Editorial

Misinterpretation of *P* Values and Statistical Power Creates a False Sense of Certainty: Statistical Significance, Lack of Significance, and the Uncertainty Challenge

Abstract: Despite great advances in our understanding of statistics, a focus on statistical significance and *P* values, or lack of significance and power, persists. Unfortunately, this dichotomizes research findings comparing differences between groups or treatments as either significant or not significant. This creates a false and incorrect sense of certainty. Statistics provide us a measure of the degree of uncertainty or random error in our data. To improve the way in which we communicate and understand our results, we must include in reporting a probability, or estimate, of our degree of certainty (or uncertainty). This will allow us to better determine the risks and benefits of a treatment or intervention. Approaches that allow us to estimate, account for, and report our degree of uncertainty include use of confidence intervals, *P*-value functions, and Bayesian inference (which incorporates prior knowledge in our analysis of new research data). Surprise values (S values, which convert *P* values to the number of successive identical results of flips of a fair coin) express outcomes in an intuitive manner less susceptible to dichotomizing results as significant or not significant. In the future, researchers may report *P* values (if they wish) but could go further and provide a confidence interval, draw a *P*-value function graph, or run a Bayesian analysis. Authors could calculate and report an S value. It is insufficient to mindlessly report results as significant versus not significant without providing a quantitative estimate of the uncertainty of the data.

"We're talking about practice, man. I mean how silly is that? We're not even talking about the game, the actual game, when it matters. We're talking about practice."—Allen Iverson, May 7, 2002[1]

Statistical significance dichotomizes research findings into significant versus not significant creating a false sense of certainty. In an epic press conference, 4 days after his Philadelphia 76ers fell to the Boston Celtics in the first round of the 2002 National Basketball Association playoffs, all-star Allen Iverson expressed amusement and bewilderment at questions regarding practice.[1] After all, there was so much more to talk about and consider when it came to his future with the team, yet there he was answering questions about practice.

The same can be said for statistical significance. Since the concept was first introduced, there's been a near century's worth of advancements in statistics and yet this method which, as above, dichotomizes research findings into significant versus not significant, remains the focus.

Near the front of this months' issue of *Arthroscopy*, readers will find the original scientific article, "The Potential Effect of Lowering the Threshold of Statistical Significance From $P < .05$ to $P < .005$ in Orthopaedic Sports Medicine" by Evans, Johnson, Anderson, Checketts, Scott, Middlemist, Fishbeck, and Vassar of the Oklahoma State University.[2] Evans et al.[2] report that some respected scholars recommend "redefining statistical significance by changing the *P* value threshold from .05 to .005" to lower the risk of that medical research studies reach "false-positive" conclusions, and their results show that lowering the *P* value threshold would dramatically reduce the number of statistically significant findings in randomized controlled trials. This is clinically relevant because if the statistical significance threshold were thus changed, evidence-based recommendations used to guide clinical decision-making would be greatly affected.

As editors, we feel that it is important to both "compliment" the effort of Evans et al. and "complement" their effort by pointing out a fundamental and additional fact: regardless of the threshold of statistical significance, bias in a study's design and conduct also needs to be considered as a potential source error. Whether the *P* value is .5, .05, .005, .0005, etc. we need to be mindful that bias is lurking. While their capable investigation of the literature brings out important points, we can't help but think of Allen Iverson. There's so much to more to consider when it comes to data, and yet we're talking about *P* values.

0749-8063/21202/$36.00
https://doi.org/10.1016/j.arthro.2021.02.010

When we finally embrace rationale statistical reasoning, we wonder if we'll look back on our obsession with *P* values and the misinterpretation that often accompanies them and ask ourselves, "What took so long?" Here, we briefly describe the genesis of null hypothesis significance testing and the potential consequences that become of it. We draw on examples from those who have previously espoused similar concerns, each taking creative approaches in an effort to stamp out the same problem—a false sense of certainty. We outline potential solutions that are easily applied to enhance sensible interpretations of data. Finally, we offer a challenge for future authors.

Random Error

Consider posterior tibial slope. There's a true average value for posterior tibial slope in the population of patients with anterior cruciate ligament (ACL) tears. The degree to which the average posterior slope in a sample of 100 patients undergoing ACL reconstruction deviates from the true population average represents random error. Of course, quantifying posterior slope measurements in all patients with ACL tears at a given moment in time is not possible, but the concept holds. For every sample measurement we make, there's a certain degree of uncertainty in terms of the degree to which the sample measurement deviates from the mean. Statistics helps us deal with uncertainty by providing tools for estimating how random error distorts measurements. The results of these analyses serve as a means for communicating the degree of uncertainty in our data.

At the turn of the twentieth century, the relationship between experimental design and statistics was of great interest. Sir Ronald A. Fisher, the statistician to whom the concept of statistical significance is most often attributed, approached this problem of uncertainty by considering the probability of observing the data at hand under some hypothesis to be nullified.[3] That probability is referred to as the level of significance and is given by the *P* value. High *P* values suggest departures from the null hypothesis that could be easily attributed to random error or chance. Small *P* values warrant close examination, as chance alone is unlikely to have produced the results. In Fisher's world, the level of significance and the null hypothesis are contextual to the subject under study, and the *P* value is the language we use to communicate the significance of the results.

Around the same time, a competing approach motivated by industrial control problems was introduced by statisticians Jerzy Neyman and Egon Pearson.[4] In the Neyman—Pearson set-up, there's a null and an alternative hypothesis as well as costs and benefits of proceeding according to each. This is often explained with an example from industry.[5,6] Say you're in the business of manufacturing nails. The ideal nail, which you make by the millions, is labeled as being 1 cm in length. Knowing that each nail will not be exactly 1 cm, you deem a standard deviation of ±0.2 cm to be acceptable. In a quality check of thousands of nails, if the average length is 1.1 cm with a standard deviation of ±0.3 cm, what can we conclude about the entire batch of a million nails? We may decide that this is evidence that most nails are faulty in length and go on to destroy the entire batch. Alternatively, we may decide that this is not sufficient evidence to determine the batch is faulty and decide to package it send it out. What do we do? This is the heart of the Neyman—Pearson approach. How often we want to wrongly destroy an otherwise-good batch (because some nails are significantly too long due to random error) is the Type I error rate or alpha level, and how often we want to wrongly assume the batch is good (no significant difference in nail lengths when in truth the batch is faulty) is the Type II error rate or beta level. The concern is not whether the hypotheses are true but rather the costs and benefits of proceeding according to each.

Much of what followed was a heated debate between the 2 camps. Unlike today, barbs were exchanged in papers and book chapters over a period of a years rather than in daily Tweets and social media posts. Amidst the grumbling, researchers begin to use a hybrid approach that borrowed elements from each to form what we know today as null hypothesis significance testing.[5] Generally speaking, it goes as follows: A null hypothesis of no difference is exclusively assumed. Data are collected and a *P* value calculated. If the *P* value is less than .05, the results, whatever they may be, are considered statistically significant and are often treated as being certain. If the *P* value is .05 or greater, then the results, whatever they may be, are discarded and the null hypothesis of no difference is considered true. However, in both cases, there is uncertainty, and the conclusions are, statistically speaking, estimates. To reiterate: the conclusions are, statistically speaking, estimates.

Ironically, this approach ignores much of what Fisher and Neyman—Pearson were advocating. Neither believed in a mechanical use of statistics that null hypothesis significance testing embodies.[5,7] Although Fisher stated that it's convenient for researchers to take 5% as the standard level of significance,[8] he also went on to say that the habitual use of 5% indicated a lack of statistical sophistication.[9] Fisher also objected to the null always being defined as zero or no difference, stating that the null is any hypothesis to be nullified, and he pointed out that the former approach was a primitive form of analysis to be used when only very little information about a problem exists.[4]

It's important to consider the time in which the Fisher and Neyman—Pearson approaches were developed. It was the early-mid twentieth century, and science and statistics were coevolving.[7] It's clear from both approaches that accounting for uncertainty was, and is, an evolving practice. Since their time, statistical thinking

has continued to evolve. Measures like confidence intervals and *P*-value functions are available that provide information on how precisely an observed difference or relationship has been estimated. Methods that were obscured by the great hypothesis debate are now used more regularly. Most notably, Bayesian inference, which incorporates pre-existing knowledge and beliefs about the world into the analysis of data, is used in many areas of medical research. Yet for many studies, we continue to resort to a method of analysis with which even those who are credited with its development appear to have disagreed. As a result, great deal of medical research insists on describing results and conclusions in most basic and misleading form: significant versus not significant.

Consider the data in Table 1 as having come from 2 hypothetical studies on anatomic failure following rotator cuff repair.

Table 1. Data From 2 Studies Comparing the Rates of Rotator Cuff Repair Retear (Torn as Compared With Intact) in Smokers Versus Nonsmokers Including Odds Ratios and *P* Values

	Torn	Intact	OR	*P* Value
Study A				
Smoker	17 (31%)	38 (69%)	2.14	.063
Nonsmoker	14 (17%)	67 (83%)		
Study B				
Smoker	19 (32%)	40 (68%)	2.38	.027
Nonsmoker	15 (17%)	75 (83%)		

In Study A, 31% of smokers have failed their repair compared to 17% of nonsmokers. This works out to an odds ratio (OR) of 2.14 (OR that smoking is associated with retear). However, the corresponding *P* value is .063, and so we fail to reject the null hypothesis and conclude that smoking makes no difference. In Study B, the rates are similar leading to an OR of 2.38; however, thanks to a few more patients, the *P* value is .027 and thus we reject the null hypothesis and thus conclude that smoking is associated with retear. Yet, whether the null hypothesis is rejected or not, we don't seem to be any closer to quantifying the effect of smoking on repair failure. In Study A, we're stuck having to ignore what appears to be a meaningful finding (OR = 2.14). In Study B, we have rejected the null hypothesis (of zero effect of smoking), yet this doesn't eliminate the uncertainty that's present in the measurements, and so we are again stuck and possibly mislead to believe that the OR of 2.38 is a measure of certainty free from random error—which it is not.

This is the central problem of the null hypothesis approach. It only addresses the probability of the results given the null hypothesis, and probability does not mean certainty. Worse, achieving statistical significance suggests random error has suddenly been ruled out—which it never is. To understand the uncertainty in our data, we need to shift our focus to methods that estimate random error rather than methods rooted in dichotomous decision making regarding statistical significance.[10]

Confidence Intervals

The data in Table 2 is the same as Table 1 but with the addition of 95% confidence intervals.

Table 2. Data From 2 Studies Comparing the Rates of Rotator Cuff Repair Retear (Torn as Compared With Intact) in Smokers Versus Nonsmokers Including Odds Ratios and *P* Values, and 95% Confidence Intervals

	Torn	Intact	OR	*P* Value	95% Confidence Interval
Smoker A					
Smoker	17 (31%)	38 (69%)	2.14	.063	(0.95-4.82)
Nonsmoker	14 (17%)	67 (83%)			
Study B					
Smoker	19 (32%)	40 (68%)	2.38	.027	(1.09-5.17)
Nonsmoker	15 (17%)	75 (83%)			

In Study A, the estimated risk (OR) of repair failure among smokers is 2.14 and the 95% confidence interval covers ORs between 0.95 and 4.82. It's important to note that the 95% confidence is in reference to the real risk of smoking on retear falling somewhere within the interval. Specifically, if this hypothetical study were repeated an infinite number of times with 95% confidence intervals being calculated for each repetition, 95% of those intervals would contain the true OR. In this regard, we are 95% confident that we have an interval that contains the true OR. The utility of the interval lies in examining its width.[11] Despite the lack of statistical significance, the vast majority of the intervals in Study A are greater than 1, providing evidence of the deleterious effects of smoking on healing. The interval extends up to almost 5, indicating that the exact OR (that smoking *is* associated with retear) has not been estimated with great precision. Interestingly, the interval for Study B is quite similar, with the 95% confidence interval in an OR between 1.09 and 5.17. Here, the entire interval rests on the side of risk for smoking (OR greater than 1), which is why the *P* value is less than .05. Like Study A, the interval is wide, with ORs as small as 1.09 and as large as 5.17. Smoking seems to matter, and more data are needed to improve the precision of our estimates. Confidence intervals are a step toward improving how we communicate our results.

P-Value Function

In 1987 epidemiologist Charles Poole suggested we go beyond confidence intervals and consider the *P*-value function.[12] In this function, *P* values for several different effects can be considered rather than a single *P* value that relates to the null hypothesis. This graph can be

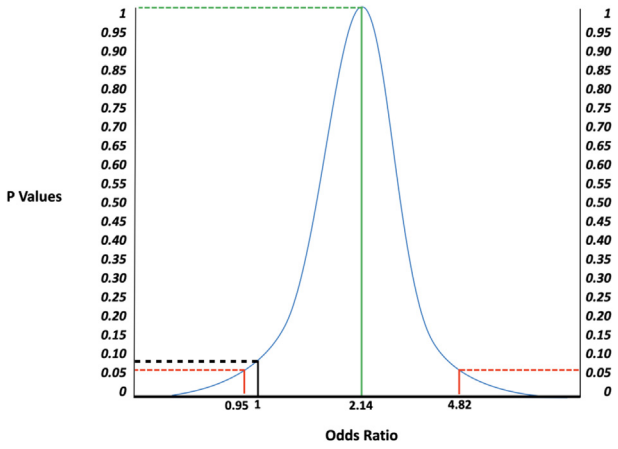

Fig 1. *P*-value function graph for a study comparing the rates of rotator cuff repair retear (torn compared with intact) in smokers versus nonsmokers. The *P* value is .063, which corresponds to an OR of 1, and at the *P* value of .05 (as noted on the dual y-axes) are limits of the 95% confidence interval (0.95 and 4.82). The values on the x-axis in between the 95% confidence limits can be considered the ORs that would not have been rejected had they served as the null hypothesis ($P > .05$). (OR, odds ratio.)

roughly sketched using the upper and lower limits of the confidence interval or generated using freely available spreadsheets.[13] Figure 1 presents the *P* value function for Study A. For ease of interpretation dual y-axes are included to represent *P* values.

In Study A, $P = .063$, which corresponds to an OR of 1, i.e., the null hypothesis. Recall, the *P* value in null hypothesis testing represents the probability of the data given the null hypothesis. At the *P* value of .05 on the y-axis, you'll find the limits of the confidence interval (0.95 and 4.82). As you move along the x-axis considering different ORs, the corresponding y-axis values provide the *P* values for these hypothesized values. No longer bound to consider the probability of the data given the null, the compatibility of data given a selected OR can be determined. Figure 2 illustrates the utility of the *P* value function when considering multiple estimates. The curves for Study A (red) and B (yellow) are near identical despite the fact that Study A, where the 95% confidence interval includes values less than 1 (Fig 1), failed to be statistically significant.

The S Value

The recently introduced S value or "Surprise Index" converts *P* values into a model of coin flipping.[14] It is defined as $-\log_2(P \text{ value})$ and is easily calculated in Excel or using an online logs calculator. Specifically, the S value converts the *P* value to the number of identical results of successive coin flips of a fair coin. While this may seem trivial, flipping a fair coin is a great demonstration of how something with presumably no underlying bias may produce results to the contrary of our expectations (e.g., the truth) on account of random error. As anyone who has flipped a coin knows, it's unlikely that heads and tails will simply alternate from flip to flip, and random outcomes will be observed despite the fact that the coin is fair. However, if the coin was gimmicked or unbalanced, i.e., biased, how many consecutive flips of heads or tails might it take before the fairness of the coin is called into question? The point at which we become alarmed, suspicious, or "surprised," the S value, provides an intuitive description.

If we consider the null hypothesis of no difference in retear rates between smokers and nonsmokers to be the same as an assumption of a fair coin, we can view the hypothetical results of studies A and B from the perspective of coin flips. The more successive coin flips are identical, the more evidence there is against the coin being fair, or in this case, the more evidence there is that the null hypothesis of no difference in retear rates in smokers is incorrect. In Study A, the OR is 2.14 and the *P* value is .063, which converts to a S value of 3.98 or roughly 4 identical flips. If there are truly no difference in retear rates in smokers, the result is as surprising as all heads in 4 flips of a presumably fair coin. In Study B, with an OR of 2.38 and a *P* value of .027, the S value is 5.2; in this case, a conclusion that smoking results in no difference in retear rates is as surprising as approximately 5 consecutive flips of a fair coin. While S values fall short when it comes to estimation, an intuitive description fosters a more rational discussion of the results.[15]

Bayesian Analyses

Up until now, the approaches we've discussed are considered "frequentists" methods. These are the classical approaches to uncertainty that consider the frequency with which an infinite number of trials result in the correct conclusion or truth. This is why the 95% confidence applies to the interval and not the OR itself. Frequentists are interested in the probability of the data

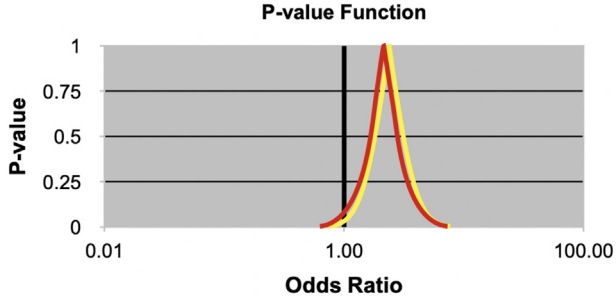

Fig 2. *P*-value functions for both Study A (red) and Study B (yellow). While it's evident that the extreme left portion of both graphs includes values <1, they are associated with small *P* values on the x-axis. Further, the vast majority of both curves reside to the right of 1, thereby providing evidence against the null hypothesis of no difference (odds ratio = 1).

in terms of the null hypothesis. In contrast, Bayesian approaches, which outdate frequentists by over 150 years, flip the problem around and start with what is already known (smoking is detrimental to healing), and seek to determine the probability of the prior hypothesis given the new data. This requires a "prior" be constructed, which represents the belief about the world prior to seeing data.[16]

To illustrate the utility of a Bayesian approach, we can analyze the data in Study A using a relatively noninformative prior, which simply means that we choose a starting point for the analysis that reflects that little is known about the problem at hand, and therefore, many possible results might be expected. The idea is to combine our belief about the world (the prior) with data from the study to arrive at an updated belief about the world, which in a Bayesian analysis is called the "posterior." The posterior is actually a distribution of values from which one can take the mean of the distribution and calculate a "credibility interval." For Study A, the mean of the posterior is an OR of 1.66 with a 95% credibility interval of 1 to 3.38. What's happened here is that the Study A data has updated our prior belief about the relationship between smoking and rotator cuff repair retear. The results indicate that given our prior belief, which in this case was to allow many possible values to reflect little was known about the problem, our updated beliefs are centered on an OR of 1.66. Further, there's a 95% probability that the true OR is between 1 and 3.38. Notice that unlike the confidence interval where the probability refers to the interval, the credibility interval reflects the probability that the parameter, in this case the OR, is in the interval.

Full presentation of Bayesian approaches is well beyond this commentary; however, the general approach warrants consideration. The ability to incorporate prior knowledge is seen by many as an advantage to the approach; however, the inherently subjective nature of this process has caused debate between so called frequentists and Bayesians. After all, 2 investigators could take the same data and, by stating different prior beliefs, produce different results. However, consider the alternative of not incorporating prior information. There are reams of biological data demonstrating the harmful effects of smoking. At the very least, we would not believe that smoking is protective regarding rotator cuff repair failure, and therefore, we wouldn't accept any value less than 1 for the OR regardless of whether it's in an interval or not. Furthermore, clinical studies have examined factors including smoking that are associated with anatomic failure. For example, the failure rates between smokers and nonsmokers reported by Neyton et al.[17] converts to an OR of 3.76, and O'Donnell et al.[18] adjusted for covariates to arrive at an OR of 1.38. We are not suggesting that these studies should be pooled, but their results give us a sense of what we might expect. When we take inventory of existing knowledge, it becomes clear that we are not information naive on smoking and repair failure, so neither should be our analysis of our data. Why start with the assumption of ignorance when we are information rich?

In This Issue

As noted previously, near the front of this issue of *Arthroscopy*, readers will find the original scientific article, "The Potential Effect of Lowering the Threshold of Statistical Significance From $P < .05$ to $P < .005$ in Orthopaedic Sports Medicine" by Evans et al.[2] In addition, near the end of this issue, readers will find the Level V evidence (expert opinion) article, "The Blight of the Type II Error: When No Difference Does Not Mean No Difference" by Domb and Sabetian of the American Hip Institute in Chicago.[19] As mentioned, Evans et al. focus on avoidance of falsely positive, albeit statistically significant, conclusions.[2] In contrast, Domb and Sabetian focus on reasons why studies could fail, errantly, to achieve statistically significant results. Domb and Sabetian introduce that "underpowered studies caused by small sample sizes are especially prevalent in the surgical literature," where power is defined as "the capacity of the study to recognize whether there is a difference (between treatment groups), given that such difference exists." Domb and Sabetian note that generally, in prospective medical research, statistical power analyses typically establish the requirement of a power of 0.8 (80%), which means we thus accept a risk of false-negative conclusions of up to 20%, and the danger of a "false-negative" conclusion (of no difference between two treatments when, if fact, a difference actually does exist) is that we "may impede incremental advances in our field, as the advantages of small improvements may go undetected."[19]

Once again, as editors, we feel that it is important to both "compliment" the effort of Domb and Sabetian and "complement" their effort by pointing out a fundamental and additional fact: regardless of the threshold of the statistical power target, bias in a study's design and conduct also needs to be considered as a potential source error. Whether the power is 0.8, 0.08, 0.008, 0.0008, etc., we need to be mindful, once again, that bias is lurking. While Domb and Sabetian bring out important points, we can't help but think once again of Allen Iverson. There's so much to more to consider when it comes to data, and yet we're still misleadingly dichotomizing study results as significant or not. In the end, readers of both Evans et al.[2] and Domb and Sabetian[19] could keep in mind that, moving forward, we need to avoid the rigid and dichotomous thought process of difference versus no difference in favor of a process that estimates uncertainty. When it comes to

medical research data, uncertainty, on some level, is omnipresent.

A Challenge for Future Authors

In December 2020, viewers of the night sky observed the planets Jupiter and Saturn moving ever closer to each other until the night of the 21st, when Jupiter eclipsed Saturn in what is known as the "Great Conjunction."[20,21] This conjunction occurs once every 20 years when the orbits of Jupiter, Saturn, and our home planet of Earth become aligned, and this one was particularly special. The last time Jupiter and Saturn were this close was in 1623, but they were near the sun and by nightfall they were no longer visible.[21] One has to go back almost 800 years ago to 1226 to find the last time Jupiter and Saturn were this close and could be seen from the ground.[20]

Perturbations in the motion of Jupiter and Saturn had long puzzled astronomers as far back as 1625.[22] In the late 18th century, French mathematician Pierre-Simon de LaPlace finally provided resolution to the disturbing observations of Jupiter and Saturn, allaying fears of Jupiter being swallowed by the Sun and Saturn migrating out of our solar system.[23] LaPlace's work on planetary motion lead him to consider how repeated measures of the same phenomena could lead to different results, each exhibiting some degree of errant deviation from the average. LaPlace connected the random errors to a bell-shaped curve and in doing so introduced the world to the normal distribution and the central limit theorem.

Perhaps it's fitting that as we settle into 2021, fresh off an epic conjunction between the 2 planets whose orbits played a pivotal role in the development of statistics that populate our literature, we revisit our approach to data analysis. We challenge future authors to move past statistical significance and use the many tools of statistics to communicate the uncertainty in the data. Go ahead and report *P* values but also provide a confidence interval, draw a *P*-value function graph, or calculate an S value. More importantly: interpret. Consider the values in the confidence interval or in the *P* value function graph and interpret how these may inform the results of the study. Have a look at Bayesian methods and consider the value of incorporating existing belief and understanding of the world into the analysis of the data. For the brave, take the leap and run a Bayesian analysis. Do anything other than mindlessly and simply reporting results as significant or not significant. We've all done it, but no longer should this be the standard. There's just so much more we could be talking about. Allen Iverson knew it, and so should we. What are we talking about? We're talking about *P* values and power. We're not even talking about the data, the actual data, and the different ways in which we can communicate our findings. We're talking about significant versus not, based on arbitrary thresholds, and we're ignoring the uncertainty. "How silly is that?"[1]

Mark P. Cote, P.T., D.P.T., M.S.C.T.R.
Associate Editor, Statistics
James H. Lubowitz, M.D.
Editor-in-Chief
Jefferson C. Brand, M.D.
Assistant Editor-in-Chief
Michael J. Rossi, M.D., M.S.
Assistant Editor-in-Chief

References

1. Swinton E. This day in sports history: Allen Iverson's practice rant. *Sports Illustrated* May 7, 2020.
2. Evans S, Johnson AL, Anderson JM, et al. The potential effect of lowering the threshold of statistical significance from $P < .05$ to $P < .005$ in orthopaedic sports medicine. *Arthroscopy* 2021;37:1068-1074.
3. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 1935.
4. Neyman J, Pearson ES. The testing of statistical hypotheses in relation to probabilities a priori. *Math Proc Cambridge Philos Soc* 1933;29:492-510.
5. Gigerenzer G. Mindless statistics. *J Socio-Econ* 2004;33:587-560.
6. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: A reassessment. *Front Hum Neurosci* 2017;11:390.
7. Kennedy-Shaffer L. Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize p values and significance testing. *Am Stat* 2019;73:82-90.
8. Fisher RA. *The design of experiments*. Edinburgh: Oliver & Boyd, 1935.
9. Fisher RA. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd, 1956.
10. Rothman KJ. Curbing type I and type II errors. *Eur J Epidemiol* 2010;25:223-224.
11. Harris JD, Brand JC, Cote MP, Faucett SC, Dhawan A. Research pearls: The significance of statistics and perils of pooling. Part 1: Clinical versus statistical significance. *Arthroscopy* 2017;33:1102-1112.
12. Poole C. Beyond the confidence interval. *Am J Public Health* 1987;77:195-199.
13. Rothman K. Episheet. Available at, http://www.krothman.org. Accessed February 6, 2021.
14. Cole SR, Edwards JK, Greenland S. Surprise! *Am J Epidemiol* 2020;190:191-193.
15. Rothman KJ. Taken by surprise. *Am J Epidemiol* 2020;190:194-195.
16. Hohmann E, Wetzler MJ, D'Agostino RB. Research pearls: The significance of statistics and perils of pooling. Part 2: Predictive modeling. *Arthroscopy* 2017;33:1423-1432.
17. Neyton L, Godenèche A, Nové-Josserand L, Carrillon Y, Cléchet J, Hardy MB. Arthroscopic suture-bridge repair for small to medium size supraspinatus tear: Healing rate and retear pattern. *Arthroscopy* 2013;29:10-17.
18. O'Donnell EA, Fu MC, White AE, et al. The effect of patient characteristics and comorbidities on the rate of

revision rotator cuff repair. *Arthroscopy* 2020;36:2380-2388.
19. Domb BG, Sabetian PW. When no difference does not mean there is no difference: The blight of the type II error. *Arthroscopy* 2021;37:1353-1356.
20. Koren M. Jupiter and Saturn are just showing off now. This year is ending with a rare cosmic alignment. The Atlantic *December* 2020;21.
21. NASA. The 'Great' Conjunction of Jupiter and Saturn. Available at, https://www.nasa.gov/feature/the-great-conjunction-of-jupiter-and-saturn. December 15, 2020. Accessed February 6, 2021.
22. Lovett EO. The great inequality of Jupiter and Saturn. *Astronomical J* 1895;15:113-127.
23. Pannekoek A. The planetary theory of LaPlace. *Popular Astronomy* 1948;56:300-311.