical rationale is the driving argument for proposing a larger $\alpha$ and not only the financial savings (as a result of smaller required sample size with a larger $\alpha$).

$P$ Values in the Group Comparison of Baseline Characteristics in Clinical Trials

Typically, primary publications of many clinical trials in the typical Table 1 a long list of baseline characteristics of the study sample and their summary statistics (eg, mean and standard deviation; median and interquartile range; or proportions). In addition, many include $P$ values associated with statistical tests comparing the groups or denote with a variety of asterisks the variables where the comparison yields $P<0.05$, $P<0.01$, and $P<0.001$. Some authors assume that the journal editors require them. In the instructions to authors of prospective New England Journal of Medicine (NEJM) manuscripts, it states under the statistical methods:

For tables comparing treatment groups in a randomized trial (usually the first table in the trial report), significant differences between or among groups (i.e. $P < 0.05$) should be identified in a table footnote and the $P$ value should be provided in the format specified in the immediately preceding paragraph. The body of the table should not include a column of $P$ values. (http://www.nejm.org/page/author-center/manuscript-submission; obtained on August 18, 2014)

Meanwhile, according to the current Consolidated Standards of Reporting Trials (CONSORT) 2010 guidelines on the publications of clinical trials:

Unfortunately significance tests of baseline differences are still common; they were reported in half of 50 RCTs trials published in leading general journals in 1997. Such significance tests assess the probability that observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance. Tests of baseline differences are not necessarily wrong, just illogical. Such hypothesis testing is superfluous and can mislead investigators and their readers. Rather, comparisons at baseline should be based on consideration of the prognostic strength of the variables measured and the size of any chance imbalances that have occurred. (http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data; obtained on August 18, 2014)

The 2 are somewhat contradictory: one (NEJM) requiring statistical tests be performed on the baseline characteristics and the other (CONSORT) discouraging such tests.

Recall that $P$ values are associated with hypothesis testing. The hypotheses that are tested for these baseline characteristics evaluate whether the differences between the groups are statistically significant, but that does not necessarily equate to clinical significance or relevance of the difference. Note that the $P$ value is partially influenced by sample size. Generally, the larger the sample size, the easier it is to obtain a smaller $P$ value from the data for the same difference. For any study with large enough sample size, statistical significance can be achieved; however, the observed mean group difference is not necessarily clinically relevant. Conversely, one may note a clinically relevant difference in a baseline characteristic, but the $P$ value from the test may not reach statistical significance with a small sample size. Therefore, an important clinical difference in a baseline characteristic may be overlooked.

Suppose in a large (n=2100 per group) clinical trial of acute stroke to detect a difference of 5% in good outcome between 2 treatment groups, the typical Table 1 shows the mean baseline systolic blood pressure of 125 and 120 mm Hg, each with standard deviation of 15 mm Hg. The difference is 5 mm Hg, and the $t$ test yields $P<0.01$. But one could hardly argue that this difference is clinically significant. In contrast, suppose a small study (say, n=40 in each group) to test intensive serum glucose control in acute stroke patients had enrolled subjects with history of diabetes mellitus: 20% in one group and 33% in the other. The $\chi^2$ test yields $P=0.20$, not a statistically significant difference at the $\alpha$ of 0.05. Nevertheless, a 13% difference in the proportion of subjects with history of diabetes mellitus is likely
to be a clinically important factor to consider in the analysis and interpretation of the primary outcome. In other words, \( P \) values are meaningless at best, and potentially misleading, to ascertain whether the treatment groups are balanced in the baseline characteristics. The same issue of seeking statistical significance without consideration for clinical relevance also applies to analyses of outcomes data. Many articles have been published in both statistical and clinical journals addressing this topic and will not be addressed further here.\(^4,5\)

**Recommendation**

So what are we to do? Should we stop using \( P \) values altogether? No, but additional information, such as the prespecified minimum clinically important difference, the observed group differences, and their confidence intervals will enable other investigators to better assess the level of evidence for or against the treatment effect because they provide a range of plausible values for the unknown true difference between the groups.

For example, in the IMS III Trial,\(^3\) the study investigators prespecified minimum clinically important difference of 10%. The reported adjusted (for baseline National Institutes of Health Stroke Scale score per the study analysis plan) difference was 1.5%, with 95% confidence interval of (−6.1%, 9.1%). Because the 95% confidence interval includes 0, the result is not statistically significant at \( \alpha=0.05 \). In addition, if the confidence interval had included 10%, the study result can be interpreted as inconclusive because 10% may be a plausible value for the true but unknown group difference; otherwise, the study could be viewed as negative. Therefore, such information will allow the readers to apply their knowledge, experience, and judgment on the importance and relevance of the study results beyond whether it is statistically significant or not.

**Conclusions**

In conclusion, R.A. Fisher did not intend for the \( P \) value, much less \( P<0.05 \), to be the be-all and end-all of an experiment (or a clinical trial). He meant it as a guide to determine whether the study result is worthy of another look through replication. In spite of increasingly vocal criticism of our sole dependence on \( P \) values by the biostatistical and even some clinical communities, it will take some time to change the culture, but the change should be embraced.

**Acknowledgments**

I thank the 2 anonymous reviewers for their thorough and constructive comments to clarify and improve the discussions in this article.

**Sources of Funding**

This work was partially supported by National Institutes of Health (NIH) U01-NS087748 and U01-NS077304.

**Disclosures**

The author is a Data Monitoring Committee member for a study of Brainsgate Ltd and for a study by Biogen Idec Inc.

**References**

2. Kelly M. Emily Dickinson and monkeys on the stair or what is the significance of 5% significance level. *Significance.* 2013;10:21–22.

**Key Words:** statistical data interpretation
Some Common Misperceptions About P Values
Yuko Y. Palesch

Stroke. 2014;45:e244-e246; originally published online November 6, 2014;
doi: 10.1161/STROKEAHA.114.006138
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2014 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/45/12/e244