Understanding the true significance of a *P* value

Andrew D. Althouse, PhD,[a] and Prem Soman, MD, PhD[a]

[a] Heart and Vascular Institute, University of Pittsburgh Medical Center, Pittsburgh

Received Jun 22, 2016; accepted Jun 22, 2016
doi:10.1007/s12350-016-0605-1

In an effort to ensure that scientific conclusions are supported by accompanying data, many journals require data to be presented with an assessment of ''statistical significance''—most commonly with a *P* value. Unfortunately, this has created a persistent dogmatic approach where *P* values are often treated as the sole determinant of ''significance''—and furthermore, reporting a *P* value <0.05 has become both a necessary and sufficient condition to make virtually any claim,[1] regardless of how well the *P* value addresses a particular question. Readers should be aware that the *P* value is just one element of data analysis, and a number of other elements should also be weighed in the interpretation of study data.

The American Statistical Association recently released a statement addressing the widespread abuse and misunderstanding of *P* values.[2] The purpose of this communication is to relay the sentiments in the ASA statement while simultaneously providing some specific context for cardiovascular researchers.

DEFINING THE *P* VALUE

Calculating any ''*P* value'' requires that we begin with a *null hypothesis* (for example, ''Treatment A offers equal benefit to Treatment B'') and collect some data that will help us confirm or refute the hypothesis. Once we have collected the data, we may compute *the probability of observing the data actually seen in the study if the null hypothesis were true.*

However, most readers believe that a *P* value represents the *probability that the null hypothesis is true given the observed data*: If this probability is small, most readers will interpret this to mean that the data provide strong evidence against the null hypothesis, and many will conclude that the null hypothesis is false (or likely to be false). That is:

Probability (Null Hypothesis True | Observed Data) = ''What Most People Want The *P* Value To Be''

However, this interpretation is incorrect. As noted in the recent ASA statement, the *P* value says nothing about the probability that the null hypothesis is true or false. In fact, the *P* value actually represents the *probability of observing the data seen in the study if the null hypothesis is actually true*:

Probability (Observed Data | Null Hypothesis True) = ''What The *P* Value Actually Represents''

To use a classic example from introductory statistics, suppose that a researcher finds a coin on the street and wishes to determine whether it is a fair coin (equal probability of landing on heads or tails). In this case, the ''null hypothesis'' is that the coin is fair, meaning that there is a 50% probability of landing heads on any single toss. If the coin holder tosses the coin five times and the coin comes up heads all five times, we can easily compute the probability that this would have occurred with a fair coin:

Probability (Five Consecutive Heads | Fair Coin) = 0.5 * 0.5 * 0.5 * 0.5 * 0.5 = 0.03125 = 3.125%

The proper interpretation of this probability says that there was a 3.125% chance of getting five straight heads on five tosses given that the coin is fair; the incorrect interpretation says that there is a 3.125% chance that the coin is fair given that we have observed five consecutive heads. And yet many readers would likely choose the second interpretation!

What does that mean in the context of cardiovascular research? Consider a parallel-group randomized trial with two treatment arms that lists a *P* value of 0.03 for the primary treatment comparison. This *P* value does *not* mean that there is only 3% probability that the two treatments are equal, given the results seen in the trial (although this is the interpretation many readers will give). Rather, this $P = 0.03$ actually says that there was *3% probability of observing the results seen in the trial if the two treatments are equal*.

Reprint requests: Andrew D. Althouse, PhD; Heart and Vascular Institute, University of Pittsburgh Medical Center, Pittsburgh; *althousead@upmc.edu*

J Nucl Cardiol 2017;24:191–4.
1071-3581/$34.00

Despite this fundamental misunderstanding, the *P* value still functions reasonably well as an assessment of the strength of evidence in properly designed randomized controlled trials of sufficient sample size. In general, the *P* value will control the risk of a false-positive result when (a) the primary hypothesis is well defined and (b) there is a commitment to publishing the results regardless of the study outcome. Most trials have prespecified primary hypotheses and some form of requirement for public reporting, minimizing the file-drawer effect (where nonsignificant results are buried while significant results are published, leading to an overestimation of the true effect) and decreasing the risk of substantial amounts of data dredging (where many associations are examined and only ''significant'' associations are presented in a paper).

However, the vast majority of published research is not performed in randomized trials; most research is performed in observational settings. Furthermore, many observational studies have more nuanced hypotheses than simple 2-group comparisons, occasionally creating situations in which *P* values are either meaningless or, at the very least, less informative than simple graphical displays and other summary statistics. We provide contextual examples to help the reader understand how to interpret data rather than merely looking for ''significant'' *P* values.

WHEN THE *P* VALUE DOESN'T REALLY ANSWER THE PERTINENT QUESTION(S): EXAMPLE 1

The misguided emphasis on finding small *P* values to highlight the presence of ''significant'' findings has led many investigators to dig around for a *P* value to ''prove'' something even when it's not really appropriate for the study question.

Suppose that a researcher has recently purchased a new camera for their cardiac testing lab. They wish to compare the new camera's measurement of left ventricular ejection fraction (LVEF) against the LVEF measured by their lab's old camera. The researcher enrolls a series of 30 patients and measures LVEF on all 30 patients with both cameras. They perform a simple linear regression with LVEF measured by the new camera as the independent variable and LVEF measured by the old camera as the dependent variable, and while digging through their statistical output, they are delighted to find a *P* value of 0.008 for the relationship between LVEF measured by the two cameras. Satisfied with the small *P* value, the author concludes that there is a ''significant'' relationship between the two cameras, and that the new camera provides similar measurements to the old camera.

What's important to understand here is that the ''null hypothesis'' most people are testing in this scenario is only testing whether the relationship between the two variables (assessed by the slope of a regression line or the correlation coefficient) is different from zero. Therefore, the *P* value for the specified model above tells us only *the probability of observing this data if the two variables are not correlated at all*. Again please note that the small *P* value does not tell us anything about the *strength* of the association between the two variables or even the probability that the two variables are correlated with any degree of strength; the small *P* value means only that ''we are unlikely to observe this data if the variables are entirely uncorrelated.''

Despite this, many investigators will report the small *P* value and conclude that the two measurements are ''significantly correlated'' with one another. Technically, that is true—there is a relationship between the two variables—but the actual *agreement* between the two cameras can be quite poor. For illustration, we present a randomly generated sequence of 30 data points which have a correlation coefficient of 0.477 and a *P* value of 0.008. As the graphical representation shows (Figure 1), the two variables are modestly *associated,* but do not actually *agree* with one another very well at all. However, the *P* value of 0.008 would suggest a degree of ''significance'' to the result that belies its clinical interpretation.

For this particular question, the more appropriate analysis would be producing Bland-Altman plots and computation of 95% limits of agreement, followed by a qualitative evaluation of whether the limits of agreement were acceptable for a difference between the two cameras. The question of what constitutes ''acceptable'' agreement between the two cameras should be a *clinically* driven decision (i.e., we will accept cameras that reliably estimate EF within 5% of one another); this is not something that can be answered by statistics alone. But that requires additional thought and effort instead of merely saying that $P < 0.05$ means significant correlation between the variables.

WHEN THE *P* VALUE DOESN'T REALLY ANSWER THE PERTINENT QUESTION(S): EXAMPLE 2

Another situation in which *P* values are often improperly used is the construction of multivariable models. Suppose that a researcher wishes to construct a risk score for major adverse cardiovascular events based on a selection of imaging parameters. The researcher records 20 different imaging parameters on all patients that pass through their clinic and follows the patients for a period of time until a sufficient proportion have

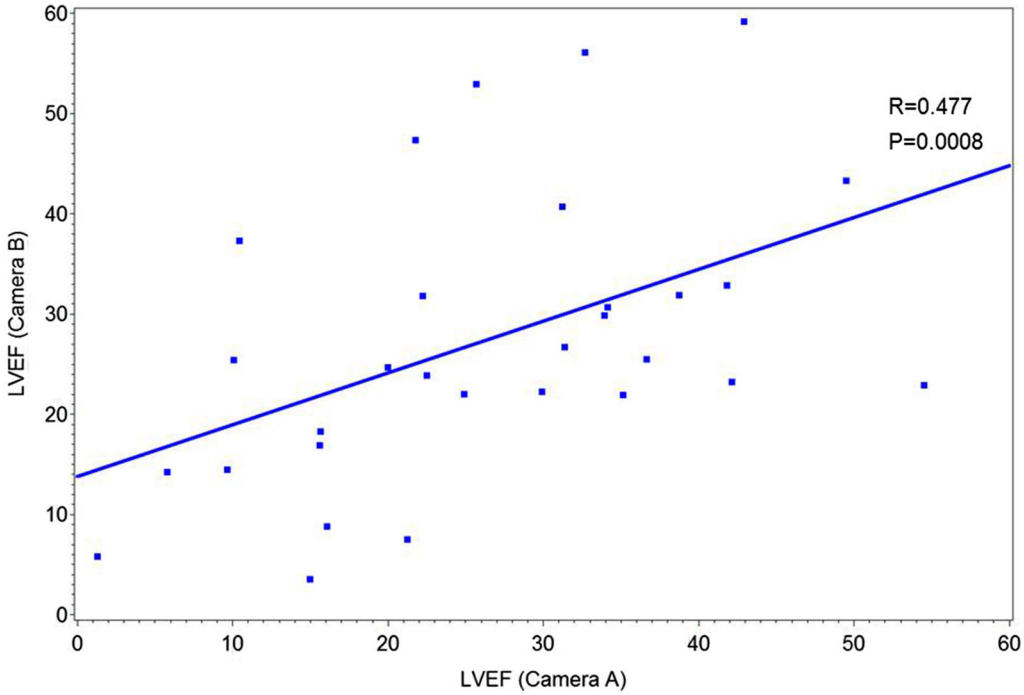

Figure 1. Scatterplot of simulated data illustrating poor agreement despite ''Significant'' *P* value.

reached the selected endpoint (i.e., major adverse cardiovascular event or death) to perform data analysis.

One common approach seen in the literature is to perform multivariable selection, often based on which variables have the smallest *P* values in univariable analysis. However, this does not necessarily lead to the best-fitting model with the greatest predictive power. In this case, the model building procedure should be informed by some index of the model's predictive capacity (something like the c-statistic or the Net Reclassification Index). Variables chosen for the multivariable model should be added based on their additive effects on the model's predictive accuracy, *not* based on their *P* values in univariable analysis.

WHEN THE *P* VALUE DOESN'T REALLY ANSWER THE PERTINENT QUESTION(S): EXAMPLE 3

The last and perhaps the most crucial thing to understand is that *P* values can be influenced by other considerations in the study design, and that these choices can create an illusion of ''statistical significance'' that has somewhat of a dubious real-world meaning.

Recall that *P* values actually have a very specific definition: the *probability of observing study data given that the null hypothesis is true*. Note that the ''null hypothesis'' is a crucial piece in determining the *P* value, and that changes to the null hypothesis will influence this computation. In many cases, the null hypothesis is straightforward to understand: ''mortality in patients treated with Drug A is equal to mortality in patients treated with Drug B.'' However, in selected cases, the null hypothesis is more complex, and subtle decisions in study design can have tremendous influence on the presented results.

An outstanding example is the ABSORB III trial[3] evaluating the effectiveness of bioresorbable scaffolds against the existing class of drug-eluting stents in patients with coronary artery disease. The trial was designed as a noninferiority study (appropriately, given the clinical context; if the scaffold is noninferior to stenting, it may be preferable for some patients).

Let us dig into the statistical weeds, though. Noninferiority studies have a different null hypothesis than the traditional superiority trial. The classic superiority trial has a null hypothesis of ''equivalence'' between the two treatments being studied, so a very small *P* value typically leads us to reject the null hypothesis, noting that it was very unlikely to observe these results if the two treatments were equal, and conclude that one treatment is superior to the other.

The noninferiority study begins with a null hypothesis that ''one treatment is not worse than the other'' which requires the designation of an allowable noninferiority margin; 4.5% was the chosen margin in ABSORB III. Therefore, the null hypothesis is no longer ''patients treated with the bioresorbable stent have equal risk of target lesion failure to patients treated

with drug-eluting stents'' but rather ''patients treated with the bioresorbable scaffold have no worse than a 4.5% greater risk of target lesion failure than patients treated with drug-eluting stents.'' The *P* value, then, will be the *probability of observing the study data if the bioresorbable scaffold is no more than 4.5% worse than the drug-eluting stent*.

The reader should now infer that the *P* value for a noninferiority study will be highly sensitive to changes in the null hypothesis. The probability of observing the data seen in ABSORB III under a null hypothesis of ''patients treated with the bioresorbable scaffold have no worse than a 4.5% greater risk of target lesion failure than patients treated with drug-eluting stents'' will be much different than the probability of observing the data under a null hypothesis that ''patients treated with the bioresorbable scaffold have no worse than a 3.5% greater risk of target lesion failure than patients treated with drug-eluting stents'' or a null hypothesis that ''patients treated with the bioresorbable scaffold have no worse than a 2.5% greater risk of target lesion failure than patients treated with drug-eluting stents.''

The actual study results show slightly poorer performance with the scaffold vs. the drug-eluting stent on all outcomes. However, with the large allowable noninferiority margin (a 4.5% margin seems quite high, given the overall anticipated event rate of 7%—this translates to nearly 1.75 times the risk of target lesion failure with the scaffold!) the authors were able to present a highly significant *P* value for noninferiority of 0.007, very impressive looking for the bioresorbable scaffold.

The highly significant *P* value for noninferiority is somewhat of an illusion, a mirage created by choosing a large enough noninferiority margin to give a ''significant'' result, rather than strong evidence that the scaffold is truly ''not inferior'' to the drug-eluting stent. An allowable noninferiority margin of 3.5% would change the result to *P* value of 0.064; a margin 2.5% would change the result to *P* value of 0.251, changing the study interpretation quite drastically, no longer allowing the authors to claim that they had significant evidence that the scaffold was noninferior.

The purpose of this passage is not to quibble specifically with the ABSORB III findings, but rather to illustrate the complexities of study design and the potential manipulations of the *P* value by tinkering with other elements of the study design. The emphasis on finding small *P* values above all else has promoted a research environment where investigators may choose to manipulate other design elements to assure that they will have a small *P* value, regardless of whether that approach is truly appropriate for the primary study question. This is obviously an undesirable outcome.

CONCLUSION

Statistical analysis remains a critical piece of the scientific process, but many researchers struggle with proper implementation and interpretation. Periodically, statisticians and/or quantitatively oriented researchers have published editorial commentary in clinical journals in an effort to target cardiovascular researchers on their home turf.[4-8] Although reading such material is not a substitute for applied statistical training, the authors typically hope that such pieces will accomplish two things: (1) improve the statistical literacy of journal readers and reviewers and (2) encourage biomedical researchers to enlist statistical support before proceeding with complex analytic efforts.

This piece is another contribution in that tradition. No doubt the dogmatic acceptance of $P < 0.05$ to evaluate all statistical findings was largely driven by an effort to simplify things for nonstatisticians; unfortunately, statistics are not simple, and rigid acceptance of the single guideline without understanding the deeper context. As for the specific question at hand, regarding *P* values, what a researcher has to do? The best advice to the investigator is to enlist a good statistician at the beginning of your research process, design the research effort appropriately for the study question, and to not rely solely on the *P* value when interpreting your findings, but also consider the research design, the effect size, the confidence interval, and graphical summaries when attempting to integrate their research findings into the broader context.

References

1. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting *P* values in the biomedical literature, 1990-2015. JAMA. 2016;315(11):1141-8.
2. Wasserstein RL, Lazar NA. The ASA's statement on *P* values: Context, process, and purpose. Am Stat. 2016;70(2):129-33.
3. Ellis SG, Kereiakes DJ, Metzger C, Caputo RP, Rizik DG, Teirstein PS, et al. Everolimus-eluting bioresorbable scaffolds for coronary artery disease. NEJM. 2015;373:1905-15.
4. Glantz SA. Biostatistics: How to detect, correct, and prevent errors in the medical literature. Circulation. 1980;61:1-7.
5. George SL. Statistics in medical journals: A survey of current policies and proposals for editors. Med Pediatr Oncol. 1985;13:109-12.
6. Glantz SA. It is all in the numbers. J Am Coll Cardiol. 1993;21(3):835-7.
7. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
8. Moye L. Statistical methods for cardiovascular researchers. Circ Res. 2016;118:439-53.