OPEN ACCESS

P values: from suggestion to superstition

John Concato,[1,2] John A Hartigan[3]

[1]Clinical Epidemiology Research Center, Cooperative Studies Program, Veterans Affairs Connecticut Healthcare System, West Haven, Connecticut, USA
[2]Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA
[3]Department of Statistics, Yale University, New Haven, Connecticut, USA

Correspondence to
Dr John Concato, Clinical Epidemiology Research Center (151B), VA Connecticut Healthcare System, 950 Campbell Ave, 151B, West Haven, CT 06516, USA;
john.concato@va.gov;
john.concato@yale.edu

Accepted 8 July 2016
Published Online First 3 August 2016

To cite: Concato J, Hartigan JA. *J Investig Med* 2016;**64**:1166–1171.

ABSTRACT

A threshold probability value of '$p \leq 0.05$' is commonly used in clinical investigations to indicate statistical significance. To allow clinicians to better understand evidence generated by research studies, this review defines the p value, summarizes the historical origins of the p value approach to hypothesis testing, describes various applications of $p \leq 0.05$ in the context of clinical research and discusses the emergence of $p \leq 5 \times 10^{-8}$ and other values as thresholds for genomic statistical analyses. Corresponding issues include a conceptual approach of evaluating whether data do *not* conform to a null hypothesis (ie, no exposure–outcome association). Importantly, and in the historical context of when $p \leq 0.05$ was first proposed, the 1-in-20 chance of a false-positive inference (ie, falsely concluding the existence of an exposure–outcome association) was offered only as a suggestion. In current usage, however, $p \leq 0.05$ is often misunderstood as a rigid threshold, sometimes with a misguided 'win' ($p \leq 0.05$) or 'lose' ($p > 0.05$) approach. Also, in contemporary genomic studies, a threshold of $p \leq 10^{-8}$ has been endorsed as a boundary for statistical significance when analyzing numerous genetic comparisons for each participant. A value of $p \leq 0.05$, or other thresholds, should *not* be employed reflexively to determine whether a clinical research investigation is trustworthy from a scientific perspective. Rather, and in parallel with conceptual issues of validity and generalizability, quantitative results should be interpreted using a combined assessment of strength of association, p values, CIs, and sample size.

INTRODUCTION

Clinicians and biomedical researchers frequently encounter reports containing results of statistical analyses (eg, $p = 0.052$). In particular, generations of practitioners and investigators have learned that a threshold probability value—or, more formally, a tail probability value—of '$p \leq 0.05$' is used commonly to define statistical significance. For those unfamiliar with the underlying mathematical principles, however, the exact meaning of such information can be elusive. In addition, the corresponding procedures and practices themselves have been criticized.[1–6] This report, in mainly non-mathematical terms, defines the p value, summarizes the historical origins of the p value approach to hypothesis testing, describes various applications of $p \leq 0.05$ in the context of clinical research, and discusses the emergence of $p \leq 5 \times 10^{-8}$ and other values as thresholds for genomic statistical analyses.

DEFINITION AND IMPLICATIONS

Studies of exposure–outcome associations typically include four stages: specifying a research question, designing a study architecture, collecting data, and conducting a statistical analysis to draw inferences from the results. (Of note, we use exposure–outcome instead of cause–effect to avoid implications regarding causality; other characterizations include independent variable(s)-dependent variable). The predominant format for conducting statistical analyses is the frequentist approach, referring to the frequency of the occurrence of outcome events in repeated samples from a source population. Introducing an example that will be referred to later, if a 'fair' coin were to be flipped 10 times, every occurrence for the possible number of heads has an expected frequency, including a maximum of 24.6% for five heads and five tails, as well as lower expectations for the other possibilities (75.4% combined).

Consider a simple two-variable clinical scenario, with exposure as the independent variable and outcome as the dependent variable. Leaving aside various details and assumptions, the frequentist researcher often examines the results with respect to a null hypothesis of no association between exposure and outcome. In this context, the probability of an association at least as strong as what is observed, by random chance, is the p value. Importantly, the p value is *not* the probability that the null hypothesis, of no association, is true. Instead, we intentionally presume the null hypothesis—as a straw man argument—is true, to indirectly assess the plausibility of the data *not* conforming to it, notwithstanding issues of measurement error or systematic bias (including the concept of 'confounding'). In more formal terms, the American Statistical Association recently published an editorial[6] stating 'a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value'.

From a practical perspective, a test statistic typically determines the probability of the observed result (or a more extreme result) occurring by random chance, if no association exists. Among various reasons for selecting a particular statistical test, one consideration involves the measurement scales of the variables describing the exposure–outcome association. A selected list of some commonly encountered tests is shown in table 1, including the χ^2 test

Table 1 Examples of type of variables and selected statistical test(s)

Bivariate (unadjusted) analysis		
First variable (*)	**Second variable (*)**	**Statistical test(s)**
Binary (unpaired)	Binary	χ^2, Fisher's exact
(paired)	Binary	McNemar χ^2
Binary (unpaired)	Continuous	Student's t-test
(paired)	Continuous	Paired t-test
Binary	'Moving' binary (survival curves)	Log-rank
Continuous	Continuous	Correlation (r), linear regression
Multivariable (adjusted) analysis		
Target variable	**Statistical test(s)**	**Other target variable(s)**
Binary	Multiple logistic regression	Ordinal
Continuous	ANOVA; ANCOVA	–
Continuous	Multiple linear regression	Ordinal, binary
Integer count	Poisson regression	(contingency tables)
'Moving' binary (eg, survival curve)	Proportional hazard function analysis (Cox regression)	–

*Independent and dependent variables, when applicable, are not distinguished in this table.
Paired indicates that the study design links (matches) particular participants in compared groups.
ANCOVA, analysis of covariance; ANOVA, analysis of variance.

for categorical variables (eg, binary–binary comparisons, in a 2×2 table), t-test for a binary-continuous comparison, and log-rank test for time-to-event or 'survival' analyses when evaluating unadjusted associations. Logistic regression for binary outcomes and proportional hazards regression for time-to-event analyses are common approaches when evaluating adjusted, or multivariable,[7 8] associations. Of note, different tests can be applied in the same situation, as with a Fisher's exact test in lieu of a χ^2 test, especially when sample sizes are small.

Regardless of which statistical test is used, a 'good' result for assessing exposure–outcome associations is a small p value representing a low probability, thereby providing statistical evidence that an exposure–outcome relationship exists. In formal terms, the null hypothesis of no association is rejected. Specifically, $p \leq 0.05$ indicates that if no association exists, then the probability of the observed or a stronger association being attributable to chance is no greater than 1-in-20. Conversely, an analysis with $p > 0.05$ is considered not statistically significant; chance is considered a plausible explanation, and the null hypothesis is not rejected (although it is never 'accepted', given that an infinitely sized source population is assumed).

HISTORICAL ORIGINS

Published work on using concepts of probability for comparing data to a scientific hypothesis can be traced back for centuries. In the early 1700s, for example, the physician John Arbuthnot analyzed data on christenings in London during the years 1629–1710 and observed that the number of male births exceeded female births in each of the years studied. He reported[9] that if one assumes a balance of male and female births is based on chance, then the probability of observing an excess of males over 82 consecutive years is $0.5^{82}=2\times10^{-25}$, or less than a one in a septillion (ie, one in a trillion-trillion) chance. As an early example of how statistical significance should not be the sole basis for interpreting results, Arbuthnot included what he called an explanatory note—with the findings linked to an assertion that 'polygamy is contrary to the law of nature and justice'.[9]

In 1900, and during the initial development of the formal discipline of statistics,[10] mathematician Karl Pearson[11] described the χ^2 statistical test, applied to topics including throws of dice and roulette balls at Monte Carlo.[12] For example, examining data for n=26,306 dice throws, Pearson compared the observed versus the expected frequencies of 5s or 6s, based on a uniform probability for each face value; the tail probability of 0.000016 indicated that the dice were biased toward the higher values.[11] Also in 1900, although not involving tests of statistical significance, work done in the late 1800s by the Austrian monk Gregor Mendel on inheritance patterns in peas was first fully appreciated.[13] Mendel had established the genetic principles of segregation and independent assortment, and the renewed interest in Mendel's research later spawned a 'Mendelian–biometrician' controversy involving statistical methods[14 15] that helped to spur development of the science of genetics.

Whether focused on games of chance, patterns of inheritance or other topics, research on statistical methods flourished in the early 20th century. In particular, the 1925 publication of *Statistical Methods for Research Workers*[16] by the mathematician and biologist R.A. Fisher is considered a landmark event in statistics. This text, and later editions, is credited with helping to have developed a formal approach to significance testing using probability, or p values.

THRESHOLD VALUES

Importantly, when deciding on what p value threshold should indicate statistical significance, Fisher and other statisticians were not dogmatic. In 1926, as one of Fisher's early statements endorsing a p value of 0.05 as a boundary, he wrote: "…*it is convenient* [emphasis added] to draw the line at about the level at which we can say: 'Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials'."[17] In 1956, Fisher wrote: "[…] no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."[18]

Despite Fisher's intent, '$p \leq 0.05$' is currently a benchmark in many domains of scientific investigation. Thus, clinicians are often taught that $p \leq 0.05$ indicates statistical significance, based on the 1-in-20 threshold described earlier. If a clinical research study has a lower (better) p value of 0.001, for example, then the probability of chance alone explaining the findings would be one in a thousand—approximately the chance, invoking the previous coin scenario, of getting 10 heads in a row if a 'fair' coin is flipped 10 times (calculated as $0.5^{10}=0.00098$, or ≈ 0.001).

CONTEMPORARY USAGE

The use of p values is now ubiquitous, but at the same time, their application has taken on excessive reverence, as if rituals are being followed in their application. For example, an editorial[19] discussing a randomized trial of a therapeutic intervention indicated that 'the trial failed to meet its goal: the P value for death for any cause was 0.052, which was higher than the pre-specified value of 0.05. All clinical trials are a gamble, and the [investigators] came close to winning but did not win. Thus, the results of the trial are difficult to interpret'. Although the assessment of results was clarified elsewhere in the editorial, these particular statements (in a high-impact journal) prioritized the p value threshold of 0.05. Readers might mistakenly view the original trial[20] as a failure—only on the basis of a p value calculated to the third decimal place.

Two scenarios can illustrate why a p value threshold does not represent a win–lose situation. As shown in figure 1, in study A with n=87 participants, the p value is 0.062, not meeting the p≤0.05 threshold for statistical significance. If only two participants are added, however, and if the additional exposed participant has the outcome, whereas the additional non-diseased participant does not, then in study B with n=89 participants, the p value is 0.037—a statistically significant result. (Of note, these results were calculated using a Fisher's exact test).

Study A: sample size=87

Exposure:	Outcome: yes	Outcome: no	
yes	13 (30%)	31 (70%)	44
no	5 (12%)	38 (88%)	43
	18	69	87

Study B: sample size=89

Exposure:	Outcome: yes	Outcome: no	
yes	14 (31%)	31 (69%)	45
no	5 (11%)	39 (89%)	44
	19	70	89

Results for Study A: relative risk = 2.5; 95% C.I. = 0.99 to 6.5; P value = 0.062

Results for Study B: relative risk = 2.7; 95% C.I. = 1.1 to 7.0; P value = 0.037

Figure 1 In the first study 'A', with n=87, the relative risk is 2.5 (95% CI 0.99 to 6.5) and the p value is 0.062. In the second study 'B', with n=89, the relative risk is 2.7 (95% CI 1.1 to 7.0) and the p value is 0.037. The two studies are quite similar from an overall perspective, but 'B' is statistically significant, whereas 'A' is not.

Most readers would agree that the distinction between the two hypothetical studies, involving n=87 or n=89 participants, is modest. The relative risks are similar, showing that both studies suggest a similar level of strength of association. The appropriateness of the study design, the quality of data, or other issues can dominate a modest distinction in calculated p values.

CONFIDENCE INTERVALS

Although beyond the scope of this paper, CIs are a more informative counterpart of p values, reporting mathematical stability in the format of the relative risk or other expressions (eg, OR, HR, risk difference) of the strength of association. From a pragmatic perspective, as shown in figure 2, and using relative risk for illustration, a 95% CI that excludes the null value is statistically significant—leading to the same conclusion with regard to statistical significance as a p≤0.05. Conversely, a 95% CI that includes the null value is not statistically significant.

Described in the 1930s[21] and endorsed later by influential papers[22 23] on this topic, CIs are now a welcomed accompaniment of p values, providing information on stability linked to information on the strength of association. Although vulnerable to the same problems as p values regarding inference, CIs can help to interpret analytic results.[24] Consider a result with a 'non-significant' result, such as relative risk=1.4, 95% CI=0.80 to 2.4 and p=0.20. In another project addressing the same question, a stronger point estimate, wider CI and significant p value are determined, such as relative risk=4.1, 95% CI=1.2 to 14.0 and p=0.02. In a side-by-side comparison, the first scenario can actually be viewed[24] as providing more trustworthy information on a possible association, given a narrower CI—despite the lack of statistical significance when judged by p≤0.05. The take-home message is that a p value alone does not provide comprehensive information on an analytic result.

SAMPLE SIZE AND STATISTICAL POWER

The general relationship between sample size and statistical significance tends to be appreciated by experienced

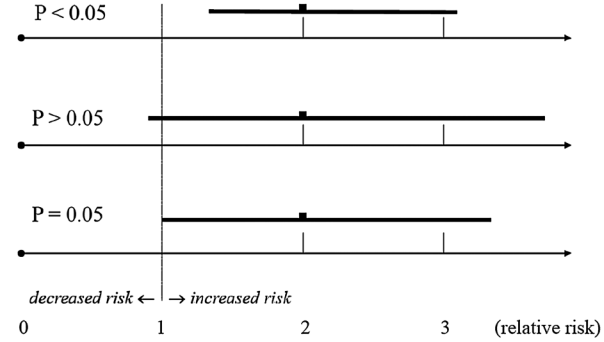

Figure 2 p Values and CIs provide concordant information regarding statistical significance. A 95% CI that excludes the null value of one, as with a p value of ≤0.05, indicates a statistically significant result. A 95% CI that includes the null value of one, as with a p value of ≥0.05, indicates a statistically non-significant result. A p=0.05 occurs when a CI ends at 1.0 and is considered statistically significant.

Table 2 Examples of large and small quantitative differences, and corresponding p values[25]

Large quantitative difference		
Outcome A=0.333	**Outcome B=0.250**	**p Value**
1/3	1/4	0.81
10/30	10/40	0.45
100/300	100/400	0.02
1000/3000	1000/4000	<0.0000001
Small quantitative difference		
Outcome A=0.288	**Outcome B=0.282**	**p Value**
288/1000	282/1000	0.77
2880/10,000	2820/10,000	0.35
28,800/100,000	28,200/100,000	0.003

p Values calculated using the χ^2 test, for demonstration purposes.

researchers, but may not be immediately apparent to clinicians. As shown in the top portion of table 2 (using data from a previously prepared example[25]), a relatively big quantitative difference (one-third vs one-quarter) can be statistically non-significant, due to small sample size, and such results should not be surprising. This situation provides an argument in favor of calculating a priori statistical power, based on the difference in outcomes that might be expected, to avoid 'underpowered' studies. Interestingly, the concept of statistical power—as with p values—was developed in the early 20th century,[26 27] but its relevance was not appreciated widely by many researchers until decades later, with recognition that numerous underpowered studies had been published in the social science[28] and medical literature.[29]

In brief, the power of a test is the probability that a statistically significant result will be detected, given the existence of an association of a certain (hypothesized) strength. Most randomized trials are designed to have at least 80% power, indicating a <20% chance of concluding that the data are not supportive of an exposure–outcome association if a specified exposure–outcome relationship were to exist (ie, making a false-negative inference). In practical terms, power should be calculated when a study is designed, and is then discussed again (but not recalculated) if the results are *not* statistically significant.

CLINICAL SIGNIFICANCE

Statistical (probabilistic) significance and clinical (quantitative) significance are different concepts. Regardless of arguments in favor or against hypothesis testing, few would argue against the claim that $p \leq 0.05$ has become a standard approach. In contrast, no single threshold for clinical significance exists—and none will likely develop—because each clinical context is different. For example, even a small improvement in survival for a new therapeutic agent for an aggressive type of cancer would likely be viewed positively; the same percentage improvement involving a benign and self-limited ailment might be met with less enthusiasm.

From a statistical perspective, the bottom portion of table 2 shows that even a modest quantitative difference (eg, 28.8% vs 28.2%) can be statistically significant with a large sample size. Analyses of information in healthcare (administrative) databases are likely contexts for this scenario to occur. Sometimes characterized as 'overpowered' analyses, the results can include clinically unimportant differences that achieve statistical significance. If the implications of $p \leq 0.05$ are misunderstood, then associations can be misinterpreted when evaluating research questions involving therapeutic effectiveness, quality improvement, and other topics.

Interestingly, an argument was made in the early 20th century that statistical significance does not confer quantitative significance,[30] yet for decades many studies only reported p values. Other studies used '*' to indicate $p \leq 0.05$ and 'NS' to indicate non-significance. Reporting p values without a relative risk, OR, etc, to describe the strength of association is inappropriate, and corresponding uncertainty in estimates should be provided (eg, using CIs). In addition, using symbols is less informative than reporting actual p values, in that the actual values quantify the probabilistic evidence against the null hypothesis. More generally, and almost 100 years later, a statement published in 1919 still applies: "[…] statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere."[30]

P VALUES IN THE GENOMIC ERA

The threshold of $p \leq 0.05$ was established when sample sizes in medical investigations tended to have a modest number of measurements per participant. In contrast, genome-wide association studies can evaluate hundreds of thousands, to several million, single nucleotide polymorphisms (SNPs) as the exposure variable, and a disease or trait as the outcome variable, for each participant. Studies involving whole-exome or whole-genome sequencing have even larger numbers to consider.

If a p value of ≤ 0.05 were to be used as the threshold for statistical significance in these situations, numerous associations would be expected by chance alone. Accordingly, and using various strategies for calculations, thresholds such as 5×10^{-8} have been proposed[31 32] to distinguish 'chance' from potentially 'real' genomic associations. This particular approach is related to pregenomic concepts of multiple comparisons.[33–36] For example, the Bonferroni correction[35] calculates a threshold p value for each comparison as 0.05/N, where N is number of comparisons. Thus, when evaluating 10 associations in a clinical study, $p \leq 0.005$ for any comparison indicates statistical significance. Using this strategy for a genome-wide association study involving 1 million SNPs, $p \leq 10^{-8}$ would be the calculated threshold for each polymorphism. The false discovery rate[37] is another strategy used for this purpose.

COMMENTS AND CAVEATS

As shown in table 2, a p value of ≤ 0.05 for a given strength of association can be achieved by enlarging sample size. Even for a fixed sample size, however, a calculated p value is not a unique assessment of any given data set. For example, using the data in figure 1 and choosing a χ^2 test instead of a Fisher's exact test, $p=0.04$ for $n=87$ and $p=0.02$ for $n=89$, suggesting that both associations are statistically significant—just by choosing a different

statistical test that analyzes the data via another conceptual approach and a different mathematical algorithm. As another example, and although other reasons for inconsistent results exist (including confounding), adding or removing several variables (ie, covariates) from a multivariable regression model can affect the p value, and possibly change the statistical significance, of a primary variable of interest.[38] Clinicians and investigators should certainly not assume that $p \leq 0.05$ implies a 'true' association, even if the non-statistical aspects of a study are conducted and reported impeccably.

To emphasize the arbitrary aspect of $p \leq 0.05$, the field of physics commonly uses a threshold p value of 3×10^{-7} for statistical significance, based on observations at least 5 SDs from the null hypothesis. Although the concept of SD is not discussed in this review, $p \leq 0.05$ corresponds to ~2 SDs, a less restrictive threshold to achieve. Other topics include issues such as one-tail versus two-tail significance testing. In brief, 'tail' or directionality refers to whether the hypothesis allows for a drug, for example, to have either a beneficial or a harmful effect, or is just expected to show benefit. Of note, a one-tail p value of 0.025 corresponds to a two-tail p value of 0.05, but two-tail (bidirectional) testing is endorsed in most situations.[39] As a separate conceptual issue, this narrative focuses on p values in settings where independence of compared groups, such as treatment and control arms in a trial, is the desired outcome. In some situations, including the Monte Carlo example mentioned earlier,[11] as well as for Mendelian genetics,[13] similarity to a specific pattern is expected. (In addition, $p > 0.05$ can even be a desired result, as with analyses to show that the observed data conform to a predicted model).

Finally, problems with the frequentist approach have been well documented, and Bayesian strategies are considered to be an appealing alternative approach.[40–45] Consistent with the explanation provided in the Definition and Implications section, the frequentist conceptual approach can be stated as: 'What is the probability that the observed data are inconsistent with the null hypothesis?' In contrast, the Bayesian conceptual approach for a typical research project can be stated as: 'Given the observed data, what is the probability that the true effect is negative (null)?'. Although the Bayesian flow of logic is more in keeping with how clinicians think, the need to specify the strength of association ahead of time—in formal terms, the 'prior probability distribution' of the effect size—seems to have made many researchers reluctant to adopt Bayesian methods.

BROAD PERSPECTIVE

This review does not discuss all issues related to p values, and the topics that are included are presented only as an overview, or in illustrative terms. For example, several distinctions that exist[46–49] between recommendations for significance testing by Ronald Fisher (focused mainly on a null hypothesis) and by Jerzy Neyman and Egon Pearson (incorporating an alternative hypothesis using power calculations) are not described. In addition, at least one health-related journal (*Basic and Applied Social Psychology*) has recently banned the 'null hypothesis significance testing procedure (NHSTP)',[50] stating specifically that prior to publication, 'authors will have to remove all vestiges of the NHSTP (p values [and] statements about "significant" differences or lack thereof, and so on)'.[50]

Notwithstanding such theoretical underpinnings and conceptual debates, authors of contemporary research articles should, at a minimum, avoid performing a perfunctory social ritual[51] involving p values. This scenario involves thoughtlessly repeating the same action (significance testing procedures), focusing on a special number ($p \leq 0.05$), fearing sanctions (by reviewers or editors) for rule violations, and thinking wishfully (seeking, and sometimes manipulating, a desired p value) while limiting critical judgment (as reflected by superficial discussion of statistical results, such as the win–lose dichotomy described in the Contemporary usage section).

As a general guideline, authors should report—and thoughtfully interpret—results describing associations in terms of strength, such as relative risk, as well as stability, including the magnitude of p values and CIs. The size of the study population is also relevant. From a more general perspective, information on the stability of results is important, but the clinical relevance of a research report is also affected by the issues of validity, confirming that the results are correct for the participants involved, and generalizability, describing to whom the results apply.[52] Box 1 provides several take-home points in this context.

Box 1 Take-home points for using probability values in clinical research

- Assuming no association exists, a test statistic determines a p value for (ie, the tail probability of) an observed result, or a more extreme result, occurring by random chance.
- The threshold p value of ≤ 0.05 for statistical significance, promoted in the early 20th century only as an informal suggestion, indicates a 1-in-20 chance of a false-positive inference (ie, assuming an exposure–outcome association when it does not exist).
- Even if a study is conducted impeccably and reported accurately, clinicians and investigators should not assume that $p \leq 0.05$ implies a 'true' association—and comparing a p value to a threshold does not represent a win–lose situation.
- In genomic studies, p value thresholds such as 5×10^{-8} reflect the extremely large number of associations (eg, alleles) being evaluated for each participant.
- In addition to p values, or CIs (as another format for expressing stability of results), a numerical result for the strength of an association (eg, relative risk) is essential information.
- Rigorous statistical analyses should be combined with relevant clinical insight regarding the corresponding research question, data collection, and study design.
- While considering the conceptual issues of validity and generalizability, interpreting the numerical results of clinical research investigations should assess the strength of association, magnitude of p values, CIs, and sample size.

CONCLUSION

Statistical testing describes the stability of quantitative results from a probabilistic perspective, but tests of significance should not be viewed as an all-or-none approach, and p values should rarely be the main focus of attention or the primary basis for evaluating a research study. At the very least, $p \leq 0.05$ or $p \leq 5 \times 10^{-8}$ should not be employed reflexively to determine whether a study is trustworthy from a scientific perspective. Along with the conceptual issues of validity and generalizability, the strength of association, magnitude of p values, width of CIs, and size of the study sample are all relevant when interpreting the results of clinical research investigations.

In 1880, the biologist T.H. Huxley stated "…it is the customary fate of new truths to begin as heresies and end as superstitions."[53] After almost a century of originally being adopted, p value thresholds have evolved into a superstition. To improve medical research and ultimately clinical care, more judgment, and less ritual, is warranted.

Acknowledgements The authors thank Peter Peduzzi for helpful comments and Sandra Augustitus for assistance in preparing the manuscript.

Contributors JC drafted the article and JAH contributed important intellectual content; both authors approved the final version.

Funding JC is supported by the VA Cooperative Studies Program.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Salsburg DS. The religion of statistics as practiced in medical journals. *Am Stat* 1985;39:220–3.
2 Carver RP. The case against statistical significance testing, revisited. *J Exp Educ* 1993;61:287–92.
3 Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. *Ann Intern Med* 1999;130:995–1004.
4 Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005;2:696–701.
5 Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol* 2010;25:225–30.
6 Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016;70:129–33.
7 Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118:201–10.
8 Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003;138:644–50.
9 Arbuthnott J. An argument for divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Phil Trans* 1710;27:186–90.
10 Stigler SM. *The history of statistics: the measurement of uncertainty before 1900*. Cambridge: The Belknap Press of Harvard University, 1986.
11 Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag* 1900;50:157–75.
12 Plackett RL. Karl Pearson and the chi-squared test. *Int Stat Rev* 1983;51:59–72.
13 Moore R. The "rediscovery" of Mendel's work. *Bioscene* 2001;27:13–24.
14 Rushton AR. Nettleship, Pearson and Bateson: the biometric-Mendelian debate in a medical context. *J Hist Med* 2000;55:134–57.
15 Magnello ME. Karl Pearson's mathematization of inheritance: from ancestral heredity to Mendelian genetics (1895–1909). *Ann Sci* 1998;55:35–94.
16 Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1925.
17 Fisher RA. The arrangement of field experiments. *J Min Agric* 1926;33:503–13.
18 Fisher RA. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd, 1956.
19 Rabe KF. Treating COPD—the TORCH trial, p values, and the Dodo. *N Engl J Med* 2007;356:851–4.
20 Calverly PMA, Anderson JA, Celli B, *et al*. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007;356:775–89.
21 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934;26:404–13.
22 Rothman KJ. A show of confidence. *N Engl J Med* 1978;299:1362–3.
23 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746–50.
24 Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* 2001;12:291–4.
25 Concato J. Overview of research design in epidemiology. *J Law and Policy* 2004;XII:489–507.
26 Neyman J, Pearson ES. On the use of interpretation of certain test criteria for purposes of statistical inferences: Parts I and II. *Biometrika* 1928;20A:175–240, 263–294.
27 Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philos T Roy Soc A* 1933;231:289–337.
28 Cohen J. The statistical power of abnormal—social psychological research: a review. *J Abnorm Soc Psychol* 1962;65:145–53.
29 Freiman JA, Chalmers TC, Smith H, *et al*. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690–4.
30 Boring EG. Mathematical vs. scientific significance. *Psychol Bull* 1919;16:335–8.
31 Panagiotou OA, Ioannidis JPA. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol* 2012;41:273–86.
32 Jannot AS, Ehret G, Perneger T. $P < 5 \times 10^{-8}$ has emerged as a standard of statistical significance for genome-wide association studies. *J Clin Epidemiol* 2015;68:460–5.
33 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;198:679–84.
34 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43–6.
35 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.
36 Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC Med Res Methodol* 2002;2:8.
37 Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 2003;164:829–33.
38 Detre KM, Peduzzi P, Chan YK. Clinical judgment and statistics. *Circulation* 1981;63:239–41.
39 Fleiss JL. Some thoughts on two-tailed tests. *Control Clin Trials* 1987; 8:394.
40 Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med* 1999;130:1005–13.
41 Goodman SN. Of P-values and Bayes: a modest proposal. *Epidemiology* 2001;12:295–7.
42 Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. *Am Stat* 2005;59:121–6.
43 Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 2006;35:765–75.
44 Hartigan JA. *Bayes theory*. New York: Springer-Verlag, 2011.
45 Greenland S, Poole C. Living the *P* values. Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology.* 2013;24:62–8.
46 Box JF. R.A. Fisher and the design of experiments, 1922–1926. *Am Stat* 1980;34:1–7.
47 Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Statist Assoc* 1993; 88:1242–9.
48 Lenhard J. Models and statistical inference: the controversy between Fisher and Neyman-Pearson. *Brit J Phil Sci* 2006;57:69–91.
49 Kyriacaou DN. The enduring evolution of the *P* Value. *JAMA* 2016;315:1113–15.
50 Trafimow D, Marks M. Editorial. *Basic Appl Soc Psych* 2015;37:1–2.
51 Gigerenzer G. Mindless statistics. *J Socio Econ* 2004;33:587–606.
52 Concato J. Study design and "evidence" in patient-oriented research. *Am J Respir Crit Care Med* 2013;187:1167–72.
53 Barr AP, ed. *The major prose of Thomas Henry Huxley*. Athens, GA: University of Georgia Press, 1997.