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.