

Bayesian Analysis Revisited: A Radiologist's Survival Guide

Paul J. Chang¹

Today's clinicians are confronted by a bewildering number of new, occasionally intimidating, diagnostic imaging tests. Increasingly, colleagues are calling on the radiologist not only to interpret imaging studies but also to provide guidance in the rational selection of appropriate tests and to offer opinions as to the clinical usefulness of newer techniques. Addressing the latter concerns will be significantly more important for the radiologist in light of present and future economic (and other external) constraints. Although thoroughly trained to perform and interpret diagnostic imaging studies, radiologists in general are not adequately trained to address these important concerns regarding analytic decisions.

Much of the radiologist's advice to the clinician (as well as our enthusiasm for newer imaging techniques) is derived from our review of the radiologic literature. Recently, there has been increasing criticism of this literature with respect to the evaluation of the clinical efficacy of our newer techniques (e.g., MR imaging, transrectal sonography); much of this criticism comes from the same clinicians who depend on radiologists to give them accurate advice [1-5]. A number of workers have raised serious concerns about the design and interpretation of radiologic experimental studies; they point out that many papers contain significant design errors/biases that make our recommendations (and enthusiasm) regarding certain newer imaging tests suspect [1, 5].

A review of the radiologic literature supports this concern; many studies reflect a basic misunderstanding of the fundamental principles of Bayesian analysis. This problem is compounded by the fact that most practicing radiologists feel uncomfortable with Bayesian analysis and, as a result, cannot

critically evaluate the "validity" of the literature. This ability is crucial for the radiologist, who must make an informed contribution to the clinical management of patients.

Thus, it is time to revisit Bayes' theorem. One reason that radiologists have trouble understanding and correctly using this analytical tool is that we were originally exposed to Bayesian analysis in a rather dry, theoretical manner, with a confusing array of contingency tables and equations that soon were forgotten. Although this approach is rigorous, it does not yield an *intuitive* understanding of the concepts. The purpose of this paper, therefore, is not to give an exhaustive, formal exposition of Bayesian analysis but to give radiologists an intuitive feel for the basic principles involved.

What the Clinician Knows and Wants to Know

Prior Probability

Before the selection or evaluation of a specific diagnostic imaging test is considered, it is important to understand clearly what confronts the referring clinician: a patient with a constellation of historical, physical, and laboratory findings, all of which suggest a tentative clinical or differential diagnosis. This leads to the important Bayesian concept of the *prior* or "*pretest*" *probability*, the clinician's initial assessment (i.e., *before* the imaging test is performed) of how likely it is that the patient truly has a specific disease.

Many workers use the term *prevalence* interchangeably with "prior probability." This is misleading because the clinician usually has significantly more information than the disease

Received June 10, 1988; accepted after revision December 14, 1988.

¹ Department of Diagnostic Radiology and Nuclear Medicine, Stanford University School of Medicine, Stanford, CA 94305. Address reprint requests to P. J. Chang.

prevalence before referral to the radiologist. However, disease prevalence is important in deriving a reasonable prior probability and, occasionally, may be the only information available before referral.

Action Threshold

Clinicians institute therapy or other intervention only if they are convinced that the probability that the patient has the disease in question is beyond an "action threshold" probability. In addition, most physicians will not exclude a diagnosis unless the probability is below an "exclusion threshold." In most cases, the prior or pretest probability will lie somewhere between these threshold values (Fig. 1). Therefore, the purpose of the diagnostic test is to obtain additional information to move the *posterior* or "*posttest*" probability either above the action threshold or below the exclusion threshold.

It is important for the radiologist to determine whether the clinician truly has an action and/or exclusion threshold before a test is recommended; if no threshold exists, the rationality of performing *any* test must be questioned (e.g., the frequently encountered scenario in which the results of a requested imaging test will *not* alter a clinician's decision to institute therapy). Similarly, if the clinician's prior probability is already beyond either the action or exclusion threshold, the usefulness of performing any additional confirmatory tests must be questioned.

This principle applies primarily to tests used for *diagnosis*. There are, of course, other valid reasons for performing imaging tests that are not strictly for diagnostic purposes, such as providing anatomic information for surgical planning, tumor staging, and evaluating treatment response. In these cases, patients would have prior probabilities beyond the action threshold.

How Good Is the Imaging Test?

After the clinician determines the prior probability and the action and exclusion threshold values, the radiologist is ready

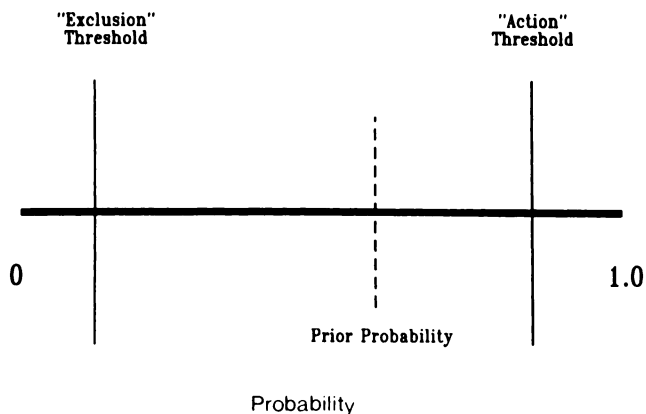


Fig. 1.—Goal of diagnostic test is to update prior or "pretest" probability above or below "action" or "exclusion" threshold, respectively. This updated "posttest" probability represents posterior probability.

to suggest an appropriate diagnostic imaging test. An important consideration in the selection of a test is the test's ability to discriminate between those who have the disease in question and those who do not. The characterization of the test's discriminating performance involves principles derived from Bayesian analysis; these principles should be explicitly addressed by any experimental study evaluating the efficacy of an imaging technique.

The Gold Standard

Determining who has the disease and who does not *must* be done on some basis other than the examination being evaluated. This gold standard may represent pathologic and/or surgical correlation; occasionally, it may correspond to the results of another diagnostic test (i.e., pulmonary angiography for pulmonary embolism or contrast venography for deep venous thrombosis). Despite its label, the gold standard is frequently "imperfect," and it is subject to both false-positive and false-negative results. In addition, one gold standard may be replaced by another as a result of such factors as improved technology or surgical or laboratory advances. The important point is that any experimental study that evaluates the efficacy of diagnostic imaging tests *must* have an explicitly defined, independently derived gold standard. For the purposes of Bayesian analysis, the gold standard is considered to be "perfect," with no false-positive or false-negative results assumed.

Discriminant Criteria

Discriminant criteria are defined specifically for each diagnostic test. For example, possible discriminant criteria for the diagnosis of a malignant liver mass by CT could include the size/configuration of the lesion, attenuation, or contrast enhancement characteristics. (For the purposes of this discussion, let us assume that there is a single discriminant criterion used in our test.) The test is then performed on a population (ideally in a prospective manner), and this discriminant criterion is used with the classification of those with and without the disease defined independently by the gold standard (Fig. 2).

Positivity Criterion

Unfortunately, an overlap almost always exists between those with the disease and those without when the defined discriminant criterion is used. Accordingly, a threshold or "positivity" criterion must be selected according to which an individual is to be called "positive" or "negative" by the test (Fig. 2). The definition of this positivity criterion should be stated explicitly when the efficacy of diagnostic tests is evaluated.

Sensitivity and Specificity

The discriminating power of the imaging test is characterized by the degree of overlap between the diseased and

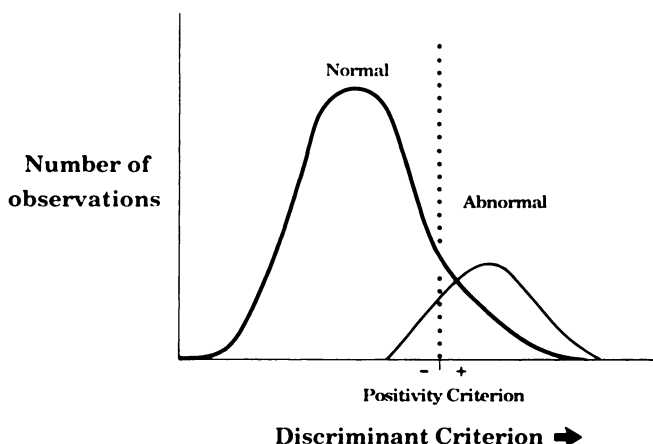


Fig. 2.—Degree of separation between abnormal and normal populations when a specific discriminant criterion is used determines discriminating ability of test. Selected positivity criterion defines a positive or negative test.

disease-free populations present when the defined discriminant and positivity criteria are used. This discriminating power can be described by the *sensitivity* and *specificity* of the test. Sensitivity is defined as the true-positive rate, the proportion of those with the disease (as defined by the gold standard) who test positive when the defined discriminant criteria and positivity criteria are used (Fig. 3A). Specificity is one minus the false-positive rate, where the false-positive rate is defined as the proportion of those without the disease who test positive (Fig. 3B).

Bayes' Theorem

Bayes' theorem models the performance of a diagnostic test as it relates to a specific clinical application by incorporating not only the discriminating power of the diagnostic test

(as measured by sensitivity and specificity) but also the prior or pretest probability to derive the posterior or posttest probability. The formula for Bayes' theorem is a familiar one (Fig. 4); unfortunately, it is also easily forgotten or blindly applied without true understanding of its underlying meaning. It is not important to memorize this formula; it can be looked up as needed. The absolutely crucial point to remember from Bayes' theorem is that *the prior probability (or pretest clinical assessment) is as important as the sensitivity and specificity of the diagnostic test in the determination of the posterior probability (the probability that the patient with a positive test truly has the disease)*.

The foregoing statement can be demonstrated by the use of dreary contingency tables and mathematical formulas; however, a more intuitive understanding of this principle can be achieved by a simple illustration:

Suppose a 25-year-old woman comes to the gynecology clinic worried that she might have syphilis. Assume the sensitivity of the VDRL is 0.90 and the specificity is 0.85. A VDRL is performed and is POSITIVE. What is the probability that this patient truly has syphilis?

An informal survey of physicians at our institution revealed that most clinicians and radiologists believed the probability that this hypothetical patient had syphilis would be relatively high, say about .75–.80. I encourage readers to estimate their own probabilities for this scenario. Now consider the following scenario:

Scenario One: The woman has a past history of multiple episodes of pelvic inflammatory disease and is an IV drug abuser.

Given this information, most physicians updated their probability upward to around .90–.99. The next scenario demonstrates the essential Bayesian principle:

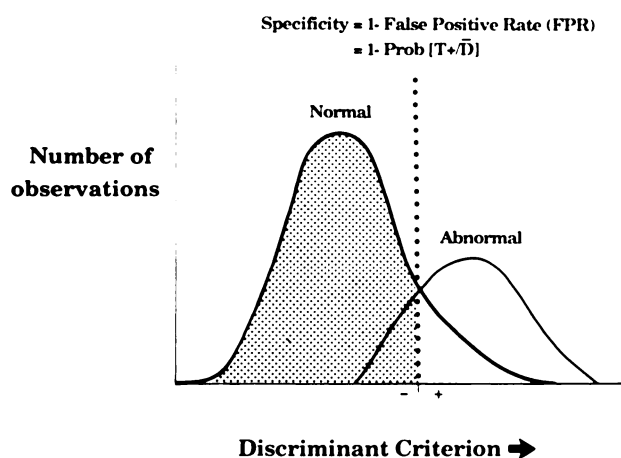
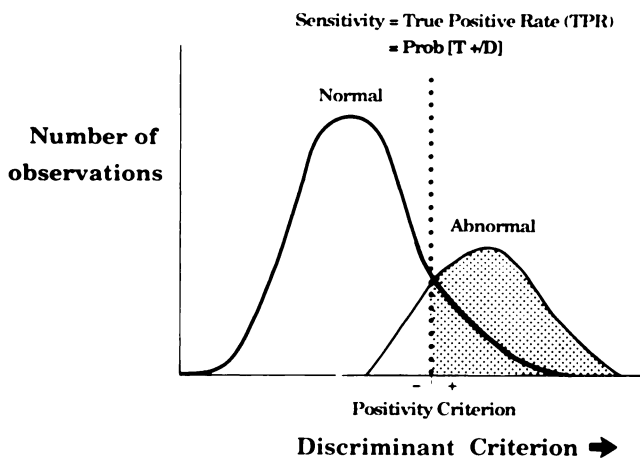


Fig. 3.—Discriminating ability of a test using defined discriminant and positivity criteria is characterized by sensitivity (A) and the specificity (B). Prob [T+|D] represents the probability of a positive test in those with the disease. Prob [T+|D̄] represents the probability of a positive test in those who do not have the disease.

Bayes' Theorem:

$$P [D|T+] = \frac{(TPR \times \text{Prior Prob})}{(TPR \times \text{Prior Prob}) + [FPR \times (1 - \text{Prior Prob})]}$$

Fig. 4.—Bayes' theorem shows how posterior probability (P[D|T+] represents probability of having disease if test is positive) is a function not only of test efficacy (as defined by sensitivity and specificity) but also of prior probability. TPR = true positive rate or sensitivity; FPR = false positive rate or [1 - specificity].

Scenario Two: The woman is a virgin studying to be a nun and is afraid that she might have caught syphilis by thinking "impure, immoral thoughts."

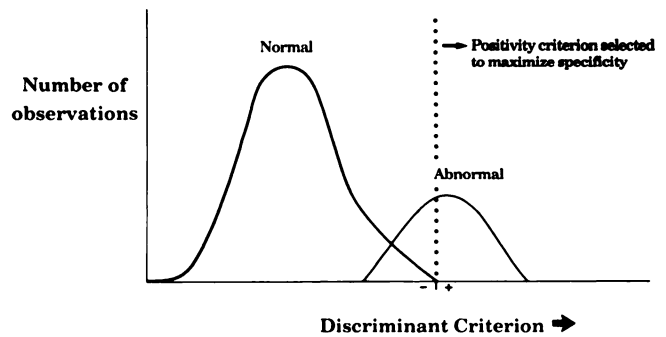
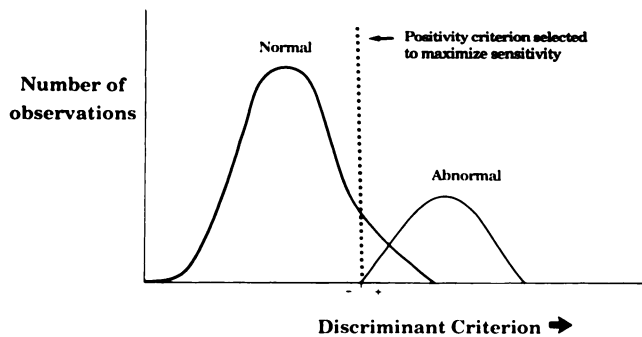
With this alternative scenario, all physicians surveyed dramatically lowered their estimate to be less than .10. The important observation is that in both scenarios, *the performance characteristics of the test (as described by the sensitivity and specificity) did not change; it was only the prior probability that changed.* Clearly, the posterior probability (the probability the patient with a positive test truly has the disease) depends not only on test efficacy but, to a high degree, on the prior probability. This simple illustration provides an intuitive understanding of the essential principle behind Bayes' theorem that is not as easily forgotten as the mathematical formula.

The Fluidity of the Positivity Criterion: The ROC Curve

Examination of Figures 2 and 3 suggests that the observed test sensitivity and specificity can be changed by moving the positivity criterion (Figs. 5A and 5B). By moving this threshold criterion, the sensitivity of our test is increased but at the expense of the specificity and vice versa. This change in test sensitivity and specificity as a function of the positivity criterion does *not* alter the intrinsic discriminating power of the examination (the degree of overlap between diseased and undiseased populations).

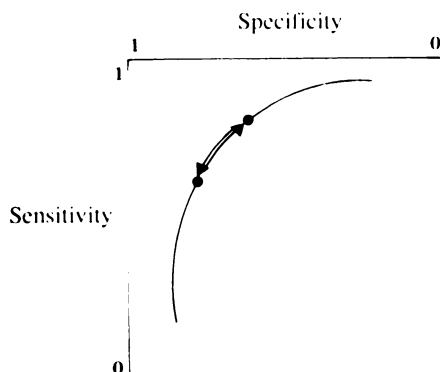
The selection of the positivity criterion can be quite fluid and specific to the application, especially when the test interpretation is subjective and observer-dependent. This can be demonstrated in any radiology department: a vague, somewhat "nodular" shadow on a chest film can be interpreted either as "worrisome for metastasis, suggest aggressive workup" (in the case of a patient with known malignancy) or as "the probable confluence of normal vascular and osseous structures" (in the case of the "routine" preemployment chest film). This familiar example represents the shifting of the positivity criterion toward improved sensitivity (in the first patient) or specificity (in the second patient).

Accordingly, a wide variation can be observed in test sensitivity and specificity corresponding to identical discriminant criteria and test efficacy. This makes the interpretation of published test performance expressed in terms of sensitivity and specificity alone sometimes difficult and misleading. Clearly, the scalar values of sensitivity and specificity do not describe adequately the intrinsic discriminating power of a test subject to variation in the positivity criterion.



A

B



C

Fig. 5.—Moving positivity criterion can optimize either sensitivity (A) or specificity (B). Reciprocal relationship between sensitivity and specificity as a function of changing positivity criterion can be expressed by receiver operating characteristic curve (C).

The reciprocal relationship between sensitivity and specificity as a function of varying the positivity criterion can be represented more effectively by the use of the receiver operating characteristic (ROC) curve (Fig. 5C). This curve is a more meaningful representation of the intrinsic discriminating power of a given test and enables the direct comparison of the relative performance of different tests, especially when these tests are subject to variation in the positivity criteria. A discussion of the power and elegance of ROC analysis is beyond the scope of this paper; the interested reader is directed to the excellent review by Metz [6].

The foregoing discussion emphasizes the importance of obtaining accurate and thorough clinical information from the referring physician before imaging tests are selected and performed. Without adequate "clinical history," the radiologist is unable to select an appropriate positivity threshold, which results in unacceptably high false-negative or false-positive interpretations. Thorough clinical information also enables the radiologist to help the clinician select an appropriate imaging test: knowledge of the prior probability and the action/exclusion thresholds can be used with Bayes' theorem to select a test with appropriate sensitivity/specificity characteristics [7].

Pitfalls

With this introduction to Bayesian analysis, one can now discuss the various misinterpretations of the theory found in our literature and in our everyday practice; the investigator as well as the practicing radiologist must be able to recognize these potential pitfalls.

No Gold Standard

One cannot meaningfully discuss "sensitivity" or "specificity" without an explicitly defined, independently derived gold standard. Unfortunately, some recent articles, especially those regarding MR imaging, have lacked an *independent* gold standard or used the imaging techniques being evaluated themselves as the gold standard. This leads to misleadingly high "sensitivities" and "specificities."

Ignoring the Importance of the Prior Probability

Bayes' theorem shows that the probability of having the disease given a positive test (the posterior probability) is not the same as the probability of a positive test given the disease (the sensitivity). The relationship between these two values is defined by Bayes' formula and is highly dependent on the prior probability. Radiologists must remind clinicians of this fact before and *after* an examination is performed. We should not be "seduced" by performance characteristics of an imaging test with "high" sensitivity and/or specificity; the interpretation and clinical usefulness of the test are highly dependent on the prior probability—a "positive" test does not necessarily confirm a diagnosis.

Failure to Take Operator-Dependence into Account

The previously discussed variation in observed test sensitivity and specificity as a function of the positivity criterion

suggests that these scalar values do not describe adequately the intrinsic discriminating power of the test when intraobserver variability exists. We must be wary of reported sensitivities and specificities with any imaging technique in which there is considerable observer dependence; your personal experience probably will be different. An ROC curve analysis should be used in the evaluation of tests subject to intraobserver variability.

Avoid "Positive Predictive Value"

In the radiologic literature, authors tend to publish the positive predictive value (PPV) in addition to sensitivity and specificity. The PPV is identical to the posterior or posttest probability discussed previously. I would encourage investigators to avoid stating this value; readers should ignore this number. I believe that the PPV has great potential to mislead. As has been shown, the PPV is highly dependent not only on the performance characteristics of the test being evaluated, but also on the prior probability. Because most studies evaluating imaging are performed at tertiary referral medical centers, the prior probability (which incorporates the prevalence seen at that center) will almost always be much higher than what the "average" practicing radiologist will see. (This is known as *referral bias* [8]. In addition, the prior probability at such centers can be overestimated by a form of *work-up bias*, in which prior test results contribute to the inclusion or exclusion of a patient [8]. As a result, patients with "negative" prior tests may not be evaluated with the imaging technique being studied, thus falsely overestimating the prior probability in retrospective studies.) The high prior probability seen by the investigator usually results in a very high PPV for almost any imaging test being evaluated. Because some will remember that the formal definition of the PPV is "the probability of a patient truly having the disease given a positive test," the practicing radiologist can be given a false sense of security and have ill-founded high expectations for test performance; the actual posttest probability of a positive test will almost always be lower for the practicing radiologist.

Ignoring the Dynamic Nature of Diseases

Occasionally, one sees in the literature a relatively wide variation in observed sensitivities and specificities when the same imaging technique is evaluated. Two "classic" Bayesian hypotheses frequently are used to explain this apparent discrepancy: (1) The various investigators were using different positivity criteria. This would result in a reciprocal relationship between sensitivity and specificity on the *same* ROC curve as seen in Figure 5C; studies claiming high sensitivities would also show lower specificities and vice versa. (2) There is great operator/machine dependence; some investigators are either "better" observers or technically "superior"; alternatively, some workers have "better" or more advanced equipment ("my machine is better than yours"). This would correspond to *different* ROC curves, with one showing superior performance in both sensitivity and specificity relative to another (Fig. 6). Another explanation is intuitively clear but frequently ignored by investigators and practicing radiologists reviewing the literature: "*sensitivity*" and "*specificity*" are actually func-

tions of time and both will change as the disease progresses with time, independent of the intrinsic "goodness" of the test. This is shown in Figures 7A and 7B. Early in almost any disease, significant overlap between diseased and disease-free populations is expected, regardless of the diagnostic test or discriminant criteria (Fig. 7A). However, as the disease continues, overlap between the two populations will be progressively less regardless of the specific imaging test used (e.g., a liver metastasis will increase in size, change attenuation values, and increasingly disrupt surrounding parenchyma as it grows, thus making it easier to detect by CT). As a result, we will observe improvement in both the sensitivity and specificity of any test over time as the disease progresses (Fig. 7B). Clearly, the observed sensitivity and specificity of any test will be highly dependent on the temporal stage of the disease.

Radiologists must be aware of this temporal dependency of sensitivity and specificity on the dynamic progression of the disease process. The extrapolation of sensitivity/specificity data derived from relatively advanced disease or from consolidating data spanning all stages of a disease (especially neoplasm) to draw inferences about the performance of an imaging test in the direction of early lesions is highly mislead-

ing. Investigators must explicitly control for the temporal stage of the disease being tested.

Comparing Techniques

An important extension of this principle involves studies that attempt to compare the relative performance of various imaging techniques (e.g., comparing CT, MR, and sonography in the detection of liver metastasis). Again, the radiologist must recognize the temporal dependency of sensitivity/specificity on the dynamic progression of the disease. If this dependency is ignored, comparative studies of imaging techniques can be misleading.

The following hypothetical example comparing CT and transrectal sonography in the detection of prostatic carcinoma illustrates the problem. We have already shown that sensitivity (and specificity) are actually functions of time, so it is reasonable to regard Figure 8A as a conceivable sensitivity function curve for CT in the detection and/or staging of prostate cancer. The S shape of the curve reflects the relatively low sensitivity of CT for detecting intraparenchymal lesions that have yet to invade the periprostatic fat or the seminal vesicles. Once the disease progresses through the capsule and in-

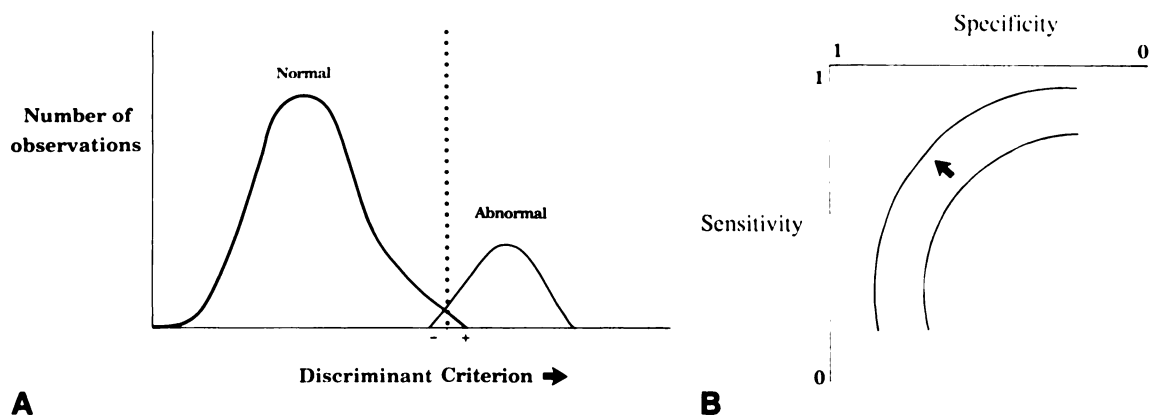


Fig. 6.—We can improve our discriminating ability by using a different test (or by using a different discriminant criterion); this improvement in test performance is due to less overlap between abnormal and normal populations (A). Improvement in both sensitivity and specificity corresponds to moving to a different receiver operating characteristic curve (B).

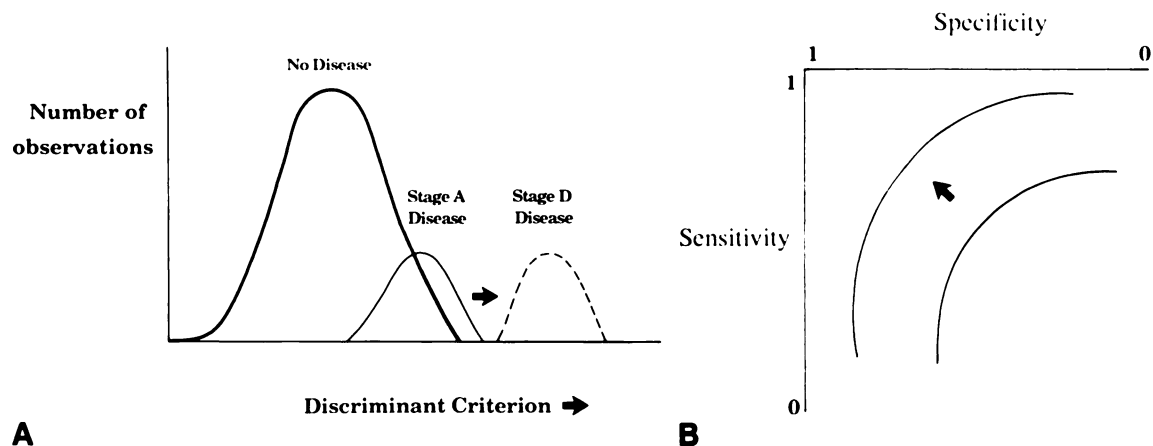
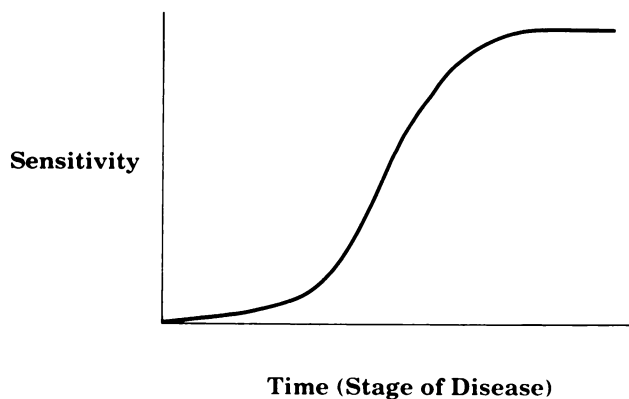
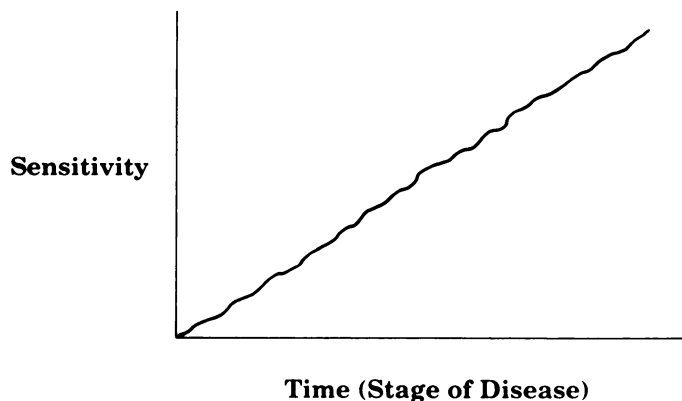


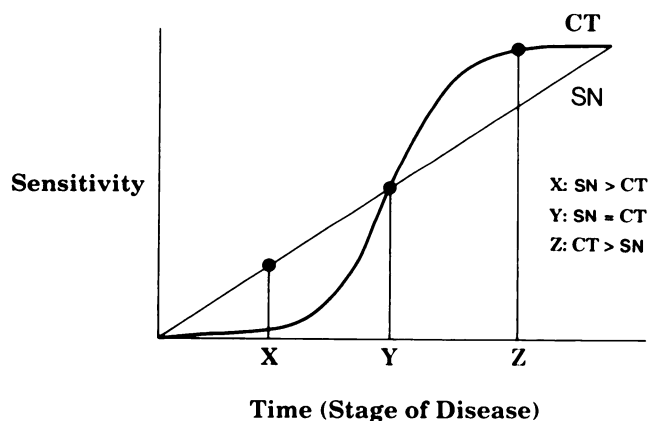
Fig. 7.—By waiting, one will observe both improved sensitivity and specificity. This results from dynamic nature of disease process, which results in progressively less overlap between populations (A) and thus better discriminating ability (B). This improvement in test performance is independent of "intrinsic goodness" of test.



A



B



C

Fig. 8.—Hypothetical sensitivity of CT (A) and transrectal sonography (B) in detection of prostatic carcinoma as a function of stage of disease. Note how relative “superiority” of one technique compared with the other is highly dependent on temporal stage (X,Y,Z) of disease (C).

vades the seminal vesicles, however, sensitivity (as well as specificity) increases rapidly. The corresponding sensitivity curve for transrectal sonography should be somewhat different (Fig. 8B). We expect this curve to increase monotonically in a more “linear” fashion as the disease progresses. This reflects the observation that sonography is better able to differentiate intraparenchymal lesions because of subtle changes in echogenicity well before capsular invasion or other morphologic distortion has occurred.

Overlapping these two curves illustrates the potential misunderstanding (Fig. 8C): *depending on the stage of the disease, we can come to three totally different conclusions regarding the performance of sonography relative to that of CT.* Failure to control carefully for disease stage will almost always guarantee misleading conclusions. Therefore, we must be wary of retrospective multiinstitution comparative studies in which one author compares results of one imaging technique obtained in one clinical setting with results evaluating another imaging test obtained at another institution. Even if disease stage is carefully controlled, such multiinstitutional studies can be misleading because controlling *precisely* for disease stage is extremely difficult. With regard to clinical stage, a wide spectrum occurs within each defined stage. This definition is frequently physician-dependent and open to subjective assessment. One clinician’s “stage B1” lesion may be “stage B2” according to another, even if both are using the same “objective” criteria. The argument can be made that the only rigorous method to compare different imaging techniques critically is to perform all tests on all patients at the same place and time.

Design Bias

Other, somewhat more subtle, experimental design biases can result in overly optimistic sensitivity and specificity claims. These include biases in selection, workup, and test interpretation (including diagnosis and test review bias). Discussion of these pitfalls is beyond the scope of this article; readers are encouraged to read the concise review by Begg and McNeil [8].

ACKNOWLEDGMENT

I am grateful for the helpful advice, assistance, and enthusiastic encouragement received from Gerald Friedland.

REFERENCES

1. Kent DL, Larson EB. Magnetic resonance of the brain and spine: is clinical efficacy established after the first decade? *Ann Intern Med* 1988;108:402-418
2. Kent DL, Larson EB. Magnetic resonance imaging of the brain and spine: health and public policy committee. *Ann Intern Med* 1988;108:474-476
3. Kent DL, Larson EB. Editorial. Diagnostic technology assessments: problems and prospects. *Ann Intern Med* 1988;108:759-761
4. Miller-Catchpole R. Diagnostic and therapeutic technology assessment. *JAMA* 1988;259:2757-2759
5. Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. *JAMA* 1988;259:3277-3280
6. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283-298
7. Doubilet P. A mathematical approach to interpretation and selection of diagnostic tests. *Med Decis Making* 1983;3:177-195
8. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* 1988;167:565-569