Check for updates

Détente: A Practical Understanding of *P* values and Bayesian Posterior Probabilities

Stephen J. Ruberg[1,*]

Null hypothesis significance testing (NHST) with its benchmark *P* value < 0.05 has long been a stalwart of scientific reporting and such statistically significant findings have been used to imply scientifically or clinically significant findings. Challenges to this approach have arisen over the past 6 decades, but they have largely been unheeded. There is a growing movement for using Bayesian statistical inference to quantify the probability that a scientific finding is credible. There have been differences of opinion between the frequentist (i.e., NHST) and Bayesian schools of inference, and warnings about the use or misuse of *P* values have come from both schools of thought spanning many decades. Controversies in this arena have been heightened by the American Statistical Association statement on *P* values and the further denouncement of the term "statistical significance" by others. My experience has been that many scientists, including many statisticians, do not have a sound conceptual grasp of the fundamental differences in these approaches, thereby creating even greater confusion and acrimony. If we let A represent the observed data, and B represent the hypothesis of interest, then the fundamental distinction between these two approaches can be described as the frequentist approach using the conditional probability pr(A | B) (i.e., the *P* value), and the Bayesian approach using pr(B | A) (the posterior probability). This paper will further explain the fundamental differences in NHST and Bayesian approaches and demonstrate how they can co-exist harmoniously to guide clinical trial design and inference.

We are now approaching a "100-year war" on the proper approach for analyzing and interpreting the results of an experiment to test a scientific hypothesis.[1–10] The centerpiece of scientific inference for nearly a century has been the development of null hypothesis significance testing (NHST) and its ultimate output, the *P* value. The emergence of $P < 0.05$ appears to have come as a matter of convenience as Sir Ronald Fisher noted, "The value for which $P = 0.05$ is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."[11] He reiterated this idea when he made statements about what he thought might be considered "statistically significant" evidence.[12] Further advances in inferential reasoning were developed by Jerzy Neyman and Egon Pearson as they established the theory for deciding whether to "accept" or reject the null hypothesis—H_0.[13] These developments are known as the frequentist school of statistical inference because they rely on the frequency (or probability) with which one might observe data (denoted A) under the assumption that H_0 is true (i.e., no treatment effect; denoted B). If that probability, the *P* value, is small, then the data is considered incompatible with H_0, and H_0 is rejected in favor of the alternative hypothesis (i.e., there is a treatment effect). Throughout this paper it will be denoted generally as pr(A | B)—pr(the data given a hypothesis).

The Bayesian approach derives its name from the ideas developed by Reverend Thomas Bayes in 1763.[14] He was interested in the inverse probability, denoted pr(B | A); that is, given the data that has been observed from an experiment (A), what probability might be assigned to the truth of a hypothesis (B) or its alternative hypothesis, denoted B^C or the complement of B. Bayes formula describes the relationship between $pr(B^C \mid A)$ and pr(A | B):

$$pr\left(B^C \mid A\right) = pr\left(A \mid B^C\right)\, pr\left(B^C\right) / \left[pr\left(A \mid B^C\right)\, pr\left(B^C\right) + pr\left(A \mid B\right)\, pr\left(B\right)\right].$$

I present the formula in the context of the alternative hypothesis B^C for reasons that will become clear in the subsequent exposition.

The history of the philosophical skirmishes between Fisher, Neyman, and Pearson and their followers are well-documented and have been recounted elsewhere.[15] Perhaps even more notable are the clashes between the frequentist and Bayesian schools of inference.[1] What is noteworthy is that warnings about the use or misuse of *P* values come from both frequentist *and* Bayesian schools of thought and span many decades.

Discussions in this arena have been heightened by the American Statistical Association statement on *P* values[16] and the further denouncement of the term "statistical significance"[17,18] as well as its rebuttal.[19] Many alternative ideas are contained in a Special Issue of *The American Statistician* (volume 73:supplement 1, 2019), including many authors urging that inference move away from frequentist approaches toward Bayesian approaches. My experience has been that many scientists, including many statisticians, do not have a sound conceptual grasp of the fundamental differences in these approaches, thereby creating even greater confusion and acrimony. Furthermore, only modest progress has been made over the past decades to resolve these schools of inferential thought.

[1]Analytix Thinking, LLC, Indianapolis, Indiana, USA. *Correspondence: Stephen J. Ruberg (analytixthinking@gmail.com)

Received April 21, 2020; accepted June 27, 2020. doi:10.1002/cpt.2004

Hopefully, this paper will help solidify concepts and understanding as well as propose a path forward for improving statistical AND thereby scientific inference.

Section "Frequentist and Bayesian Inference Made Simple" provides a simple illustrative example that, in my experience, has been very enlightening for distinguishing the fundamental differences between frequentist and Bayesian inference. A novel explanation of the *P* value is presented to extend the notion of the *P* value fallacy.[6] Section "Clinical Trials as Diagnostic Tests" describes a common, well-accepted, real-world application of Bayesian thinking and generalizes to clinical trials and more broadly research experiments. Section "Détente: The Peaceful Co-Existence of P-values and Bayesian Probabilities" describes how to integrate frequentist and Bayesian thinking. Section "A Path Forward" recommends a path forward, and section "Summary" is a summary and call for change.

FREQUENTIST AND BAYESIAN INFERENCE MADE SIMPLE

A thought experiment

Suppose a bag of 10,000 coins contains 9,999 coins balanced for heads (H) and tails (T) and 1 biased coin with H on both sides. An experimenter will draw one coin at random from the well-mixed bag of coins but does not show the coin to an observer. The experimenter flips the selected coin repeatedly and tells the observer the result, H or T. The observer is to declare when s/he is willing to bet that the experimenter drew the biased coin. Of course, if the result of any flip is T, then the observer would immediately know that a fair coin was drawn. So, the intriguing thought experiment is to suppose there is a sequence of H's. Stated simply, the question is "How many consecutive H's are needed before one should be willing to bet that the biased coin was selected?" When I have posed this thought experiment to clinicians, scientists, and statisticians (henceforth called my students), most of them answer in the range of 6–10 H's being sufficient to make the bet.

There are two perspectives for answering the question. The first perspective is the NHST or frequentist approach, which defines a simple null hypothesis H_0: pr(H) = 0.50 (i.e., a fair coin) and a simple alternative H_a: pr(H) = 1.0 (i.e., the biased coin). Suppose the observer's prespecified decision rule for this NHST is to reject H_0 when N consecutive H's are observed, where N is the observer's choice. In this case, the probability of a type 1 error (probability of rejecting H_0 if indeed H_0 is true) for such a decision rule is simply:

$$\text{pr}\left(N \text{ consecutive } H\text{'s} \mid H_0 \text{ is true}\right) = 0.5^N.$$

Note that N is a fixed value that defines the decision rule, and the choice of N can be made based on how small the observer wants to make the type 1 error (uniformly denoted as α). Once again, if A represents the observed data and B represents H_0, then the above can be written as pr(A | B). This is a conditional probability because the probability of A is computed under the condition or assumption that B has occurred or that B is true.

Once the experiment proceeds, this same calculation can be done for any *observed* value n and is known as the *P* value, which is presented in **Table 1** for various values of n. Of course, small *P* values are interpreted as *evidence* against the null hypothesis (i.e., Fisher's notion of "significant" used in a colloquial sense) or lead to a *decision* to reject H_0 whenever $n \geq N$ (i.e., Neyman–Pearson decision theory).

The second perspective follows a Bayesian approach. Some have argued that scientists are more interested in the likelihood of whether a hypothesis is true or false given the results of an experiment (i.e., the observed data).[20] In particular, the greatest interest is in pr(H_0 is false | data) or pr(B^C | A). Equivalently, this is the pr(H_a is true | data). As noted previously, this inverse probability calculation requires the use of Bayes formula (see **Supplementary Materials** for calculations pertinent to this example). For this thought experiment, Bayesian inference calculates:

$$\text{pr}(\text{the biased coin was drawn} \mid n \text{ consecutive heads are observed})$$

which is also displayed in **Table 1** for values of n = 1, 20. This probability is directly related to the question at hand for making a bet.

If we were to repeat the thought experiment with a bag of 100 coins, with 99 fair coins and 1 biased coin, the same betting question can be asked. Most of my students rightly guess that fewer consecutive H's are needed to make the bet. The precise Bayesian probabilities are also presented in **Table 1**.

Table 1 ***P* values and Bayesian probabilities for the coin toss thought experiment**

Number of consecutive Heads (n)	*P* value[a]	Prior = 1/10,000 pr(biased coin)[b]	Prior = 1/100 pr(biased coin)[b]
1	0.5	0.000200	0.019802
2	0.25	0.000400	0.038835
3	0.125	0.000799	0.074766
4	0.0625	0.001598	0.139130
5	0.03125	0.003190	0.244275
6	0.015625	0.006360	0.392638
7	0.0078125	0.012639	0.563877
8	0.0039063	0.024963	0.721127
9	0.0019531	0.048711	0.837971
10	0.0009766	0.092897	0.911843
11	0.0004882	0.170001	0.953889
12	0.0002441	0.290600	0.976400
13	0.0001220	0.450333	0.988059
14	0.0000610	0.621006	0.993994
15	0.0000305	0.766198	0.996988
16	0.0000153	0.867624	0.998492
17	0.0000076	0.929121	0.999245
18	0.0000038	0.963258	0.999622
19	0.0000019	0.981285	0.999811
20	0.0000010	0.990554	0.999906

[a]The *P* value is calculated as pr(n consecutive H's | H_0 is true) = 0.5^n. [b]The pr(biased coin) is calculated using Bayes formula (see **Supplementary Information**) with the stated prior in each column.

Several observations in **Table 1** are warranted to understand the fundamental differences in these two perspectives and to assess when it is appropriate to bet that the experimenter selected the biased coin.

1. Many "*P* value minded" students intuitively know that a large string of consecutive *H*'s is unlikely using a fair coin. Although they may not make the 0.5^n calculation in their minds, they have a sense that the *P* value is rapidly getting small as n gets large. Interestingly, for 6–10 consecutive *H*'s (the common range of answers), the *P* values range from 0.016 to 0.001. Many students find this gratifying or confirmation of their sense of "statistically significant."
2. The *P* values in **Table 1** remain the same, regardless of which bag of coins is used in the thought experiment. This is because the calculation is based on a conditional probability assuming the truth is known (i.e., H_0 is true) or said differently, is independent of the prior probability of selecting the biased coin.
3. Using the Bayesian calculations for the bag with 10,000 coins, several remarks are noteworthy.
   a. Before the coin is ever flipped (i.e., generate any data), the probability that the biased coin was selected is 1 in 10,000 or 0.0001. As seen in **Table 1**, when one *H* is observed, there is one small piece of evidence that the coin may be biased, and the probability is now 0.0002 (i.e., 2/10,000). With each consecutive head, or piece of data, that probability is getting larger, but even at 10 consecutive *H*'s, the probability that the biased coin was selected is < 0.10, which appears to be greatly at odds with any interpretation based on the *P* value.
   b. Note that with 13 consecutive *H*'s, the probability that the biased coin was selected is 0.45, and with 14 consecutive *H*'s that probability is 0.62. Thus, with 14 consecutive *H*'s, the observer would calculate that the odds are in his/her favor that the experimenter has the biased coin, and s/he would make the bet.
   c. Now note that the *P* value for 13 consecutive *H*'s is 0.0001220, which is slightly *larger* than 1 in 10,000. The *P* value for 14 consecutive *H*'s is 0.0000610, which is slightly *smaller* than 1 in 10,000. That is, if the observer were to think rightly about this evidence, the observer would realize that 13 consecutive *H*'s using a fair coin is *more likely* to occur than the experimenter pulling the one biased coin from the bag of 10,000 coins. Furthermore, 14 consecutive *H*'s is *less likely* to occur with a fair coin than for the experimenter to pull the one biased coin from the bag of 10,000 coins. Again, this is the point at which the observer should make the bet. This demonstrates that the *P* value can *only* be rightly interpreted in the context of our prior knowledge (i.e., the prior probability of pulling the biased coin from the bag).
4. In the experiment with the bag of 100 coins, the same observations can be made. At the outset, the probability of pulling the biased coin is 0.01, and with the first flip of the coin being *H*, that probability increases slightly to about 0.02. Furthermore, with 6 consecutive *H*'s, the *P* value is slightly greater than 1 in 100 (0.016) and the Bayesian probability is 0.393 (i.e., < 0.50). With 7 consecutive *H*'s, the *P* value is slightly < 1 in 100 (0.008) and the Bayesian probability is 0.564. Thus, in the context of this bag of coins, 7 consecutive heads are more likely to occur under H_a than H_0, representing sufficient evidence to bet that the biased coin was drawn from the bag of 100 coins.

This example illustrates the relationship between the Bayesian probability, the prior used for that Bayesian calculation and the *P* value. Observing a rare event, such as 10 consecutive heads, must only be evaluated in the context of prior knowledge, in this case, how many coins are in the bag. Although observing 10 consecutive heads with a fair coin is quite unusual (*P* value about 1/1,000), it is not nearly as unusual as the experimenter pulling THE one biased coin from a bag having 9,999 other fair coins (1/10,000). Although the relationships between the Bayesian probability, the prior and the *P* value for this example do not necessarily carry over precisely to more complex hypotheses and experimental data, the concepts do. One should only consider how extreme an observed set of data are in the context of prior knowledge or belief.

As noted previously, the *P* values remain the same, regardless of which bag of coins is used in the thought experiment. However, as demonstrated in **Table 1**, its proper interpretation depends on the experimental setting and is conditional on the prior. *P* values and Bayesian posterior probabilities give the same insight *provided* the calculations are seen in the context of the specific experiment (i.e., the size of the bag of coins and the prior probability it confers). Because the *P* value calculations are the same irrespective of the prior, their interpretation must be done *a posteriori*, whereas the Bayesian posterior probabilities explicitly incorporate prior knowledge.

Too often, the debate about the veracity of H_0 vs. H_a is conducted after the experiment is completed based on the *P* value, but its interpretation requires consideration of the experimental setting and how a finding compares to the extant body of scientific knowledge. Was the experiment well-designed? Does the laboratory or clinic have credibility? Are the findings a surprise or consistent with other studies? In contrast, the Bayesian approach forces such considerations prior (I use this term intentionally) to the conduct of the study and requires thoughtful assimilation of information and quantification of belief. Both paradigms require subjective judgments, but the difficult work is done *a priori* in the Bayesian framework and *a posteriori* (or *post hoc*) in the frequentist approach.

Last, prespecification is an important principle in scientific research (e.g., prespecification of the study design, the patient population, the data collection, and the analysis approach). This is critical in confirmatory research but also valuable in early clinical development or exploratory research. Greater credibility of findings is achieved when prespecified research protocols and analyses are executed as planned. In this way, the Bayesian approach has an epistemological advantage over the frequentist approach in that (i) it forces prespecification of a prior that is used directly in the analysis and subsequent interpretation of results, and (ii) the resulting posterior probability statement is *directly* related to the question at hand: "Given an observed number of

consecutive *H*'s, what is the probability that the biased coin was drawn?"

P values: A closer look

NHST is so strongly embedded in the process of scientific research that it is worth reviewing the basic elements in its construction. First, one must declare a hypothesis of interest, most often "no effect." Second, an alternative hypothesis is declared and usually captures what a scientist wants to prove—that there is a difference in experimental groups. Third, a test statistic is defined that measures the difference in response between the experimental and control groups relative to the underlying variability in the data. Fourth, a clear definition of the magnitude of the type 1 error probability that is allowable is defined, that is the maximum probability of rejecting H_0 when, in fact, it is true (α). This is also known as controlling the probability of a false-positive (FP) finding or the significance level of the test. The key is that this statistical machinery must be prespecified; statistical principles and scientific validity for producing credible findings do not allow for one to observe data and then decide what hypotheses or test statistics or α-level to use.

The *P* value is the ultimate summary statistic because it is a single number distilled from the data (no matter how much) through a model (no matter how complex) that includes a probability distribution for capturing the uncertainty in the data. That is, a *P* value contains the same information as the test statistic on which it is based because there is a 1-to-1 mapping of the test statistic to the (0,1) scale through an inverse probability function.[21] Many forget this. I suspect we gravitate to *P* values because they exist on this common scale, and it is much easier to quickly judge the importance of $P = 0.03$, for example, than reporting a χ^2 statistic with 6 degrees of freedom equaling 13.97.

The problem with interpreting *P* values arises when they are construed as the smallest significance level for which H_0 could have been rejected,[22] as is often done in practice. However, this is a *post hoc* assessment that uses the observed data to define the significance level of the hypothesis test the researcher wishes they would have used. The scientific community would never allow a researcher to change their hypothesis or test statistic or the critical value for that test statistic based on observed data, and yet this is done routinely when interpreting the *P* value *statistic*. The researcher can use the *P* value to decide to reject H_0 whenever the observed *P*-value is less than the predefined α-level, but one should not make the *post hoc* interpretation that the observed *P* value is the type 1 error rate for the prespecified hypothesis test. It may seem to be a subtle distinction in inference and interpretation, but when the *P* value is seen for what it is—the transformation of a test statistic based on observed data to the interval (0,1)—then it is easy to appreciate the readily accepted principle that the hypothesis testing procedure should not be changed based on observed data.

Confusion arises when the α-level of the test used for setting a decision rule (N in the thought experiment) and the observed *P* value statistic (based on the observed n in the thought experiment) are both called the "significance level." Informally, the term "significance" has been used in a colloquial sense, following Fisher's original statements and his view that a *P* value may be indicative of an important finding. However, formally, the term "significance level" should only be used to describe the probability of making a FP decision (i.e., the α-level) using the NHST procedure chosen for the analysis. The conflating of the prespecified α-level, which is a characteristic of the hypothesis test, and the observed *P* value, which is a characteristic of the data, has been called the "silent hybrid solution."[23–25] It stems, in part, from the notion that evidence is assessed on a continuous scale, or gradations of probability, but decision making is discrete (i.e., false/true). In that sense, evidence is judged on the continuous interval (0, 1), and decision-making is judged on the set {0, 1}, but they both use the same *P* value.[26] In section "Détente: The Peaceful Co-Existence of *P* values and Bayesian Probabilities," the relationship between a *P* value and *a posterior* probability will be explored in more detail to understand the degree of evidence against H_0.

CLINICAL TRIALS AS DIAGNOSTIC TESTS

The Merriam–Webster definition of a bioassay is the "determination of the relative strength of a substance (such as a drug) by comparing its effect on a test organism with that of a standard preparation." In this regard, a clinical trial is nothing more than a very sophisticated bioassay, and, indeed, any research experiment may be considered as an assay. It is the attempt to quantify an unknown characteristic of a substance or organism through chemical or biological analysis.

More specifically, a clinical trial is a diagnostic. Diagnostic tests are often based on a biochemical assay. Even those that are not, such as electrocardiograms, are still identical in structure to a biochemical diagnostic test—there is some "machinery" for making measurements that is calibrated to distinguish a patient with or without an unknown characteristic. When decision making is involved, that calibration includes some cutoff value that determines the operating characteristics (i.e., sensitivity and specificity) of the diagnostic test. The design, operating characteristics, and interpretation of a diagnostic test are well-known and serve as an excellent analogy for clinical trials[27] and experimental research in general. This is clearly depicted in the familiar 2 × 2 table shown in **Table 2**.

In diagnostic testing and clinical research, the goal is to make inference about an unknown characteristic of the patient (i.e., Is the patient pregnant?) or an unknown truth about the state of nature (i.e., Does this drug work?), respectively. The diagnostic test is designed to have suitable sensitivity (i.e., ability to identify patients with the characteristic) and specificity (i.e., ability to identify those without the characteristic). In the NHST paradigm, a statistical test is designed to have adequate power (i.e., ability to detect an effect if it exists) while controlling the type 1 error (i.e., limit FP findings). The probability of a type 2 error (denoted β) is (1-power) and such an error in decision making is referred to as a false–negative finding. The sensitivity and specificity that constitutes an acceptable diagnostic test depends on the costs or consequences of the probability of FP and false–negative (FN) findings. At its inception, NHST was also to "decide about α, β, and sample size before the experiment,

Table 2 The analogy of diagnostic testing (bold font) and NHST (italicized font)

		Unknown characteristic/truth *Hypothesis*		
		Present (B^C) *Alternative (H_a)*	**Absent (B)** *Null (H_0)*	
Diagnostic *Statistical* **Test Result**	**Positive (A)** *Significant*	**Sensitivity** *Power – true positive*	*Type 1 error (α)*	**PPV**
	Negative *Not significant*	*Type 2 error (β)*	**Specificity** *True negative*	**NPV**

NHST, null hypothesis significance testing; NPV, negative predictive value; PPV, positive predictive value.

based on subjective cost-benefit considerations"[28] (my emphasis added). This seems to have been lost in the modern rote application of NHST whereby $\alpha = 0.05$ and $\beta = (0.80, 0.90)$ for confirmatory clinical trials, regardless of the medical context or societal circumstance.

In both diagnostic testing and NHST, the *design* of the tests depends on the conditional existence or nonexistence of the characteristic or truth. Sensitivity, specificity, type 1 error, power (and equivalently type 2 error) are conditional probabilities that operate in the "vertical direction" of **Table 2**. That is, sensitivity and power start with an assumption of the characteristic being present or a positive treatment effect, whereas specificity and type 1 error start with the assumption of the characteristic being absent or a null effect.

In diagnostic testing, the sensitivity and specificity can be tuned by changing the cutoff value of the biochemical assay for defining a positive/negative test result, and most often various cutoffs are evaluated to optimize sensitivity and specificity appropriately for a medical condition. For example, if human chorionic gonadotrophin exceeds 50 milli-international units per milliliter of urine in a urine sample from the patient, then it is predicted with 99% probability that she is pregnant. That 50 milli-international units per milliliter of urine can be changed to alter the operating characteristics (i.e., sensitivity and specificity) of the diagnostic test.

In NHST, the sample size may be calculated to meet $\alpha = 0.05$ and $\beta = (0.80, 0.90)$, but rarely are α and β changed. Furthermore, α is used solely to define the critical value (i.e., cutoff) that demarcates the rejection region for the hypothesis test (e.g., 1.96 for $\alpha = 0.05$). Thus, the concepts underlying the *design* of a diagnostic test are identical to those for designing a clinical trial or research experiment under the NHST paradigm.

So, how should one interpret the outcome of a single experiment or clinical trial? Let's continue with the diagnostic testing analogy. The important quantity is the positive predictive value (PPV), which is the conditional probability of the patient having the characteristic given (i.e., assuming) the test result is positive. This conditional probability operates in the "horizontal direction" of **Table 2** and is computed by evaluating the fraction true positives (TPs) relative to all positive findings (i.e. TPs and FPs). Using **Table 2**, if A represents "the diagnostic test is positive," and B^C represents "the patient characteristic is present," then sensitivity is the following.

$pr(A \mid B^C)$ = pr(the diagnostic test is positive | the patient has the characteristic), and PPV is the inverse probability

$pr(B^C \mid A)$ = pr(the patient has the characteristic | the diagnostic test is positive).

Similar statements can be written for negative predictive value (NPV). Algebraically, the PPV is written as

$$\begin{aligned} PPV &= pr(B^C \mid A) = TP/(TP+FP) \\ &= pr(A \mid B^C)pr(B^C) / \left[pr(A \mid B^C)pr(B^C) + pr(A \mid B)\, pr(B)\right]. \end{aligned}$$

This is *precisely* Bayes formula presented previously!

A key element of the PPV formula is $pr(B^C)$, which is known as the prevalence. One way to conceptualize the prevalence of a patient characteristic in a population of interest is to think of it as the probability of a randomly selected patient from that population having the characteristic. In a very real sense, it is the likelihood or belief that a patient has the characteristic of interest *prior* to performing the diagnostic test. It is well known in the diagnostic testing arena that the PPV decreases as the prevalence decreases. For example, for a diagnostic test with 95% sensitivity and 95% specificity, the PPV is given in **Table 3** for various levels of prevalence (i.e., prior probability that a patient has the characteristic of interest).

Neither the physician nor the patient knows with certainty whether the patient has the characteristic or not. The only evidence they have is a diagnostic test result. Thus, when interpreting the diagnostic test result for an individual patient, sensitivity and specificity have little meaning, and the PPV and NPV are the *only* meaningful probabilities to the physician and the patient. Sensitivity and specificity are pertinent when designing a diagnostic test, but PPV or NPV are the quantities of primary interest when interpreting a diagnostic result for an individual patient. As in the pregnancy test example, the 99% probability that the woman is pregnant is the PPV, not the sensitivity or specificity of the diagnostic test.

The analogy with NHST is quite direct for interpretation of clinical trial results as well. In NHST, prevalence is directly analogous to the likelihood that H_0 is false—or equivalently H_a is true—in a

Table 3 PPV as a function of prevalence for a diagnostic test with 95% sensitivity and 95% specificity

Prevalence	PPV
0.50	95%
0.20	90%
0.10	68%
0.05	50%
0.01	16%

PPV, positive predictive value.

clinical trial (i.e., the prior). Furthermore, at the end of a clinical trial, the researcher can only observe whether the statistical test rejects H_0 or not, and the analogous question to diagnostic testing must be asked: "Given that H_0 has been rejected by a suitable statistical test for a given clinical trial design, what is the probability that, in truth, H_0 is false?" The answer is, of course, the Bayesian posterior probability, analogous to PPV and NPV. However, Bayesian posterior probabilities are rarely considered in the inferential paradigm for interpreting clinical trial results. Yet, no one with any understanding of diagnostic testing would ever conclude that there is a 95% probability that a patient has a characteristic of interest if they tested positive when using a diagnostic test with 95% sensitivity. They would use the PPV, which is dependent on the prevalence (**Table 3**). Why then, when we do a sophisticated "diagnostic test," such as a clinical trial, do we conclude we have a positive finding if the observed *result* from a single clinical trial is a P value < 0.05 just because we *designed* the statistical test with a significance level of 0.05? Although the design of a clinical trial uses α and power (1−β) analogous to sensitivity and specificity, the interpretation of an individual clinical trial must be based on the posterior probability of H_0 being false (alternatively H_a being true) in the same way that the interpretation of any individual result from a diagnostic test can only be interpreted using PPV or NPV.

DÉTENTE: THE PEACEFUL CO-EXISTENCE OF *P* VALUES AND BAYESIAN PROBABILITIES

As noted in the above, there have been long-standing debates on the use of frequentist and Bayesian approaches to inference. At times, the arguments on both sides have been quite philosophical or abstract or mathematical, leaving many practitioners and consumers of statistical information confused and alienated, even to the point of abandoning their use.[29]

In fact, the frequentist and Bayesian approaches could be harmonized by taking a diagnostic testing mindset. The design of a clinical trial can be done using the interplay between the α-level and power of the statistical test with resulting sample size and critical value to optimize the performance of the trial. However, when interpreting the results of the trial, a Bayesian evaluation is most appropriate. The additional requirements and complexity of the Bayesian approach lies in the quantification of a prior for H_0 being false, or equivalently H_a being true.

In its simplest form, the prior could be stated as a point probability—a single number in the interval (0,1). For example, for a phase II clinical trial of a new treatment, one may argue the probability that the new treatment works is 0.30. This may be derived from historical data on such treatments in the therapeutic class,[30,31] preclinical models of disease, pharmacokinetic/pharmacodynamic models, the success/failure of other treatments in the same mechanistic class, or other sources of scientific knowledge.[32] There is a full literature on rigorous, scientific elicitation and construction of a prior for a hypothesis of interest.[33]

Using a point prior, there is a simple approximation for computing the probability of H_0 being false using the Bayes Factor Bound (BFB), which is based on reasonable, practical assumptions.[20,34] Let p_0 be the prior probability that H_0 is false and let p = P value from the test of H_0 from the current experiment. Then the BFB is

$$\mathrm{BFB} = 1/\left[-\mathrm{e} * \mathrm{p} * \ln\left(\mathrm{p}\right)\right],$$

and the *upper bound* on the posterior probability that H_0 is false (p_1) given the observed data is

$$p_1 \le p_0 * \mathrm{BFB}/\left(1 - p_0 + p_0 * \mathrm{BFB}\right) \quad (1)$$

In words, this formula contains the prior probability that H_0 is false and the current level of evidence against H_0 to update the probability of H_0 being false after the experiment. This posterior probability is directly related to our belief against H_0, which is decidedly NOT what a P value is, and is more understandable and interpretable. The significance level and the posterior probability are as distinct from each other as sensitivity and PPV. Finally, note that this posterior can be used in constructing a prior for subsequent experiments/trials for hypotheses of the same or similar nature.

To complete the example, with a prior probability of 0.30 that an experimental drug works in a clinical trial (i.e., H_0 is false), suppose our hypothetical phase II clinical trial produces a P value of 0.05. Using Eq. 1, the posterior probability that H_0 is false is ≤ 0.513, perhaps surprisingly less certainty than might be conveyed by a significant P value. This also illustrates why some have noted that a P value of 0.05 is not very strong evidence against H_0.[4] In this case, the increase in probability against H_0 has moved from a prior of 0.30 to a posterior of about 0.50 or 0.20 increase. Thus, there is some modest movement of the evidentiary needle against H_0, but I suspect many would a see a significant phase II result as a reason to believe that a phase III study would quite likely be successful when that simply is not supported by the evidence.

There are more sophisticated approaches that use a full probability distribution of effect size rather than a point probability as a prior, but that is beyond this tutorial. However, the concepts and principles are identical. Equation 1 may serve as a quick or approximate assessment of the likelihood of H_0 being false, but my experience has indicated that it is a particularly good guidepost. Individual scientists may have different priors based on their knowledge, experience, or even bias, leading to different levels of posterior belief. That is OK. What is important is to discuss the sources of prior data and information rather than fixate solely on a single P value from the current clinical trial.

A PATH FORWARD

A logical paradigm for making inference that integrates statistical and scientific thinking can be summarized in three successive questions,[35] each one building on the previous:

1. What do the data say?
2. What do we believe about a hypothesis based on that data?
3. What should we decide?

This section will follow this line of inquiry considering the previous sections.

What do the data say?

In many situations, a P value is a reasonable summary measure of evidence from data collected in a single experiment about a specific null hypothesis. As a piece of evidence, it should be reported as a continuous measure, not dichotomized and declared significant or nonsignificant with the implication that a hypothesis is true or false.[17] Hiding behind labels, such as "$P < 0.05$," or terms, such as "significant," obscures the level of evidence conveyed by the data. Certainly, a P value = 0.001 is greater evidence against H_0 than a P value = 0.04, although both could be reported as "$P < 0.05$" or "statistically significant." The key is that they need to be calibrated using a posterior probability, or at least its upper bound using Eq. 1.

The misinterpretation of P values, which is exacerbated by the dichotomization findings as significant/nonsignificant around $P = 0.05$, has been documented almost since their inception and continues today.[36] They are the probability of observing the data given H_0, rather than what scientists are most interested in understanding—the likelihood of H_0 or H_a given the data.[20] When described in these terms, one can plainly see that when a P value is the final arbiter for making inference and decisions, it is a precise answer to the wrong question. Such over-reliance on P values has been suggested as a contributor to lack of reproducibility of research.[9]

> Recommendation 1: Do not abandon P values; abandon dichotomous labels and conclusions or decisions based solely on P values.

What do we believe?

In the thought experiment, the probability of the biased coin being selected depends on the "prevalence" of biased coins in the bag. Observing 10 consecutive H's results in a probability of having the biased coin of 0.093 and 0.912, depending on whether the prior probability is 1 in 10,000 or 1 in 100, respectively. Thus, for the same observed data or level of evidence ($P = 0.00098$), the observer can have quite different degrees of certainty about whether to make a bet that the biased coin was selected from the bag of coins.

This is not a radical notion. All scientists read new research findings and put them in the context of what they already know. For example, when reading about a new potential treatment for Alzheimer's disease being successful in an animal study or even an early phase human trial, do we not apply some subjective discounting of the results and have lower expectations for success in a large-scale phase III trial? Experience and data tell us that the probability of success in Alzheimer's disease is low from the outset, and, therefore, "significant" findings must always be weighed in that context.

A simple, small step forward in formally calculating our belief in a hypothesis can be achieved using the upper bound on the posterior probability of H_0 being false, given in (Eq. 1). Bayarri *et al*. note this upper bound on H_0 being false is quite reasonable in practice.[37] Using this approach requires no new software or advanced computing but does require the prespecification of a prior probability of the null hypothesis being false. Although this may seem daunting, Wacholder *et al*. state, "The practice of choosing a prior probability may not be quite as unfamiliar as it seems. Investigators already informally use prior probability to decide whether to launch a study, which genes to study, and how to interpret the results. We believe that formally developing prior probabilities before seeing study results can, in itself, lead to a substantial improvement in interpreting study findings over current scientific practice."[38]

Although using Eq. 1 to compute the upper bound on the posterior probability of H_0 being false would be extremely useful for reporting scientific results (far better than a P value alone), there needs to be systematic social change for registering some representation of prior belief as well as updating that belief as scientific knowledge increases during the course of a long trial. Fortunately, clinical trials are registered in clinicatrials.gov, and there should be an extra requirement to state a prior for H_0 being false quantitatively (as a point probability or a full probability distribution) as well as its justification.

In the absence of such registries, as is the case in most preclinical research, such a statement should be a necessary part of the statistical methods section of a publication. This is another small step forward because most research papers have an introduction or background section to describe the history and justification for the current research work. It may be difficult to certify whether the stated priors were done *a priori*, but this simple step could have outsized benefit because editors, reviewers, and the ultimate readership could decide for themselves on the credibility of the stated prior in their overall evaluation of the research findings. Finally, when *post hoc* or exploratory analyses are presented in a paper, they too should be considered in the context of a formal quantification of a prior. Any analyses that are not prespecified would automatically start with an extremely low prior.

As a parallel benefit to stating priors for hypotheses at the outset of a study, there would be less debate whether a study or a finding within a study is "exploratory" or "confirmatory."[39,40] Many studies, especially clinical trials, involve both confirmatory and exploratory analyses, the former being prespecified and the latter being of secondary interest or led by the observed data. In many cases, these labels are used to assess whether a "statistically significant" finding is credible or not. In some cases, journals even prohibit reporting a P value of secondary or exploratory findings,[41] implying that no inference is possible or reasonable. Certainly, researchers will draw conclusions or form some level of belief based on the data, even in the absence of a P value. Why not provide some quantitative description, such as the upper bound on the probability of H_0 being false, as a benchmark for starting the discussion? With a stated prior in place, the labels "confirmatory" and "exploratory" lose their meaning and utility. Equation 1 contains the elements of interest—a synthesis of prior belief and current evidence.

To illustrate further, a so-called confirmatory clinical trial will also have exploratory elements within it. Assume that a new treatment in drug development has had a successful phase II trial and proceeds to phase III with a prior probability of 0.70 that the treatment works, a reasonable assumption for many drug-development programs. Furthermore, there is some interest in exploring the possibility that the treatment effect is more pronounced in a subgroup of patients. The prior for an exceptional treatment effect in that subgroup is based on literature and biological mechanism but was not studied explicitly in phase II. Thus, assume the prior for the hypothesis that the treatment works better in the subgroup is 0.20. Now suppose the

results from the trial produce a P value = 0.03 for the overall "confirmatory" treatment effect and a P value = 0.001 for the "exploratory" subgroup analysis. Using Eq. 1, the upper bound on the posterior probability of H_0 being false is 0.89 for the confirmatory analysis and is 0.93 for the exploratory analysis. Thus, the so-called "confirmatory" result is slightly less convincing than the "exploratory" result, and perhaps the result in the overall trial population is driven by a large effect size in the "exploratory" subgroup. Traditionally, the exploratory subgroup would not be given credence, and recommendations would be made for further clinical trials, a time-consuming and expensive endeavor. With the elimination of such dichotomous labels and the use of quantitative priors at the outset of the trial, interpretation of results can stand on their own based on a meaningful probability assessment of each hypothesis. As noted earlier, the difficult work of prespecifying hypotheses and their related priors is necessary, but such thoughtful work during the conception of the clinical trial can make results more interpretable and potentially save substantial time and money thereafter.

Pharmaceutical companies and regulators sometimes debate whether a clinical trial is a confirmatory trial or not (perhaps based on the strength of phase II data) when, in fact, what is needed is an agreement on what prior should be assigned to whether a drug works.[32] Thus, the practice of stating a prior probability related to H_0 being false eliminates yet another arbitrary and vague dichotomization - confirmatory versus exploratory - that muddles the interpretation of scientific research. Stating the implicit priors, which we all have in our minds, explicitly will make our intentions, beliefs, and subsequent inferences more transparent and perhaps lessen *post hoc* debate about whether a finding is dichotomized as credible or spurious. It is preferable to have a quantified level of belief from which decisions can be made.

> Recommendation 2: Prespecify a prior probability of H_0 being false. State the posterior probability of H_0 being false using, *at a minimum*, Eq. 1. Replace the dichotomous labels of "confirmatory" or "exploratory" with a prior for each prespecified hypothesis and use a posterior probability for evaluating the credibility or strength of evidence for each hypothesis.

What do we decide?

The thought experiment provides another useful insight into this three-question paradigm. When presented with this thought experiment and asked how many H's one needs to see before betting that the biased coin has been selected, some clever students ask, "How much is the bet?"

This is precisely the right question! What we decide depends on a utility function—the cost-benefit of our decision. As these clever students point out, if there is a $5 bet involved, they are willing to make the bet as soon as the Bayesian probability exceeds 50% (i.e., the odds are in their favor). If the bet is for a million dollars, even a 99% Bayesian probability of having drawn the biased coin may not be enough. They reason that even a 1% chance of such a devastating loss is not worth a highly probable windfall.

The same is true in diagnostic testing. What level of cost for FP findings and subsequent actions taken for patients are worth the cost of FN findings? Answering this question can be exceedingly difficult as in the case of screening mammography in asymptomatic women. Every patient has a different risk-benefit calculus, and the US Preventative Services Task Force, which recommends routine mammography screening to begin at age 50, rightly notes, "Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years."[42]

This thinking is generally not an explicit element of the medical literature. Even in trials where the analytical approach is Bayesian, there is rarely formal evaluation of the benefits and harms and their associated costs for making treatment decisions. For example, in the study of therapeutic hypothermia in newborns with hypoxic-ischemic encephalopathy, Bayesian posterior probabilities are calculated as to whether hypothermia treatment is superior to non-cooling treatment in terms of survival or disability.[43] The results showed that the hypothermia treatment had a 76% posterior probability of being superior to non-cooling treatment and a 64% posterior probability of having a clinically meaningful 2% absolute reduction in mortality or disability. This positive effect is in the context of additional adverse events in the hypothermia treatment group. The discussion notes, "A decision to use hypothermia ... will need to consider the probability of benefit, the frequency of adverse events, and the availability of evidence-based alternative treatments." If an institution or a healthcare system decided to implement hypothermia treatment routinely, they would have to calculate the benefits of improved outcomes and the costs of the procedure itself with its increased side effects. The probability that this is a TP finding (76% or 64% depending on the perspective) would then be used to balance this decision against the probability that this is a FP finding (i.e., a 24% chance or 36% chance depending on the perspective) that hypothermia treatment is worse than non-cooling treatment, including the cost of doing the hypothermia procedure unnecessarily with no benefit and costs related to the procedure and its associated adverse events.

In the original design of NHST *for use in decision making*, α and β were to be defined based on the costs of these erroneous decisions. In the quest for simplicity and through sheer force of tradition, the scientific world has settled on nearly uniform values for α and β, regardless of the scientific problem or the societal context in which subsequent decisions are made.

> Recommendation 3: Instead of a clinical study being declared positive/negative based solely on a P value, probabilities against H_0 should be stated[44] with a thoughtful quantification and discussion about the consequences of such a decision being a FP or FN finding.

SUMMARY

One notable dimension of the reproducibility crisis is the confusion and consternation over statistical inference and its impact on scientific findings.[36,45] This stems from (i) the confounding of the significance level of a NHST procedure and the resulting P

value, (ii) the unavoidable fact that evidence is continuous whereas decisions are dichotomous, and (iii) the lack of understanding of conditional probability. As for the latter, one would never confuse pr(rain | cloudy) with pr(cloudy | rain), and yet for nearly a century, the scientific community has used the pr(A | B), the *P* value, as a substitute for pr(B | A), the Bayesian posterior probability, because of the ease and convenience of computing a *P* value. Advances in statistical methodology and computing power now make that substitution obsolete.

Appropriate statistical analysis for obtaining an accurate *P* value is critical for answering "What do the data say?" Additional scientific knowledge—defining priors as objectively as possible—is essential for answering "What do we believe?" Any answers to these questions should exist on a continuum of probability to convey the strength of evidence against H_0. Only then, in the realm of decision making with an appropriate utility function that weighs the cost of FP and FN findings against the benefits of TP and true negative findings should the evidence be distilled into a dichotomous choice for answering "What do we decide?"

The eminent statistician John Tukey wrote, "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."[46] A Bayesian posterior probability may be viewed as subjective or "vague" due to the definition of the prior, but I agree with Tukey that a very precise *P* value, which can also have subjective assumptions, models, and interpretations, fundamentally addresses the wrong question. If we are to make meaningful improvements in our scientific inference, then our statistical inference must be directed at the right question and quantified using Bayesian approaches.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The author wishes to thank two anonymous reviewers for their thoughtful comments, which brought greater clarity to the exposition of the concepts in this tutorial.

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

The author declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

S.J.R. is solely responsible for the content of this paper.

© 2020 The Authors. *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. McGrayne, S.B. *The Theory that Would not Die.* (Yale University Press, New Haven, CT, 2011).
2. Cutler, S.J., Greenhouse, S.W., Cornfield, J. & Schneiderman, M.A. The role of hypothesis testing in clinical trials. *J. Chron. Dis.* **19**, 857–882 (1966).
3. Carver, R.P. The case against statistical significance testing. *Harvard Educ. Rev.* **48**, 378–399 (1978).
4. Berger, J.O. & Sellke, T. Testing a point null hypothesis: the irreconcilability of p values and evidence. *J. Am. Stat. Assoc.* **82**, 112–122 (1987).
5. Goodman, S.N. A comment on replication, P-values and evidence. *Stat. Med.* **11**, 875–879 (1992).
6. Goodman, S.N. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann. Intern. Med.* **130**, 995–1004 (1999).
7. Senn, S.J. Two cheers for P-value. *J. Epidemiol. Biostat.* **6**, 193–204 (2001).
8. Senn, S.J. A comment on replication, p-values and evidence, S.N.Goodman, Statistics in Medicine 1992;11:875-879. *Stat. Med.* **21**, 2437–2444 (2002); author reply 2445–2447.
9. Nuzzo, R. Scientific method: statistical errors. *Nature* **506**, 150–152 (2014).
10. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
11. Fisher, R.A. *Statistical Methods for Research Workers.* (Oliver & Boyd, Edinburgh, 1925).
12. Fisher, R.A. The arrangement of field experiments. *J. Ministry Agric. Great Britain* **33**, 503–513 (1926).
13. Neyman, J. & Pearson, E. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R Soc. A* **231**, 289–337 (1933).
14. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R Soc. London* **53**, 370–418 (1763).
15. Kennedy-Shaffer, L. Before p < 0.05 to beyond p < 0.05: using history to contextualize p-values and significance testing. *Am. Stat.* **73** (suppl. 1), 82–90 (2019).
16. Wasserstein, R.L. & Lazar, N.A. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* **70**, 129–133 (2016).
17. Amrhein, V., Greenland, S. & McShane, B. Retire statistical significance. *Nature* **567**, 305–307 (2019).
18. Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. Moving to a world beyond "p<0.05". *Am. Stat.* **73** (suppl. 1), 1–19 (2019).
19. Ioannidis, J.P.A. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA* **321**, 2067–2068 (2019).
20. Benjamin, D. & Berger, J. Three recommendations for improving the use of p-values. *Am. Stat.* **73** (suppl. 1), 186–191 (2019).
21. Kuffner, T.A. & Walker, S.G. Why are p-values controversial? *Am. Stat.* **73**, 1–3 (2019).
22. Lehmann, E.L. *Testing Statistical Hypotheses.* (Chapman and Hall, New York, 1994).
23. Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Kruger, L. *The Empire of Chance.* (Cambridge University Press, Cambridge, UK, 1989).
24. Hubbard, R. Alphabet soup: blurring the distinctions between p's and α's in psychological research. *Theory Psych.* **14**, 295–327 (2004).
25. Greenland, S. Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values. *Am. Stat.* **73** (suppl. 1), 106–114 (2019).
26. Lew, M.J. Three inferential questions, two types of p-value. *Am. Stat.* **70**, 1–2 (2016).
27. Browner, W.S. & Newman, T.B. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. *JAMA* **257**, 2459–2463 (1987).
28. Gigerenzer, G. Mindless statistics. *J. Socio. Econ.* **33**, 587–606 (2004).
29. Trafimow, D. Editorial. *Basic Appl. Social Psych.* **36**, 1–2 (2014).
30. Hay, M., Thomas, D.W., Craighead, J.L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotech.* **32**, 40–51 (2014).
31. Patel, D.D., Antoni, C., Freedman, S.J., Levesque, M.C. & Sundy, J.S. Phase 2 to phase 3 clinical trial transitions: reasons for

success and failure in immunologic diseases. *J. Allergy Clin. Immun.* **140**, 685–687 (2017).
32. Ruberg, S.J. *et al*. Inference and decision making for 21st-century drug development and approval. *Am. Stat.* **73** (suppl. 1), 319–327 (2019).
33. O'Hagan, A. Expert knowledge elicitation: subjective but scientific. *Am. Stat.* **73** (suppl. 1), 69–81 (2019).
34. Sellke, T., Bayarri, M.J. & Berger, J.O. Calibration of p values for testing precise null hypotheses. *Am. Stat.* **55**, 62–71 (2001).
35. Royall, R.M. *Statistical Evidence: A Likelihood Paradigm.* (Chapman & Hall, London, 1997).
36. McShane, B.B. & Gal, D. Statistical significance and the dichotomization of evidence. *J. Am. Stat. Assoc.* **112**, 885–895 (2017).
37. Bayarri, M.J., Benjamin, D., Berger, J. & Sellke, T. Rejection odds and rejection ratios: a proposal for statistical practice in testing hypotheses. *J. Math. Psych.* **72**, 90–103 (2016).
38. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J. Nat. Cancer Inst.* **96**, 434–442 (2004).
39. Kimmelman, J., Mogil, J.S. & Dirnagl, U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol.* **12**, e1001863 (2014).
40. Schwab, S. & Held, L. Different worlds: confirmatory versus exploratory research. *Signif. (Oxf).* **17**(2), 8–9 (2020).
41. Harrington, D. *et al*. New guidelines for statistical reporting in the journal. *N. Engl. J. Med.* **381**, 286 (2019).
42. Siu, A.L., on behalf of the U.S. Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann. Int. Med.* **164**, 279–296 (2016).
43. Laptook, A.R. *et al*. Effect of therapeutic hypothermia initiated after 6 hours of age on death or disability among newborns with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA* **318**, 1550–1560 (2017).
44. Goodman, S. How sure are you of your result? Put a number on it. *Nature* **564**, 7 (2018).
45. Localio, A.R. *et al*. Inappropriate statistical analysis and reporting in medical research: perverse incentives and institutional solutions. *Ann. Intern. Med.* **169**, 577–578 (2018).
46. Tukey, J.W. The future of data analysis. *Ann. Math. Stat.* **33**, 1–67 (1962).