The American Journal of Pathology, Vol. 191, No. 5, May 2021

ELSEVIER

The American Journal of PATHOLOGY

ajp.amjpathol.org

See related article on page 784

COMMENTARY

Rigor, Reproducibility, and the *P* Value

Timothy J. O'Leary

From the Office of Research and Development, Veterans Health Administration, Washington, DC; and the Department of Pathology, University of Maryland School of Medicine, Baltimore, MD

Although concerns about rigor and reproducibility in preclinical research are not new, a highly cited 2005 article by Ioannidis[1], entitled "Why Most Published Research Findings Are False," attracted attention by the *New York Times* (*https://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html?*, last accessed March 16, 2021), was condemned by some (*https://replicationindex.com/2020/12/24/ioannidis-is-wrong-most-of-the-time*, last accessed March 16, 2021), and has been followed by both special collections of journal articles, such as *Nature*'s "Challenges in Irreproducible Research" collection (*https://www.nature.com/collections/prbfkwmwvz*, last accessed March 16, 2021), and by increased attention from funding agencies. Whether or not the title to his article is "true," Ioannidis[1] correctly points to several issues contributing to erroneous scientific conclusions, including bias, flawed experimental designs, and the chase for statistical significance (as measured by the *P* value). These points were emphasized by Hsieh et al[2] in a 2018 article in *The American Journal of Pathology* (*AJP*), where they noted that blinding and randomization are often "overlooked in preclinical science, even though they are considered routine in clinical trials." They pointed out that these two techniques help to guard against confirmation bias—the unconscious tendency to interpret experimental results in a manner that confirms one's preconceived notions of what they should be, and of imbalance in variables (other than the experimental intervention) that may affect outcome. This, in turn, reduces the probability that an investigator will reach false conclusions about the meaning of his/her experiments. In this issue of *AJP*, Mata and Milner[3] show that statistical testing that gives rise to *P* values has become a nearly ubiquitous element of experimental pathology research.

Null Hypothesis Statistical Testing: Chasing *P* Values?

Mata and Milner[3] report that, for authors publishing in *AJP*, most statistical testing falls into a group of approaches that give rise to *P* values, often referred to as "null hypothesis statistical testing." Null hypotheses can be formulated in many ways, so this group includes studies that are designed simply to find differences (if any) between two groups, studies intended to see if one group differs from another in a specific direction (eg, superior response to therapy), and designs that are intended to demonstrate noninferiority of one treatment approach to another that is already in widespread use. To a large extent, null hypothesis statistical testing has devolved into a "cookbook approach to statistical/scientific inference, rote and accessible to all,"[4] with the result that, as Mata and Milner's work suggests, experimental pathologists would benefit from a better understanding of both some fundamental statistical principles and of the strengths and limitations of some commonly used statistical tests. To put this effort into context (and make explicit several points that are implicitly included in the article by Mata and Milner[3]), it is worth considering what the *P* value means (and does not mean). Wasserstein and Lazar,[5] writing for the American Statistical Association, note:

A "*P*-value is the probability under a specified statistical model that a statistical summary of the data (such as the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

Wasserstein and Lazar[5] further point out that:

"*P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."

"A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result."

"By itself, *P*-value does not provide a good measure of evidence regarding a model or hypothesis."

Accepted for publication March 2, 2021.

Disclosures: T.J.O. has consulted for MDisrupt.

Address correspondence to Timothy J. O'Leary, M.D., Ph.D., Office of Research and Development, Veterans Health Administration, Washington, DC 21201. E-mail: timothy.oleary@va.gov.

https://doi.org/10.1016/j.ajpath.2021.03.001

To rephrase these statements, a "*P*-value is simply the probability that in a research study the data would be at least as extreme as those observed, if the null hypothesis were true."[6] The *P* value gives us a summary of the data associated with a specific experiment. "It cannot work backwards and make statements about the underlying reality. That requires another piece of information: the odds that a real effect was there in the first place."[7] One implication of these observations is that "*P*-values test not only the null hypothesis, but everything else in the experiment" (*http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/P-values9.pdf*, last accessed March 16, 2021).

Ioannidis[1] builds on these principles, reminding us that the actual probability that a tested hypothesis is true is dependent on the *a priori* probability that it was true, and that this *a priori* probability goes down as the flexibility of designs, analytical approaches, and outcomes within a field increases. For statistical tests, just as for diagnostic tests, the predictive value depends not only on sensitivity, but on prevalence—in the case of statistical tests, the prevalence of potential alternative hypotheses.

Despite these significant limitations, statistical techniques are valuable tools that, when properly employed and understood, can reduce the probability that scientists will draw false conclusions. Unfortunately, they may have precisely the opposite effect when used or interpreted improperly. Mata and Milner,[3] after careful analysis, concluded that statistical methods are not always employed correctly in the experimental pathology literature, in many cases because the experimental data may not conform to the statistical model used for testing. They found that few studies adjusted *P* values to account for testing multiple hypotheses, for example, and that there was significant overutilization of parametric tests when nonparametric statistical tests would have been more appropriate. When the wrong statistical model is used, *P* does not mean what we think it does. The *P* value may be artificially reduced, for example, when small samples that are not taken from normally distributed populations are compared using parametric methods Mata and Milner[3] corrected this problem by providing an extensive discussion of statistical testing strategies useful for many of the kinds of experimental studies that appear in *AJP*. They effectively illustrate the strengths and weaknesses of many statistical tools and point out the value of obtaining statistical consultation when planning studies. More importantly, they emphasized that, before performing any experiments, researchers should generate "an *a priori* statistical analysis plan, including anticipated data types, sample size calculations, and test selection."

Planning for Power and Replicability

Development of a statistical analysis plan begins with recognizing that scientific inference is much broader than statistical inference,[4,8] a point that sometimes seems forgotten. To be useful, the results of an experiment should make sense in the context of the broader knowledge base. Statistical analysis must also make sense of the data themselves. Mata and Milner's thoughtful overview of descriptive statistics, and the need to truly examine and think about experimental data, serves to point out that the purpose of both experiments and of statistics is insight, not *P* values. Nevertheless, as the authors note, care must be taken to ensure that such examinations do not introduce bias into later analyses, which are often best conducted using methods that do not assume anything about the underlying data (by using nonparametric or distribution-independent methods). Thus, insistence that a statistical analysis plan be generated before the onset of experiments is an important safeguard against bias and resulting error.

Mata and Milner's emphasis on the importance of determining sample size in advance of performing an experiment is similarly well considered. This process, typically referred to as power analysis, has been outlined previously,[9] and many computer programs are available to assist with power analyses appropriate for particular testing strategies. Power analysis requires choice of a statistical model, definition of a meaningful minimum effect size, and identification of researcher tolerance for type I (falsely rejecting the null hypothesis) or type II (falsely accepting the null hypothesis) errors. Examination of the scientific literature (including that in *AJP*) shows that researchers outside the realm of genetics/genomics almost invariably choose threshold *P* values, $\alpha \leq 0.05$, for rejecting the null hypothesis (with the unstated justification being that everybody does it, or referees will accept it). This particular threshold originally arose because it was usually easy to compute in the era before digital computers; it is totally arbitrary today. Authors of preclinical investigations seldom justify their choice of α, then ignore the risk of type II error and leave β, the threshold for falsely accepting the null hypothesis, unspecified. It seems likely that the power $(1-\beta)$ associated with a particular study may in many cases be limited less by scientific carelessness than by funding-driven feasibility—increasing power invariably requires increasing sample size, and thus cost. Nevertheless, failure to ensure adequate sample sizes and statistical power when trying to replicate previously reported results can be a major driver for replication failure. A study designed with statistical power of 0.3 cannot reasonably be expected to confirm an earlier finding, even if that earlier finding is true. In contrast to basic science studies, power is typically set in the range of 0.8 to 0.9 when designing clinical trials.

Power analysis must be performed before an experiment rather than afterwards, because the observed significance level of a *post hoc* analysis also determines the observed power, which is thus biased by the results themselves.[10] Sample sizes that result from power analysis should be regarded as estimates—large deviations from sample sizes resulting from power analysis should probably be considered meaningful, but small deviations probably should not; similarly, the estimate of statistical power is itself a highly educated guess.

There Is No Universal Threshold for Statistical Significance

The use of a *P*-value threshold for the assessment of significance, although required to conduct a power analysis, is not without its problems. There is probably no single appropriate threshold, and determination of a threshold (if one is to be used at all) is non-trivial.[11] Should *P* values of 0.051 and 0.049 really be treated as "categorically different?"[12] If many of the results reported using a *P*-value threshold of 0.05 are false, should we require that it be arbitrarily lowered, as suggested by some authors?[13,14] These are difficult questions that engender several differing opinions, many of which are found in a single supplemental issue of the *American Statistician* published in 2019 (*https://www.tandfonline.com/toc/utas20/73/sup1?nav=toclist*, last accessed March 16, 2021). One approach worthy of consideration is computation of the "false positive report probability (FPRP)"[15] to guide selection of α. The challenge with this approach is that it requires an estimate of the prior probability that the tested hypothesis is actually true. When testing a previously unexplored hypothesis, it may be reasonable to assume the prior probability to be low, with the result that false-positive report probability is likely to be high even for $\alpha \leq 0.01$,[15] necessitating replication of the finding (preferably by others), or by a different experimental approach.

One situation in which virtually every statistician advocates for reducing the *P*-value threshold for significance is that of multiple comparisons conducted within a single study. If one conducts a series of properly chosen statistical tests, each with a probability α of giving a false-positive result, then the probability P_{fp} of obtaining at least one false-positive result (ignoring the impact of the *a priori* probability) scales roughly as $(1-\alpha)^n$, where n is the number of tests conducted. For $n = 13$, $P_{fp.}$ is nearly 0.5. Mata and Milner[3] point out the importance of correcting for multiple-hypothesis testing both directly and when considering the use of post-hoc analytical methods, noting that "studies are rarely adequately powered for definitive post hoc testing," and comment that such results should usually be considered "exploratory" and subjected to a new round of experiments. This process of exploratory testing followed by additional, independent experiments is a powerful way to increase the a priori probability of a particular conclusion, the predictive value of a statistical test intended to test that conclusion, and the likelihood that a conclusion supported by that test is actually true. Mata and Milner[3] also warn researchers against using post-hoc analysis to "data dredge" subgroups if their original hypothesis was not supported by their experiment. Both here and in their discussion of stepwise regression, the authors appropriately warn of the risks associated with data dredging practices, which significantly increase the odds that a researcher will arrive at (and publish) conclusions that are actually false.

Statistics Is About Understanding, Not About *P* Values

As noted above, a *P* value itself says little about the magnitude of an effect; nevertheless, knowing the magnitude of an effect is usually critical to understanding its biological importance. Mata and Milner[3] note the importance of effect size estimates when referring to the predictive value of +/− diagnostic tests (analysis of categorical data), as well as by their inclusion of regression and life table methods among the statistical techniques they survey. Understanding effect size also requires understanding the degree of certainty around the effect size estimate; many software packages, such as MedCalc (*https://www.medcalc.org/calc/diagnostic_test.php*, last accessed March 16, 2021), provide confidence limits (typically 95%) around their effect size estimates. Combining effect size and confidence limits with determination of a *P* value may be expected in many cases to improve the assessment of significance.[16]

Some of the methods included by Mata and Milner[3]—all well suited for the experimental designs generally employed by investigative pathologists—may not be appropriate for the kinds of diagnostic test comparisons typically performed in the clinical laboratory. As an example, the Fisher exact test is appropriate for assessing differences between two groups of subjects, such as a group of patients who received a medication versus a group who did not. It is not, however, appropriate for comparing two different diagnostic tests for a single disease, both of which were run on each member of a group of subjects. Statistical considerations for diagnostic tests are described by Biswas[9] in an article that in some ways serves as a companion to that of Mata and Milner.[3]

The discussion above (together with the article by Mata and Milner[3]) is based on the use of a frequentist approach to statistical inference. Although this approach to determining truth dominates biological sciences, it is much less frequently employed in disciplines such as physics and chemistry. Frequentist statistical analysis is not the only statistical approach applicable to biological science; one statistician lamented that "We teach it because it's what we do; we do it because it's what we teach."[5] Other approaches, such as Bayesian analysis, or combinations of Bayesian and frequentist approaches, are more appropriate (although perhaps less accessible and less used) for many scientific investigations.[17] Some people would argue that large and consistent biological effects do not even benefit from statistical inference; it would be hard to advocate for requiring statistical validation before using transfusion to treat exsanguinating hemorrhage.

Following Mata and Milner's guidance on using correct statistical approaches, together with improved understanding of the meaning and limits of the *P* value, will help to improve rigor, but neither these steps not arbitrary reduction in the threshold for significance will solve the reproducibility in science. We can only see what we are looking for, and if we are not careful, overreliance on statistical analysis alone can bias what we decide to look for.

Many of the challenges associated with rigor and reproducibility in the conduct of scientific research can be improved by changes in how authors and reviewers think about experiments, write articles, and evaluate the work of others. When statistical methods are used, we must ensure that they are both appropriately employed and properly interpreted—an effort that will be aided by Mata and Milner's efforts. We should make publication decisions not on the basis of *P* values, but on the presentation of well-developed hypotheses and adequately powered experiments that correctly interpret these hypotheses. We ought to be careful about how we present and interpret *P* values, given their limited meaning in determining whether conclusions are actually true or not, considering statistical tests in the context of their overall predictive value. Finally, we should remember that careful application of statistical thinking and analysis cannot guarantee that we will interpret an experiment correctly, particularly if we have not attempted to look for other equally compatible explanations.

Replication failure based on a failure to understand our own experiments is hardly a new phenomenon, as demonstrated by articles on "polywater"[18] and "cold fusion"[19]. As scientists, our first obligation is to be open to the notion that we may be entirely on the wrong track, and ready to change course when the data are compatible with explanations other than the one we have favored. Statistical analysis, when properly employed, can improve the quality of the scientific work; chasing *P* values is likely to have precisely the opposite effect.

References

1. Ioannidis JPA: Why most published research findings are false. PLoS Med 2005, 2:e124
2. Hsieh T, Vaickus MH, Remick DG: Enhancing scientific foundations to ensure reproducibility: a new paradigm. Am J Pathol 2018, 188:6−10
3. Mata DA, Milner DA: Statistical methods in experimental pathology: a review and primer. Am J Pathol 2021, 191:784−794
4. Hubbard R, Haig BD, Parsa RA: The limited role of formal statistical inference in scientific inference. Am Stat 2019, 73(Suppl):91−98
5. Wasserstein RL, Lazar NA: The ASA's statement on p-values: context, process, and purpose. Am Stat 2016, 70:129−133
6. Kmetz JL: Correcting corrupt research: recommendations for the profession to stop misuse of p-values. Am Stat 2019, 73(Suppl):36−45
7. Nuzzo R: Scientific method: statistical errors. Nature 2014, 506: 150−152
8. Hubbard R: Will the ASA's efforts to improve statistical practice be successful? some evidence to the contrary. Am Stat 2019, 73 (Suppl): 31−35
9. Biswas B: Clinical performance evaluation of molecular diagnostic tests. J Mol Diagn 2016, 18:803−812
10. Hoenig JM, Heisey DM: The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001, 55:19−24
11. Betensky RA: The p-value requires context, not a threshold. Am Stat 2019, 73(Suppl):115−117
12. Wasserstein RL, Schirm AL, Lazar NA: Moving to a world beyond "p $<$ 0.05." Am Stat 2019, 73(Suppl):1−19
13. Janiaud P, Serghiou S, Ioannidis JPA: New clinical trial designs in the era of precision medicine: an overview of definitions, strengths, weaknesses, and current use in oncology. Cancer Treat Rev 2019, 73: 20−30
14. Benjamin DJ, Berger JO: Three recommendations for improving the use of p -values. Am Stat 2019, 73(Suppl):186−191
15. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004, 96:434−442
16. Goodman WM, Spruill SE, Komaroff E: A proposed hybrid effect size plus p-value criterion: empirical evidence supporting its use. Am Stat 2019, 73(Suppl):168−185
17. Krueger JI, Heck PR: Putting the p-value in its place. Am Stat 2019, 73(Suppl):122−128
18. Rousseau DL: "Polywater" and sweat: similarities between the infrared spectra. Science 1971, 171:170−172
19. Pool R: Cold fusion: only the grin remains: a year and a half after the original report of "fusion in a jar," a few dogged researchers are still hoping to confirm the existence of a low-level, neutron-producing nuclear process. Science 1990, 250:754−755