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.

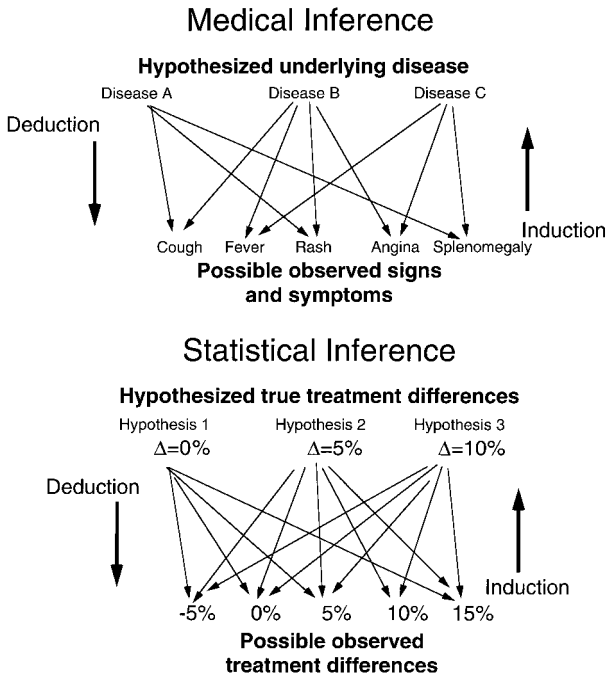

Figure 1. The parallels between the processes of induction and deduction in medical inference (*top*) and statistical inference (*bottom*). Δ = treatment difference.

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 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).

Prior odds of hypothesis being true (before seeing data)	**Subjective component**
×	
Bayes Factor	**Data component** *("evidence")*
⇓	
Final ("posterior") odds of null hypothesis being true	**Probability of truth**

Figure 2. Bayes theorem, in words.

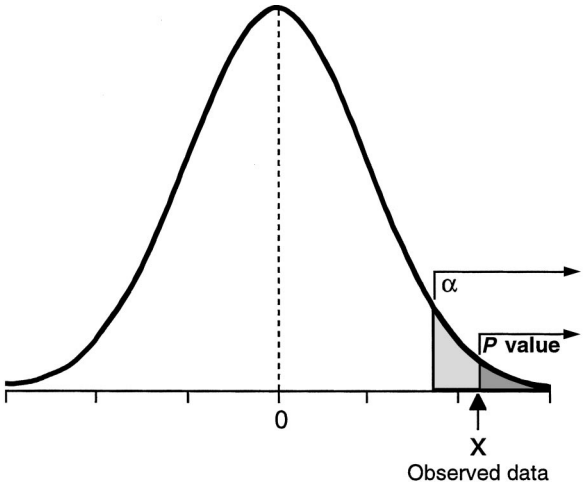

Figure 3. The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both α (the type I error rate) and the *P* value are "tail areas" under this curve. The tail area for α is set before the experiment, and a result can fall anywhere within it. The *P* value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

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 α *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 β *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 α 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 (α) 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

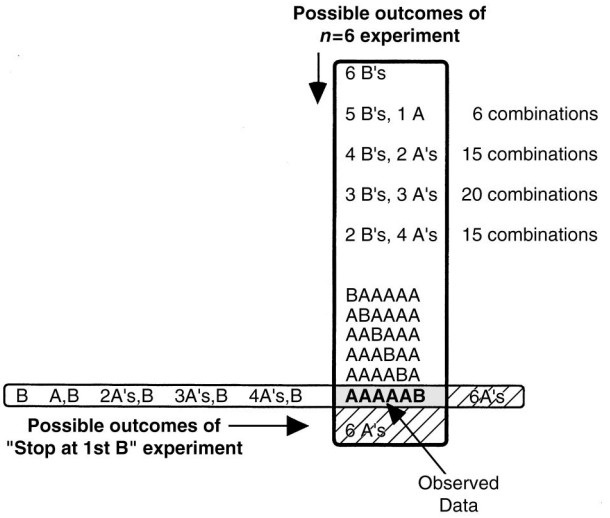

Figure 4. Possible outcomes of two hypothetical trials in six patients (Appendix). The only possible overlapping results are the observed data and the result in which treatment A was preferred in all patients.

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 prespecified 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 occurring 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 ≥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_o [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:

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

"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:

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

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, Hinckley 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.

Toward Evidence-Based Medical Statistics. 2: The Bayes Factor

Steven N. Goodman, MD, PhD

Bayesian inference is usually presented as a method for determining how scientific belief should be modified by data. Although Bayesian methodology has been one of the most active areas of statistical development in the past 20 years, medical researchers have been reluctant to embrace what they perceive as a subjective approach to data analysis. It is little understood that Bayesian methods have a data-based core, which can be used as a calculus of evidence. This core is the Bayes factor, which in its simplest form is also called a *likelihood ratio*. The minimum Bayes factor is objective and can be used in lieu of the *P* value as a measure of the evidential strength. Unlike *P* values, Bayes factors have a sound theoretical foundation and an interpretation that allows their use in both inference and decision making. Bayes factors show that *P* values greatly overstate the evidence against the null hypothesis. Most important, Bayes factors require the addition of background knowledge to be transformed into inferences—probabilities that a given conclusion is right or wrong. They make the distinction clear between experimental evidence and inferential conclusions while providing a framework in which to combine prior with current evidence.

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

Ann Intern Med. 1999;130:1005-1013.

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

In the first of two articles on evidence-based statistics (1), I outlined the inherent difficulties of the standard frequentist statistical approach to inference: problems with using the *P* value as a measure of evidence, internal inconsistencies of the combined hypothesis test–*P* value method, and how that method inhibits combining experimental results with background information. Here, I explore, as nonmathematically as possible, the Bayesian approach to measuring evidence and combining information and epistemologic uncertainties that affect all statistical approaches to inference. Some of this presentation may be new to clinical researchers, but most of it is based on ideas that have existed at least since the 1920s and, to some extent, centuries earlier (2).

The Bayes Factor Alternative

Bayesian inference is often described as a method of showing how belief is altered by data. Because of this, many researchers regard it as nonscientific; that is, they want to know what the data say, not what our belief should be after observing them (3). Comments such as the following, which appeared in response to an article proposing a Bayesian analysis of the GUSTO (Global Utilization of Streptokinase and tPA for Occluded Coronary Arteries) trial (4), are typical.

> When modern Bayesians include a "prior probability distribution for the belief in the truth of a hypothesis," they are actually creating a metaphysical model of attitude change . . . The result . . . cannot be field-tested for its validity, other than that it "feels" reasonable to the consumer. . . .
>
> The real problem is that neither classical nor Bayesian methods are able to provide the kind of answers clinicians want. That classical methods are flawed is undeniable—I wish I had an alternative . . . . (5)

This comment reflects the widespread misperception that the only utility of the Bayesian approach is as a belief calculus. What is not appreciated is that Bayesian methods can instead be viewed as an evidential calculus. Bayes theorem has two components—one that summarizes the data and one that represents belief. Here, I focus on the component related to the data: the Bayes factor, which in its simplest form is also called a *likelihood ratio.* In Bayes theorem, the Bayes factor is the index through which the data speak, and it is separate from the purely subjective part of the equation. It has also been called the relative betting odds, and its logarithm is sometimes referred to as the *weight of the evidence* (6, 7). The distinction between evidence and error is clear when it is recognized that the Bayes factor (evidence) is a measure of how much the probability of truth (that is, 1 − prob(error), where *prob* is probability) is altered by the data. The equation is as follows:

$$\begin{matrix}\textit{Prior Odds}\\ \textit{of Null Hypothesis}\end{matrix} \times \begin{matrix}\textit{Bayes}\\ \textit{Factor}\end{matrix} = \begin{matrix}\textit{Posterior Odds}\\ \textit{of Null Hypothesis}\end{matrix}$$

where Bayes factor =

$$\frac{\text{Prob}(\textit{Data}, \textit{given the null hypothesis})}{\text{Prob}(\textit{Data}, \textit{given the alternative hypothesis})}$$

The Bayes factor is a comparison of how well two hypotheses predict the data. The hypothesis that predicts the observed data better is the one that is said to have more evidence supporting it. Unlike the *P* value, the Bayes factor has a sound theoretical foundation and an interpretation that

See related article on pp 995-1004 and editorial comment on pp 1019-1021.

Table 1. Final (Posterior) Probability of the Null Hypothesis after Observing Various Bayes Factors, as a Function of the Prior Probability of the Null Hypothesis

Strength of Evidence	Bayes Factor	Decrease in Probability of the Null Hypothesis	
		From	To No Less Than
		%	
Weak	1/5	90	64*
		50	17
		25	6
Moderate	1/10	90	47
		50	9
		25	3
Moderate to strong	1/20	90	31
		50	5
		25	2
Strong to very strong	1/100	90	8
		50	1
		25	0.3

* Calculations were performed as follows:
A probability (Prob) of 90% is equivalent to an odds of 9, calculated as Prob/(1 − Prob).
Posterior odds = Bayes factor × prior odds; thus, (1/5) × 9 = 1.8.
Probability = odds/(1 + odds); thus, 1.8/2.8 = 0.64.

allows it to be used in both inference and decision making. It links notions of objective probability, evidence, and subjective probability into a coherent package and is interpretable from all three perspectives. For example, if the Bayes factor for the null hypothesis compared with another hypothesis is 1/2, the meaning can be expressed in three ways.

1. *Objective probability:* The observed results are half as probable under the null hypothesis as they are under the alternative.

2. *Inductive evidence:* The evidence supports the null hypothesis half as strongly as it does the alternative.

3. *Subjective probability:* The odds of the null hypothesis relative to the alternative hypothesis after the experiment are half what they were before the experiment.

The Bayes factor differs in many ways from a *P* value. First, the Bayes factor is not a probability itself but a ratio of probabilities, and it can vary from zero to infinity. It requires two hypotheses, making it clear that for evidence to be *against* the null hypothesis, it must be *for* some alternative. Second, the Bayes factor depends on the probability of the observed data alone, not including unobserved "long run" results that are part of the *P* value calculation. Thus, factors unrelated to the data that affect the *P* value, such as why an experiment was stopped, do not affect the Bayes factor (8, 9).

Because we are so accustomed to thinking of "evidence" and the probability of "error" as synonymous, it may be difficult to know how to deal with a measure of evidence that is not a probability. It is helpful to think of it as analogous to the concept of energy. We know that energy is real, but because it is not directly observable, we infer the meaning of a given amount from how much it heats water, lifts a weight, lights a city, or cools a house. We begin to understand what "a lot" and "a little" mean through its effects. So it is with the Bayes factor: It modifies prior probabilities, and after seeing how much Bayes factors of certain sizes change various prior probabilities, we begin to understand what represents strong evidence, and weak evidence.

Table 1 shows us how far various Bayes factors move prior probabilities, on the null hypothesis, of 90%, 50%, and 25%. These correspond, respectively, to high initial confidence in the null hypothesis, equivocal confidence, and moderate suspicion that the null hypothesis is not true. If one is highly convinced of no effect (90% prior probability of the null hypothesis) before starting the experiment, a Bayes factor of 1/10 will move one to being equivocal (47% probability on the null hypothesis), but if one is equivocal at the start (50% prior probability), that same amount of evidence will be moderately convincing that the null hypothesis is not true (9% posterior probability). A Bayes factor of 1/100 is strong enough to move one from being 90% sure of the null hypothesis to being only 8% sure.

As the strength of the evidence increases, the data are more able to convert a skeptic into a believer or a tentative suggestion into an accepted truth. This means that as the experimental evidence gets stronger, the amount of external evidence needed to support a scientific claim decreases. Conversely, when there is little outside evidence supporting a claim, much stronger experimental evidence is required for it to be credible. This phenomenon can be observed empirically, in the medical community's reluctance to accept the results of clinical trials that run counter to strong prior beliefs (10, 11).

Bayes Factors and Meta-Analysis

There are two dimensions to the "evidence-based" properties of Bayes factors. One is that they are a proper measure of quantitative evidence; this issue will be further explored shortly. The other is that they allow us to combine evidence from different experiments in a natural and intuitive way. To understand this, we must understand a little more of the theory underlying Bayes factors (12–14).

Every hypothesis under which the observed data are not impossible can be said to have some evidence for it. The strength of this evidence is proportional to the probability of the data under that hypothesis and is called the *likelihood* of the hypothesis. This use of the term "likelihood" must not be confused with its common language meaning of

probability (12, 13). Mathematical likelihoods have meaning only when compared to each other in the form of a ratio (hence, the *likelihood ratio*), a ratio that represents the comparative evidential support given to two hypotheses by the data. The likelihood ratio is the simplest form of Bayes factor.

The hypothesis with the most evidence for it has the maximum mathematical likelihood, which means that it predicts the observed data best. If we observe a 10% difference between the cure rates of two treatments, the hypothesis with the maximum likelihood would be that the true difference was 10%. In other words, whatever effect we are measuring, the best-supported hypothesis is always that the unknown true effect is equal to the observed effect. Even when a true difference of 10% gets more support than any other hypothesis, a 10% observed difference also gives a true difference of 15% some support, albeit less than the maximum (**Figure**).

This idea—that each experiment provides a certain amount of evidence for every underlying hypothesis—is what makes meta-analysis straightforward under the Bayesian paradigm, and conceptually different than under standard methods. One merely combines the evidence provided by each experiment for each hypothesis. With log Bayes factors (or log likelihoods), this evidence can simply be added up (15–17).

With standard methods, quantitative meta-analysis consists of taking a weighted average of the observed effects, with weights related to their precision. For example, if one experiment finds a 10% difference and another finds a 20% difference, we would average the numbers 10% and 20%, pool their standard errors, and calculate a new *P* value based on the average effect and pooled standard error. The cumulative evidence (*P* value) for the meta-analytic average has little relation to the *P* values for the individual effects, and averaging the numbers 10% and 20% obscures the fact that both experiments actually provide evidence for the *same* hypotheses, such as a true 15% difference. Although it might be noted that a 15% difference falls within the confidence intervals of both experiments, little can be done quantitatively or conceptually with that fact. So while meta-analysts say they are combining evidence from similar studies, standard methods do not have a measure of evidence that is directly combined.

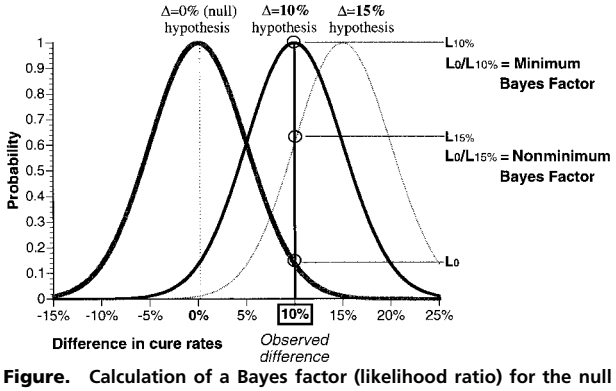

Figure. Calculation of a Bayes factor (likelihood ratio) for the null hypothesis versus two other hypotheses: the maximally supported alternative hypothesis (change Δ = 10%) and an alternative hypothesis with less than the maximum support (Δ = 15%). The likelihood of the null hypothesis (L_0) divided by the likelihood of the best supported hypothesis ($L_{10\%}$), is the minimum likelihood ratio or minimum Bayes factor, the strongest evidence against the null hypothesis. The corresponding ratio for the hypothesis Δ = 15% results in a larger ratio, which means that the evidence against the null hypothesis is weaker.

Of Bayes Factors and *P* Values

If we are to move away from *P* values and toward Bayes factors, it is helpful to have an "exchange rate"—a relation between the new unit of measurement and the old. With a few assumptions, we can make this connection. First, to compare like with like, we must calculate the Bayes factor for the same hypothesis for which the *P* value is being calculated. The *P* value is always calculated by using the observed difference, so we must calculate the Bayes factor for the hypothesis that corresponds to the observed difference, which we showed earlier was the best-supported hypothesis. Second, because a smaller *P* value means less support for the null hypothesis (or more evidence against it), we must structure the Bayes factor the same way, so that a smaller Bayes factor also means less support for the null hypothesis. This means putting the likelihood of the null hypothesis in the numerator and the likelihood of an alternative hypothesis in the denominator. (Whether the null hypothesis likelihood is in the top or bottom of the ratio depends on the context of use.) If we put the evidence for the best-supported hypothesis in the denominator, the resulting ratio will be the smallest possible Bayes factor with respect to the null hypothesis. This reciprocal of the maximum likelihood ratio is also called the *standardized likelihood*. The minimum Bayes factor (or minimum likelihood ratio) is the smallest amount of evidence that can be claimed for the null hypothesis (or the strongest evidence against it) on the basis of the data. This is an excellent benchmark against which to compare the *P* value.

The simplest relation between *P* values and Bayes factors exists when statistical tests are based on a Gaussian approximation, which is the case for most statistical procedures found in medical journals. In that situation, the minimum Bayes factor (the minimum likelihood ratio) is calculated with the same numbers used to calculate a *P* value (13, 18, 19). The formula is as follows (see Appendix I for derivation):

$$\text{Minimum Bayes factor} = e^{-Z^2/2}$$

Table 2. Relation between Fixed Sample Size *P* Values and Minimum Bayes Factors and the Effect of Such Evidence on the Probability of the Null Hypothesis

P Value (Z Score)	Minimum Bayes Factor	Decrease in Probability of the Null Hypothesis, %		Strength of Evidence
		From	To No Less Than	
0.10 (1.64)	**0.26** (1/3.8)	75 50 17	44 21 5	Weak
0.05 (1.96)	**0.15** (1/6.8)	75 50 26	31 13 5	Moderate
0.03 (2.17)	**0.095** (1/11)	75 50 33	22 9 5	Moderate
0.01 (2.58)	**0.036** (1/28)	75 50 60	10 3.5 5	Moderate to strong
0.001 (3.28)	**0.005** (1/216)	75 50 92	1 0.5 5	Strong to very strong

where z is the number of standard errors from the null effect. This formula can also be used if a t-test (substituting t for Z) or a chi-square test (substituting the chi-square value for Z^2) is done. The data are treated as though they came from an experiment with a fixed sample size.

This formula allows us to establish an exchange rate between minimum Bayes factors and *P* values in the Gaussian case. **Table 2** shows the minimum Bayes factor and the standard *P* value for any given *Z* score. For example, when a result is 1.96 standard errors from its null value (that is, $P = 0.05$), the minimum Bayes factor is 0.15, meaning that the null hypothesis gets 15% as much support as the best-supported hypothesis. This is threefold higher than the *P* value of 0.05, indicating that the evidence against the null hypothesis is not nearly as strong as "$P = 0.05$" suggests.

Even when researchers describe results with a *P* value of 0.05 as being of borderline significance, the number "0.05" speaks louder than words, and most readers interpret such evidence as much stronger than it is. These calculations show that *P* values of 0.05 (corresponding to a minimum Bayes factor of 0.15) represent, at best, moderate evidence against the null hypothesis; those between 0.001 and 0.01 represent, at best, moderate to strong evidence; and those less than 0.001 represent strong to very strong evidence. When the *P* value becomes very small, the disparity between it and the minimum Bayes factor becomes negligible, confirming that strong evidence will look strong regardless of how it is measured.

The right-hand part of **Table 2** uses this relation between *P* values and Bayes factors to show the maximum effect that data with various *P* values would have on the plausibility of the null hypothesis. If one starts with a chance of no effect of 50%, a result with a minimum Bayes factor of 0.15 (corresponding to a *P* value of 0.05) can reduce confidence in the null hypothesis to no lower than 13%. The last row in each entry turns the calculation around, showing how low initial confidence in the null hypothesis must be to result in 5% confidence after seeing the data (that is, 95% confidence in a non-null effect). With a *P* value of 0.05 (Bayes factor ≥ 0.15), the prior probability of the null hypothesis must be 26% or less to allow one to conclude with 95% confidence that the null hypothesis is false. This calculation is not meant to sanctify the number "95%" in the Bayesian approach but rather to show what happens when similar benchmarks are used in the two approaches.

These tables show us what many researchers learn from experience and what statisticians have long known; that the weight of evidence against the null hypothesis is not nearly as strong as the magnitude of the *P* value suggests. This is the main reason that many Bayesian reanalyses of clinical trials conclude that the observed differences are not likely to be true (4, 20, 21). They conclude this not always because contradictory prior evidence outweighed the trial evidence but because the trial evidence, when measured properly, was not very strong in the first place. It also provides justification for the judgment of many experienced meta-analysts who have suggested that the threshold for significance in a meta-analysis should be a result more than two standard errors from the null effect rather than two (22, 23).

The theory underlying these ideas has a long history. Edwards (2) traces the concept of mathematical likelihood into the 18th century, although the name and full theoretical development of likelihood didn't occur until around 1920, as part of R.A. Fisher's theory of *maximum likelihood*. This was a frequentist theory, however, and Fisher did not acknowledge the value of using the likelihood directly for inference until many years later (24). Edwards (14) and Royall (13) have built on some of Fisher's ideas, exploring the use of likelihood-based measures of evidence outside of the Bayesian paradigm. In the Bayesian realm, Jeffreys (25) and Good (6) were among the first to develop the theory behind Bayes factors, with the most comprehensive recent summary being that of Kass (26). The suggestion that the minimum Bayes factor (or minimum likelihood ratio) could be used as a reportable index appeared in the biomedical literature at least as early as 1963 (19). The settings in which Bayes factors differ from likelihood ratios are discussed in the following section.

Bayes Factors for Composite Hypotheses

Bayes factors larger than the minimum values cited in the preceding section can be calculated (20, 25–27). This is a difficult technical area, but it is important to understand in at least a qualitative way what these nonminimum Bayes factors measure and how they differ from simple likelihood ratios.

The definition of the Bayes factor is the probability of the observed data under one hypothesis divided by its probability under another hypothesis. Typically, one hypothesis is the null hypothesis of no difference. The other hypothesis can be stated in many ways, such as "the cure rates differ by 15%." That is called a *simple hypothesis* because the difference (15%) is specified exactly. The null hypothesis and best-supported hypothesis are both simple hypotheses.

Things get more difficult when we state the alternative hypothesis the way it is usually posed: for example, "the true difference is not zero" or "the treatment is beneficial." This hypothesis is called a *composite hypothesis* because it is composed of many simple hypotheses ("The true difference is 1%, 2%, 3%. . . ,"). This introduces a problem when we want to calculate a Bayes factor, because it requires calculating the probability of those data under the hypothesis, "The true difference is 1%, 2%, 3%. . . ." This is where Bayes factors differ from likelihood ratios; the latter are generally restricted to comparisons of simple hypotheses, but Bayes factors use the machinery of Bayes theorem to allow measurement of the evidence for composite hypotheses.

Bayes theorem for composite hypotheses involves calculating the probability of the data under each simple hypothesis separately (difference = 1%, difference = 2%, and so on) and then taking an average. In taking an average, we can weight the components in many ways. Bayes theorem tells us to use weights defined by a prior probability curve. A prior probability curve represents the plausibility of every possible underlying hypothesis, on the basis of evidence from sources other than the current study. But because prior probabilities can differ between individual persons, different Bayes factors can be calculated from the same data.

Different Questions, Different Answers

It may seem that the fact that the same data can produce different Bayes factors undermines the initial claim that Bayesian methods offer an objective way to measure evidence. But deeper examination shows that this fact is really a surrogate for the more general problem of how to draw scientific conclusions from the totality of evidence. Applying different weights to the hypotheses that make up a composite hypothesis does not mean that different answers are being produced for the same evidential question; it means that different questions are being asked. For example, in the extreme, if we put all of the weight on treatment differences near 5%, the question about evidence for a nonzero treatment difference becomes a question about evidence for a 5% treatment difference alone. An equal weighting of all hypotheses between 5% and 20% would provide the average evidence for a difference in that range, an answer that would differ from the average evidence for all hypotheses between 1% and 25%, even though all of these are nonzero differences.

Thus, the problem in defining a unique Bayes factor (and therefore a unique strength of evidence) is not with the Bayesian approach but with the fuzziness of the questions we ask. The question "How much evidence is there for a nonzero difference?" is too vague. A single nonzero difference does not exist. There are many nonzero differences, and our background knowledge is usually not detailed enough to uniquely specify their prior plausibility. In practical terms, this means that we usually do not know precisely how big a difference to expect if a treatment or intervention "works." We may have an educated guess, but this guess is typically diffuse and can differ among individuals on the basis of the different background information they bring to the problem or the different weight that they put on shared information. If we could come up with generally accepted reasons that justify a unique plausibility for each underlying truth, these reasons would constitute a form of explanation. Thus, the most fundamental of statistical questions—what is the strength of the evidence?—is related to the fundamental yet most uncertain of scientific questions—how do we explain what we observe?

This fundamental problem—how to interpret and learn from data in the face of gaps in our substantive knowledge—bedevils all technological approaches to the problem of quantitative reasoning. The approaches range from evasion of the problem by considering results in aggregate (as in hypothesis testing), solutions that leave background information unquantified (Fisher's idea for *P* values), or representation of external knowledge in an idealized and imperfect way (Bayesian methods).

Proposed Solutions

Acknowledging the need for a usable measure of evidence even when background knowledge is incomplete, Bayesian statisticians have proposed many approaches. Perhaps the simplest is to conduct a sensitivity analysis; that is, to report the Bayes factors produced by a range of prior distributions, representing the attitudes of enthusiasts to skeptics (28, 29). Another solution, closely related, is to report the smallest Bayes factor for a broad class of prior distributions (30), which can have a one-to-one re-

lation with the *P* value, just as the minimum Bayes factor does in the Gaussian case (31). Another approach is to use prior distributions that give roughly equal weight to each of the simple hypotheses that make up the composite hypothesis (25, 26, 32), allowing the data to speak with a minimal effect of a prior distribution. One such index, the Bayesian information criterion, for which Kass (26) makes a strong case, is closely related to the minimum Bayes factor, with a modification for the sample size. Finally, there is the approach outlined here: not to average at all, but to report the strongest Bayes factor against the null hypothesis.

Beyond the Null Hypothesis

Many statisticians and scientists have noted that testing a hypothesis of exact equivalence (the null hypothesis) is artificial because it is unlikely to be exactly true and because other scientific questions may be of more interest. The Bayesian approach gives us the flexibility to expand the scope of our questions to, for example, "What is the evidence that the treatment is harmful?" instead of "What is the evidence that the treatment has no effect?" These questions have different evidential answers because the question about harm includes all treatment differences that are not beneficial. This changes the null hypothesis from a simple hypothesis (a difference of 0) into a composite hypothesis (a difference of zero or less). When this is done, under certain conditions, the one-sided *P* value can reasonably approximate the Bayes factor (33, 34). That is, if we observe a one-sided *P* value of 0.03 for a treatment benefit and give all degrees of harm the same initial credibility as all degrees of benefit, the Bayes factor for treatment harm compared with benefit is approximately 0.03. The minimum Bayes factor for no treatment effect compared with benefit would still be 0.095 (**Table 2**).

Objectivity of the Minimum Bayes Factor

The minimum Bayes factor is a unique function of the data that is at least as objective as the *P* value. In fact, it is more objective because it is unaffected by the hypothetical long-run results that can make the *P* value uncertain. In the first article (1), I presented an example in which two different *P* values (0.11 and 0.03) were calculated from the same data by virtue of the different mental models of the long run held by two researchers. The minimum Bayes factor would be 0.23, identical for both scientists' approaches (Appendix 2). This shows us again how *P* values can overstate the evidence, but more important, it vindicates our intuition that the identical data should produce identical evidence.

This example is important in understanding two problems that plague frequentist inference: multiple comparisons and multiple looks, or, as they are more commonly called, *data dredging* and peeking at the data. The frequentist solution to both problems involves adjusting the *P* value for having looked at the data more than once or in multiple ways. But adjusting the measure of evidence because of considerations that have nothing to do with the data defies scientific sense (8, 35–41), belies the claim of "objectivity" that is often made for the *P* value, and produces an undesirable rigidity in standard trial design. From a Bayesian perspective, these problems and their solutions are viewed differently; they are caused not by the reason an experiment was stopped but by the uncertainty in our background knowledge. The practical result is that experimental design and analysis is far more flexible with Bayesian than with standard approaches (42).

External Evidence

Prior probability distributions, the Bayesian method for representing background knowledge, are sometimes derided as representing opinion, but ideally this opinion should be evidence-based. The body of evidence used can include almost all of the factors that are typically presented in a discussion section but are not often formally integrated with the quantitative results. It is not essential that an investigator know of all of this evidence before an experiment. This evidence can include the following:

1. The results of similar experiments.
2. Experiments studying associations with similar underlying mechanisms.
3. Laboratory experiments directly studying the mechanism of the purported association.
4. Phenomena seen in other experiments that would be explained by this proposed mechanism.
5. Patterns of intermediate or surrogate end points in the current experiment that are consistent with the proposed mechanism.
6. Clinical knowledge based on other patients with the same disease or on other interventions with the same proposed mechanism.

Only the first of these types of evidence involves a simple comparison or summation of results from similar experiments, as in a meta-analysis. All of the others involve some form of extrapolation based on causal reasoning. The use of Bayes factors makes it clear that this is necessary in order to draw conclusions from the statistical evidence.

Use of the Bayes Factor

We will now use two statements from the results sections of hypothetical reports to show the mini-

mum Bayes factor can be used to report and interpret evidence.

Hypothetical Statement 1

The difference in migraine relief rates between the experimental herbal remedy and placebo groups (54% compared with 40% [CI for difference, −2% to 30%]) was not significant ($P = 0.09$).

Bayesian data interpretation 1: The P value of 0.09 ($Z = 1.7$) for the difference in migraine relief rates corresponds to a minimum Bayes factor of $e^{-1.7^2/2} =$ ¼ for the null hypothesis. This means that these data reduce the odds of the null hypothesis by at most a factor of 4, fairly modest evidence for the efficacy of this treatment. For these data to produce a final null hypothesis probability of 5%, the external evidence supporting equivalence must justify a prior probability of equivalence less than 17%. But no mechanism has been proposed yet for this herbal migraine remedy, and all previous reports have consisted of case studies or anecdotal reports of relief. This a priori support is weak and does not justify a prior probability less than 50%. The evidence from this study is therefore insufficient for us to conclude that the proposed remedy is effective.

Bayesian data interpretation 2: . . . For these data to produce a final null hypothesis probability of 5%, the external evidence supporting equivalence must justify a prior probability of equivalence less than 17%. However, the active agent in this remedy is in the same class of drugs that have proven efficacy in migraine treatment, and this agent has been shown to have similar vasoactive effects both in animal models and in preclinical studies in humans. Three uncontrolled studies have all shown relief rates in the range seen here (50% to 60%), and the first small randomized trial of this agent showed a significant effect (60% compared with 32%; $P = 0.01$). The biological mechanism and observed empirical evidence seem to justify a prior probability of ineffectiveness of 15% to 25%, which this evidence is strong enough to reduce to 4% to 8%. Thus, the evidence in this trial, in conjunction with prior findings, is strong enough for us to conclude that this herbal agent is likely to be effective in relieving migraine.

Hypothetical Statement 2

Among the 50 outcomes examined for their relation with blood transfusions, only nasopharyngeal cancer had a significantly elevated rate (relative risk, 3.0; $P = 0.01$).

Bayesian data interpretation: The minimum Bayes factor for relative risk of 1.0 compared with a relative risk not equal to 1.0 for nasopharyngeal cancer is 0.036. This is strong enough to reduce a starting probability on the null hypothesis from at most 59% to 5%. However, there is no previous evidence for such an association or of a biological mechanism to explain it. In addition, rates of cancers with similar risk factor profiles and molecular mechanisms were not elevated, meaning that blood transfusion would have to produce its effect by means of a mechanism that differs from any other previously identified causes of this cancer. Previous studies of blood transfusions have not reported this association, and there have been no reports of increased incidence of nasopharyngeal cancer among populations who undergo repeated transfusions. Therefore, prior evidence suggests that the probability of the null hypothesis is substantially higher than 60%. A minimum Bayes factor of 0.036 means that this result can reduce a 85% prior probability to no lower than 17% and a 95% prior probability to no lower than 41%. Therefore, more evidence than that provided by this study is needed to justify a reliable conclusion that blood transfusion increases the risk for nasopharyngeal cancer. However, future studies should explore this relation and its potential mechanisms.

Discussion

The above examples do not nearly represent full Bayesian interpretation sections, which might use a range of prior distributions to define a range of Bayes factors, or use priors that have been elicited from experts (29, 43, 44). These scenarios do, however, illustrate a few essential facts. First, this measure of evidence can usually be easily calculated from the same information used to calculate a P value or confidence interval and thus can be implemented without specialized software or extensive statistical expertise. Some expertise is needed to assure that the Gaussian approximation underlying the formula applies in a particular situation. When it doesn't apply, many standard software programs report some function of the exact likelihood (typically, its logarithm), from which it is not hard for a statistician to calculate the minimum Bayes factor. Its independence from prior probabilities can also help overcome the reluctance of many investigators to abandon what they regard as objective statistical summaries.

More important, these examples highlight how this index can help keep the statistical evidence distinct from the conclusions, while being part of a calculus that formally links them. The first example showed how the same quantitative results could be included in discussions that came to different conclusions. The explicitness of this process encourages debate about the strength of the supporting evidence. As outlined in the first article, standard methods discourage this because they offer no way to combine supporting evidence with a study's P values or confidence intervals.

These examples demonstrate how the minimum

Bayes factor enables simple threshold Bayesian analyses to be performed without a formal elicitation of prior probability distributions. One merely has to argue that the prior probability of the null hypothesis is above or below a threshold value, on the basis of the evidence from outside the study. If the strongest evidence against the null hypothesis (the minimum Bayes factor) is not strong enough to sufficiently justify a conclusion, then the weaker evidence derived from a Bayes factor from a full Bayesian analysis will not be either.

The use of the minimum Bayes factor does not preclude a formal Bayesian analysis and indeed might be an entrée to one. Recent reviews and books outline how full Bayesian analyses can be conducted and reported (21, 29, 45–50). Bayesian results can also be extended into formal decision analyses (51). The availability of user-friendly software for Bayesian calculations (52) makes implementation of this method more practicable now than in the past.

In not using a proper Bayesian prior probability distribution, the minimum Bayes factor represents a compromise between Bayesian and frequentist perspectives, which can be criticized from both camps. Some statisticians might deride the minimum Bayes factor as nothing more than a relabelled *P* value. But as I have tried to show, *P* values and Bayes factors are far more than just numbers, and moving to Bayes factors of any kind frees us from the flawed conceptual framework and improper view of the scientific method that travels with the *P* value.

The Bottom Line: Both Perspectives Are Necessary, but *P* Values Are Not

Standard frequentist methods are most problematic when used to draw conclusions in a single experiment. Their denial of a formal role for external information in inference poses serious practical and logical problems. But Bayesian methods, designed for inductive inference in single experiments, do not guarantee that in the long run, conclusions in which we have 95% confidence will turn out to be true 95% of the time (53). This is because Bayesian prior probability distributions are not ideal quantitative descriptors of what we know (or what we don't know) (54, 55), and Bayes theorem is an imperfect model for human learning (54, 56). This means that the frequentist, long-run perspective cannot be completely ignored, leading many statisticians to emphasize the importance of using frequentist criteria in the evaluation of Bayesian and likelihood methods (6, 13, 32, 53), which these methods typically fulfill quite well.

In the end, we must recognize that there is no automatic method in statistics, as there is not in life, that allows us both to evaluate individual situations and know exactly what the long-run consequences of that evaluation will be. The connection between inference in individual experiments and the number of errors we make over time is not found in the *P* value or in hypothesis tests. It is found only in properly assessing the strength of evidence from an experiment with Bayes factors and uniting this with a synthesis of all of the other scientific information that bears on the question at hand. There is no formula for performing that synthesis, nor is there a formula for assigning a unique number to it. That is where room for meaningful scientific discourse lies.

Sir Francis Bacon, the writer and philosopher who was one of the first inductivists, commented on the two attitudes with which one can approach nature. His comment could apply to the perspectives contrasted in these essays: "If we begin with certainties, we shall end in doubts; but if we begin with doubts, and are patient with them, we shall end with certainties" (57). Putting *P* values aside, Bayesian and frequentist approaches each provide an essential perspective that the other lacks. The way in which we balance their sometimes conflicting demands is what makes the process of learning from nature creative, exciting, uncertain, and, most of all, human.

Appendix I

Derivation of the minimum Bayes factor under a Gaussian distribution: The likelihood of a hypothesis given an observed effect, x, is proportional to the probability of x under that hypothesis. For a Gaussian distribution, the hypothesis typically concerns the mean. The probability of x under a Gaussian distribution with true mean = μ and standard error = σ, is (where the symbol "|" is read as "given"):

$$\Pr(x|\,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

Because the exponent is negative, the above probability is maximized when the exponent is zero, which occurs when $\mu = x$ (that is, the true mean μ equals the observed effect, x). The likelihood ratio for the null hypothesis ($\mu = 0$) versus the maximally supported hypothesis ($\mu = x$) is the minimum Bayes factor:

$$\frac{\Pr(x|\,\mu = 0,\sigma)}{\Pr(x|\,\mu = x,\sigma)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-0}{\sigma}\right)^2/2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-x}{\sigma}\right)^2/2}} = e^{-\left(\frac{x}{\sigma}\right)^2/2}$$

Because the Z-score is the observed effect, x, divided by its standard error, σ, the final term in the above equation is:

$$e^{-\left(\frac{x}{\sigma}\right)^2/2} = e^{-Z^2/2}$$

Appendix II

In the example posed in the first article (1), two treatments, called A and B, were compared in the same patients, and the preferred treatment in each patients was chosen. The two experimenters had different mindsets while conducting the experiment: one planned to study all six patients, whereas the other planned to stop as soon as treatment B was preferred. The first five patients preferred treatment A, and the sixth preferred treatment B.

The probability of the data under the two hypotheses is as follows.

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

Alternative hypothesis: Probability that treatment A is preferred = 5/6

In the "$n = 6$" experiment, this ratio is:

$$6\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 \Big/ 6\left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right)^1 = 0.23$$

The "6" appears above because the preference for treatment B could have occurred in any of the first five patients or in the sixth patient without a change in the inference.

In the "stop at first preference for treatment B" experiment, the ratio is:

$$\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 \Big/ \left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right)^1 = 0.23$$

Acknowledgments: The author thanks Dan Heitjan, Russell Localio, Harold Lehmann, and Michael Berkwitz for helpful comments on earlier versions of this article. The views expressed are the sole responsibility of the author.

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

References

1. **Goodman SN.** Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med. 1999;130:995-1004.
2. **Edwards A.** A History of Likelihood. International Statistical Review. 1974; 42:9-15.
3. **Fisher LD.** Comments on Bayesian and frequentist analysis and interpretation of clinical trials. Control Clin Trials. 1996;17:423-34.
4. **Brophy JM, Joseph L.** Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA. 1995;273:871-5.
5. **Browne RH.** Bayesian analysis and the GUSTO trial. Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Coronary Arteries [Letter]. JAMA. 1995;274:873.
6. **Good I.** Probability and the Weighing of Evidence. New York: Charles Griffin; 1950.
7. **Cornfield J.** The Bayesian outlook and its application. Biometrics. 1969;25: 617-57.
8. **Berger JO, Berry DA.** Statistical analysis and the illusion of objectivity. American Scientist. 1988;76:159-65.
9. **Berry D.** Interim analyses in clinical trials: classical vs. Bayesian approaches. Stat Med. 1985;4:521-6.
10. **Belanger D, Moore M, Tannock I.** How American oncologists treat breast cancer: an assessment of the influence of clinical trials. J Clin Oncol. 1991;9:7-16.
11. **Omoigui NA, Silver MJ, Rybicki LA, Rosenthal M, Berdan LG, Pieper K, et al.** Influence of a randomized clinical trial on practice by participating investigators: lessons from the Coronary Angioplasty Versus Excisional Atherectomy Trial (CAVEAT). CAVEAT I and II Investigators. J Am Coll Cardiol. 1998;31:265-72.
12. **Goodman SN, Royall R.** Evidence and scientific research. Am J Public Health. 1988;78:1568-74.
13. **Royall R.** Statistical Evidence: A Likelihood Primer. Monographs on Statistics and Applied Probability, #71. London: Chapman and Hall; 1997.
14. **Edwards A.** Likelihood. Cambridge, UK: Cambridge Univ Pr; 1972.
15. **Goodman SN.** Meta-analysis and evidence. Control Clin Trials. 1989;10:188-204, 435.
16. **Efron B.** Empirical Bayes methods for combining likelihoods. Journal of the American Statistical Association. 1996;91:538-50.
17. **Hardy RJ, Thompson SG.** A likelihood approach to meta-analysis with random effects. Stat Med. 1996;15:619-29.
18. **Berger J.** Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag; 1985.
19. **Edwards W, Lindman H, Savage L.** Bayesian statistical inference for psychological research. Psychol Rev. 1963;70:193-242.
20. **Diamond GA, Forrester JS.** Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med. 1983;98:385-94.
21. **Lilford R, Braunholtz D.** The statistical basis of public policy: a paradigm shift is overdue. BMJ. 1996;313:603-7.
22. **Peto R.** Why do we need systematic overviews of randomized trials? Stat Med. 1987;6:233-44.
23. **Pogue J, Yusuf S.** Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998;351:47-52.
24. **Fisher R.** Statistical Methods and Scientific Inference. 3d ed. New York: Macmillan; 1973.
25. **Jeffreys H.** Theory of Probability. 2d ed. Oxford: Oxford Univ Pr; 1961.
26. **Kass R, Raftery A.** Bayes Factors. Journal of the American Statistical Association. 1995;90:773-95.
27. **Cornfield J.** A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. Journal of the American Statistical Association. 1966;61:577-94.
28. **Kass R, Greenhouse J.** Comments on "Investigating therapies of potentially great benefit: ECMO" (by JH Ware). Statistical Science. 1989;4:310-7.
29. **Spiegelhalter D, Freedman L, Parmar M.** Bayesian approaches to randomized trials. Journal of the Royal Statistical Society, Series A. 1994;157:357- 87.
30. **Berger J, Sellke T.** Testing a point null hypothesis: the irreconcilability of p-values and evidence. Journal of the American Statistical Association. 1987; 82:112-39.
31. **Bayarri M, Berger J.** Quantifying surprise in the data and model verification. Proceedings of the 6th Valencia International Meeting on Bayesian Statistics, 1998. 1998:1-18.
32. **Carlin C, Louis T.** Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman and Hall; 1996.
33. **Casella G, Berger R.** Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association. 1987;82:106-11.
34. **Howard J.** The 2 × 2 table: a discussion from a Bayesian viewpoint. Statistical Science. 1999;13:351-67.
35. **Cornfield J.** Sequential trials, sequential analysis and the likelihood principle. American Statistician. 1966;20:18-23.
36. **Savitz DA, Olshan AF.** Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol. 1995;142:904-8.
37. **Perneger T.** What's wrong with Bonferroni adjustments. BMJ. 1998;316: 1236-8.
38. **Goodman SN.** Multiple comparisons, explained. Am J Epidemiol. 1998;147: 807-12.
39 **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.
40. **Greenland S, Robins JM.** Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology. 1991;2:244-51.
41. **Rothman KJ.** No adjustments are needed for multiple comparisons. Epidemiology. 1990;11:43-6.
42. **Berry DA.** A case for Bayesianism in clinical trials. Stat Med. 1993;12:1377-93.
43. **Chaloner K, Church T, Louis T, Matts J.** Graphical elicitation of a prior distribution for a clinical trial. The Statistician. 1993;42:341-53.
44. **Chaloner K.** Elicitation of prior distributions. In: Berry D, Stangl D, eds. Bayesian Biostatistics. New York: Marcel Dekker; 1996.
45. **Freedman L.** Bayesian statistical methods [Editorial]. BMJ. 1996;313:569-70.
46. **Fayers PM, Ashby D, Parmar MK.** Tutorial in biostatistics: Bayesian data monitoring in clinical trials. Stat Med. 1997;16:1413-30.
47. **Etzioni RD, Kadane JB.** Bayesian statistical methods in public health and medicine. Annu Rev Public Health. 1995;16:23-41.
48. **Berry DA.** Benefits and risks of screening mammography for women in their forties: a statistical appraisal. J Natl Cancer Inst. 1998;90:1431-9.
49. **Hughes MD.** Reporting Bayesian analyses of clinical trials. Stat Med. 1993; 12:1651-64.
50. **Berry DA, Stangl D, eds.** Bayesian Biostatistics. New York: Marcel Dekker; 1996.
51. **Berry DA.** Decision analysis and Bayesian methods in clinical trials. Cancer Treat Res. 1995;75:125-54.
52. **Spiegelhalter D, Thomas A, Best N, Gilks W.** BUGS: Bayesian Inference Using Gibbs Sampling. Cambridge, UK: MRC Biostatistics Unit; 1998. Available at www.mrc-bsu.cam.ac.uk/bugs.
53. **Rubin D.** Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics. 1984;12:1151-72.
54. **Shafer G.** Savage revisited. Statistical Science. 1986;1:463-501.
55. **Walley P.** Statistical Reasoning with Imprecise Probabilities. London: Chapman and Hall; 1991.
56. **Tversky A, Kahneman D.** Judgment under uncertainty: heuristics and biases. In: Slovic P, Tversky A, Kahneman D, eds. Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge Univ Pr; 1982:1-20.
57. **Bacon F.** De Augmentis Scientarium, Book I (1605). In: Curtis C, Greenslet F, eds. The Practical Cogitator. Boston: Houghton Mifflin; 1962.

Standing Statistics Right Side Up

During the years I taught students about diagnostic reasoning, I would begin by explaining that the sensitivity of a diagnostic test for disease X is found by measuring how often the test result is positive in a population of patients, all of whom are known (by some independent and definitive criterion, the "gold standard") to have disease X: that is, by measuring the frequency of true-positive results in that population. A test that yields positive results in 95 of 100 diseased patients, for example, has a sensitivity of 0.95. We would then talk about test specificity—the likelihood that the same test would have a false-positive result in a population of patients known by the gold standard not to have the disease. A test that yields positive results in 10 of 100 nondiseased patients has a specificity of 0.90.

I would then ask the students to imagine that in working up a new patient, they have gotten back a positive result from a test with the above sensitivity and specificity. What would they tell the patient about his or her probability of having disease X? Their answer was almost always "95%." On the face of it, that answer seems pretty reasonable: Isn't that what you'd expect if a test were capable of detecting 95% of diseased patients? The problem is, it's wrong; worse, it actually stands diagnostic reasoning on its head.

In fact, test sensitivity and specificity are *deductive* measurements; they reason down from hypothesis (we assume the truth of the hypothesis that the patient being tested does, or does not, have the disease) to data (the likelihood that we will get a positive test result). The students' reasoning is upside down because what clinicians and patients really need to know is exactly the inverse. In short, they need an *inductive* measurement, a reasoning up from data (the test result) to hypothesis (that the patient has the disease).

Stated differently, what clinicians and patients need is a way to calculate the probability that any particular test result, positive or negative, is a true result. It is possible to make that inductive calculation, but doing so requires combining sensitivity and specificity to create something called a *likelihood ratio,* which is an overall measure of the "evidence" provided by the test result (positive or negative) itself. The likelihood ratio is then used to modify the pretest estimate (the "prior probability") that

---

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

the patient has the disease, thereby creating a new and better post-test estimate—sometimes known as the test's *predictive value*—of the chance that the patient has the disease. (For obvious reasons, predictive values are also known as *posterior probabilities.* Positive predictive values express the post-test likelihood that disease is present after a positive test result; negative predictive values indicate the post-test likelihood that disease is absent after a negative test result.)

Although the deductive inference in a test's sensitivity and specificity differs profoundly from the inductive inference in its predictive values, that difference is also an extremely subtle one; it was not widely appreciated in the biomedical literature until the mid-1970s (1). In a two-part article in this issue (2, 3), Goodman demonstrates how the standard statistical methods (sometimes called "frequentist" statistics) used in analyzing biomedical research, which we have come to accept as a kind of revealed truth, also stand statistical inference on its head in much the same way that students' initial attempts at diagnostic reasoning do.

The article by Goodman is not light reading. He is, however, a true hermeneut, a venerable word meaning "one who is skilled at interpretation." Those who make the effort to understand him will be rewarded with a number of important, if disconcerting, insights. Thus, just as clinicians need to know the likelihood that a particular patient has a disease given a certain test result, researchers (and those who read papers describing research) need to know the likelihood that a hypothesis is true given the data actually obtained in a particular trial or experiment. Both of these are inductive inferences. But, as Goodman points out, researchers generally resort to an inverse, deductive calculation. That is, they calculate the probability of finding the results they actually obtained, plus any more extreme results, on the assumption that a certain hypothesis is true (usually the "null hypothesis"—the assumption that the comparison groups do not differ), a concept expressed in the all-too-familiar *P* value.

The *P* value has been the subject of much criticism because a *P* value of 0.05 has been frequently and arbitrarily misused to distinguish a true effect from lack of effect. Although Goodman does not disagree with that criticism, his real concerns lie deeper, and he catalogues for us several more serious and more convoluted misinterpretations of the concepts of evidence, error, and testing. These misconceptions are particularly troubling because they confuse our ability to judge whether, over the long run of experience with many studies, "we shall not often be wrong" with our ability to judge the likelihood that each separate hypothesis tested in an individual study is true or false.

Enter Bayes theorem. Unfortunately, those ominous words, with their associations of hazy prior probabilities and abstruse mathematical formulas, strike fear into the hearts of most of us, clinician, researcher, and editor alike. But Bayesian inference immediately loses much of its menace once we realize that it is, in fact, the exact equivalent of predictive value, a concept now familiar from its wide use in diagnostic reasoning. It also helps to understand that, mathematical niceties aside, Bayes theorem is essentially a quantitative description of what we do, qualitatively, every minute of the day: use new information inductively to refine our judgments about the correctness of what we already know. In fancier language, Bayesian inference says that the most effective way to develop a new and better degree of confidence (posterior odds) in our knowledge is to combine our previous confidence, derived from sources outside a particular test or study (the prior odds), with the "evidence" from that test or study itself (the Bayes factor).

The importance of information from outside sources becomes particularly clear in considering the impact of a single diagnostic test across the full spectrum of clinical situations. Thus, the positive predictive value (posterior probability of disease) of even a fairly sensitive and specific test might be only 0.1 or 0.2 when that test is used in the "screening mode," that is, when the patient being tested is very unlikely to have the disease in the first place. In this situation, combining the "evidence"—the likelihood ratio for a positive test result—with outside information—a very low pretest (prior) probability—changes that probability relatively little, unless the specificity of the test involved is almost perfect. In contrast, the positive predictive value of the *very same test* might be 0.90 to 0.95 or higher when testing in the "confirmatory mode," that is, when testing is done in a patient who is already strongly suspected of having the disease. Here, a relatively high pretest (prior) probability can become substantially higher when it is combined with the evidence from a test that has even relatively modest specificity. The *same test* can produce intermediate positive predictive values when testing is done in the "diagnostic mode," that is, when the pretest (prior) suspicion of disease is moderate to begin with.

In like fashion, the use of prior knowledge is critical in interpreting biomedical studies, and failure to take it into account can easily lead to serious misinterpretation of the "evidence." For example, a recent meta-analysis found an odds ratio of 1.66 in favor of the beneficial effects of homeopathic therapies over placebos. The associated 95% CI of 1.33 to 2.08, taken by itself, was interpreted as evidence that is "not compatible with the hypothesis that the clinical effects of homeopathy are due to placebo"

(4). If, however, that evidence is combined with the minimal plausibility (extremely low prior probability) that clinically meaningful biological activity can result from small doses of pure water, even water that is shaken in a special way, the resulting posterior level (posterior probability) of confidence in biological activity remains very low. Explanations other than biological efficacy are thus likely to account for the results actually observed (5). Conversely, in view of the existing evidence that vitamin E may protect against coronary heart disease, the finding, reported in this issue (6), that vitamin E appears statistically not to prevent ischemic stroke should be interpreted as ratcheting down the probability of stroke prevention slightly, rather than flatly ruling out the possibility of such activity.

Figuring out the best way to combine the evidence from a trial with prior information from sources outside the trial is an important challenge. It is also a very difficult one, because we often weigh outside information subjectively. Goodman has therefore chosen to focus his discussion primarily on the less controversial and more objective core of Bayesian inference: the measure of "the evidence" from a trial or study. This measure is expressed by the Bayes factor, a metric already familiar to many readers in the form of the likelihood ratio, and one that, in itself, provides logically sound and statistically meaningful information (3). An important lesson from this element of his discussion is that the statistical evidence against a null hypothesis is usually weaker when the data are interpreted by using the Bayes factor than when the same data are interpreted by using the *P* value approach.

Convinced that inductive inference is both useful and feasible in interpreting scientific studies, in 1997 we began encouraging authors of manuscripts submitted to *Annals* to include Bayesian interpretation of their results (7). Few have done so, probably both because frequentist methods are universally taught, enshrined in statistical software, and expected by biomedical journals and because researchers are generally not familiar with alternative methods. Researchers will be particularly interested in Goodman's essay, therefore, because Bayesian principles can contribute importantly to the design of biomedical studies. These principles include the importance of an exhaustive search of the existing, prior evidence, a step that is now often omitted (8), and calculation of a minimum Bayes factor from the data. But others stand to benefit as well from working their way through his analysis. This includes clinicians, who are increasingly required to interpret the strength of evidence from individual studies in making decisions at the bedside, and medical reporters, who are quick to seize on the latest individual trial without considering other available studies, thereby creating a great deal of unnecessary confusion.

Frequentist statistics can serve a useful purpose, but their limitations are many and serious. Some members of the biostatistical community have therefore worked long and hard to encourage the medical researchers and readers to use the Bayesian approach to statistical inference in the design and interpretation of their studies. Goodman's article is an elegant reflection of those efforts, providing both an explication of underlying theory and solid suggestions for practice. In our view, this article will contribute importantly to the task of standing statistical inference right side up. We recommend it to our readers' most serious attention.

Frank Davidoff, MD
Editor

Ann Intern Med. 1999;130:1019-1021.

References

1. **Galen RS, Gambino SR.** Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses. New York: Wiley; 1975.
2. **Goodman SN.** Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med. 1999;130:995-1004.
3. **Goodman SN.** Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13.
4. **Linde K, Clausius N, Ramirez G, Melchart D, Eitel F, Hedges LV, et al.** Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. Lancet. 1997;350:834-43.
5. **Vandenbroucke JP.** 175th anniversary lecture. Medical journals and the shaping of medical knowledge. Lancet. 1998;352:2001-6.
6. **Ascherio A, Rimm EB, Hernán MA, Giovannucci E, Kawachi I, Stampfer MJ, et al.** Relation of consumption of vitamin E, vitamin C, and carotenoids to risk for stroke among men in the United States. Ann Intern Med. 1999;130: 963-70.
7. Information for authors. Ann Intern Med. 1997;127:I-15.
8. **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.