JAMA Cardiology | Special Communication

Understanding the Role of *P* Values and Hypothesis Tests in Clinical Research

Daniel B. Mark, MD, MPH; Kerry L. Lee, PhD; Frank E. Harrell Jr, PhD

P values and hypothesis testing methods are frequently misused in clinical research. Much of this misuse appears to be owing to the widespread, mistaken belief that they provide simple, reliable, and objective triage tools for separating the true and important from the untrue or unimportant. The primary focus in interpreting therapeutic clinical research data should be on the treatment ("oomph") effect, a metaphorical force that moves patients given an effective treatment to a different clinical state relative to their control counterparts. This effect is assessed using 2 complementary types of statistical measures calculated from the data, namely, effect magnitude or size and precision of the effect size. In a randomized trial, effect size is often summarized using constructs, such as odds ratios, hazard ratios, relative risks, or adverse event rate differences. How large a treatment effect has to be to be consequential is a matter for clinical judgment. The precision of the effect size (conceptually related to the amount of spread in the data) is usually addressed with confidence intervals. *P* values (significance tests) were first proposed as an informal heuristic to help assess how "unexpected" the observed effect size was if the true state of nature was no effect or no difference. Hypothesis testing was a modification of the significance test approach that envisioned controlling the false-positive rate of study results over many (hypothetical) repetitions of the experiment of interest. Both can be helpful but, by themselves, provide only a tunnel vision perspective on study results that ignores the clinical effects the study was conducted to measure.

JAMA Cardiol. 2016;1(9):1048-1054. doi:10.1001/jamacardio.2016.3312
Published online October 12, 2016.

Editor's Note page 1055

Author Affiliations: Duke Clinical Research Institute, Duke University School of Medicine, Durham, North Carolina (Mark, Lee); Department of Biostatistics, Duke University School of Medicine, Durham, North Carolina (Lee); Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee (Harrell).

Corresponding Author: Daniel B. Mark, MD, MPH, Duke Clinical Research Institute, Duke University School of Medicine, 2400 Pratt St, Room 0311, Durham, NC 27705 (daniel.mark@duke.edu).

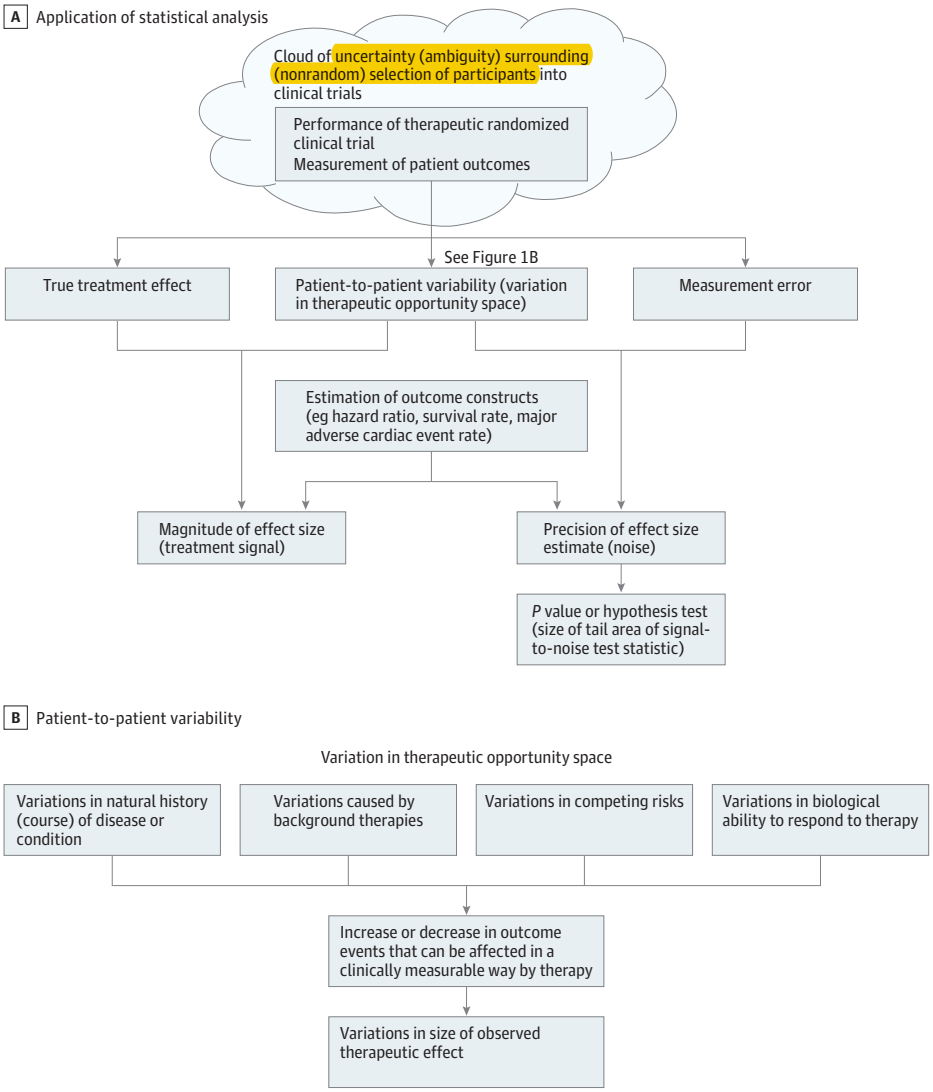
P values (the product of significance tests) and hypothesis testing methods are frequently misunderstood and often misused in clinical research.[1-3] Despite a large body of literature advising otherwise, reliance on these tools to characterize and interpret scientific findings continues to increase.[4] The American Statistical Association[5] recently published a consensus white paper attempting to promote a more limited, rational role for the *P* value in science. That consensus statement was accompanied by 21 individual commentaries from members of the panel, each adding his or her own caveats to the discussion. Our justification for writing yet another article on this surprisingly controversial subject lies in the hope that, by taking a substantially different approach that is more conceptual and less technical, we can enhance understanding of the roles that *P* values and hypothesis tests are best suited to fill.

Science as Measurement: Understanding the Nature of Clinical Evidence

When we do science, in most situations that means we measure something.[6-8] In the context of a therapeutic clinical trial, we can think of the treatment as a metaphorical therapeutic force, an "oomph" effect that pushes the intervention group away from the control group, which creates distance or separation between the 2 groups with respect to outcomes of interest.[9,10] An ineffective treatment, therefore, has no oomph. To measure the consequences of this therapeutic force, we use 2 complementary concepts, namely, magnitude and precision. Outcomes measured in individual patients are used to estimate the magnitude or size of the treatment effects produced, both for benefit and for harm as the circumstances dictate. The further the therapy in question "pushes" the treatment group away from its peers in the control group, the larger the average treatment effect is. How large a treatment effect has to be to be consequential is a matter for clinical judgment. One of the added complexities of clinical sciences is that, when a treatment saves a life or prevents a stroke or heart attack, for example, the prevented event is invisible clinically and can only be measured indirectly using appropriate controls. We usually measure therapeutic and other clinical effects in groups of patients using quantitative constructs, such as odds ratios, hazard ratios, relative risk reductions, and survival or adverse event rate differences. Statisticians often use the term *estimation* when summarizing the measurement of cohort-level clinical therapeutic effects.

The second key concept involves the idea of "precision," the amount of spread or variability in the data. At the level of individual patients, precision can be understood in terms of measure-remeasure variability of clinical variables. The tighter measurements cluster around each other, the more precise (free from "random" or patternless error) the measurement process is thought to be. In much of clinical research, however, variability of the measure-

Figure 1. *P* Values and Hypothesis Tests in Context



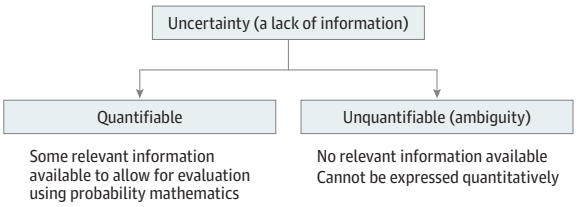
A, In a clinical trial, patients are selected for enrollment (in an undefined, nonrandom fashion), randomized treatment strategies are administered, and outcomes are measured. From these patient-level measurements, outcome constructs are estimated that convey the primary trial findings, typically in terms of effect size magnitude (treatment signal) and precision (noise). *P* values (and hypothesis tests) provide ancillary information regarding how unexpected the study results are if one assumes the null hypothesis (no treatment effect) is correct. B, "Therapeutic opportunity space" is a metaphor we use to draw attention to the different dimensions of patient-to-patient variability that collectively affect individual patient responses to treatment and the result on observed therapeutic effect size. When the size of therapeutic opportunity space is closely coupled to overall patient risk level, our ability to select the patients who will demonstrate benefit from therapy is often good. When this relationship becomes weak or uncoupled, trials of therapies may demonstrate perplexing variability.

remeasure type has quantitatively much less influence on the precision of estimated outcomes than patient-to-patient variability. In clinical research, when we refer to the *precision* of the effect size, we are using a concept that blends the effects of uncertainty due to patternless measurement errors with the uncertainty (often nonrandom) resulting from variations among patients. The most commonly used precision assessment tools in clinical research are confidence intervals. All other things being equal, studies with greater precision in the estimation of a nonzero effect size are also more likely to have a smaller *P* value, but the relationship between precision and *P* values is indirect, as will be discussed below.

The treatment effects we report in clinical research thus reflect the admixture of at least 4 distinct but intertwined factors (**Figure 1**A), including the true size of the treatment effect under ideal circumstances (which is unknowable) and the following 3 sources of uncertainty: (1) the selection (almost always nonrandom) of participants into the trial from the universe of nominally eligible individuals, (2) the amount of measurement error in the effect size estimate (eg, the difference in event rates reported in a clinical trial by the enrolling sites vs the clinical events committee), and (3) the variability in the potential for study participants to demonstrate a real therapeutic effect (which we have termed variations in the *therapeutic opportunity space* (Figure 1B). Statistical analysis uses probability tools to attempt to untangle the true value from the measurement error and the uncertainty created by study population variability. However, the uncertainties involved in some parts of the research process, particularly unmeasured factors exerting an important (and unsuspected) influence on the creation of the study cohort or the effectiveness of the therapy delivered to patients (depicted by the cloud in Figure 1A), can influence whether a study shows a significant treatment benefit or no treatment effect at all and are not controlled with statistical or probability tools.

The recent Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial[11] illustrates some of the complex ways in which unquantifiable uncertainties (**Figure 2**) can produce confusing, unexpected trial results. TOPCAT tested spi-

Figure 2. Two Main Types of Uncertainty in Science



Uncertainty, as it is encountered in science, reflects a lack of information and can be usefully divided into 2 subtypes: quantifiable uncertainties and unquantifiable uncertainties (ambiguity). Many times, our uncertainty involves a situation where some information is available, enough that the uncertainty can be quantified with probability mathematics. Significance testing and hypothesis testing were derived to address this form of uncertainty. However, another important form of uncertainty (ambiguity) is one where no relevant past information is available, and thus this form of uncertainty cannot be quantified.

ronolactone vs placebo in 3445 patients who have heart failure with preserved systolic function. The trial was officially interpreted as "negative" (hazard ratio, 0.89; 95% CI, 0.77-1.04), but prespecified subgroup analyses revealed an 18% reduction in the primary end point (composite that included cardiovascular mortality and hospitalization for heart failure) in the 1767 patients enrolled in the Americas (ie, United States, Canada, Brazil, and Argentina) (hazard ratio, 0.82; 95% CI, 0.69-0.98) and no effect in the 1678 patients enrolled in Russia and Georgia (hazard ratio, 1.10; 95% CI, 0.79-1.51).[12] In addition, patients who qualified for the study based on an elevated brain-type natriuretic peptide (BNP) level showed a larger treatment benefit (hazard ratio, 0.65; 95% CI, 0.49-0.87), while the individuals who qualified based on a prior heart failure hospitalization showed no effect (hazard ratio, 1.01; 95% CI, 0.84-1.21). Further analysis demonstrated the following 2 notable factors: (1) only 11% of the Russia and Georgia patients were enrolled in the BNP stratum of the trial vs 45% of the Americas patients and (2) the Russia and Georgia patients had a much lower event rate in the placebo arm than the Americas patients (8% vs 32%). Therefore, an unexpectedly heterogeneous admixture of responders (meeting BNP entry criteria, high placebo event rate, and mostly enrolled in the Americas) and nonresponders (BNP status unknown, very low placebo event rate, and predominantly enrolled in Russia and Georgia) into TOPCAT is a plausible explanation for the overall negative results. A much less likely and less plausible interpretation is that spironolactone had no therapeutic benefit in the target population and that the appearance of benefit in the BNP and regional subgroups was "the play of chance."

In clinical trials, investigators often rely on the convenient but implausible assumption that they somehow have obtained a sufficiently representative sample of some theoretical underlying population of interest. In other words, investigators regard their experimental data as if they were generated by some unseen random sampling process, a "metaphorical lottery."[13] This assumption facilitates the expectation that statistical probability tools, along with the correct form of analysis, are sufficient to understand the outcome of the study and generalize usefully from it. One danger in this assumption is that, when probability concepts (eg, the play of chance) are used to explain unexpected or discordant trial results, further research work on the treatment may be prematurely abandoned.

The *P* Value and the Hypothesis Test: Using the Tail to Learn About the Dog?

The *P* Value or Significance Test

Technically, the "*P* value" is the product of a statistical procedure, the "significance test." In this discussion, we will use the 2 expressions synonymously. Understanding the *P* value may be easier if we start with the related idea used in engineering of a signal-to-noise ratio.[14,15] Statistics commonly uses the ratio between the effect size of the treatment or exposure being studied (the "signal") and the precision in the estimation of that effect size (the "noise") to create useful statistical tools (eg, *t* score or *z* score). Summarizing the precision component using standardized precision units (termed *standard errors*), a treatment (oomph) effect that displaces the intervention group by a distance equal to 2 or more standardized precision units away from the control group would generally produce a "statistically significant" *P* value. Therefore, statistically significant in this context simply means that the data had a sufficiently large signal-to-noise ratio, which is a ratio between the effect size and the precision units of at least 2.

Moving from the ratio between the effect size and the precision units (test statistic) concept to the now familiar *P* value was an innovation most closely associated with Sir Ronald Fisher, a British scientist and one of the founders of modern statistics.[16,17] As part of this innovation, Fisher used the idea of the null hypothesis, a statistical straw man intended to reflect what the outcome data would look like (how it would be distributed) if the treatment or intervention had no actual effect. In a fully deterministic world (no uncertainty), "no effect" would register as a 0 treatment effect size, and the matter would end there. Because of the uncertainty inherent in the natural world, there is almost always some nonzero measured effect size even if the treatment has no biological effects at all. Hence, the effect size needs to be "big enough" that it reasonably exceeds what might be observed from the underlying noise in the data. For Fisher, the *P* value reflected the degree to which the observed data were incompatible with the hypothetical null hypothesis. If the calculated *P* value was less than .05 (eg, effect size estimate ≥2 precision units or standard errors away from the position of the null hypothesis), Fisher proposed that either (1) a rare event had occurred (ie, the appearance of a seemingly meaningful pattern or causal relationship that was actually due to purposeless variation in the data) or (2) the null hypothesis (no effect and no difference) was false; in other words, the data were (probably) showing a real meaningful pattern or causal relationship. Therefore, the Fisher *P* value is an "unexpectedness" test in the sense that a small *P* value is an unexpected outcome if the null hypothesis is correct (**Table 1**).

Fisher did not anticipate or endorse the enshrinement of the $P < .05$ criterion, which he had proposed informally in early writings as an example rather than a standard. He also did not need or postulate the idea of large numbers of identical (hypothetical) repetitions of the experiment to interpret the *P* value.[16,18] In fact, he did not appear to give much consideration to interpretation of the actual *P* value number.[19] Aside from recommending multiple repetitions of each experiment, Fisher bypassed the ambiguity problem in statistics by assuming that the sample data were randomly drawn from a hypothetical infinite population.[19]

Table 1. Comparison of Tests by Fisher and Neyman-Pearson

Fisher "Unexpectedness" Test (*P* Value Explained in 6 Steps)	Neyman-Pearson Hypothesis Test ("Fluke Detection" in 7 Steps)
Formulate relevant null hypothesis to be used as a straw man (eg, no treatment effect, hazard ratio of 1).	Formulate straw man null hypothesis plus alternate hypothesis (usually anything different from the null).
Calculate the test statistic, a measure of the distance between the observed experimental results and the postulated null hypothesis results (roughly the "signal" of interest) adjusted for the amount of variation or uncertainty (the "noise" in the data) in the experimental results.	Choose desired α level (probability of false-positive result in testing the null hypothesis). This is not the *P* value!
Use an appropriate mathematical or statistical model for interpreting the test statistic (eg, normal distribution).	Use an appropriate mathematical or statistical model for interpreting the test statistic (eg, normal distribution).
Use the statistical model and experimental test statistic to gauge how unusual or unexpected the experimental results would be if the null hypothesis was true, typically expressed as *P* value (also called significance test but perhaps clearer if renamed unexpectedness test).	Use the statistical model (including Neyman-Pearson calipers set at fixed detection width based on α level chosen) and calculated test statistic to test the null hypothesis based on the distance the observed experimental results are from the hypothetical null (signal), adjusting for data variability (noise).
Assuming the null hypothesis was true (the straw man), recognize that the experimental results (test statistics) that correspond with small *P* values signify either something unexpected has happened or the null hypothesis is not true or not supported by the data.	If within fixed Neyman-Pearson data calipers (defining the null hypothesis), do not reject the null hypothesis.
Repeat the experiment multiple times to decide which of those 2 options to believe.	If the experimental results are outside (beyond) fixed Neyman-Pearson data calipers, reject the null hypothesis.
	Imagine repeating the identical (or almost identical) experiment a large number of times (possibly approaching infinite repetitions).

Table 2. Common Misconceptions About *P* Value

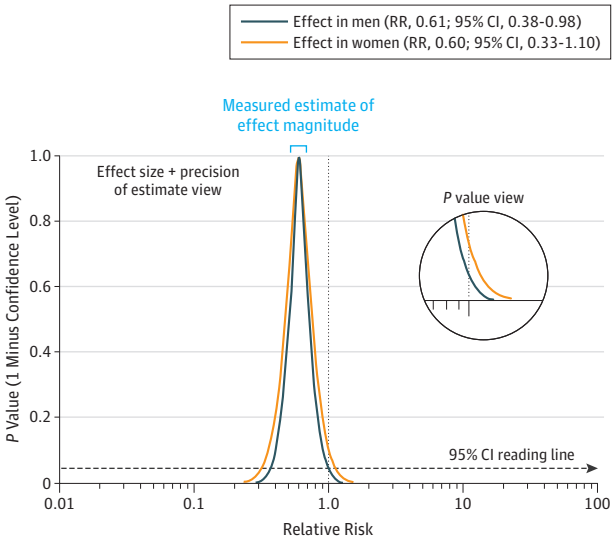
Misconception	Comment
P value equals the probability that the null hypothesis is true.	*P* value is computed by assuming the null hypothesis is true.
P value equals the probability that the observed effect is due to "the play of chance."	*P* value is defined as the probability of a difference (effect) as large as that observed or larger if the null hypothesis is true. Even if the difference observed is consistent with a simple chance mechanism, other more complex explanations are also possible, and nothing in *P* value calculation allows one to conclude that this is the best or most likely explanation for the observed differences.
P value ≤.05 means the null hypothesis is false. *P* value >.05 means the null hypothesis is true.	*P* value is computed assuming the null hypothesis is true. It is not the probability that the null hypothesis is either true or false.
P value ≤.05 identifies a clinically or scientifically important difference (effect). *P* value >.05 rules out a clinically or scientifically important difference (effect).	Clinical or scientific importance of study results is a judgment integrating multiple elements, including effect size (expected and observed), precision of estimate of effect size, and knowledge of prior relevant research. At best, *P* value has a minor role in shaping this judgment.
A small *P* value indicates study results are reliable and likely to replicate.	*P* value provides no information about whether a given study result can be reproduced in a second, replication experiment. There are many other factors that must be considered in judging the reliability of study results. Understanding what works in medicine is a process and not the product of any single experiment.

However, Fisher envisioned that experiments would be repeated until the investigator was reasonably sure that he or she had learned how to use the experimental intervention to get a predictable and desired result. One or 2 experiments usually could not settle any complex scientific question. This insight is as relevant to modern clinical trials as it was to Fisher's small agricultural experiments. In some areas of cardiovascular medicine, such as secondary prevention trials with β-blockers, angiotensin-converting enzyme inhibitors, antiplatelet agents, and statins, the ability to perform multiple, reasonably similar trials has been pivotal in providing the consistent evidence necessary to achieve widespread clinical acceptance and to support incorporation into clinical guidelines. However, even repeated huge modern clinical trials sometimes do not provide everything we need to know to understand a therapy. One of the best examples involves the thrombolytic therapy megatrials that laid the foundations of modern reperfusion therapy for acute myocardial infarction.[20,21] Two independent trials (Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico [GISSI 2][22] with 12 490 patients and the Third International Study of Infarct Survival [ISIS 3][20] with 41 299 patients) failed to find any of the expected clinical advantages suggested by earlier mechanistic investigations for tissue plasminogen activator over streptokinase.[20] While many considered the matter settled based on this substantial body of trial evidence, the unwillingness of some others to dismiss the cognitive dissonance produced by this result led to a third megatrial (Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries [GUSTO][21] with 41 021 patients), which by modifying some aspects of the tissue plasminogen activator therapy from the earlier trials demonstrated the precise mortality benefit of tissue plasminogen activator over streptokinase predicted at the outset of the trial.[23]

The *P* value has inspired an impressive amount of passionate editorializing in the scientific literature.[5,24-35] Much of the negativity expressed about *P* values seems due to 2 problems. First, the *P* value has nothing to say about the issue of how different the 2 cohorts are (the treatment oomph in the data).[10,36] Fisher likely would have argued that it is the responsibility of the investigator to make that assessment and that it is outside the intended scope of significance testing. Second, the *P* value is sample size sensitive.[37,38] With a large enough sample, any nonzero treatment difference will yield a "significant" *P* value.[39] Conversely, with small samples, only very large treatment effects are likely to generate significant *P* values.

If the *P* value has such important limitations, why is it still so incredibly popular with researchers? We believe that several factors help explain this phenomenon. First, the *P* value appears to offer a simple, objective triage tool: $P < .05$ means "pay attention," while $P \geq .05$ is "safe to ignore."[40] Second, as this triage use of the *P* value illustrates, misinterpretation and misapplication of the *P* value are common (**Table 2**).[5,32] The most frequently offered statistical textbook definition is not much help, namely, that the *P* value is the probability of observing a difference (effect size) as large as was measured or larger under the assumption that the null hypothesis (no difference) is actually correct. Viewing the data from the perspective of *P* values is a form of statistical tunnel vision (**Figure 3**), focusing on the tail (of the test statistic distribution where it crosses the

Figure 3. Comparison of the "*P* Value View" of Data With the "Effect Size or Precision View"



The results are shown from a hypothetical study comparing the effect of a treatment in men vs women. The y-axis shows the full range of possible *P* values (0 to 1.0), and the x-axis shows relative risk (RR) (values <1 indicate better outcomes with treatment relative to control). This type of display is called a *P* value function (and can also be shown as a confidence interval function by changing the y-axis to 1 minus the *P* value). It is explained in greater detail in the *Epidemiology* textbook by Rothman.[41] Typically, we are most interested in the *P* values calculated against the null hypothesis (in this case represented by an RR of 1, shown as a solid vertical line). Three key concepts are shown. First, the *P* value view of the data, shown in the inset on the right, focuses completely on where the tails of the *P* value functions cross the null position (RR of 1). The *P* value view does not include any direct assessment of the size of the treatment effect produced. Second, even within the 95% CIs, the possible values of RR are not all equally likely. The values most compatible with the data collected are those at or close to the point estimates of effect size. Third, the narrowness or wideness of the *P* value function, reflecting precision of the estimate of effect size, can perversely affect interpretation using *P* values. In this example, it is clear that the effect sizes are essentially identical, but there is less precision in the estimate for the women, leading to a nonsignificant result. A *P* value–centric interpretation might lead to the misguided conclusion that the therapy works in men but not in women. The figure was generated using a program created by Kenneth Rothman, DMD, DrPH (http://www.krothman.org/episheet.xls) and used with his permission.

null point) and ignoring the dog (the effect size the experiment was conducted to measure).

The hazard of this practice is well illustrated by the Surgical Treatment for Ischemic Heart Failure (STICH) trial,[42] which randomized 1212 patients with an ejection fraction of 35% or less and coronary artery disease amenable to coronary artery bypass grafting (CABG) to CABG or medical therapy alone. With a median follow-up of 56 months, the CABG-medicine treatment effect on all-cause mortality (the primary end point) was a hazard ratio of 0.86 with *P* = .12, and the trial was interpreted as negative. After an additional 5 years of follow-up, the CABG effect size for all-cause mortality was unchanged (0.84 vs 0.86); however, the precision of the estimate was improved (95% CI, 0.73-0.97 vs 0.72-1.04), and the *P* value was now significant (*P* = .02), a "positive trial."[43] All that really changed between the first (negative trial) report and the second (positive trial) was the position of the upper tail of the test statistic relative to the null position. Such overreliance on *P* values (the tails) relative to the oomph effect of treatment (the dog), is common[4] but makes no sense.

The Hypothesis Test

Jerzy Neyman, a Polish mathematician and founder of the American school of mathematical statistics, working in collaboration with the British statistician Egon Pearson, proposed "hypothesis testing" to "improve" some of the mathematically fuzzy parts of the work by Fisher, particularly its informal (heuristic) interpretation.[44] Neyman and Pearson argued that the null hypothesis required an alternative hypothesis, and together they could be used to define the now familiar type I and type II errors that may occur in testing a hypothesis.[16] It is important to note that *hypothesis* here refers to a statistical, not a scientific, one. Fisher believed that the *P* value or significance test was most useful to help generalize from experiments to the world outside his experiments (the process of inductive inference). Neyman explicitly rejected that idea. Instead, he proposed that the hypothesis test could serve as a guide to what he called "inductive behavior," which in the context of an experiment was to either accept the null hypothesis or the alternative hypothesis (essentially anything that was not the null hypothesis). Neyman did not care what the scientist thought about the evidence and instead believed that the experiment should tell the scientist what to do, an idea Fisher rejected as "unscientific." The decision by Neyman to use Fisher's *P* < .05 heuristic for the default type I error rate (α level) has probably led to much added confusion.[18]

To explain what the hypothesis test and the associated error rates meant, Neyman postulated that an experiment would undergo a large series of identical or almost identical hypothetical repetitions (*large* is undefined here but may approach infinity).[9,19] This device (imagining the results of a "long run" of hypothetical experimental repetitions) is not controversial in mathematics but has caused major headaches for scientists, who have great difficulty understanding what it actually means in empirical terms.[45] Neyman and Pearson never actually specified what constituted an experimental repetition in this context.[46] Two key factors spring from the hypothetical long-run repetitions concept. First, nothing could be concluded about whether the experiment the investigator actually did was correct or not.[18] The hypothesis test procedure had nothing to say about that or about the evidence in the data set. Second, the hypothesis testing procedure was meant to ensure that, in the long run, the probability (frequency) of errors would be controlled at an acceptable level. That was all that was possible, Neyman thought. Yet, the long run was all hypothetical. Goodman[47] has likened this formulation to a justice system that does not care about the correctness of verdicts (guilty or innocent) for individual defendants and instead is completely focused on controlling the long-run proportion of mistaken verdicts.

To understand Neyman's notion of the type I (false-positive) error rate, one can imagine a set of data calipers, like an electrocardiogram caliper, designed to measure the distance the observed data have been pushed away from the hypothetical null hypothesis data position by the therapy under study (which so far is similar to Fisher's approach [Table 1]). However, Fisher's data calipers were fully adjustable, capable of measuring distances corresponding to *P* values between 0 and 1, whereas Neyman conceived of calipers that

were preset in the design phase of the research to the desired width, which turned out to be almost universally at an α level (false-positive rate) of .05. A hypothesis test, then, uses this fixed-width caliper to judge the distance between the observed treatment effect and the null hypothesis (ie, where the results are vs where they would be expected to be if the therapy had no effect). If the treatment effect size is sufficiently far away—outside the prongs of the caliper—one can reject the null hypothesis; if not, one cannot reject the null hypothesis. Therefore, from a hypothesis testing point of view, $P \leq .04$ means reject the null hypothesis, whereas $P \geq .06$ means do not reject the null hypothesis. This seemingly arbitrary inflexibility of the hypothesis test tool, and specifically its inability to appreciate degrees of incompatibility with the null hypothesis, has caused much confusion and led some to view the hypothesis test as a sort of statistical fluke detector (Table 1).

Because of the strong influence of the Neyman-Pearson hypothesis testing concepts in science, clinicians have come to accept, often without objection, the idea that if one performs 20 significance or hypothesis tests with the nominal threshold for "significance" of .05, then on average one can expect at least 1 to be a false positive (with no way to tell which one is the bad apple). When applied abstractly to games of chance, such concepts—which are based on simple probability mathematics and assumptions about independent equally likely outcomes—may be a reasonable approximation. Clinical experiments are vastly more complex than games of chance, however, and do not conform to the assumptions required for those false-positive error rate calculations. Doing many statistical tests undoubtedly raises the likelihood of 1 or more "false-positive" results. However, the probability of that occurrence depends on a host of factors and is almost never uniform across the tests performed (thus violating a key assumption of the 1 in 20 error rate rule). Therefore, invoking the play of chance as a cause for some unexpected outcome in clinical research is never the best first explanation to consider, although it may be what one is left with when all reasonable, and some unreasonable, possibilities have been excluded. That said, if statistical tests are used as a kind of data beachcombing tool unguided by clear (and ideally prospective) specification of what findings are expected and why, much that is nonsense will be "discovered" and added to the peer-reviewed literature. Concern about such "data dredging" has promoted requirements for prespecification as a part of the triage use of *P* values referred to earlier. However, prespecification without a strong, plausible underlying rationale is simply guessing and does not enhance the validity or credibility of findings.

The Modern Hybrid Null Hypothesis Significance Testing Procedure

In recent years, some researchers have begun to use both the Fisher *P* value or significance test and the Neyman-Pearson hypothesis test together, despite the very different intentions of the originators, as discussed above.[48] The most common hybrid procedure is to use the Neyman-Pearson structure to help design a clinical trial (including setting the desired type I error rate and rules for interim data and safety monitoring board looks at the outcome data, estimating the needed sample size and the expected power) but then to use the Fisher *P* value in the analysis phase to test the null hypothesis.[18,49] The main problem this hybrid approach creates is that it encourages the unwary to conflate the idea of the α level or type I error (conventionally set at .05) with the observed *P* value (which is almost always interpreted using a threshold benchmark of .05 to denote significance). Viewing the *P* value as a roving type I error probability for each completed study amplifies its apparent importance and spawns misguided arguments about the need to adjust the observed *P* value for "multiplicity" or "multiple comparisons," the total number of statistical tests performed on the data.[50,51] Once it is understood that the *P* value is not the probability that the researchers have made a type I error, then adjusting the *P* value for the number of significance tests performed loses its rationale.[52]

Conclusions

In this article, we have presented a view of clinical research centered on measuring and understanding something of clinical or scientific importance, the treatment oomph effect of interest. Figuring out how best to conceptualize and measure clinical "effects," as well as finding the most insightful ways to analyze and interpret the data using modern statistical tools, is difficult, complex, and often messy. The primary goal of the research process is not to generate *P* values or perform hypothesis tests. Those tools, like the many others in the statistical toolbox, can be helpful but must be applied thoughtfully and with full appreciation for their assumptions and limitations. Unfortunately, no statistical tool or technique can guarantee access to the shortest path to the truth. Repeated experiments, as Fisher recognized years ago, may be the best way to tame much of the messiness. If that is not feasible, and it often is not, medicine can still accomplish much by making pragmatic, well-reasoned use of the evidence it does have.

ARTICLE INFORMATION

Accepted for Publication: July 22, 2016.

Published Online: October 12, 2016. doi:10.1001/jamacardio.2016.3312

Author Contributions: Dr Mark had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: All authors.
Drafting of the manuscript: Mark.
Critical revision of the manuscript for important intellectual content: All authors.
Administrative, technical, or material support: Mark, Lee.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.

Funding/Support: This publication was supported in part by award UL1TR000445 from the Clinical and Translational Science Awards Program, National Center for Advancing Translational Sciences (Dr Harrell).

Role of the Funder/Sponsor: The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The contents of this publication are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

Additional Contributions: Melanie R. Daniels, BA (Howl, Inc, Durham, North Carolina), provided research support and critical review of the manuscript. Anastasios Tsiatis, PhD (North Carolina

State University), Kevin Anstrom, PhD (Duke University), Kevin Weinfurt, PhD (Duke University), Howard Rockman, MD (Duke University), and Abhina Sharma, MD (Duke University), read and commented on early versions of the manuscript. None received compensation.

REFERENCES

1. Kyriacou DN. The enduring evolution of the *P* value. *JAMA*. 2016;315(11):1113-1115.
2. Cohen HW. *P* values: use and misuse in medical literature. *Am J Hypertens*. 2011;24(1):18-23.
3. Goodman S. A dirty dozen: twelve *P*-value misconceptions. *Semin Hematol*. 2008;45(3):135-140.
4. 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.
5. Wasserstein RL, Lazar NA. The ASA's statement on *P*-values: context, process, and purpose. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 8, 2016.
6. Willink R. *Measurement Uncertainty and Probability*. New York, NY: Cambridge University Press; 2013.
7. Kline M. *Mathematics for the Nonmathematician*. Mineola, NY: Dover Publications; 1985.
8. Mandel J. *The Statistical Analysis of Experimental Data*. Mineola, NY: Dover Publications; 1984.
9. Salsburg DS. *The Use of Restricted Significance Tests in Clinical Trials*. New York, NY: Springer-Verlag; 1992.
10. Ziliak ST, McCloskey DN. *The Cult of Statistical Significance*. Ann Arbor: University of Michigan Press; 2008.
11. Pitt B, Pfeffer MA, Assmann SF, et al; TOPCAT Investigators. Spironolactone for heart failure with preserved ejection fraction. *N Engl J Med*. 2014;370(15):1383-1392.
12. Pfeffer MA, Claggett B, Assmann SF, et al. Regional variation in patients and outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial. *Circulation*. 2015;131(1):34-42.
13. Weisberg HI. *Willful Ignorance: The Mismeasure of Uncertainty*. Hoboken, NJ: John Wiley & Sons; 2014.
14. Box GEP, Hunter JS. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2005.
15. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 3rd ed. Hamilton, ON: BC Decker; 2008.
16. Salsburg D. *The Lady Tasting Tea*. New York, NY: WH Freeman and Co; 2001.
17. Fisher RA. *Statistical Methods and Scientific Inference*. 3rd ed. New York, NY: Hafner Press; 1973.
18. Hubbard R, Bayarri MJ. Confusion over measures of evidence (*P*'s) vs errors (*α*'s) in classical testing. *Am Stat*. 2003;57(3):171-178.
19. Lehmann EL. *Fisher, Neyman and the Creation of Classical Statistics*. Norwell, MA: Springer; 2011.
20. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41,299 cases of suspected acute myocardial infarction. *Lancet*. 1992;339(8796):753-770.
21. GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med*. 1993;329(10):673-682.
22. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico. GISSI-2: a factorial randomised trial of alteplase versus streptokinase and heparin versus no heparin among 12,490 patients with acute myocardial infarction. *Lancet*. 1990;336(8707):65-71.
23. Gersh BJ, Anderson JL. Thrombolysis and myocardial salvage: results of clinical trials and the animal paradigm: paradoxic or predictable? *Circulation*. 1993;88(1):296-306.
24. Leek JT, Peng RD. Statistics: *P* values are just the tip of the iceberg. *Nature*. 2015;520(7549):612.
25. Abelson RP. On the surprising longevity of flogged horses: why there is a case for the significance test. *Psychol Sci*. 1997;8(1):12-15.
26. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psych*. 2015;37(1):1-2.
27. Senn S. Two cheers for *P*-values? *J Epidemiol Biostat*. 2001;6(2):193-204.
28. Benjamini Y. It's not the *P*-values' fault. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
29. Berry DA. *P*-values are not what they're cracked up to be. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
30. Gelman A. The problems with *P*-values are not just with *P*-values. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
31. Greenland S. The ASA guidelines and null bias in current teaching and practice. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
32. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P*-values, confidence intervals, and power: a guide to misinterpretations. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
33. Rothman KJ. Disengaging from statistical significance. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
34. Senn SJ. Are *P*-values the problem? 2016. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
35. Stark PB. The value of *P*-values. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108. Published 2016. Accessed April 9, 2016.
36. Abelson RP. *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1995.
37. Royall RM. The effect of sample size on the meaning of significance tests. *Am Stat*. 1986;40(4):313-315.
38. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *Am Stat*. 1966;20(2):18-23.
39. Lin M, Lucas HC, Shmueli G. Too big to fail: large samples and the *P*-value problem. *Inf Syst Res*. 2013;24(4):906-917.
40. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Sci*. 1988;76:159-165.
41. Rothman KJ. *Epidemiology: An Introduction*. New York, NY: Oxford University Press; 2002.
42. Velazquez EJ, Lee KL, Deja MA, et al; STICH Investigators. Coronary-artery bypass surgery in patients with left ventricular dysfunction. *N Engl J Med*. 2011;364(17):1607-1616.
43. Velazquez EJ, Lee KL, Jones RH, et al; STICHES Investigators. Coronary-artery bypass surgery in patients with ischemic cardiomyopathy. *N Engl J Med*. 2016;374(16):1511-1520.
44. Pearson ES. Some thoughts on statistical inference. *Ann Math Stat*. 1962;33(2):394-403.
45. Jaynes ET. *Probability Theory: The Logic of Science*. New York, NY: Cambridge University Press; 2003.
46. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc*. 1993;88(424):1242-1249.
47. Goodman SN. Toward evidence-based medical statistics, 1: the *P* value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
48. Royall RM. *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman & Hall; 1997.
49. Berger JO. Could Fisher, Jeffreys, and Neyman have agreed on testing? *Stat Sci*. 2003;18(1):1-32.
50. Goodman SN. *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol*. 1993;137(5):485-496.
51. Hubbard R. Blurring the distinctions between *P*'s and *α*'s in psychological research. *Theory Psychol*. 2004;14(3):295-327.
52. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43-46.