Are All Significant *P* Values Created Equal?

The Analogy Between Diagnostic Tests and Clinical Research

Warren S. Browner, MD, MPH, Thomas B. Newman, MD, MPH

Just as diagnostic tests are most helpful in light of the clinical presentation, statistical tests are most useful in the context of scientific knowledge. Knowing the specificity and sensitivity of a diagnostic test is necessary, but insufficient: the clinician must also estimate the prior probability of the disease. In the same way, knowing the *P* value and power, or the confidence interval, for the results of a research study is necessary but insufficient: the reader must estimate the prior probability that the research hypothesis is true. Just as a positive diagnostic test does not mean that a patient has the disease, especially if the clinical picture suggests otherwise, a significant *P* value does not mean that a research hypothesis is correct, especially if it is inconsistent with current knowledge. Powerful studies are like sensitive tests in that they can be especially useful when the results are negative. Very low *P* values are like very specific tests; both result in few false-positive results due to chance. This Bayesian approach can clarify much of the confusion surrounding the use and interpretation of statistical tests.

(*JAMA* 1987;257:2459-2463)

IN THE four ORIGINAL CONTRIBUTIONS in this issue of THE JOURNAL, the authors report the results of statistical tests of 76 hypotheses.[1-4] Of these, 32 had significant *P* values ($P<.05$). But do these *P* values imply that the 32 hypotheses are true? Or that 95% of them are true? Are all significant *P* values created equal?

The answer to these questions is "No!" What then is a *P* value? It is the likelihood of observing the study results under the assumption that the null hypothesis of no difference is true. Probably because this definition is elusive and intimidating, understanding *P* values (and other statistical concepts like power, confidence intervals, and multiple hypothesis testing) is often left to experts in the field. It is easier just to check whether a *P* value is .05 or less, call the result "statistically significant," regard the tested hypothesis as probably true, and move on to the next paragraph.

Readers of medical literature need not give up quite so quickly, however. As Diamond and Forrester[5] pointed out, many statistical concepts have remarkably similar analogues in an area familiar to clinicians—the interpretation of diagnostic tests. In the diagnosis of Cushing's syndrome, for example, most clinicians recognize that an elevated serum cortisol level is more useful than an elevated blood glucose level, and that an elevated cortisol level is more likely to be due to Cushing's syndrome in a moon-faced patient with a buffalo hump and abdominal striae than in an overweight patient with hypertension.[6,7] Why? Because the interpretation of a test result depends on the characteristics of both the test and the patient being tested.[8-13]

The same type of reasoning—called *Bayesian analysis* after Thomas Bayes, the mathematician who developed it more than 200 years ago[14]—can also be used to clarify the meaning of the *P* value and other statistical terms. Although this application of Bayes' ideas has been discussed in epidemiologic and statistical literature,[15-18] it has received less attention in the journals read by clinicians. In this article, we begin with the basic aspects of the analogy between research studies and diagnostic tests, such as the similarity between the power of a study and the sensitivity of a test, and then examine more challenging issues, such as how a study with multiple hypotheses resembles a serum chemistry panel.

THE ANALOGY

An overview of the analogy between research studies and diagnostic tests is shown in Table 1. In this analogy, a clinician obtains diagnostic data to test for the presence of a disease, such as breast cancer, and an investigator collects study data to determine the truth of a *research hypothesis*, such as that the efficacies of two drugs differ in the treatment of peptic ulcer disease. (The research hypothesis is often called the *alternative hypothesis* in standard terminology.) The absence of a *disease* (no breast cancer) is like the *null hypothesis* of no difference in the efficacy of the two drugs.

The term "positive" is used in its usual sense: to refer to diagnostic tests that are consistent with the presence of the disease and to studies that have statistically significant results. Similarly, "negative" refers to diagnostic tests consistent with the absence of disease and research results that fail to reach statistical significance. Thus there are four possible results whenever a patient undergoes a diagnostic test. Consider the use of fine-needle aspiration in the evaluation of a breast mass, for example (Table 2). If the patient has breast cancer, there are two possibilities: the test result can either be correctly positive or incorrectly negative. On the other hand, if the patient actually does not have cancer, then the result will either be correctly negative or incorrectly positive. Similarly, there are four possible results whenever an investigator studies a research hypothesis (Table 3). If the efficacies of the two drugs really do differ, there are two possibilities: the study can be correctly positive if it finds a difference or incorrectly negative if it

Table 1.—The Analogy Between Diagnostic Tests and Research Studies

Diagnostic Test	Research Study
Absence of disease	Truth of null hypothesis
Presence of disease	Truth of research (alternative) hypothesis
Positive result (outside normal limits)	Positive result (reject null hypothesis)
Negative result (within normal limits)	Negative result (fail to reject null hypothesis)
Sensitivity	Power
False-positive rate (1 − specificity)	*P* value
Prior probability of disease	Prior probability of research hypothesis
Predictive value of a positive (or negative) test result	Predictive value of a positive (or negative) study

From the Departments of Medicine (Dr Browner), Pediatrics (Dr Newman), and Epidemiology and International Health (Drs Browner and Newman), School of Medicine, University of California at San Francisco, and the Clinical Epidemiology Program, Institute for Health Policy Studies, San Francisco (Drs Browner and Newman).

Reprint requests to the Division of General Internal Medicine 111A1, Veterans Administration Medical Center, San Francisco, CA 94121 (Dr Browner).

Table 2.—The Four Possible Results of a Diagnostic Test

		If Breast Mass Is Actually:	
		Malignant	**Benign**
And Result of Fine-Needle Aspirate Is:	**Positive**	This is a true-positive test: result is correct	This is a false-positive test: result is incorrect
	Negative	This is a false-negative test: result is incorrect	This is a true-negative test: result is correct

Table 3.—The Four Possible Results of a Research Study

		If Research Hypothesis is Actually:	
		True (Efficacy of Drug A and Drug B Differ in Treatment of Ulcer Disease)	**False (Drug A Has Same Efficacy as Drug B in Treatment of Ulcer Disease)**
And Result of Study Is:	**Positive**	This is a true-positive study: result is correct	This is a false-positive study: result is incorrect
	Negative	This is a false-negative study: result is incorrect	This is a true-negative study: result is correct

misses the difference. If the two drugs actually have the same efficacy, then the study can either be correctly negative if it finds no difference or incorrectly positive if it does find one.

The relationships between the four possible outcomes of a diagnostic test are usually expressed as the *sensitivity* and *specificity* of the test, which are determined by assuming that the presence or absence of the disease is known. Sensitivity is the likelihood that a test result will be positive in a patient with the disease. Specificity is the likelihood that a test result will be negative in a patient without the disease. If the result from a fine-needle aspiration is positive in 80 of 100 women with breast cancer, and negative in 95 of 100 women without cancer, the test would have a sensitivity of 80% and a specificity of 95%. There is another term that is useful in the analogy: the false-positive rate (1 – specificity), which is the likelihood that a test result will be (falsely) positive in someone without the disease. In this example, the false-positive rate is 5%: of 100 women without breast cancer, five will have falsely positive test results.

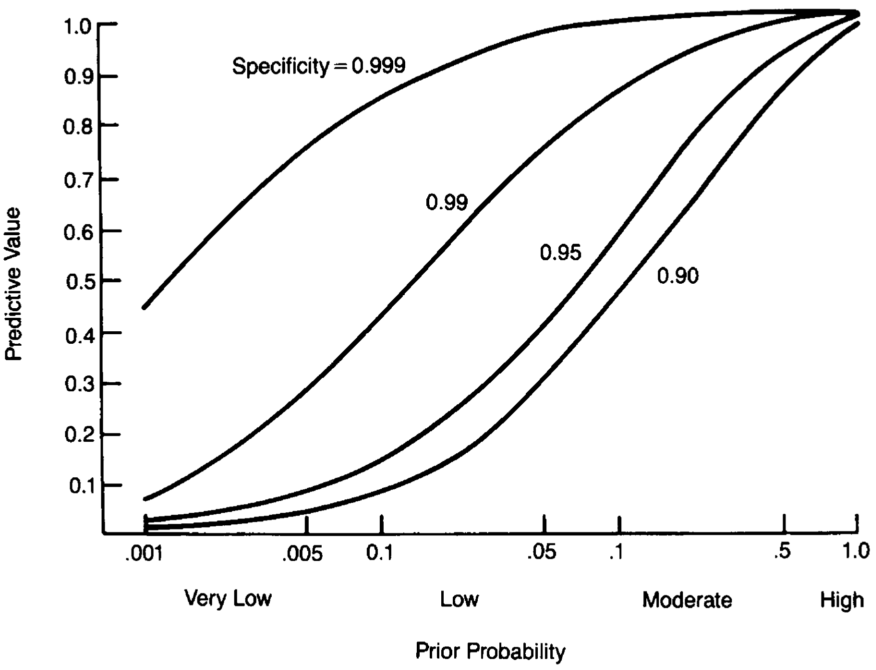
Similarly, the relationships between the four possible outcomes of a research study are usually expressed as the *power* and *P value* of the study, which are determined by assuming that the truth or falsity of the null hypothesis is known. Power is the likelihood of a study being positive if the research hypothesis is true (and the null hypothesis is false); it is analogous to the sensitivity of a diagnostic test. The *P* value is the likelihood of a study being positive when the null hypothesis is true; it is analogous to the false-positive rate (1 – specificity) of a diagnostic test. A study comparing two drugs in the treatment of ulcers that has an 80% chance of being correctly positive if there really is a difference in their efficacies would have a power of 0.80. A study with a 5% chance of being incorrectly positive if there is no difference between the drugs would have a *P* value of .05. (Conventionally, when the *P* value is less than a certain predetermined "level of statistical significance," usually .01 or .05, the results are said to be "statistically significant.")

Knowing the sensitivity and specificity of a test is not sufficient, however, to interpret its results: that interpretation also depends on the characteristics of the patient being tested. If the patient is a 30-year-old woman with several soft breast masses, a positive result from a fine-needle aspiration (even with a false-positive rate of only 5%) would not suffice to make a diagnosis of cancer. Similarly, if the patient is a 60-year-old woman with a firm solitary breast mass, a negative aspirate result (with a sensitivity of 80%) would not rule out malignancy.[19] Clinicians use these sorts of patient characteristics to estimate the *prior probability* of the disease—the likelihood that the patient has the disease, made prior to knowing the test results. The prior probability of a disease is based on the history and physical findings, previous experience with similar patients, and knowledge of alternative diagnostic explanations. It can be very high (breast cancer in the 60-year-old woman with a single firm mass), very low (breast cancer in the younger woman), or somewhere in between. Although they may not realize it, clinicians express prior probabilities when using phrases such as "a low index of suspicion" or "a strong clinical impression."

In the same way, knowing the power and the *P* value of a study is not sufficient to determine the truth of the research hypothesis. That determination also depends on the characteristics of the hypothesis being studied. Suppose one drug is diphenhydramine hydrochloride (Benadryl) and the other is chlorpheniramine maleate (Chlor-Trimeton): a positive study (at $P = .05$) would not ensure that one of the drugs is effective in the treatment of ulcers. Similarly, if one drug was ranitidine hydrochloride (Zantac) and the other a placebo, a negative study (even with power of 0.80) would not establish the ineffectiveness of ranitidine. The characteristics of a research hypothesis determine its prior probability—an estimate of the likelihood that the hypothesis is true, made prior to knowing the study results. The prior probability of a hypothesis is based on biologic plausibility, previous experience with similar hypotheses, and knowledge of alternative scientific explanations. Analogous to the situation with diagnostic tests, the prior probability of a research hypothesis can be very high (that an H_2-blocker, such as ranitidine, is more effective than placebo in the treatment of ulcers), very low (that the efficacies of two H_1-blockers, such as diphenhydramine and chlorpheniramine, differ in the treatment of ulcer disease), or somewhere in between. Authors of research reports indicate prior probabilities with terms like "unanticipated" or "expected" when they discuss their results.

The advantage of Bayesian analysis in interpreting diagnostic tests is that it can determine what the clinician really wants to know—the likelihood that the patient has the disease, given a certain test result. Bayesian analysis combines the characteristics of the patient (expressed as the prior probability of disease), the characteristics of the test (expressed as sensitivity and specificity), and the test result (positive or negative) to determine the predictive value of a test result. The *predictive value of a positive diagnostic test* is the probability that given a positive result, the patient actually has the disease. (The *predictive value of a negative test* is the probability that given a negative result, the patient does not have the disease.)

As an example, recall the 60-year-old woman with a firm breast mass. The prior probability that the mass is malignant is moderate, say 50%. A positive result from a fine-needle aspirate (with a specificity of 95% and a sensitivity of 80% for cancer) results in a very high predictive value for malignancy, about 94% (Figure). Next, consider the 30-year-old woman with multiple soft masses. The prior probability of cancer is low, say 1%. Even given a positive aspirate result, the likelihood that she has breast cancer is still small (about 14%).



Relationship between prior probability and predictive value of positive result of diagnostic test with sensitivity of 0.80, at several specificities. Figure can also be used to estimate predictive value of positive research study with power of 0.80 by substituting (1 − *P* value) for specificity, and prior probability of hypothesis for prior probability of disease (see "Limitations" section).

A Bayesian approach can also be used to determine what the reader of a research study really wants to know—the likelihood that the research hypothesis is true, given the study results. It combines the characteristics of the hypothesis (expressed as prior probability), the characteristics of the study (expressed as power and the *P* value), and the study results (positive or negative) to determine the predictive value of a study. The *predictive value of a positive study* is the probability that given a positive result, the research hypothesis is actually true. (The *predictive value of a negative study* is the probability that given a negative result, the research hypothesis is false.)

The predictive value of a research study, however, is usually harder to estimate than the predictive value of a diagnostic test (see "Limitations" section). Nonetheless, the basic analogy remains valid: the prior probability of the hypothesis must be combined with the power and the *P* value of the study to determine the likelihood that the research hypothesis is true. In the next section, we discuss how this analogy can be used to understand several statistical concepts.

IMPLICATIONS

Specificity and the *P* Value

How low must a *P* value be for it to be accepted as evidence of the truth of a research hypothesis? This question is analogous to asking: how high must the specificity of a test be to accept a positive test result as evidence of a disease? Requiring that a *P* value be less than .05 before it is "significant" is as arbitrary as requiring that a diagnostic test have a specificity of at least 95%. A more important criterion, but one that is not as easy to quantitate, is whether the results of the study combined with the prior probability of the research hypothesis are sufficient to suggest that the hypothesis is true. Consider the hypothesis, tested in the Lipid Research Clinics Primary Prevention Trial,[20] that cholestyramine resin decreases the incidence of coronary heart disease in hypercholesterolemic men. This research hypothesis had at least a low to moderate prior probability, based on previous evidence. Even with a "nonsignificant" *P* value of .094 (the two-sided equivalent of the controversial one-sided $P = .047$ reported by the investigators), the hypothesis is likely to be true.

It is also a mistake to *believe* a research hypothesis just because a *P* value *is* statistically significant. Consider a study that found that drinking two or more cups of coffee a day was associated with pancreatic cancer ($P < .05$).[21] This hypothesis had a very low prior probability: the authors called the association "unexpected." Thus, finding a significant *P* value did not establish the truth of the hypothesis; subsequent studies, including one by the same authors, failed to confirm the association.[22-27]

Of course, many diagnostic test results are not simply reported as "positive"; they also indicate how abnormal the result is. The more abnormal that result, the less likely that it is just a chance finding in a normal person. If the upper limit of normal for a serum thyroxine level at a specificity of 95% is 12.0 μg/dL (154 nmol/L), then a thyroxine level of 18.0 μg/L (232 nmol/L) is almost certainly abnormal. The question becomes whether it represents hyperthyroidism, another disease, or a laboratory error. By analogy, if the cutoff for calling a study positive is a *P* value less than .05, then a *P* value of .0001 means chance is an extremely unlikely explanation for the findings. The question becomes whether the results indicate the truth of the research hypothesis or are a result of confounding or bias (see "Laboratory Error and Bias" and "Alternative Diagnoses and Confounding Explanations" sections). Because the *P* value is analogous to the false-positive rate (1 − specificity), a study with a very low *P* value is like a test with very high specificity: both give few false-positive results due to chance, but may require careful consideration of other possible explanations.

Sensitivity and Power

When the result of a diagnostic test that has a high sensitivity is negative, such as a urinalysis in the diagnosis of pyelonephritis, it is especially useful for ruling out a disease. Similarly, when a powerful research study is negative, it strongly suggests that the research hypothesis is false. However, if the sensitivity of a test is low, such as a sputum smear in a patient with possible tuberculosis, then a negative result does not rule out the disease.[9] In the same way, a negative study with inadequate power cannot disprove a research hypothesis.[28,29]

Laboratory Error and Bias

When unexpected or incredible results on a diagnostic test are found, such as a serum potassium level of 9.0 mEq/L (mmol/L) in an apparently well person, the first possibility to consider is laboratory error: Was the test adequately performed? Did the sample hemolyze? Was the specimen mislabeled? Similarly, readers of a research study, such as a trial of biofeedback in the treatment of hypertension, must always consider the possibility of bias, especially if the study yields surprising results: Was the study adequately designed and executed? Did the investigators assign subjects randomly? Was blood pressure measured blindly?[30] Improperly performed tests and biased studies do not yield reliable information, no matter how specific or significant their results.

Alternative Diagnoses and Confounding Explanations

Even if a diagnostic test is adequately performed, there may be several explanations for the result. An elevated serum amylase level, for example, has a high specificity to distinguish patients who have pancreatitis from those with nonspecific abdominal pain. However, there are extrapancreatic diseases (such as bowel infarction) that elevate the amylase level and that must be considered in the differential diagnosis. In the same way, although a low *P* value may indicate an association between an exposure and a disease (like the association between carrying matches and lung cancer), a confounder (cigarette smoking) may actually be responsible. Readers of research studies should always keep in mind potential confounding explanations for significant *P* values.

Better Tests and Bigger Studies

Increasing the sample size in a research study is similar to using a better diagnostic test. Better diagnostic tests can have more sensitivity or specificity or both; large studies can have greater power or lower levels of statistical significance or both. Often the choice of a diagnostic test is a matter of practicality: biopsies are not feasible in every patient for every disease. Similarly, power or the significance level may be determined by practical considerations, since studies of 20 000 or more subjects cannot be done for every research question. Of course, bigger studies may find smaller differences, just like better tests may detect less advanced cases of a disease. A small but statistically significant difference in a research study is like a subtle but definite abnormality on a diagnostic test; its importance is a matter of judgment.

Intentionally Ordered Tests and Prospective Hypotheses

A positive result on a single intentionally ordered test is more likely to indicate disease than the same result that turns up on a set of routine admission laboratory tests. Similarly, the *P* value for a research hypothesis stated in advance of a study is usually more meaningful than the same *P* value for a hypothesis generated by the data. The reason is that clinicians usually order tests and investigators state hypotheses in advance when the prior probability is moderate or high. Thus the predictive values of positive results are generally greater for intentionally ordered tests and prospectively stated hypotheses.

Not all unexpected results, however, have low prior probabilities. Occasionally, clinicians or investigators are just not smart or lucky enough to consider the diagnosis or hypothesis in advance. For example, a house officer caring for a patient with fatigue and vague abdominal symptoms might ignore a serum calcium level of 10.5 mg/dL (2.62 mmol/L) until the attending physician mentions the possibility of hyperparathyroidism in rounds the next morning. Similarly, researchers might disregard the association between smoking and cervical cancer until a plausible biologic explanation is suggested.[31-34] Estimating the prior probability of a hypothesis on the basis of whether it was considered prospectively is a useful, but not infallible, method. The truth, elusive though it sometimes may be, does not depend on when a hypothesis is first formulated.

Multiple Tests and Multiple Hypotheses

Most of us are intuitively skeptical when one of 50 substances on a checklist is associated with a disease at $P<.05$ because of the likelihood of finding such an association by chance alone. A standard technique for dealing with this problem of testing multiple hypotheses is to use a more stringent level of statistical significance, thus requiring a lower *P* value.[35,36] This approach is simple and practical, but it leads to some unsatisfying situations. It seems unfair, for example, to reduce the required significance level for a reasonable hypothesis just because other, perhaps ridiculous, hypotheses were also tested. What if the disease was mesothelioma and one of the exposures was asbestos: should a more stringent level of statistical significance be required because 49 other substances were also included? Should the level of significance be reduced when testing the main hypothesis of a study whenever additional hypotheses are considered? Need statistical adjustments for multiple hypothesis testing be made only when reporting all of the hypotheses in a single publication?

This vexing problem of multiple hypothesis testing resembles the interpretation of a serum chemistry panel. When a clinician evaluates a patient with a swollen knee, a serum uric acid level of 10.0 mg/dL (0.6 mmol/L) has the same meaning no matter how many other tests were also performed on the specimen by the autoanalyzer. However, an unanticipated abnormal value on another test in the panel is likely to be a false-positive: that is because the diseases it might represent usually have low prior probabilities, *not* because several tests were performed on the same sample of serum. Similarly, testing multiple hypotheses in a single study causes problems because the prior probabilities of such hypotheses tend to be low: when investigators are not sure of what they are looking for, they test many possibilities. The solution is to recognize that it is not the number of hypotheses tested, but the prior probability of each of them, that determines whether a result is meaningful.[37]

Confirmatory Tests and Pooled Studies

When a single diagnostic test is insufficient to make a diagnosis, additional tests are often ordered, some results of which may be positive and some negative. The clinician revises the probability of the disease by combining these results, often weighting them by the tests' characteristics. In a patient with a swollen leg, for example, a normal result from a Doppler study would lower the probability of deep venous thrombosis, but an abnormal result of a fibrinogen scan might raise it sufficiently to make the diagnosis. In the same way, it may be necessary to combine the results of several research studies, weighting them by the characteristics of each study. This process, known as *pooling*, allows studies with both significant and nonsignificant *P* values to change incrementally the likelihood that a research hypothesis is true. However, just as only those tests that are relevant to the diagnosis in question should be combined, only those research studies that address the same research hypothesis should be pooled.

Confidence Intervals

There is no ready diagnostic test analogy for confidence intervals from research studies (the concept of test precision comes closest). But because confidence intervals are commonly misinterpreted as expressions of predictive value, they merit a short discussion. The term "confidence interval" is unfortunate, because it leads many people to believe that they can be confident that the interval contains the true value being estimated. Actually, confidence intervals are determined entirely by the study data: the prior probability that the true value lies within that interval is not at all considered in the calculations. A 95% confidence interval is simply the range of values that would *not* differ from the estimate provided by the study at a statistical significance level of .05.[38,39]

Confidence intervals are useful because they define the upper and lower limits consistent with a study's data. But they do not estimate the likelihood that the results of the research are correct. A confidence interval provides no more information about the likelihood of chance as an explanation for a finding than does a P value.[40] As an example, suppose a well-designed study finds that joggers are twice as likely as nonjoggers to develop coronary heart disease, with a 95% confidence interval for the relative risk of 1.01 to 3.96. (This is equivalent to rejecting the null hypothesis of no association between jogging and heart disease at $P = .05$.) Despite a 95% confidence interval that excludes 1.0, there is obviously not a 95% likelihood that joggers are at an increased risk of coronary heart disease. There are many other studies that have found that exercise is associated with a reduced risk of heart disease. Given the low prior probability of the hypothesis that jogging increases the risk of coronary heart disease, chance (or perhaps bias) would be a more likely explanation for the results.

LIMITATIONS

While it provides several useful insights, the analogy between diagnostic tests and clinical research is not perfect. It is easier to determine the prior probability of a disease, based on the prevalence of the disease in similar patients, than the prior probability of a hypothesis, based on the prevalence of the truth of similar hypotheses. Similarity in patients can be defined by characteristics known to be associated with a disease, such as age, sex, and symptoms.[11] But what defines similar hypotheses? Thus the prior probability of most research hypotheses tends to be a subjective estimate (although, in practice, estimates of the prior probability of a disease are generally subjective as well).

Second, as long as there is a gold standard for its diagnosis, a disease is either present or absent: there are only these two possibilities. If a group of patients known to have the disease is assembled, a single value for the sensitivity of a test can be determined empirically. But there is no single value for the power of a research study: it depends on the sample size, as well as the magnitude of the actual difference between the groups being compared. A study comparing IQ in internists and surgeons, for example, might have a power of only 50% to detect a difference between them if surgeons actually scored five points higher than internists, but a power of 98% if surgeons actually scored ten points higher. Since the actual difference is unknown, a unique value for power cannot be calculated.

CONCLUSIONS

Clinicians do not simply decide that a patient has a disease when a diagnostic test result is positive or rule out the disease when the test result is negative. They also consider the sensitivity and specificity of the test and the characteristics of the patient being tested. In the same way, readers should not believe or disbelieve the research hypothesis of a study on the basis of whether the results were statistically significant. They should also take into account the study's power and P value and the characteristics of the hypothesis being tested.

Thus, all significant P values are not created equal. Just as the accuracy of a diagnosis depends on how well the clinician has estimated the prior probability and considered alternative diagnoses and laboratory errors, the interpretation of a research study depends on how well the reader has estimated the prior probability and considered confounders and biases. Knowing the power and P value (or the confidence interval) for a study's results, like knowing the sensitivity and specificity of a diagnostic test, is necessary but not sufficient. This Bayesian approach requires the active participation of the reader and emphasizes the importance of scientific context in the interpretation of research.

This project was supported by a grant from the Andrew W. Mellon Foundation.

References

1. Schade DS, Mitchell WJ, Griego G: Addition of sulfonylurea to insulin treatment in poorly controlled type II diabetes: A double-blind, randomized clinical trial. *JAMA* 1987;257:2441-2445.
2. Cramer DW, Goldman MB, Schiff I, et al: The relationship of tubal infertility to barrier method and oral contraceptive use. *JAMA* 1987;257:2446-2450.
3. Bennett KJ, Sackett DL, Haynes RB, et al: A controlled trial of teaching critical appraisal of the clinical literature to medical students. *JAMA* 1987;257:2451-2454.
4. Chaiken BP, Williams NM, Preblud SR, et al: The effect of a school entry law on mumps activity in a school district. *JAMA* 1987;257:2455-2458.
5. Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 1983;98:385-394.
6. Nugent CA, Warner HR, Dunn JT, et al: Probability theory in the diagnosis of Cushing's syndrome. *J Clin Endocrinol Metab* 1964;24:621-627.
7. Crapo L: Cushing's syndrome: A review of diagnostic tests. *Metabolism* 1979;28:955-977.
8. Vecchio TJ: Predictive value of a single diagnostic test in unselected populations. *N Engl J Med* 1966;274:1171-1173.
9. Boyd JC, Marr JJ: Decreasing reliability of acid-fast smear techniques for detection of tuberculosis. *Ann Intern Med* 1975;82:489-492.
10. Jones RB: Bayes' theorem, the exercise ECG, and coronary artery disease. *JAMA* 1979;242:1067-1068.
11. Diamond GA, Forrester JS: Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med* 1979;300:1350-1358.
12. Griner PF, Mayewski RJ, Mushlin AI, et al: Selection and interpretation of diagnostic tests and procedures: Principles and applications. *Ann Intern Med* 1981;94:553-600.
13. Havey RJ, Krumlovsky F, delGreco F, et al: Screening for renovascular hypertension: Is renal digital-subtraction angiography the preferred noninvasive test? *JAMA* 1985;254:388-393.
14. Bayes T: An essay towards solving a problem in the doctrine of chances. *Philos Trans R Soc Lond* 1763;53:370-418.
15. Phillips LD: *Bayesian Statistics for Social Scientists.* New York, Crowell, 1974.
16. Donner A: A Bayesian approach to the interpretation of subgroup results in clinical trials. *J Chronic Dis* 1982;35:429-435.
17. Pater JL, Willan AR: Clinical trials as diagnostic tests. *Controlled Clin Trials* 1984;5:107-113.
18. Thomas DC, Siemiatycki J, Dewar R, et al: The problem of multiple inferences in studies designed to generate hypotheses. *Am J Epidemiol* 1985;122:1080-1095.
19. Mushlin AI: Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. *Ann Intern Med* 1985;103:79-85.
20. Lipid Research Clinics Program: The Lipid Research Clinics Coronary Primary Prevention Trial results: I. Reduction in incidence of coronary heart disease. *JAMA* 1984;251:351-364.
21. MacMahon B, Yen S, Trichopolous D, et al: Coffee and cancer of the pancreas. *N Engl J Med* 1981;304:630-633.
22. Jick H, Dinan BJ: Coffee and pancreatic cancer. *Lancet* 1981;2:92.
23. Goldstein HR: No association found between coffee and cancer of the pancreas. *N Engl J Med* 1982;306:997.
24. Wynder EL, Hall NEL, Polansky M: Epidemiology of coffee and pancreatic cancer. *Cancer Res* 1983;43:3900-3906.
25. Kinlen LJ, McPherson K: Pancreas cancer and coffee and tea consumption: A case-control study. *Br J Cancer* 1984;49:93-96.
26. Gold EB, Gordis L, Diener MD, et al: Diet and other risk factors for cancer of the pancreas. *Cancer* 1985;55:460-467.
27. Hsieh C, MacMahon B, Yen S, et al: Coffee and pancreatic cancer (chapter 2). *N Engl J Med* 1986;315:587-589.
28. Frieman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, type II errors and sample size in the randomized control trial: Survey of 71 'negative' trials. *N Engl J Med* 1978;299:690-694.
29. Young MJ, Bresnitz EA, Strom BL: Sample size nomograms for interpreting negative clinical studies. *Ann Intern Med* 1983;99:248-251.
30. Sackett DL: Bias in analytic research. *J Chronic Dis* 1979;32:51-63.
31. Winkelstein W Jr: Smoking and cancer of the uterine cervix: Hypothesis. *Am J Epidemiol* 1977;106:257-259.
32. Wright NH, Vessey MP, Kenward B, et al: Neoplasia and dysplasia of the cervix uteri and contraception: A possible protective effect of the diaphragm. *Br J Cancer* 1978;38:273-279.
33. Harris RWC, Brinton LA, Cowdell RH, et al: Characteristics of women with dysplasia or carcinoma in situ of the cervix uteri. *Br J Cancer* 1980;42:359-369.
34. Lyon JL, Gardner JW, West DW, et al: Smoking and carcinoma in situ of the uterine cervix. *Am J Public Health* 1983;73:558-562.
35. Godfrey K: Comparing the means of several groups. *N Engl J Med* 1985;313:1450-1456.
36. Cupples LA, Heeren T, Schatzkin A, et al: Multiple testing of hypotheses in comparing two groups. *Ann Intern Med* 1984;100:122-129.
37. Cole P: The evolving case-control study. *J Chronic Dis* 1979;32:15-27.
38. Fleiss JL: *Statistical Methods for Rates and Proportions*, ed 2. New York, John Wiley & Sons Inc, 1981, p 14.
39. Rothman K: A show of confidence. *N Engl J Med* 1978;299:1362-1363.
40. Browner WS, Newman TB: Confidence intervals. *Ann Intern Med* 1986;105:973-974.