GUEST EDITORIAL

Brad Verhulst, PhD

In Defense of *P* Values

P values have become the scapegoat for a wide variety of problems in science. P values are generally overemphasized, often incorrectly applied, and in some cases even abused. However, alternative methods of hypothesis testing will likely fall victim to the same criticisms currently leveled at P values if more fundamental changes are not made in the research process. Increasing the general level of statistical literacy and enhancing training in statistical methods provide a potential avenue for identifying, correcting, and preventing erroneous conclusions from entering the academic literature and for improving the general quality of patient care.

Keywords: *P*-hacking, *P* values, statistics.

Academics love to hate *P* values. Recently, dozens of commentaries in prestigious academic journals and well-thought-out position papers in academic blogs have criticized the use or misuse of *P* values in biomedical research, and in science more broadly.[1-3] In fact, one journal (*Basic and Applied Social Psychology*) recently banned the use of hypothesis testing,[4,5] stating that there is "no inferential statistical procedure that has elicited widespread agreement". Although this aversion is justified to a limited extent, it is doubtful whether this hostility should be specifically targeted at *P* values, rather than a more diffuse sense of negativity toward absolutist thinking about hypothesis testing, a general fear of statistics and mathematics, and uncertainties associated with conducting and publishing research. If so, the solution is not as simple as doing away with *P* values, but rather teaching researchers to more effectively communicate and evaluate empirical results. This requires that people are able to effectively communicate and critically evaluate the various decisions made in the analytical pipeline so that they can identify the validity of the hypothesis being tested. That said, if the scientific community collectively decides to move away from *P* values, then logically these values must be replaced with some other construct, such as confidence intervals, Bayesian credible intervals, or perhaps another yet-to-be-discovered statistic. The catch is that simply replacing *P* values with another statistic does not address the negativity directed at *P* values.

P values are statistical tools that guide hypothesis testing and scientific decision making. However, *P* values are only one tool in a much larger statistical toolbox that includes many other implements: correlations, regression coefficients, odds ratios, relative risk ratios, standard errors, effect sizes, and so on. To conduct an analysis and make any conclusions about a research question, we must simultaneously use several procedures. It is important to stress that nurse anesthetists should not base clinical decisions too heavily on *P* values and overlook other equally important statistics, but we must also be cautious not to act too rashly by prematurely discarding potentially useful information. There are positive and negative aspects of most tools, and whether the positives outweigh the negatives often lies in the expertise of the user, not in the tool itself.

One likely reason that we emphasize *P* values over other statistics is the extreme simplicity, and near-universal acceptance, of the "rule" that a hypothesis test must have a *P* less than .05 to be deemed statistically significant. This arbitrary yet dogmatically accepted rule fools us into believing that we can compare disparate statistics by reducing everything to *P* values. In doing so, we obscure the difference between statistical significance and clinical relevance. *P* values are influenced by a plethora of scientific decisions that are made in the process of conducting a study, such as who or what should be measured, how people should be assessed, and how the data should be analyzed. Some of these decisions are made intentionally after much deliberation, whereas others are made reflexively, or because that is the way that it is always done. Many clinicians, healthcare providers, and even research scientists do not appreciate the subtleties of *P* values and are unaware of just how many factors influence, and potentially invalidate, them.

Invalidating *P* Values

Before a discussion of the potential perils of *P* values, it is necessary to define what they actually are. A *P* value is the probability of observing a hypothetical parameter (eg, an odds ratio) at least as extreme as the one observed due to chance alone. *P* values do not imply that the effect is real or important; nor do they give any indication of the magnitude of the effect.[2] Accordingly, *P* values should be interpreted in conjunction with other statistics, such as effect size, to infer not only the statistical reliability of the hypothesis test but also the clinical relevance of the result.

Although *P* values can be quite useful, it is easy for preconceived expectations to affect data analysis decisions, which subsequently make any conclusions based on the analysis invalid. Importantly, the invalidation or misuse of *P* values can be either intentional or unintentional. Although intentionally misusing *P* values is obviously dubious (and potentially fraudulent), in many ways unintentional misuse has a larger impact on the literature. This is complicated by the fact that unintentional misapplications of *P* values are subtle and are nearly impossible to distinguish from diligent, methodical, careful statistical analysis in any specific instance. In an ideal situation, researchers will have specified the exact methodological and analytical steps necessary to conduct a study before any data collection or analysis. Reality is murky, however, and unanticipated factors often arise that must be addressed. Here I discuss a few common situations that affect the validity of *P* values.

- ***"P-hacking."*** One factor that invalidates *P* values is *P*-hacking.[6,7] *P*-hacking occurs when an analysis is conducted and then reconducted with minor alterations until a statistically significant finding is obtained. Stating *P*-hacking in such stark terms gives the illusion that it is easy to identify, but in reality, in any given study, the *P*-hacking behavior could be completely justifiable and even encouraged. Three common methods of *P*-hacking include excluding selective observations, transforming the data, and adding control variables.

- *Selective Exclusions.* One potential misuse of *P* values relates to the decisions about whether to include or exclude observations from an analysis. Obviously, excluding specific observations with the sole purpose of inflating the significance is ethically problematic. However, it is common to make post hoc exclusions that may have the same nefarious impact on the results. Although excluding outlying observations based on sex, age, family history, or numerous other factors may have a basis in the literature, if the specific criteria are not clearly enumerated before data collection, researchers may be seduced by post hoc exclusion criteria that could inflate the level of statistical significance. Ideally, the exclusion criteria are articulated before the article is written, but all too often writing and analysis are conducted in an iterative manner, in which the analysis informs the writing and the writing subsequently informs the analysis. This seduction is amplified when a researcher reads a variety of articles, each using different exclusion criteria, when writing the introduction to the manuscript. These newfound criteria, which can seem obvious in hindsight, may not represent an unbiased set of additional exclusion criteria, but instead be guided by likely unconscious motivations to find significant results.

  In some cases, however, excluding observations is inherently justifiable, and to complicate matters, the optimal strategy for dealing with outliers and influential observations is unclear. Some statisticians believe that excluding an observation is inherently problematic, whereas others have more liberal criteria for justifying the removal of observations. This ambiguity epitomizes the challenges involved in delineating clear rules for exclusion criteria, knowing that whatever choice is made will influence the reported *P* value.

- *Transformations.* Another potential misuse of *P* values occurs when people transform their data. Most statistical techniques assume that the underlying data are normally distributed, but with real data the best that we can hope for is that the data are approximately normally distributed. Many common statistical techniques are robust to minor deviations from normality, but more egregious deviations can interfere with the validity of the test statistic and subsequently the *P* value. One common solution to extreme deviations from normality is to transform the variable, for example, by taking the logarithm or square root of the outcome variable. Transformations may minimize deviations from normality, but they can fundamentally change the interpretation of the results and capitalize on the oddities of a specific dataset. Although there is nothing inherently wrong with transforming variables, at minimum, researchers should be aware of the impact that it has on the *P* values they report and be appropriately skeptical of large changes in statistical significance. If this difference is large, an alternative to transforming variables is to actively account for the specific nonnormal distribution. Such methods, however, may be more technically sophisticated and might be unfamiliar to anesthesia researchers, reviewers, and clinicians.

- *Control Variables.* Including control variables can also affect the *P* values for the association of interest. This is one of the "fuzziest" methods of *P*-hacking. Many people encourage including control variables as a method of ruling out alternative explanations, reducing the negative impact of nonrandom sampling, making the results more

generalizable, or reducing the random noise in the model. Adding (or removing) post hoc control variables, however, may also capitalize on chance, especially when multiple control variables are examined (and potentially abandoned). Again, all attempts should be made to specify the appropriate control variables before conducting the study, with the knowledge that some unanticipated factors may arise during the study that need to be dealt with analytically. Thus, researchers must be careful when they decide to include control variables in their analyses.

What is clear from the discussion of *P*-hacking is that no firm rules can be gleaned that can be applied in all circumstances. Each behavior can appear appropriate and well intentioned, but can also have profound implications on the results, the *P* values, and the conclusions that are drawn.

• ***Multiple Testing.*** One of the most common ways that *P* values are invalidated is by ignoring the impact of conducting multiple hypothesis tests. The more hypotheses that a researcher tests, the more likely that one will be significant because of chance alone. Importantly, this occurs much faster than most people realize. By the time that a researcher conducts 10 independent tests using a *P* of .05 for each test, there is a 40% probability that at least 1 test will be significant. Although it is unlikely that researchers intentionally conduct a large number of tests with the explicit goal of capitalizing on chance, unintentionally conducting multiple tests is probably more common than anyone would like to admit.

It is expensive to collect data, so when data are collected, a large number of potential dependent and independent variables are typically measured. The various permutations of each potential analysis can result in an enormous number of possible hypothesis tests. To complicate things, different combinations of variables from the same study are published in different articles, with multiple sets of analyses being conducted for each article. In addition, the number of analyses conducted for each article is often not recorded. Even though it is relatively easy to apply a simple multiple testing correction to an analysis (routinely discussed in statistics textbooks), most reviewers would not think to inquire on this point, and it is not common for authors to voluntarily apply this "penalty" to their own research. Accordingly, it is likely that many of the *P* values reported in the literature are misleading because they fail to take multiple testing into account.

Alternatives to *P* Values

It is popular to criticize *P* values, but it is essential to note that no one is advocating for moving away from empirical evidence as the basis for clinical research. If we accept the premise that *P* values should be jettisoned, we must decide what metric should be used in their stead. Several possibilities exist that deserve elaboration.

The most common suggestion is to replace *P* values with confidence intervals. The primary benefit of confidence intervals relative to *P* values is that the interpretation of the confidence interval includes the parameter of interest within the interval. The problem with this suggestion is that it is possible to conceptualize confidence intervals as a transformation of *P* values. The hypothesis test for confidence intervals revolves around whether the confidence interval includes the null value (eg, a correlation of 0, or an odds ratio of 1). Therefore, the obvious danger of replacing *P* values with confidence intervals is switching from dogmatically requiring $P < .05$ to dogmatically requiring that the confidence interval does not include the null value. Other, more advanced methods such as bootstrapped or likelihood-based confidence intervals,[8,9] Bayesian credible intervals,[10] or false discovery rates[11,12] exist that relax the some of the assumptions of standard confidence intervals. However, acquiring the necessary statistical expertise can take substantial effort, and the impact on the statistical significance is often minimal.

Thus, although there are several established alternative scientific decision-making procedures that do not directly involve looking at *P* values, most of these alternative procedures are intimately related to *P* values. Accordingly, there are limited reasons to believe that adopting a new procedure will solve the problem of dogmatically looking at one factor to determine whether an analysis is important.

Solutions

As was alluded to earlier, the solution to the "*P*-value problem" in many ways revolves around increasing the general level of statistical literacy of students, researchers, reviewers, and clinicians. A firm understanding of the meaning and limitations of *P* values, and of other statistical techniques for that matter, would go a long way toward addressing the underlying malaise surrounding the communication of empirical results. Although increasing general levels of statistical expertise is no panacea, a better understanding of the problem is the first step toward a solution. If people are aware of the various ways that *P* values can be unintentionally manipulated, they can be more cognizant of these factors when conducting their own analyses, reviewing articles, and critically evaluating the existing literature and applying it in the clinic.

One possible first step toward addressing the *P*-value problem is to improve the statistical training in clinical doctorate programs. The goal of education is, broadly, to train people to critically evaluate infor-

mation, and this includes statistical concepts. The clinical training in most nurse anesthesiology programs is designed to be more rigorous than the associated methodological training. This is reasonable given that students (and their faculty) are primarily interested in mastering the specific procedures that they will routinely use in the operating room. Therefore, the students (and faculty) put substantially more emphasis on clinical classes than statistics classes. This is not to say that students should pay less attention to their clinical classes or that the statistical rigor of anesthesiology programs should rival doctoral research in statistics. Rather, the incentives could be structured in a way that motivates students to increase the attention that they pay to their methodological classes without negatively affecting the clinical aspects of their training. This is easy to suggest but difficult to implement.

The importance of this increased methodological training in clinical doctorate programs is exacerbated by the fact that for many nurse anesthetists, statistical training ends when they finish their doctorate programs. Accordingly, when they graduate, the newly minted practitioner is able to evaluate the statistical analyses in most articles, but with the lack of continued statistical practice, their methodological knowledge recedes, and the gap between their level of statistical literacy and the published literature increases. This inhibits a researcher's ability to understand the strengths and weaknesses of the appropriate statistical methods for testing hypotheses and affects the quality of research. Similarly, receding levels of statistical literacy in reviewers will prevent them from effectively evaluating articles and from catching potential errors before they are published. Finally, a better understanding of statistical concepts will allow clinicians to evaluate research methods, in order to ultimately translate novel research findings into standard care procedures.

New statistical techniques are constantly being developed and integrated into the literature. Thus, the gap between a researcher's or practitioner's level of statistical literacy and the knowledge necessary to effectively evaluate published research will grow unless they actively invest effort to stay up-to-date with methodological advancements. This is similar to the continuing education requirements necessary for medical practice. Therefore, focusing on students is not broad enough to address the current issues with statistical literacy.

Implementing this solution and changing the culture around research will be long, involved, and unpopular for several reasons. First, everyone is busy, and implementation will add an additional time burden. Second, many people (including many scientists) are scared of math, and this aversion will lead people to avoid taking methods classes or staying abreast of statistical developments. Third, many people will not see an inherent problem with the current system and may prefer a quick fix that does not fully address the underlying problem.

If we are committed to improving the quality and rigor of research, before discarding the *P* value entirely, we must address this underlying problem, by teaching people to more effectively communicate and evaluate empirical results and to increase the general level of statistical literacy. Doing so provides a potential avenue for identifying, correcting, and preventing erroneous conclusions from entering the academic literature and for improving the general quality of patient care.

REFERENCES

1. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA*. 2016;315(11):1141-1148.
2. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Statistician*. 2016;70(2):129-133.
3. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716.
4. Trafimow D. Editorial. *Basic and Applied Social Psychology*. 2014;36(1):1-2.
5. Trafimow D, Marks M. Editorial. *Basic and Applied Social Psychology*. 2015;37(1):1-2.
6. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*. 2011;22(11):1359-1366.
7. Ulrich R, Miller J. p-hacking by post hoc selection with multiple opportunities: detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *J Exp Psychol Gen*. 2015;144(6):1137-1145.
8. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman Hall/CRC; 1993.
9. Neale MC, Miller MB. The use of likelihood-based confidence intervals in genetic models. *Behav Genet*. 1997;27(2):113-120.
10. Bernardo JM. Intrinsic credible regions: an objective Bayesian approach to interval estimation. *Test*. 2005;14(2):317-384.
11. Storey JD. A direct approach to false discovery rates. *J Roy Stat Soc Ser B (Statistical Methodology)*. 2002;64(3):479-498.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B (Methodological)*. 1995;57(1):298-300.

AUTHOR

Brad Verhulst, PhD, is a research associate in statistical genetics at the Virginia Institute for Psychiatric and Behavioral Genetics in the Department of Psychiatry at Virginia Commonwealth University, Richmond, Virginia. Email: bverhulst@vcu.edu.

DISCLOSURES

The author has declared he has no financial relationships with any commercial interest related to the content of this article. The author did not discuss off-label use within the article.

ACKNOWLEDGMENTS

The author is grateful for comments from Drs Shaunna Clark and Michael Neale. This research was supported by NIDA Grants R01DA-018673 and R25DA-26119.