

## Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy

Steven N. Goodman, MD, PhD

An important problem exists in the interpretation of modern medical research data: Biological understanding and previous research play little formal role in the interpretation of quantitative results. This phenomenon is manifest in the discussion sections of research articles and ultimately can affect the reliability of conclusions. The standard statistical approach has created this situation by promoting the illusion that conclusions can be produced with certain "error rates," without consideration of information from outside the experiment. This statistical approach, the key components of which are *P* values and hypothesis tests, is widely perceived as a mathematically coherent approach to inference. There is little appreciation in the medical community that the methodology is an amalgam of incompatible elements, whose utility for scientific inference has been the subject of intense debate among statisticians for almost 70 years. This article introduces some of the key elements of that debate and traces the appeal and adverse impact of this methodology to the *P* value fallacy, the mistaken idea that a single number can capture both the long-run outcomes of an experiment and the evidential meaning of a single result. This argument is made as a prelude to the suggestion that another measure of evidence should be used—the Bayes factor, which properly separates issues of long-run behavior from evidential strength and allows the integration of background knowledge with statistical findings.

The past decade has seen the rise of evidence-based medicine, a movement that has focused attention on the importance of using clinical studies for empirical demonstration of the efficacy of medical interventions. Increasingly, physicians are being called on to assess such studies to help them make clinical decisions and understand the rationale behind recommended practices. This type of assessment requires an understanding of research methods that until recently was not expected of physicians.

These research methods include statistical techniques used to assist in drawing conclusions. However, the methods of statistical inference in current use are not "evidence-based" and thus have contributed to a widespread misperception. The misperception is that absent any consideration of biological plausibility and prior evidence, statistical methods can provide a number that by itself reflects a probability of reaching erroneous conclusions. This belief has damaged the quality of scientific reasoning and discourse, primarily by making it difficult to understand how the strength of the evidence in a particular study can be related to and combined with the strength of other evidence (from other laboratory or clinical studies, scientific reasoning, or clinical experience). This results in many knowledge claims that do not stand the test of time (1, 2).

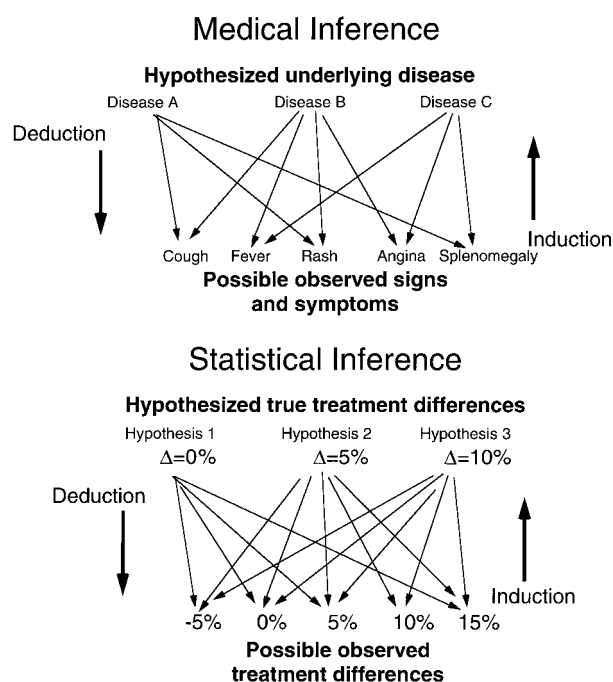
A pair of articles in this issue examines this problem in some depth and proposes a partial solution. In this article, I explore the historical and logical foundations of the dominant school of medical statistics, sometimes referred to as *frequentist statistics*, which might be described as error-based. I explicate the logical fallacy at the heart of this system and the reason that it maintains such a tenacious hold on the minds of investigators, policymakers, and journal editors. In the second article (3), I present an evidence-based approach derived from Bayesian statistical methods, an alternative perspective that has been one of the most active areas of biostatistical development during the past 20 years. Bayesian methods have started to make inroads into medical

This paper is also available at <http://www.acponline.org>.

*Ann Intern Med.* 1999;130:995-1004.

From Johns Hopkins University School of Medicine, Baltimore, Maryland. For the current author address, see end of text.

**See related article on pp 1005-1013 and editorial comment on pp 1019-1021.**



**Figure 1.** The parallels between the processes of induction and deduction in medical inference (top) and statistical inference (bottom).  $\Delta$  = treatment difference.

journals; *Annals*, for example, has included a section on Bayesian data interpretation in its Information for Authors section since 1 July 1997.

The perspective on Bayesian methods offered here will differ somewhat from that in previous presentations in other medical journals. It will focus not on the controversial use of these methods in measuring “belief” but rather on how they measure the weight of quantitative evidence. We will see how reporting an index called the *Bayes factor* (which in its simplest form is also called a *likelihood ratio*) instead of the *P* value can facilitate the integration of statistical summaries and biological knowledge and lead to a better understanding of the role of scientific judgment in the interpretation of medical research.

### An Example of the Problem

A recent randomized, controlled trial of hydrocortisone treatment for the chronic fatigue syndrome showed a treatment effect that neared the threshold for statistical significance,  $P = 0.06$  (4). The discussion section began, “. . . hydrocortisone treatment was associated with an improvement in symptoms . . . This is the first such study . . . to demonstrate improvement with a drug treatment of [the chronic fatigue syndrome]” (4).

What is remarkable about this paper is how unremarkable it is. It is typical of many medical research reports in that a conclusion based on the findings is stated at the beginning of the discussion.

Later in the discussion, such issues as biological mechanism, effect magnitude, and supporting studies are presented. But a conclusion is stated before the actual discussion, as though it is derived directly from the results, a mere linguistic transformation of  $P = 0.06$ . This is a natural consequence of a statistical method that has almost eliminated our ability to distinguish between statistical results and scientific conclusions. We will see how this is a natural outgrowth of the “*P* value fallacy.”

### Philosophical Preliminaries

To begin our exploration of the *P* value fallacy, we must consider the basic elements of reasoning. The process that we use to link underlying knowledge to the observed world is called *inferential reasoning*, of which there are two logical types: *deductive inference* and *inductive inference*. In deductive inference, we start with a given hypothesis (a statement about how nature works) and predict what we should see if that hypothesis were true. Deduction is objective in the sense that the predictions about what we will see are always true if the hypotheses are true. Its problem is that we cannot use it to expand our knowledge beyond what is in the hypotheses.

Inductive inference goes in the reverse direction: On the basis of what we see, we evaluate what hypothesis is most tenable. The concept of evidence is inductive; it is a measure that reflects back from observations to an underlying truth. The advantage of inductive reasoning is that our conclusions about unobserved states of nature are broader than the observations on which they are based; that is, we use this reasoning to generate new hypotheses and to learn new things. Its drawback is that we cannot be sure that what we conclude about nature is actually true, a conundrum known as the *problem of induction* (5–7).

From their clinical experience, physicians are acutely aware of the subtle but critical difference between these two perspectives. Enumerating the frequency of symptoms (observations) given the known presence of a disease (hypothesis) is a deductive process and can be done by a medical student with a good medical textbook (Figure 1, top). Much harder is the inductive art of differential diagnosis: specifying the likelihood of different diseases on the basis of a patient’s signs, symptoms, and laboratory results. The deductions are more certain and “objective” but less useful than the inductions.

The identical issue arises in statistics. Under the assumption that two treatments are the same (that is, the hypothesis of no difference in efficacy is true), it is easy to calculate deductively the fre-

quency of all possible outcomes that we could observe in a study (**Figure 1, bottom**). But once we observe a particular outcome, as in the result of a clinical trial, it is not easy to answer the more important inductive question, “How likely is it that the treatments are equivalent?”

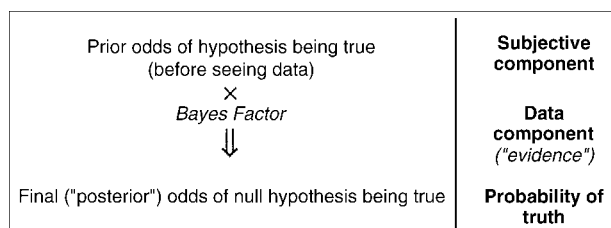
In this century, philosophers have grappled with the problem of induction and have tried to solve or evade it in several ways. Karl Popper (8) proposed a philosophy of scientific practice that eliminated formal induction completely and used only the deductive elements of science: the prediction and falsification components. Rudolf Carnap tried an opposite strategy—to make the inductive component as logically secure as the deductive part (9, 10). Both were unsuccessful in producing workable models for how science could be conducted, and their failures showed that there is no methodologic solution to the problem of fallible scientific knowledge.

Determining which underlying truth is most likely on the basis of the data is a problem in inverse probability, or inductive inference, that was solved quantitatively more than 200 years ago by the Reverend Thomas Bayes. He withheld his discovery, now known as *Bayes theorem*; it was not divulged until 1762, 20 years after his death (11). **Figure 2** shows Bayes theorem in words.

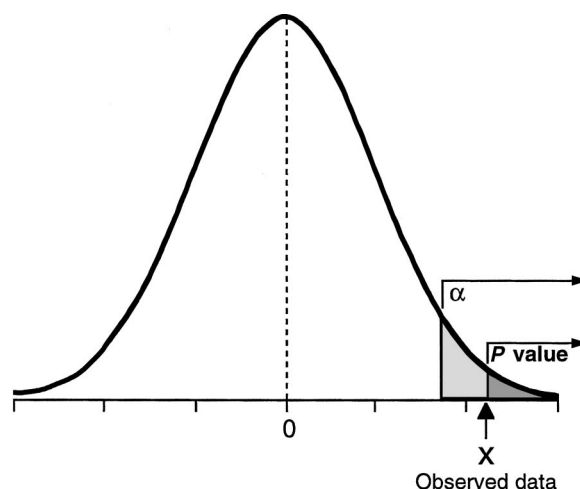
As a mathematical equation, Bayes theorem is not controversial; it serves as the foundation for analyzing games of chance and medical screening tests. However, as a model for how we should think scientifically, it is criticized because it requires assigning a prior probability to the truth of an idea, a number whose objective scientific meaning is unclear (7, 10, 12). It is speculated that this may be why Reverend Bayes chose the more dire of the “publish or perish” options. It is also the reason why this approach has been tarred with the “subjective” label and has not generally been used by medical researchers.

### Conventional (Frequentist) Statistical Inference

Because of the subjectivity of the prior probabilities used in Bayes theorem, scientists in the 1920s and 1930s tried to develop alternative approaches to



**Figure 2.** Bayes theorem, in words.



**Figure 3.** The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both  $\alpha$  (the type I error rate) and the  $P$  value are “tail areas” under this curve. The tail area for  $\alpha$  is set before the experiment, and a result can fall anywhere within it. The  $P$  value tail area is known only after a result is observed, and, by definition, the result will always lie on the border of that area.

statistical inference that used only deductive probabilities, calculated with mathematical formulas that described (under certain assumptions) the frequency of all possible experimental outcomes if an experiment were repeated many times (10). Methods based on this “frequentist” view of probability included an index to measure the strength of evidence called the  $P$  value, proposed by R.A. Fisher in the 1920s (13), and a method for choosing between hypotheses, called a hypothesis test, developed in the early 1930s by the mathematical statisticians Jerzy Neyman and Egon Pearson (14). These two methods were incompatible but have become so intertwined that they are mistakenly regarded as part of a single, coherent approach to statistical inference (6, 15, 16).

### The $P$ Value

The  $P$  value is defined as the probability, under the assumption of no effect or no difference (the *null hypothesis*), of obtaining a result equal to or more extreme than what was actually observed (**Figure 3**). Fisher proposed it as an informal index to be used as a measure of discrepancy between the data and the null hypothesis. It was not part of a formal inferential method. Fisher suggested that it be used as part of the fluid, non-quantifiable process of drawing conclusions from observations, a process that included combining the  $P$  value in some unspecified way with background information (17).

It is worth noting one widely prevalent and particularly unfortunate misinterpretation of the  $P$  value (18–21). Most researchers and readers think that a  $P$  value of 0.05 means that the null hypothesis has a probability of only 5%. In my experience teaching many academic physicians, when physi-

cians are presented with a single-sentence summary of a study that produced a surprising result with  $P = 0.05$ , the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect. This is an understandable but categorically wrong interpretation because the  $P$  value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true. Innumerable authors have tried to correct this misunderstanding (18, 20). Diamond and Forrester (19) reanalyzed several large clinical trials, and Brophy and Joseph (22) revisited the GUSTO (Global Use of Streptokinase and tPA for Occluded Coronary Arteries) trial to show that the final probability of no effect, which can be calculated only with Bayesian methods, can differ greatly from the  $P$  value. However, serious as that issue is, this article will focus on the subtler and more vexing problems created by using the  $P$  value as it was originally intended: as a measure of inductive evidence.

When it was proposed, some scientists and statisticians attacked the logical basis and practical utility of Fisher's  $P$  value (23, 24). Perhaps the most powerful criticism was that it was a measure of evidence that did not take into account the size of the observed effect. A small effect in a study with large sample size can have the same  $P$  value as a large effect in a small study. This criticism is the foundation for today's emphasis on confidence intervals rather than  $P$  values (25–28). Ironically, the  $P$  value was effectively immortalized by a method designed to supplant it: the hypothesis testing approach of Neyman and Pearson.

### Hypothesis Tests

Neyman and Pearson saw Fisher's  $P$  value as an incomplete answer to the problem of developing an inferential method without Bayes theorem. In their hypothesis test, one poses *two* hypotheses about nature: a null hypothesis (usually a statement that there is a null effect) and an alternative hypothesis, which is usually the opposite of the null hypothesis (for example, that there is a nonzero effect). The outcome of a hypothesis test was to be a behavior, not an inference: to reject one hypothesis and accept the other, solely on the basis of the data. This puts the researcher at risk for two types of errors—behaving as though two therapies differ when they are actually the same (also known as a *false-positive result*, a *type I error*, or an  $\alpha$  error [Figure 3]) or concluding that they are the same when in fact they differ (also known as a *false-negative result*, a *type II error*, or a  $\beta$  error).

This approach has the appeal that if we assume an underlying truth, the chances of these errors can be calculated with mathematical formulas, deductively and therefore “objectively.” Elements of judgment were intended to be used in the hypothesis test: for example, the choice of false-negative and false-positive error rates on the basis of the relative seriousness of the two types of error (12, 14, 29). Today, these judgments have unfortunately disappeared.

The hypothesis test represented a dramatic change from previous methods in that it was a procedure that essentially dictated the actions of the researcher. Mathematically and conceptually, it was an enormous step forward, but as a model for scientific practice, it was problematic. In particular, it did not include a measure of evidence; no number reflected back from the data to the underlying hypotheses. The reason for this omission was that any inductive element would inevitably lead back to Bayes theorem, which Neyman and Pearson were trying to avoid. Therefore, they proposed another goal of science: not to reason inductively in single experiments but to use deductive methods to limit the number of mistakes made over many different experiments. In their words (14),

no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.

But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong.

It is hard to overstate the importance of this passage. In it, Neyman and Pearson outline the price that must be paid to enjoy the purported benefits of objectivity: We must abandon our ability to measure evidence, or judge truth, in an individual experiment. In practice, this meant reporting only whether or not the results were statistically significant and acting in accordance with that verdict. Many might regard this as profoundly nonscientific, yet this procedure is often held up as a paradigm of the scientific method.

Hypothesis tests are equivalent to a system of justice that is not concerned with which individual defendant is found guilty or innocent (that is, “whether each separate hypothesis is true or false”) but tries instead to control the overall number of incorrect verdicts (that is, “in the long run of experience, we shall not often be wrong”). Controlling mistakes in the long run is a laudable goal, but just as our sense of justice demands that individual persons be correctly judged, scientific intuition says that we should try to draw the proper conclusions from individual studies.



The hypothesis test approach offered scientists a Faustian bargain—a seemingly automatic way to limit the number of mistaken conclusions in the long run, but only by abandoning the ability to measure evidence and assess truth from a single experiment. It is doubtful that hypothesis tests would have achieved their current degree of acceptance if something had not been added that let scientists mistakenly think they could avoid that trade-off. That something turned out to be Fisher's "*P* value," much to the dismay of Fisher, Neyman, Pearson, and many experts on statistical inference who followed.

### The *P* Value "Solution"

How did the *P* value seem to solve an insoluble problem? It did so in part by appearing to be a measure of evidence in a single experiment that did not violate the long-run logic of the hypothesis test. **Figure 3** shows how similar the *P* value and the  $\alpha$  value (the false-positive error rate) appear. Both are tail-area probabilities under the null hypothesis. The tail area corresponding to the false-positive error rate ( $\alpha$ ) of the hypothesis test is fixed before the experiment begins (almost always at 0.05), whereas the *P* value tail area starts from a point determined by the data. Their superficial similarity makes it easy to conclude that the *P* value is a special kind of false-positive error rate, specific to the data in hand. In addition, using Fisher's logic that the *P* value measured how severely the null hypothesis was contradicted by the data (that is, it could serve as a measure of evidence against the null hypothesis), we have an index that does double duty. It seems to be a Neyman-Pearson data-specific, false-positive error rate and a Fisher measure of evidence against the null hypothesis (6, 15, 17).

A typical passage from a standard biostatistics text, in which the type I error rate is called a "significance level," shows how easily the connection between the *P* value and the false-positive error rate is made (30):

The statement " $P < 0.01$ " indicates that the discrepancy between the sample mean and the null hypothesis mean is significant even if such a conservative significance level as 1 percent is adopted. The statement " $P = 0.006$ " indicates that the result is significant at any level up to 0.6 percent.

The plausibility of this dual evidence/error-rate interpretation is bolstered by our intuition that the more evidence our conclusions are based on, the less likely we are to be in error. This intuition is correct, but the question is whether we can use a single number, a probability, to represent both the strength of the evidence against the null hypothesis

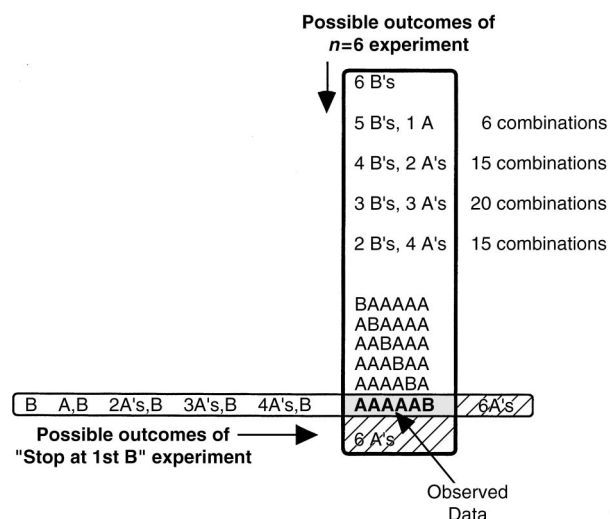
and the frequency of false-positive error under the null hypothesis. If so, then Neyman and Pearson must have erred when they said that we could not both control long-term error rates and judge whether conclusions from individual experiments were true. But they were not wrong; it is not logically possible.

### The *P* Value Fallacy

The idea that the *P* value can play both of these roles is based on a fallacy: that an event can be viewed simultaneously both from a long-run and a short-run perspective. In the long-run perspective, which is error-based and deductive, we group the observed result together with other outcomes that might have occurred in hypothetical repetitions of the experiment. In the "short run" perspective, which is evidential and inductive, we try to evaluate the meaning of the observed result from a single experiment. If we could combine these perspectives, it would mean that inductive ends (drawing scientific conclusions) could be served with purely deductive methods (objective probability calculations).

These views are not reconcilable because a given result (the short run) can legitimately be included in many different long runs. A classic statistical puzzle demonstrating this involves two treatments, A and B, whose effects are contrasted in each of six patients. Treatment A is better in the first five patients and treatment B is superior in the sixth patient. Adopting Royall's formulation (6), let us imagine that this experiment were conducted by two investigators, each of whom, unbeknownst to the other, had a different plan for the experiment. An investigator who originally planned to study six patients would calculate a *P* value of 0.11, whereas one who planned to stop as soon as treatment B was preferred (up to a maximum of six patients) would calculate a *P* value of 0.03 (Appendix). We have the same patients, the same treatments, and the same outcomes but two very different *P* values (which might produce different conclusions), which differ only because the experimenters have different mental pictures of what the results could be if the experiment were repeated. A confidence interval would show this same behavior.

This puzzling and disturbing result comes from the attempt to describe long-run behavior and short-run meaning by using the same number. **Figure 4** illustrates all of the outcomes that could have occurred under the two investigators' plans for the experiment: that is, in the course of the long run of each design. The long runs of the two designs differ greatly and in fact have only two possible results in common: the observed one and the six treatment A



**Figure 4.** Possible outcomes of two hypothetical trials in six patients (Appendix). The only possible overlapping results are the observed data and the result in which treatment A was preferred in all patients.

preferences. When we group the observed result with results from the different long runs, we get two different  $P$  values (Appendix).

Another way to explain the  $P$  value fallacy is that a result cannot at the same time be an anonymous (interchangeable) member of a group of results (the long-run view) and an identifiable (unique) member (the short-run view) (6, 15, 31). In my second article in this issue, we will see that if we stick to the short-run perspective when measuring evidence, identical data produce identical evidence regardless of the experimenters' intentions.

Almost every situation in which it is difficult to calculate the "correct"  $P$  value is grounded in this fundamental problem. The *multiple comparisons debate* is whether a comparison should be considered part of a group of all comparisons made (that is, as an anonymous member) or separately (as an identifiable member) (32–35). The controversy over how to cite a  $P$  value when a study is stopped because of a large treatment effect is about whether we consider the result alone or as part of all results that might have arisen from such monitoring (36–39). In a trial of extracorporeal membrane oxygenation in infants, a multitude of  $P$  values were derived from the same data (40). This problem also has implications for the design of experiments. Because frequentist inference requires the "long run" to be unambiguous, frequentist designs need to be rigid (for example, requiring fixed sample sizes and pre-specified stopping rules), features that many regard as requirements of science rather than as artifacts of a particular inferential philosophy.

The  $P$  value, in trying to serve two roles, serves neither one well. This is seen by examining the statement that "a result with  $P = 0.05$  is in a group of outcomes that has a 5% chance of oc-

curring under the null hypothesis." Although that is literally the case, we know that the result is not just *in* that group (that is, anonymous); we know where it is, and we know that it is the most probable member (that is, it is identifiable). It is *in* that group in the same way that a student who ranks 10 out of 100 is *in* the top 10% of the class, or one who ranks 20th is *in* the top 20% (15). Although literally true, these statements are deceptive because they suggest that a student could be anywhere in a top fraction when we know he or she is at the lowest level of that top group. This same property is part of what makes the  $P$  value an inappropriate measure of evidence against the null hypothesis. As will be explored in some depth in the second article, the evidential strength of a result with a  $P$  value of 0.05 is actually much weaker than the number 0.05 suggests.

If the  $P$  value fallacy were limited to the realm of statistics, it would be a mere technical footnote, hardly worth an extended exposition. But like a single gene whose abnormality can disrupt the functioning of a complex organism, the  $P$  value fallacy allowed the creation of a method that amplified the fallacy into a conceptual error that has profoundly influenced how we think about the process of science and the nature of scientific truth.

### Creation of a Combined Method

The structure of the  $P$  value and the subtlety of the fallacy that it embodied enabled the combination of the hypothesis test and  $P$  value approaches. This combination method is characterized by setting the type I error rate (almost always 5%) and power (almost always  $\geq 80\%$ ) before the experiment, then calculating a  $P$  value and rejecting the null hypothesis if the  $P$  value is less than the preset type I error rate.

The combined method appears, completely deductively, to associate a probability (the  $P$  value) with the null hypothesis within the context of a method that controls the chances of errors. The key word here is *probability*, because a probability has an absoluteness that overwhelms caveats that it is not a probability of truth or that it should not be used mechanically. Such features as biological plausibility, the cogency of the theory being tested, and the strength of previous results all become mere side issues of unclear relevance. None of these change the probability, and the probability does not need them for interpretation. Thus, we have an objective inference calculus that manufactures conclusions seemingly without paying Neyman and Pearson's price (that is, that it not be used to draw conclusions from individual studies) and without

Fisher's flexibility (that is, that background knowledge be incorporated).

In didactic articles in the biomedical literature, the fusion of the two approaches is so complete that sometimes no combination is recognized at all; the *P* value is identified as equivalent to the chance of a false-positive error. In a tutorial on statistics for surgeons, under the unwittingly revealing subheading of "Errors in statistical inference," we are told that "Type I error is incurred if  $H_0$  [the null hypothesis] is falsely rejected, and the probability of this corresponds to the familiar P-value" (41).

The originators of these approaches—Fisher, Neyman, and Pearson—were acutely aware of the implications of their methods for science, and while they each fought for their own approaches in a debate characterized by rhetorical vehemence and sometimes personal attacks (15, 16), neither side condoned the combined method. However, the two approaches somehow were blended into a received method whose internal inconsistencies and conceptual limitations continue to be widely ignored. Many sources on statistical theory make the distinctions outlined here (42–45), but in applied texts and medical journals, the combined method is typically presented anonymously as an abstract mathematical truth, rarely with a hint of any controversy. Of note, because the combined method is not a coherent body of ideas, it has been adapted in different forms in diverse applied disciplines, such as psychology, physics, economics, and genetic epidemiology (16).

A natural question is, What drove this method to be so widely promoted and accepted within medicine and other disciplines? Although the scholarship addressing that question is not yet complete, recent books by Marks (46), Porter (47), Matthews (48), and Gigerenzer and colleagues (16) have identified roles for both scientific and sociologic forces. It is a complex story, but the basic theme is that therapeutic reformers in academic medicine and in government, along with medical researchers and journal editors, found it enormously useful to have a quantitative methodology that ostensibly generated conclusions independent of the persons performing the experiment. It was believed that because the methods were "objective," they necessarily produced reliable, "scientific" conclusions that could serve as the bases for therapeutic decisions and government policy.

This method thus facilitated a subtle change in the balance of medical authority from those with knowledge of the biological basis of medicine toward those with knowledge of quantitative methods, or toward the quantitative results alone, as though the numbers somehow spoke for themselves. This is manifest today in the rise of the evidence-based medicine paradigm, which occasionally raises hackles by suggesting that information about biological

mechanisms does not merit the label "evidence" when medical interventions are evaluated (49–51).

### Implications for Interpretation of Medical Research

This combined method has resulted in an automaticity in interpreting medical research results that clinicians, statisticians, and methodology-oriented researchers have decried over the years (18, 52–68). As A.W.F. Edwards, a statistician, geneticist, and protégé of R.A. Fisher, trenchantly observed,

What used to be called judgment is now called prejudice, and what used to be called prejudice is now called a null hypothesis... it is dangerous nonsense (dressed up as the 'scientific method') and will cause much trouble before it is widely appreciated as such (69).

Another statistician worried about the "unintentional brand of tyranny" that statistical procedures exercise over other ways of thinking (70).

The consequence of this "tyranny" is weakened discussion sections in research articles, with background information and previous empirical evidence integrated awkwardly, if at all, with the statistical results. A recent study of randomized, controlled trials reported in major medical journals showed that very few referred to the body of previous evidence from such trials in the same field (71). This is the natural result of a methodology that suggests that each study alone generates conclusions with certain error rates instead of adding evidence to that provided by other sources and other studies.

The example presented at the start of this article was not chosen because it was unusually flawed but because it was a typical example of how this problem manifests in the medical literature. The statement that there was a relation between hydrocortisone treatment and improvement of the chronic fatigue syndrome was a knowledge claim, an inductive inference. To make such a claim, a bridge must be constructed between " $P = 0.06$ " and "treatment was associated with improvement in symptoms." That bridge consists of everything that the authors put into the latter part of their discussion: the magnitude of the change (small), the failure to change other end points, the absence of supporting studies, and the weak support for the proposed biological mechanism. Ideally, all of this other information should have been combined with the modest statistical evidence for the main end point to generate a conclusion about the likely presence or absence of a true hydrocortisone effect. The authors did recommend against the use of the treatment, primarily because the risk for adrenal suppression could outweigh the small beneficial effect, but the claim for the benefit of hydrocortisone remained.

Another interesting feature of that presentation was that the magnitude of the  $P$  value seemed to play almost no role. The initial conclusion was phrased no differently than if the  $P$  value had been less than 0.001. This omission is the legacy of the hypothesis test component of the combined method of inference. The authors (and journal) are to be lauded for not hewing rigidly to hypothesis test logic, which would dismiss the  $P$  value of 0.06 as nonsignificant, but if one does not use the hypothesis test framework, conclusions must incorporate the graded nature of the evidence. Unfortunately, even Fisher could offer little guidance on how the size of a  $P$  value should affect a conclusion, and neither has anyone else. In contrast, we will see in the second article how Bayes factors offer a natural way to incorporate different grades of evidence into the formation of conclusions.

In practice, what is most often done to make the leap from evidence to inference is that different verbal labels are assigned to  $P$  values, a practice whose incoherence is most apparent when the “significance” verdict is not consistent with external evidence or the author’s beliefs. If a  $P$  value of 0.12 is found for an a priori unsuspected difference, an author often says that the groups are “equivalent” or that there was “no difference.” But the same  $P$  value found for an expected difference results in the use of words such as “trend” or “suggestion,” a claim that the study was “not significant because of small sample size,” or an intensive search for alternative explanations. On the other hand, an unexpected result with a  $P$  value of 0.01 may be declared a statistical fluke arising from data dredging or perhaps uncontrolled confounding. Perhaps worst is the practice that is most common: accepting at face value the significance verdict as a binary indicator of whether or not a relation is real. What drives all of these practices is a perceived need to make it appear that conclusions are being drawn directly from the data, without any external influence, because direct inference from data to hypothesis is thought to result in mistaken conclusions only rarely and is therefore regarded as “scientific.” This idea is reinforced by a methodology that puts numbers—a stamp of legitimacy—on that misguided approach.

Many methodologic disputes in medical research, such as those around multiple comparisons, whether a hypothesis was thought of before or after seeing the data, whether an endpoint is primary or secondary, or how to handle multiple looks at accumulating data, are actually substantive scientific disagreements that have been converted into pseudostatistical debates. The technical language and substance of these debates often exclude the investigators who may have the deepest insight into the biological issues. A vivid example is found in a recent series of

articles reporting on a U.S. Food and Drug Administration committee debate on the approval of carvedilol, a cardiovascular drug, in which the discussion focused on whether (and which) statistical “rules” had been broken (72–74). Assessing and debating the cogency of disparate real-world sources of laboratory and clinical evidence are the heart of science, and conclusions can be drawn only when that assessment is combined with statistical results. The combination of hypothesis testing and  $P$  values offers no way to accomplish this critical task.

## Proposed Solutions

Various remedies to the problems discussed thus far have been proposed (18, 52–67). Most involve more use of confidence intervals and various allotments of common sense. Confidence intervals, derived from the same frequentist mathematics as hypothesis tests, represent the range of effects that are “compatible with the data.” Their chief asset is that, ideally, they push us away from the automaticity of  $P$  values and hypothesis tests by promoting a consideration of the size of the observed effect. They are cited more often in medical research reports today than in the past, but their impact on the interpretation of research is less clear. Often, they are used simply as surrogates for the hypothesis test (75); researchers simply see whether they include the null effect rather than consider the clinical implications of the full range of likely effect size. The few efforts to eliminate  $P$  values from journals in favor of confidence intervals have not generally been successful, indicating that researchers’ need for a measure of evidence remains strong and that they often feel lost without one (76, 77). But confidence intervals are far from a panacea; they embody, albeit in subtler form, many of the same problems that afflict current methods (78), the most important being that they offer no mechanism to unite external evidence with that provided by an experiment. Thus, although confidence intervals are a step in the right direction, they are not a solution to the most serious problem created by frequentist methods. Other recommended solutions have included likelihood or Bayesian methods (6, 19, 20, 79–84). The second article will explore the use of Bayes factor—the Bayesian measure of evidence—and show how this approach can change not only the numbers we report but, more important, how we think about them.

## A Final Note

Some of the strongest arguments in support of standard statistical methods is that they are a great improvement over the chaos that preceded them



and that they have proved enormously useful in practice. Both of these are true, in part because statisticians, armed with an understanding of the limitations of traditional methods, interpret quantitative results, especially *P* values, very differently from how most nonstatisticians do (67, 85, 86). But in a world where medical researchers have access to increasingly sophisticated statistical software, the statistical complexity of published research is increasing (87–89), and more clinical care is being driven by the empirical evidence base, a deeper understanding of statistics has become too important to leave only to statisticians.

### Appendix: Calculation of *P* Value in a Trial Involving Six Patients

*Null hypothesis:* Probability that treatment A is better = 1/2

*The n = 6 design:* The probability of the observed result (one treatment B success and five treatment A successes) is  $6 \times (1/2) \times (1/2)^5$ . The factor “6” appears because the success of treatment B could have occurred in any of the six patients. The more extreme result would be the one in which treatment A was superior in all six patients, with a probability (under the null hypothesis) of  $(1/2)^6$ . The one-sided *P* value is the sum of those two probabilities:

$$\underbrace{6 \frac{1^5}{2} \frac{1}{2}}_{\text{Probability of observed data}} + \underbrace{\frac{1^6}{2}}_{\text{Probability of "more extreme" data}} = 0.11$$

*“Stop at first treatment B preference” design:* The possible results of such an experiment would be either a single instance of preference for treatment B or successively more preferences for treatment A, followed by a case of preference for treatment B, up to a total of six instances. With the same data as before, the probability of the observed result of 5 treatment A preferences – 1 treatment B preference would be  $(1/2)^5 \times (1/2)$  (without the factor of “6” because the preference for treatment B must always fall at the end) and the more extreme result would be six preferences for treatment As, as in the other design. The one-sided *P* value is:

$$\underbrace{\frac{1^5}{2} \frac{1}{2}}_{\text{Probability of observed data}} + \underbrace{\frac{1^6}{2}}_{\text{Probability of "more extreme" data}} = 0.03$$

*Requests for Reprints:* Steven Goodman, MD, PhD, Johns Hopkins University, 550 North Broadway, Suite 409, Baltimore, MD 21205; e-mail, sgoodman@jhu.edu.

## References

1. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology [Editorial]. *Br J Cancer*. 1994;69:979-85.
2. Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *J Natl Cancer Inst*. 1996;88:206-7.
3. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130:1005-13.
4. McKenzie R, O'Fallon A, Dale J, Demitrack M, Sharma G, Deloria M, et al. Low-dose hydrocortisone for treatment of chronic fatigue syndrome: a randomized controlled trial. *JAMA*. 1998;280:1061-6.
5. Salmon WC. *The Foundations of Scientific Inference*. Pittsburgh: Univ of Pittsburgh Pr; 1966.
6. Royall R. *Statistical Evidence: A Likelihood Primer*. Monographs on Statistics and Applied Probability #71. London: Chapman and Hall; 1997.
7. Hacking I. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge, UK: Cambridge Univ Pr; 1975.
8. Popper K. *The Logic of Scientific Discovery*. New York: Harper & Row; 1934:59.
9. Carnap R. *Logical Foundations of Probability*. Chicago: Univ of Chicago Pr; 1950.
10. Howson C, Urbach P. *Scientific Reasoning: The Bayesian Approach*. 2d ed. La Salle, IL: Open Court; 1993.
11. Stigler SM. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard Univ Pr; 1986.
12. Oakes M. *Statistical Inference: A Commentary for the Social Sciences*. New York: Wiley; 1986.
13. Fisher R. *Statistical Methods for Research Workers*. 13th ed. New York: Hafner; 1958.
14. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*. 1933;231:289-337.
15. Goodman SN. *p* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol*. 1993;137:485-96.
16. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. *The Empire of Chance*. Cambridge, UK: Cambridge Univ Pr; 1989.
17. Fisher R. *Statistical Methods and Scientific Inference*. 3d ed. New York: Macmillan; 1973.
18. Browner W, Newman T. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA*. 1987;257:2459-63.
19. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med*. 1983;98:385-94.
20. Lilford RJ, Braunholtz D. For debate: The statistical basis of public policy: a paradigm shift is overdue. *BMJ*. 1996;313:603-7.
21. Freeman PR. The role of *p*-values in analysing trial results. *Stat Med*. 1993;12:1442-552.
22. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA*. 1995;273:871-5.
23. Berkson J. Tests of significance considered as evidence. *Journal of the American Statistical Association*. 1942;37:325-35.
24. Pearson E. 'Student' as a statistician. *Biometrika*. 1938;38:210-50.
25. Altman DG. Confidence intervals in research evaluation. *ACP J Club*. 1992; Suppl 2:A28-9.
26. Berry G. Statistical significance and confidence intervals [Editorial]. *Med J Aust*. 1986;144:618-9.
27. Braitman LE. Confidence intervals extract clinically useful information from data [Editorial]. *Ann Intern Med*. 1988;108:296-8.
28. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med*. 1986;105:429-35.
29. Pearson E. Some thoughts on statistical inference. *Annals of Mathematical Statistics*. 1962;33:394-403.
30. Colton T. *Statistics in Medicine*. Boston: Little, Brown; 1974.
31. Seidenfeld T. *Philosophical Problems of Statistical Inference*. Dordrecht, the Netherlands: Reidel; 1979.
32. Goodman S. Multiple comparisons, explained. *Am J Epidemiol*. 1998;147:807-12.
33. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol*. 1995;142:904-8.
34. Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol*. 1985;122:1080-95.
35. Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*. 1991;2:244-51.
36. Anscombe F. Sequential medical trials. *Journal of the American Statistical Association*. 1963;58:365-83.
37. Dupont WD. Sequential stopping rules and sequentially adjusted *P* values: does one require the other? *Controlled Clin Trials*. 1983;4:3-10.
38. Cornfield J, Greenhouse S. On certain aspects of sequential clinical trials. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: Univ of California Pr; 1977;4:813-29.
39. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *American Statistician*. 1966;20:18-23.
40. Begg C. On inferences from Wei's biased coin design for clinical trials. *Biometrika*. 1990;77:467-84.
41. Ludbrook J, Dudley H. Issues in biomedical statistics: statistical inference. *Aust N Z J Surg*. 1994;64:630-6.

42. **Cox D, Hinkley D.** Theoretical Statistics. New York: Chapman and Hall; 1974.
43. **Barnett V.** Comparative Statistical Inference. New York: Wiley; 1982.
44. **Lehmann E.** The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association.* 1993;88:1242-9.
45. **Berger J.** The frequentist viewpoint and conditioning. In: LeCam L, Olshen R, eds. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* vol. 1. Belmont, CA: Wadsworth; 1985:15-43.
46. **Marks HM.** The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990. Cambridge, UK: Cambridge Univ Pr; 1997.
47. **Porter TM.** Trust In Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton, NJ: Princeton Univ Pr; 1995.
48. **Matthews JR.** Quantification and the Quest for Medical Certainty. Princeton, NJ: Princeton Univ Pr; 1995.
49. **Feinstein AR, Horwitz RI.** Problems in the "evidence" of "evidence-based medicine." *Am J Med.* 1997;103:529-35.
50. **Spodich DH.** "Evidence-based medicine": terminologic lapse or terminologic arrogance? [Letter] *Am J Cardiol.* 1996;78:608-9.
51. **Tonelli MR.** The philosophical limits of evidence-based medicine. *Acad Med.* 1998;73:1234-40.
52. **Feinstein AR.** *Clinical Biostatistics.* St. Louis: Mosby; 1977.
53. **Mainland D.** The significance of "nonsignificance." *Clin Pharmacol Ther.* 1963;12:580-6.
54. **Morrison DE, Henkel RE.** *The Significance Test Controversy: A Reader.* Chicago: Aldine; 1970.
55. **Rothman KJ.** Significance questing [Editorial]. *Ann Intern Med.* 1986;105:445-7.
56. **Rozeboom W.** The fallacy of the null hypothesis significance test. *Psychol Bull.* 1960;57:416-28.
57. **Savitz D.** Is statistical significance testing useful in interpreting data? *Reprod Toxicol.* 1993;7:95-100.
58. **Chia KS.** "Significant-itis"—an obsession with the P-value. *Scand J Work Environ Health.* 1997;23:152-4.
59. **Barnett ML, Mathisen A.** Tyranny of the p-value: the conflict between statistical significance and common sense [Editorial]. *J Dent Res.* 1997;76:534-6.
60. **Bailar JC 3d, Mosteller F.** Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. *Ann Intern Med.* 1988;108:266-73.
61. **Cox DR.** Statistical significance tests. *Br J Clin Pharmacol.* 1982;14:325-31.
62. **Cornfield J.** The bayesian outlook and its application. *Biometrics.* 1969;25:617-57.
63. **Mainland D.** Statistical ritual in clinical journals: is there a cure?—I. *Br Med J (Clin Res Ed).* 1984;288:841-3.
64. **Mainland D.** Statistical ritual in clinical journals: is there a cure?—II. *Br Med J (Clin Res Ed).* 1984;288:920-2.
65. **Salsburg D.** The religion of statistics as practiced in medical journals. *American Statistician.* 1985;39:220-3.
66. **Dar R, Serlin RC, Omer H.** Misuse of statistical tests in three decades of psychotherapy research. *J Consult Clin Psychol.* 1994;62:75-82.
67. **Altman D, Bland J.** Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society, Series A.* 1991;154:223-67.
68. **Pocock SJ, Hughes MD, Lee RJ.** Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med.* 1987;317:426-32.
69. **Edwards A.** *Likelihood.* Cambridge, UK: Cambridge Univ Pr; 1972.
70. **Skellam J.** Models, inference and strategy. *Biometrics.* 1969;25:457-75.
71. **Clarke M, Chalmers I.** Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? *JAMA.* 1998;280:280-2.
72. **Moyé L.** End-point interpretation in clinical trials: the case for discipline. *Control Clin Trials.* 1999;20:40-9.
73. **Fisher LD.** Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Control Clin Trials.* 1999;20:16-39.
74. **Fisher L, Moyé L.** Carvedilol and the Food and Drug Administration (FDA) approval process: an introduction. *Control Clin Trials.* 1999;20:1-15.
75. **Poole C.** Beyond the confidence interval. *Am J Public Health.* 1987;77:195-9.
76. **Lang JM, Rothman KJ, Cann CI.** That confounded P-value [Editorial]. *Epidemiology.* 1998;9:7-8.
77. **Evans SJ, Mills P, Dawson J.** The end of the p value? *Br Heart J.* 1988;60:177-80.
78. **Feinstein AR.** P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol.* 1998;51:355-60.
79. **Freedman L.** Bayesian statistical methods [Editorial]. *BMJ.* 1996;313:569-70.
80. **Etzioni RD, Kadane JB.** Bayesian statistical methods in public health and medicine. *Annu Rev Public Health.* 1995;16:23-41.
81. **Kadane JB.** Prime time for Bayes. *Control Clin Trials.* 1995;16:313-8.
82. **Spiegelhalter D, Freedman L, Parmar M.** Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Series A.* 1994;157:357-87.
83. **Goodman SN, Royall R.** Evidence and scientific research. *Am J Public Health.* 1988;78:1568-74.
84. **Barnard G.** The use of the likelihood function in statistical practice. In: *Proceedings of the Fifth Berkeley Symposium.* v 1. Berkeley, CA: Univ of California Pr; 1966:27-40.
85. **Wulff HR, Anderson B, Brandenhoff P, Guttler F.** What do doctors know about statistics? *Stat Med.* 1987;6:3-10.
86. **Borak J, Veilleux S.** Errors of intuitive logic among physicians. *Soc Sci Med.* 1982;16:1939-47.
87. **Concato J, Feinstein AE, Holford TR.** The risk of determining risk with multivariable models. *Ann Intern Med.* 1993;118:201-10.
88. **Altman DG, Goodman SN.** Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *JAMA.* 1994;272:129-32.
89. **Hayden G.** Biostatistical trends in pediatrics: implications for the future. *Pediatrics.* 1983;72:84-7.