PSYCHOLOGICAL SCIENCE CAN IMPROVE DIAGNOSTIC DECISIONS

John A. Swets,[1] Robyn M. Dawes,[2] and John Monahan[3]

[1]BBN Technologies (emeritus), Cambridge, Massachusetts; Radiology Department, Brigham and Women's Hospital, and Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, [2]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania, and [3]School of Law, University of Virginia, Charlottesville, Virginia

INTRODUCTION AND SCOPE

Diagnostic problems abound for individuals, organizations, and society. The stakes are high, often life and death. Such problems are prominent in the fields of health care, public safety, business, environment, justice, education, manufacturing, information processing, the military, and government.

Particular diagnostic questions are raised repetitively, each time calling for a positive or negative decision about the presence of a given condition or the occurrence (often in the future) of a given event. Consider the following illustrations: Is a cancer present? Will this individual commit violence? Are there explosives in this luggage? Is this aircraft fit to fly? Will the stock market advance today? Is this assembly-line item flawed? Will an impending storm strike? Is there oil in the ground here? Is there an unsafe radiation level in my house? Is this person lying? Is this person using drugs? Will this applicant succeed? Will this book have the information I need? Is that plane intending to attack this ship? Is this applicant legally disabled? Does this tax return justify an audit? Each time such a question is raised, the available evidence is assessed by a person or a device or a combination of the two, and a choice is then made between the two alternatives, yes or no. The evidence may be a x-ray, a score on a psychiatric test, a chemical analysis, and so on.

In considering just yes–no alternatives, such diagnoses do not exhaust the types of diagnostic questions that exist. Other questions, for example, a differential diagnosis in medicine, may require considering a half dozen or more possible alternatives. Decisions of the yes–no type, however, are prevalent and important, as the foregoing examples suggest, and they are the focus of our analysis. We suggest that diagnoses of this type rest on a general process with common characteristics across fields, and that the process warrants scientific analysis as a discipline in its own right (Swets, 1988, 1992).

The main purpose of this article is to describe two ways, one obvious and one less obvious, in which diagnostic performance can be improved. The more obvious way to improve diagnosis is to improve its accuracy, that is, its ability to distinguish between the two diagnostic alternatives and to select the correct one. The less obvious way to improve diagnosis is to increase the utility of the diagnostic decisions that are made. That is, apart from improving accuracy, there is a need to produce decisions that are in tune both with the situational probabilities of the alternative diagnostic conditions and with the benefits and costs, respectively, of correct and incorrect decisions.

Methods exist to achieve both goals. These methods depend on a measurement technique that separately and independently quantifies the two aspects of diagnostic performance, namely, its accuracy and the balance it provides among the various possible types of decision outcomes. We propose that together the method for measuring diagnostic performance and the methods for improving it constitute the fundamentals of a science of diagnosis. We develop the idea that this incipient discipline has been demonstrated to improve diagnosis in several fields, but is nonetheless virtually unknown and unused in others. We consider some possible reasons for the disparity between the general usefulness of the methods and their lack of general use, and we advance some ideas for reducing this disparity.

To anticipate, we develop two successful examples of these methods in some detail: the prognosis of violent behavior and the diagnosis of breast and prostate cancer. We treat briefly other successful examples, such as weather forecasting and admission to a selective school. We also develop in detail two examples of fields that would markedly benefit from application of the methods, namely the detection of cracks in airplane wings and the detection of the virus of AIDS. Briefly treated are diagnoses of dangerous conditions for in-flight aircraft and of behavioral impairments that qualify as disabilities in individuals.

Enhancing the Accuracy of Decisions

As implied, there are four possible decision outcomes in the two-alternative diagnostic task under consideration: two correct and two incorrect. In one kind of correct outcome, the condition of interest is present, or "positive," and the decision is correspondingly positive. Such an outcome is termed a "true-positive" outcome. For example, cancer is present and the radiologist says it is. In the other kind of correct outcome, the condition is absent, or "negative," and the decision is properly negative. It is called a "true-negative" outcome. For example,

Address correspondence to John A. Swets, 8921 S.E. Riverfront Terrace, Tequesta FL 33469; email: swets@bbn.com.

cancer is not present and the radiologist says it is not. Similarly, of the two incorrect outcomes, one is "false-positive" (condition absent, decision positive) and the other is "false-negative" (condition present, decision negative). Accuracy may be increased by increasing the relative frequency of one or the other of the two types of correct decisions, or equivalently, by decreasing the relative frequency of one or the other of the two types of errors.

Let us be explicit about why both correct and erroneous decisions occur. The reason is that the evidence available for a decision is usually ambiguous. According to a well-established model of the process, we may think of the degree of evidence as being represented by a value along a single dimension, with high values tending to be associated with the positive diagnostic alternative and low values tending to be associated with the negative alternative. For example, a high pressure in the eye is usually associated with glaucoma, and a low one not. But the tendencies are merely that; low values of evidence can nonetheless arise from the positive alternative and high values from the negative alternative. The distribution of the degrees of evidence produced by the positive condition overlaps the distribution of the degrees produced by the negative condition. Hence, the accuracy of a series of diagnoses depends on the amount of overlap between the two distributions—that is, the inherent confusability of the two alternatives. In sum, diagnoses are not certain; errors will occur.

This conception of the diagnostic process is shown pictorially in Figure 1. Concretely in the figure, the problem is to distinguish eyes with glaucoma from normal eyes and the evidence is the amount of pressure measured in the eye. The figure shows the pressure values to vary along the evidence continuum from 0 to 50. The two distribution curves show the probability (the height of the curve) that each value will occur in connection with each of the diagnostic alternatives. The figure reflects (only for our illustrative purposes) a distribution of pressure values at the left, observed for normal eyes, ranging from 0 to 40. Meanwhile, pressure values associated with glaucoma vary from 10 to 50. Hence, the two distributions of values overlap between 10 and 40. Those values are inherently problematic; they can arise from either a diseased or normal eye.

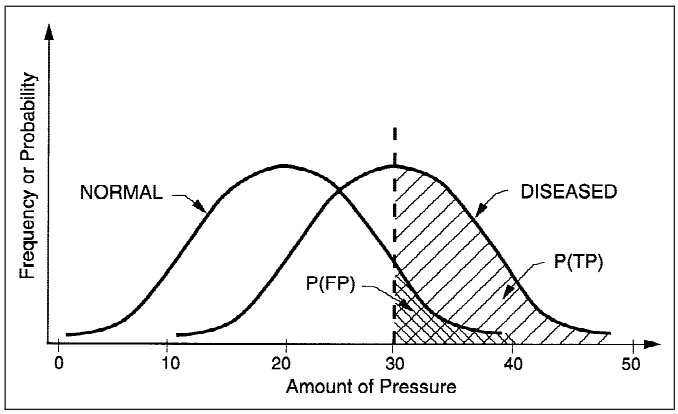

Fig. 1. Probability distributions of amounts of evidence (here, units of pressure as measured in the eye) for the negative (normal) and positive (diseased) diagnostic alternatives. Each value on the evidence continuum occurs for the negative and positive diagnostic alternatives with a probability equal to the height of the curve for that alternative. An illustrative decision threshold is shown at 30 on the evidence scale, meaning that values of 30 or greater will elicit a positive decision. That threshold yields the false-positive and true-positive probabilities, P(FP) and P(TP), as indicated by hatch marks. The two probabilities are equal, respectively, to the proportions of area under the curves that lie to the right of the decision threshold. They vary together when the threshold is varied.

We describe a class of computer-based decision-support methods that increase accuracy by providing a better quality of evidence—distributions with less overlap—by developing for the diagnostician statistical combinations and implications of relevant diagnostic data. Called *actuarial techniques,* or *statistical prediction rules* (SPRs), they use statistical analyses of cases with known outcomes to determine which pieces of diagnostic information, or which "predictor variables," are relevant to a given diagnostic decision and to what extent (most diagnoses depend on more than one variable as in our glaucoma example). As applied to each case in subsequent diagnoses, a SPR accepts case-based values of the variables and combines them, with appropriate weight given to each, to give the best possible assessment or summary of the available evidence. Many SPRs are programmed to evaluate and express the evidence as an estimate of the probability that the diagnostic condition of interest is present.

Enhancing the Utility of Decisions

Even though the accuracy of a given type of diagnosis in a particular setting may be constant, depending on the quality of evidence available and the ability of the diagnostician, the utility of the decisions can vary. In some situations, positive decisions should be more frequent than negative decisions, perhaps (1) because the probability of the positive diagnostic alternative is high (for example, most patients at a certain clinic will have glaucoma), or perhaps (2) because the value of being correct when the positive alternative is present is very high (treating glaucoma immediately may be very efficacious). Alternatively, some situations will favor negative decisions (glaucoma screening in a broad population without symptoms may not turn up many instances of the disease). We assume, however, that the diagnostician will not achieve more decisions of one or the other kind when more are indicated simply by making more of them on randomly chosen trials, irrespective of the degree of evidence on a trial. Rather, the diagnostician will strive for consistency, making the same decision each time for any given evidence value, and hence vary deliberately the amount of positive evidence that will be required to issue a positive decision.

Given the picture in Figure 1, the decision maker can make decisions most consistently by setting a cutpoint on the con-

tinuum of evidence values, such that values above the cutpoint lead always to a positive decision and values below it lead always to a negative decision. The cutpoint is called a *decision threshold.* A decision threshold is illustrated in Figure 1 by the dashed vertical line, at 30 pressure units. That cutpoint can be adjusted up or down to produce more or fewer positive decisions in a rational way, e.g., to make additional positive decisions in connection with higher amounts of positive evidence. In the preferred diagnostic procedure, adjustments of the decision threshold are made to produce the best ratio of positive to negative decisions and ultimately to produce the best balance among the four possible decision outcomes for the situation at hand, and hence to maximize the utility of the set of decisions made over time.

We must make a fine distinction, but a critical one. We have spoken of events called "decision outcomes," which are *joint* occurrences of a particular diagnostic alternative and a particular decision—for example, a positive alternative and a positive decision occur together. In addition, we need the concept of a *conditional* event, which (for our purposes) is a particular decision made when, or given that, a particular diagnostic alternative occurs (past or future). Both joint and conditional events will occur in four ways, depending on the combination of positive and negative alternatives and decisions; each way will have a probability associated with it. There are two central probabilities for us, as will be seen shortly: the conditional probability that a decision is positive given that the positive diagnostic alternative occurs, which we call simply the "true-positive probability," and denote P(TP); and the conditional probability that a decision is positive given that the negative alternative occurs, which we call the "false-positive probability," and denote P(FP).

It is now widely recognized, for a diagnostic process of constant accuracy, that adjusting the decision threshold will exhibit a fundamental correlation between P(FP) and P(TP). If the threshold is made more "lenient" (requiring less evidence for a positive decision to be made) to increase P(TP), then P(FP) will also inevitably increase. More "yes" decisions will be correct/true with the more lenient threshold, but more also will be incorrect/false. Alternatively, if the threshold is made more "strict" to decrease P(FP), then P(TP) will necessarily go down.

In Figure 1, P(TP) is equal to the proportion of the area under the positive (right) distribution to the right of the decision threshold (as hatched) and P(FP) is equal to the proportion of the area under the negative (left) distribution to the right of the threshold (cross hatched). It is clear that those proportionate areas will increase or decrease together as the threshold point moves. The conditional probabilities of the other two decision outcomes, true-negative and false-negative, are the complements of P(TP) and P(FP), respectively, and equal the proportionate areas under the distributions to the left of the threshold. Hence, they will also vary when the threshold is moved. However, because they are complements of the main two probabilities, they offer only redundant information and we shall generally not attend to them.

We describe later a formula that shows where the decision threshold should be set to maximize the utility of a decision process, to maximize a decision's benefits, on average, relative to its costs. The formula takes into account the relative probabilities that the positive and negative diagnostic alternatives will occur in the situation at hand, independent of the decision process. In general, the higher the probability of the positive alternative, the more lenient the best setting of the decision threshold (and alternatively). The formula also takes into account the benefits of being correct (in both ways) and the costs of being incorrect (in both ways). In general, the more important it is to be correct when the positive alternative occurs, the more lenient the decision threshold should be (and alternatively).

Although several experimental uses and some routine uses of statistical prediction rules exist to demonstrate increases in diagnostic accuracy, there have been relatively few attempts to evaluate methods for choosing the best decision threshold. However, analyses of certain diagnostic tasks described in this article make clear the large potential gain to be provided by threshold-setting methods.

Scope of Our Discussion

We begin by further identifying the two particular diagnostic tasks that will be presented in detail to illustrate improvements in accuracy and also the two tasks chosen to suggest improvements in decision utility that can stem from appropriate setting of the decision threshold. We proceed to describe some general characteristics of diagnostic tasks and how both objective and subjective data are used in compiling the evidence for a diagnostic decision. There follows a review of the measures of accuracy and decision threshold that are used throughout to evaluate those aspects of performance. These measures are based on the now-common "ROC" technique, the term abbreviated from "receiver operating characteristic" as used in signal detection theory, a theory developed for electronic signals (Peterson, Birdsall, & Fox, 1954) and then widely applied in psychology and in diagnostics generally (Swets, 1996).

We proceed to describe the functioning of SPRs: first, how statistical methods determine which pieces of information or which predictor variables are relevant for any given task; and second, how SPRs combine case-specific values of the variables to estimate a diagnostic probability for any given case. We then discuss optimal ways to set a decision threshold. Diagnostic illustrations follow. Our concluding sections, as suggested earlier, present possible reasons for the limited use of these decision-support methods and possible ways to extend their use.

Our intent is to promote a national awareness among the public and its policy makers of the potential for these decision-support methods in many important areas. Our orientation is

toward affecting policy and practice. We do not treat inherently statistical or psychological topics, so we do not make a comparative analysis of different types of SPRs nor an analysis of various techniques people use in making inferences and decisions.

Four Illustrative Diagnostic Tasks

Increasing accuracy

Two diagnostic tasks will illustrate in some detail the capability of SPRs to increase diagnostic accuracy: (1) A psychiatrist or clinical psychologist seeks to determine whether a particular patient in a mental health facility will, if discharged, engage in violent behavior; (2) A radiologist must determine whether or not a woman being examined with mammography has breast cancer. A parallel task is considered in which magnetic resonance (MR) imaging is used to determine the extent to which prostate cancer has advanced. Although these examples may illustrate improvements in setting decision thresholds as well as improvements in diagnostic accuracy, they are focused here on experimental demonstrations of increased accuracy.

Increasing utility

Two other diagnostic tasks will provide analyses of the benefits of optimizing the placement of the decision threshold: (1) A blood test is used to screen individuals for the presence of the human immunodeficiency virus (HIV) of the acquired immunodeficiency syndrome (AIDS); (2) An imaging test, e.g., an ultrasound display, is used by an electronics technician to detect cracks in airplane wings.

COMPONENTS OF DIAGNOSTIC DECISION MAKING

Characteristics of Diagnostic Tasks

Several or many pieces of relevant information

A fundamental characteristic of most diagnostic tasks is that several variables enter the evidence for a decision. Some variables are "objective," i.e., demonstrable facts; others are "subjective," i.e., include at least an element of a human diagnostician's judgment. In the prognosis of violence, for example, objective items with predictive power include (a) prior arrests, (b) employment status, and (c) age. Subjective items relevant to violence diagnosis include clinical judgments about (d) psychopathy, (e) schizophrenia, and (f) substance abuse.

Similarly for breast cancer, roughly 20 variables may weigh in a radiologist's diagnosis. Again, some are subjective, such as the interpretation from visual impressions of the mammogram of features that may seem to be present. Examples of such features include (a) a "mass" (tumor or a cyst); (b) "calcifications" (sand−like grains); or (c) a change from a previous mammogram. Objective variables include demographic, clinical, and biological data, such as: (d) patient's age, (e) having a previous biopsy or not, and (f) presence of a malfunctioning cancer-suppressing gene. Clearly, all relevant pieces of information must be combined in a constructive way to make their appropriate contributions to the sum of evidence.

Merging information into a diagnostic probability

The sum of evidence often, and with the SPRs of interest to us, boils down to an estimate of a probability: the probability that a given condition exists (e.g., breast cancer) or that a given event will occur (e.g., a violent act). Statistical procedures may be used to ensure that each piece of evidence contributes to the over-all probability estimate in proportion to its diagnostic weight or predictive power. These procedures also ensure that an item of data contributes to the degree that its information is independent of that in other items, because including the same information twice would artificially double its impact and so distort the final estimate of probability.

Setting a decision threshold on the probability continuum

A probability estimate, of course, is merely that. It is a continuous variable that must be converted into a decision about the case at hand, usually a choice between two diagnostic alternatives, such as cancer present or cancer absent. A threshold probability of .05 that cancer is present may be deemed appropriate in diagnosing breast cancer. A higher probability of cancer will then lead to some action or actions, such as additional imaging examinations (perhaps enlarged X-rays or ultrasound), earlier than usual re-examination by routine mammography, or biopsy. A more lenient threshold, such as a probability of cancer of .02, will find more cancers but at the expense of telling more women who do not have cancer that they might have cancer, which will cause duress and likely require further costly examinations. This vignette illustrates our point that adjusting the decision threshold will affect both P(FP) and P(TP) while a constant accuracy is maintained. By way of example, those two probabilities may be .20 and .80, respectively, for the stricter threshold at $p = .05$; and .50 and .95 for the more lenient threshold at $p = .02$. The more lenient threshold therefore detects correctly an additional 15 of 100 present cancers (from 80 to 95) at the price of "detecting" incorrectly an additional 30 of 100 cancers not there (from 20 to 50). The question of which threshold is better calls for an analysis of costs and benefits of the sort we discuss later.

As a refinement of the foregoing point, note that the diagnostic question may be framed as a choice between a specified action and no action, rather than as the choice between the presence and absence of a condition. Also, there may be more than one action available, perhaps a set of graded actions correlated with the size of the probability estimate. In such a setting, one might simultaneously set two or three decision thresholds on the probability variable.

Merging Objective Data and Subjective Judgments

Objective data

For several decades in certain diagnostic fields, relevant items of objective data (numerical or categorical) have been combined by statistical methods to provide an estimate of a diagnostic probability. In some cases, the method of combination is relatively simple; such a method might consist simply of counting how many of a set of listed symptoms of a disease are present, for example. In other cases, multivariate statistical analysis may merge predictor variables in a more sophisticated way, taking quantitatively into account their predictive power and the degree to which they provide predictive power independent of the other variables considered.

Subjective judgments

For several decades, the validity of a SPR's probability estimates, or the accuracy of its diagnoses, as based solely on objective data has been compared and contrasted to the estimates or accuracy of decisions based largely on subjective judgment. In some settings, principally in psychiatry and clinical psychology, actuarial instruments have been shown repeatedly to be more accurate than clinical judgment (Meehl, 1954; Dawes and Corrigan, 1974; Dawes et al., 1989), leading some investigators to recommend that the clinician's judgment/diagnosis be totally supplanted by the statistical device (e.g., Grove and Meehl, 1996; Quinsey et al., 1998). In other settings, such as clinical medicine or weather forecasting, there has been less of a tendency to consider objective methods and human judgments as competing alternatives. In such settings, the prevailing practice is to supply the objective method's output to the human diagnostician who then makes the final decision.

Subjective data

In a parallel to the procedure of calculating a probability based on objective data, one can merge judgments about subjective variables by using the same statistical apparatus used with objective data. This procedure might begin with the human suggesting as relevant certain items of evidence, e.g., perceptual features of a medical image, that may turn out by analysis of outcomes in proven cases to be of diagnostic importance. Diagnosticians and methodologists can then work together to devise apt names and rating scales for the items retained. Finally, the human diagnostician supplies the ratings for each item for a given case under diagnosis that are then merged statistically into a probability estimate (e.g., Getty et al., 1988; Seltzer et al., 1997). A similar approach has been taken to decisions about prison parole (Gottfredson et al., 1978).

Combining objective and subjective data statistically

Still a third possibility is to combine both objective and subjective evidence in a single SPR and either use directly the probability estimate it provides or supply the estimate to the human diagnostician for final judgment (Getty et al., 1997). In the latter event, the SPR's result can be considered by the diagnostician as a kind of "second opinion." The exact nature of what the human can or should add at that point may not be totally clear. However, the human might opt for a higher or lower probability than the rule estimates, depending on what he or she knows about the composition of the rule and the particulars of the case at hand. We will return to this question, with specifics developed in our illustration of violence prognosis.

The diagnosis of breast cancer by mammography seems to be consistent with constructing an SPR on the basis of all of the evidence while leaving the human in control. The human, in this instance, is often essential to supply perceptual data to the SPR that is beyond the capability of machines to generate objectively; humans continue to see certain critical features of medical images that computers fail to see. But humans must concede to computers superior ability to see and assess certain other features in an image and further, superior persistence to call exhaustively for examination of every relevant feature for every case, without succumbing to "satisfaction of search" after a few salient features are noticed. Computers, of course, retain more precisely the numerical values assigned to the several items of information that are considered for each case, including those assigned by the human. And the computer's statistical algorithm exceeds by far the human's capability to calculate weights and to merge the calculated values optimally. In contrast to the position sometimes taken that the SPR should supplant the clinician, however, such as in the prediction of violence, we sense little sentiment to replace the radiologist by a SPR or to suppress any further, final opinions he or she may have after the SPR has had its say.

Balancing objective and subjective contributions

The view governing this article is that the appropriate role of the SPR vis a vis the diagnostician will vary from one context to another, and will be disputed in some. Nonetheless, the ability of SPR techniques to increase diagnostic accuracy should be ascertained in diagnostic settings for which they seem to have promise and their enhanced accuracy should be used wherever it can be supplied in a cost-effective manner. The roles and interactions of human and computer that are most appropriate can be determined for each diagnostic setting in accordance with the accumulated evidence about what works best.

STATISTICAL MACHINERY

Diagnosis, as we have seen, is intrinsically probabilistic or statistical. We describe next the probability theory and statistical tools that are used to measure and to enhance diagnostic performance. Though not essential to an appreciation of this

article's theme, an appendix presents basic probability concepts in somewhat greater depth for the interested reader.

Measures of Accuracy and the Decision Threshold

The two independent aspects of diagnostic performance, accuracy and decision threshold, should of course be reflected in separate, independent measures. Two such measures are provided by a ROC graph where "ROC," as mentioned, stands for "receiver operating characteristic." The ROC's origin in electronic signal detection theory, its wide use in psychology for sensory and cognitive processes, and its wide and growing use in diagnostic fields are described elsewhere; described in the same publication are the inadequacies of measures not derived from a ROC (Swets, 1996). Suffice it to say here that the half dozen or so measures developed in various diagnostic fields are intended to be measures of accuracy; however they vary in a predictable, but unappreciated, way with changes in the decision threshold. Because the decision threshold is not assessed in connection with these accuracy measures, they are confounded by changes in the threshold and are unreliable to that extent. Their lack of a companion measure of decision threshold ignores an important aspect of performance.

The ROC graph

The ROC graph is a plot of the two basic probabilities we have emphasized in the previous discussion—the probabilities that the decision is positive when the condition of interest is present, or positive, and that the decision is positive when the condition is absent, or negative—denoted P(TP) and P(FP). They are calculated from proportions of observed frequencies as displayed in a two-by-two table of data, as described in the Appendix.

The ROC graph, specifically, is a plot of P(TP) on the y-axis and P(FP) on the x-axis, and shows how the two quantities vary together as the decision threshold is varied for a given accuracy. An example of a ROC is shown in Figure 2. The two probabilities vary together from the lower left corner of the graph in the form of a curved arc to the upper right corner. At the far lower left both probabilities are near 0, as they would be for a very strict decision threshold, under which the diagnostician rarely makes a positive decision. At the far upper right both probabilities are near 1.0, as they would be for a very lenient decision threshold, under which the diagnostician almost always makes a positive decision. In between the curve rises smoothly, with a smoothly decreasing slope, to represent all of the possible decision thresholds (for a given accuracy). Hence, the position of the curve (in the square) is independent of whatever decision threshold is chosen in a particular task. It should be noted that the curve shown in the figure is idealized. Actual, empirical ROCs will vary somewhat in form, though usually not by much (Swets, 1996, chapter 2; Hanley, 1988).

Note, fundamentally, that if an empirical ROC is not available, one would not know whether two different observed pairs of P(FP) and P(TP) represent the same or different accuracies. For example, earlier we exemplified in the context of mammography one pair of these values as .20 and .80 and another pair as .50 and .95. Do those two pairs represent different accuracies (as well as different decision thresholds) or only different decision thresholds with the same accuracy? The ROC is needed to show if they lie on the same ROC, for a given accuracy, or on different curves, representing higher and lower accuracy as described shortly.

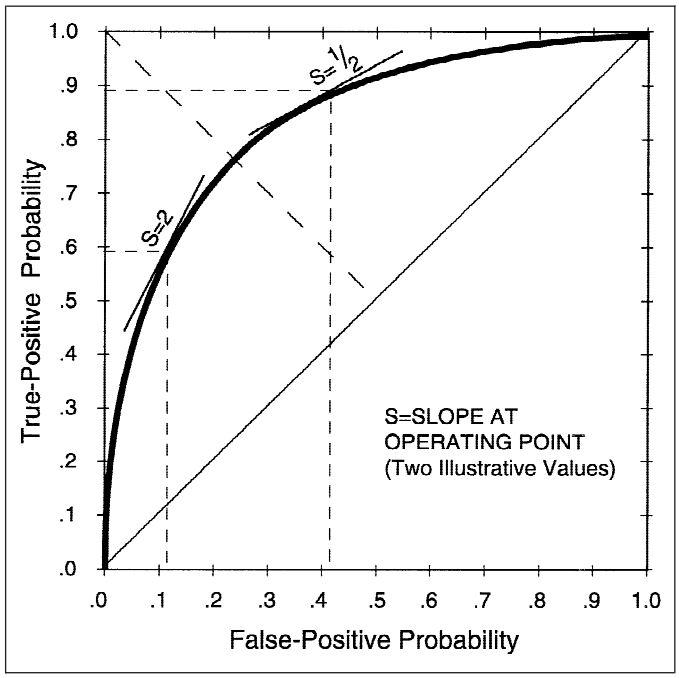

Fig. 2. Illustrative ROC (receiver operating characteristic), for a particular accuracy. True-positive probability, or P(TP), is plotted against false-positive probability, or P(FP). The curve extends from the lower left corner in an arc of decreasing slope to the upper right corner, as the decision threshold is varied from strict to lenient. Two selected points on the curve, where the curve has slopes of 2 and 1/2, respectively, are identified to indicate how the slope of the curve at any point, symbolized as *S*, may be used as a measure of the decision threshold that produced the point.

Measure of the decision threshold, S

Because the ROC's curve rises with smoothly decreasing slope, the slope of the curve at any point along the curve will serve as a measure of the decision threshold that produces that point. This measure is denoted *S*. The slope approaches infinity at the strictest threshold, or lower left corner, and 0 at the most lenient threshold, or upper right corner. Practically observed decision thresholds, however, are well within the graph and vary from about 5 for a relatively strict threshold to 1/5 for a relatively lenient threshold. Illustrative thresholds at values of *S* equal to 2 and 1/2 are shown in Figure 2. It can be seen that the threshold at $S = 2$ is relatively strict; the positive decision is made rarely and both ROC probabilities are relatively small (see dashed

lines). A threshold at $S = 1/2$, on the other hand, is lenient and produces fairly large probabilities of a positive decision.

Note that any value of *S* can be referred back to a cutpoint on the evidence or decision variable, however that variable is expressed, e.g., as an observer's rating, a probability, or some physical quantity. Indeed, the value of *S* is identical to a value of a point along the evidence continuum when the continuum is quantified as the so-called "likelihood ratio." This is the ratio of the probability that a given degree of evidence will arise when the diagnostic alternative is positive to the probability that the degree of evidence will arise when the diagnostic alternative is negative. It is thus a ratio of the heights of the two overlapping distributions of degrees of evidence as shown in Figure 1. (Notice in the figure that the positive distribution is roughly twice the height of the negative distribution at 30 on the pressure variable, and so $S = 2$ corresponds to a threshold set at 30. The height of the positive distribution is roughly one-half that of the negative distribution at 20 on the pressure variable, and so $S = 1/2$ corresponds to a threshold at 20.) Other measures of the decision threshold are sometimes preferred to *S*, but it, as will be seen, lends itself to calculation of the optimal threshold.

Measure of accuracy, A

Figure 3 shows several illustrative ROCs, which represent different levels of accuracy. The higher the curve, the greater the accuracy. That is, the accuracy is greater when P(TP) is higher for a given P(FP). Referring to Figure 1, the curve will be higher and the accuracy greater when the overlap between the two probability distributions is less, when the diagnostic alternatives are less confusable. Hence, the accuracy of diagnosis is conveniently represented by the proportion of the graph's area that lies beneath a given curve. This area measure, denoted *A*, ranges from .50 for accuracy equal to chance, up to 1.0 for perfect accuracy. Specifically, $A = .50$ for a ROC lying along the diagonal that runs from lower left to upper right, which is a ROC signifying accuracy no better than chance performance; that is, P(TP) is no higher anywhere than P(FP). $A = 1.0$ for a ROC that follows the left and upper axes, which is a curve that signifies perfect accuracy; that is, P(TP) is 1.0 for all values of P(FP), including 0. Some intermediate values of *A* are shown in Figure 3. (The reader can determine visually whether the two pairs of ROC probabilities mentioned earlier, (.20, .80) and (.50, .95), lie on the same curve or different curves.)

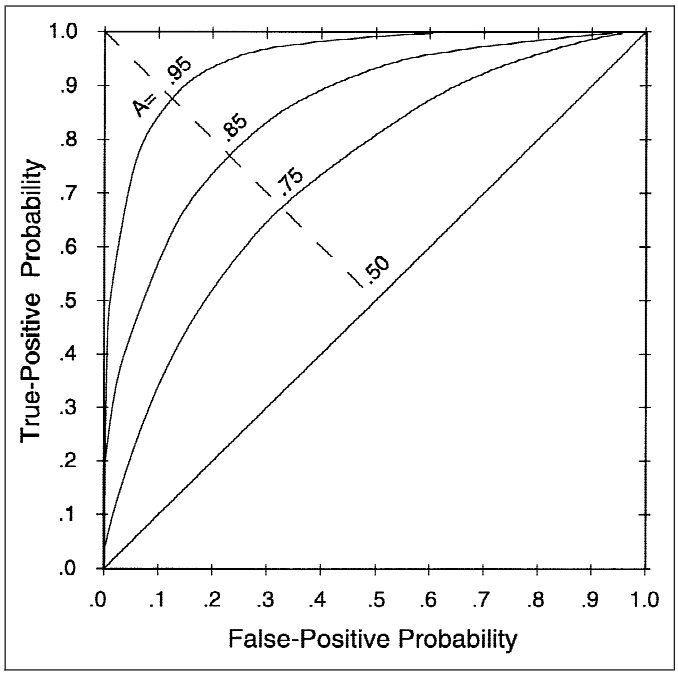

Fig. 3. Illustrative ROCs (receiver operating characteristics), for four levels of accuracy. Each curve is labeled by its area measure of accuracy, *A*. The measure *A* is defined as the proportion of the graph's area that lies below a given curve. Values of *A* range from .50, at the (solid) diagonal line that corresponds to chance accuracy, up to 1.0, for perfect accuracy.

Another way to think of *A* may subjectively help to calibrate its various values. Consider a series of trials in which a randomly selected pair of diagnostic alternatives is presented on each trial—one alternative always positive and the other always negative (e.g., a mammogram from a patient with proven cancer and another from a patient without cancer). The decision maker is asked to make a "paired-comparison" and is instructed to identify the (more likely) positive alternative on each trial. Then the quantity *A* is equal to the proportion of times the decision maker (correctly) selects the positive alternative. Hence, $A = .75$ means that in a paired-comparison task, the radiologist can correctly identify the mammogram that is associated with disease on 75% of the trials.

Other ROC measures of accuracy are sometimes used, such as the distance of the curve from the positive diagonal. One virtue of the area measure is that it is relatively insensitive to small variations in the shape of empirical ROCs. That *A* is the measure of modern choice is indicated by the ROC's listing as a keyword in over 700 articles per year in the medical literature—almost all of these articles' using *A*. Computer programs are available to fit ROCs to data and give values of the measures *A* and *S*, along with their confidence limits. A variety of useful ROC programs can be accessed via the website: http://www.radiology.arizona.edu/~mo-/rocprog.atm.

Constructing an empirical ROC

The simplest and most efficient way to construct a ROC to represent the accuracy of a given diagnostic system is to work directly with the graded or continuous output of the system. Thus, a radiologist may give a rating of confidence that a lesion/disease is present, say, on a 5-category scale ranging from "highly confident positive" to "highly confident negative." Or the radiologist, or SPR, may give an estimate of the probability of a positive condition (effectively a 100-category scale). Most diagnostic systems give such a continuous out-

put—a pressure test for glaucoma, for example, or a probability of precipitation, or a likelihood of violence. Using defined categories of judgments, or breaking a continuum into arbitrary categories, approximates the simultaneous setting of a range of thresholds. One can picture several thresholds set simultaneously and spread along the evidence variable in Figure 1.

In analysis, the investigator can adopt the multiple decision thresholds afforded by the categories used by the diagnostician, or taken from the categories defined just for purposes of analysis. So, if the variable is probability, the investigator can set thresholds for analysis, say, at $p = .90$, $p = .80$, $p = .70$, and so forth, and compute the ROC coordinates P(FP) and P(TP) for outputs that exceed each successive threshold: the strict one at .90, the less strict one at .80, and so on. In this way, a series of ROC points march up along the curve and define a ROC for the diagnostic system or task at hand.

How Statistical Prediction Rules (SPRs) are Developed

A SPR is constructed by means of statistical analysis to quantify the power of candidate predictive variables to discriminate between the positive and negative instances of the diagnostic alternatives under study. Though not true of all methods, variables may be added to a SPR and assigned their respective weights in a stepwise fashion; that is, a particular candidate variable is selected next for the mix if it adds the largest increment to the power of the variables already selected, and then it is weighted according to the size of that increment. In this way inter-correlations among variables are neutralized; each subsequent variable is assessed according to its independent, additional predictive power. The stepwise procedure is repeated until the investigator decides that additional variables are adding little of practical value and then the existing set can be regarded as necessary and sufficient.

The SPR, as a set of variables and weights, is able to function as an aid to diagnosticians because for each new case submitted to it, essentially as a collection of values for the SPR's variables, it gives an estimate of the probability of occurrence of the positive instance of the two diagnostic alternatives. (This is an "inverse" or "Bayesian" probability as defined in the Appendix.) The SPR can function directly as a decision maker if supplied a decision threshold. (Note that some computer programs for developing an SPR supply only a categorical decision, rather than a probability, that is based on some often unexplained decision threshold, perhaps one that maximizes the number of correct decisions. Such programs show an insensitivity to the need to set different thresholds appropriate to different settings.)

Alternative methods

Several statistical methods have been used to develop SPRs, including "discriminant analysis" (e.g., Lachenbruch, 1975), "logistic regression" (e.g., Hosmer & Lemeshow, 1989), and "artificial neural nets" (e.g., Hertz et al., 1991). They can be thought of as choosing predictor variables and weights to maximize the discrimination between diagnostic alternatives—as measured, for example, by an index similar to the ROC area index *A* as defined in preceding paragraphs. In general, there is little to choose among the several methods as to the goodness with which they select variables and weights, and hence their accuracy as a decision maker or aid, so the choice among them often devolves to their relative effectiveness in different problem settings and the convenience with which they are handled (Gish, 1990; Richard & Lippman, 1991).

Validating statistical prediction rules

The accuracy or predictive validity of SPRs can be assessed in two ways. In what is termed "cross validation," the SPR is "trained" on one set of cases and then "tested" on a separate set. The goal is to make sure that the SPR works well for other (similar) cases than just those on which it was built. If, for example, 200 qualified mammograms are available, the SPR might be trained on 100 of them and tested on the other 100.

However, the sample sizes in this example are small enough to make marginal both the reliability with which the SPR will operate and the reliability with which it can be assessed, and investigators can often only obtain smaller samples than they would like. Under "statistical" validation, modern computer techniques are used to permit all 200 cases (in this example) to enter both training and testing of the SPR, whereas approximating the results of cross validation. These techniques include a resampling method called "bootstrapping," which is conducted to estimate the standard deviation or confidence interval of a SPR's accuracy (Efron, 1982; Gong, 1986). In our example, 50 to 200 random samples of size 200 would be taken, with replacement of each item before another draw, from the set of 200 cases.

An alternative to a sampling procedure is to be systematic and exhaustive in varying the cases that enter training and testing. In the method called "leave-one-out," for example, 200 different SPRs would be constructed, each SPR based on 199 cases and leaving out a different, single case. Each SPR is then applied to give a diagnostic probability for the single case left out of its own construction. In our experience with medical images, the SPRs developed and tested on twice the number of cases (not saving half for testing) are appreciably more robust on several statistical dimensions. The attrition in the index *A* stemming from a statistical validation procedure has been about two or three percent. (One should remember, however, that application of an SPR to samples differing in substantial respects from the original sample will produce lowered accuracy.)

Determination of truth

Clearly, a valid and accurate SPR will rely on adequately valid assessments of the occurrence, or not, of the condition of interest on each diagnostic trial. The adequacy of these so-called "truth data" will also affect the validity of evaluations of

diagnostic accuracy. That is to say, ideally one should know with certainty for every case whether it is positive or negative; otherwise the score assigned to a diagnosis, right or wrong, will not always be correct. Incorrectly classifying cases in the sample will depress a SPR's accuracy and measures of accuracy in general. However, diagnostic settings vary a good deal in the validity of the truth determinations they can supply, ranging from good to poor.

Medical diagnosis gives truth determinations generally regarded as quite good. The "gold standard" is surgery or autopsy followed by analysis of tissue. Still, surgery and pathology are not perfectly matched in space or time: The image interpreter and the pathologist may look at different locations and a pathological result observed might not have been present when the image was taken. Moreover, the pathologist's code or language for describing lesions differs from the radiologist's. Further, this pathology standard is applied primarily to positive cases (the cases that tend to reach that point); negative truth is often based necessarily not on pathology but rather on years of follow-up without related symptoms.

Aptitude testing also gives reasonably good truth: One can determine reliably whether the student graduates or whether the employee stays on the job. In weather forecasting, one can measure amount of precipitation in limited areas, but not know if the weather event occurred throughout the area of the forecast. Panel judgments of the relevance of documents retrieved by a library query system may be generally adequate, but their validity may depend somewhat on how technical the language is in the field of the query. Truth in the field of polygraph lie detection is surely problematic: Judicial outcomes may categorize incorrectly and even confessions may not be true.

We mention some other issues related to truth determination to indicate further the need for care. Truth determination should not be affected by the diagnostic system under test. If, for example, results of MR imaging help determine the positive or negative statuses of medical cases when MR is under evaluation, because one wishes to use all available evidence in an effort get the best truth estimates, then the MR result for any case will be scored against itself, in effect, and its measured accuracy will be inflated. Also, procedures used to establish truth should not affect the selection of cases for training or testing SPRs; if pathology is the sole standard for selecting medical cases, then the case sample will tend to be made up of cases that reach that advanced stage (quite possibly cases that show lesions relatively clearly on diagnostic imagery), which will tend to be the easier cases. More detail on issues concerning the definition of truth data, and the improperly selective sampling of cases, is given elsewhere (Swets, 1988).

Methods for Optimizing the Decision Threshold

We discussed earlier how the slope *S* at any point along an empirical ROC can be taken as a measure of the decision threshold that produced that point. Here we observe that the best decision threshold for a given diagnostic task in a particular setting can be specified by computing the optimal value of *S*. "Optimal" means the best threshold for a given, well-defined, decision goal. Computing the optimal value of *S* is a concept that should advise diagnostic decision making in general, but is little known or used.

Alternative decision goals

Several different decision goals can be defined, all seeking to maximize some quantity or other. One simple goal is to maximize the over-all percentage of correct diagnostic decisions, without regard to the balance of true-positive and true-negative decisions (not a very useful goal for having ignored that balance). Another simple goal is to maximize P(TP) for a fixed P(FP); this goal may be used when it is possible to state that P(FP) of some value, e.g., .10, is acceptable and that a greater value is intolerable.

A general decision goal

The most general decision goal is defined in terms of two situational variables: (1) the prior probabilities of positive and negative diagnostic alternatives, and (2) the benefits of the two types of correct decision outcomes and the costs of the two types of incorrect outcomes. This decision goal attempts to maximize the "expected value" of a decision, i.e., to maximize its payoff in the currency of the benefits and costs. It can be expressed in a formula, as seen below.

To develop needed notation, we speak of the presence or not of the diagnostic condition of interest as the "truth" about the condition and designate the two truth states as T+ (condition present) and T− (condition absent). The prior probabilities of those truth states, i.e., probabilities before a decision or "base rates," are denoted P(T+) and P(T−). The positive and negative decisions are symbolized as D+ and D−.

Let us denote benefits of decision outcomes as B, and costs as C. Now, for the benefits and costs associated with the joint occurrence of a particular truth state and a particular decision, we have B(T+ & D+) and B(T− & D−) for benefits, when truth and decision agree, and C(T− & D+) and C(T+ & D−) for costs, when they differ.

The formula to specify the optimal decision threshold for this general goal of maximizing expected value was derived in the context of signal detection theory by Peterson, Birdsall, and Fox (1954). They showed (with more algebra than we want to repeat) that the value of *S* that maximizes expected value is expressed as the product of (1) the ratio of prior probabilities and (2) a ratio of benefits and costs, as follows:

$$S(optimal) = \frac{\mathrm{P(T-)}}{\mathrm{P(T+)}} \times \frac{\mathrm{B(T-\ \&\ D-) + C(T-\ \&\ D+)}}{\mathrm{B(T+\ \&\ D+) + C(T+\ \&\ D-)}}.$$

Hence, when one knows the quantities involved or is willing to estimate them, the optimal "operating point" on the ROC, and

hence the optimal decision threshold, can be determined. When lacking the ability or desire to estimate individual benefits and costs, one can settle for taking their ratio. Note that the numerator of the equation refers to negative diagnostic alternatives, T−, and the denominator to positive diagnostic alternatives, T+. So it is possible, for example, that we would twice as rather be right when a positive alternative occurs as when a negative one does, perhaps in predicting severe weather. Then the ratio of benefits and costs is 1/2. For equal probabilities in this case, *S(optimal)* = 1/2 (and the decision threshold is rather lenient). In equation form:

$$S(optimal) = \frac{.50}{.50} \times \frac{1}{2} = 1/2.$$

If all benefits and costs are considered equal, then their ratio is 1.0 and the prior probabilities alone determine the optimal threshold; for example, if P(T+) = .33 and P(T−) = .67, then *S(optimal)* = 2. Figure 2 shows where these last−mentioned values of *S* fall on a ROC.

EXAMPLES OF ENHANCED DECISION MAKING

We proceed now to two prominent examples of how accuracy in diagnosis has been increased by application of statistical decision rules: first prognosis of violence committed by individuals, and then image-based diagnosis of breast and prostate cancer. Also in this section, we analyze for two diagnostic fields how decision utility could be increased by quantitative consideration of the decision threshold: first in the detection of the virus of AIDS, and then in the detection of flaws in metal structures, especially cracks in airplane wings.

Increased Accuracy

Predicting violence

Violence risk assessment is a critical and expanding part of the practice of psychiatry and clinical psychology. "Dangerousness to others" replaced "need for treatment" as a pivotal criterion for involuntary hospitalization of people with mental disorders in the 1960s. Tort liability was first imposed on clinicians who negligently failed to predict their patients' violence in the 1970s. Statutes authorizing involuntary treatment in the community for otherwise "dangerous" patients were enacted in many states in the 1980s. Risk assessments of violence were explicitly mandated during the 1990s in the Americans with Disabilities Act, which protects the employment rights of people with disabilities unless those disabilities result in an employee becoming a "direct threat" of violence to co-workers or customers.

Despite the pervasiveness of violence risk assessment, the research literature on the validity of clinical prediction has been disconcerting for decades and remains so. The most sophisticated study of clinicians' unstructured violence risk assessments, for example, found them to be modestly more accurate than chance among male patients and no more accurate than chance among female patients (Lidz et al., 1993). It was in response to such findings of low validity that many have called for the use of statistical prediction in violence risk assessment and in recent years a number of relevant SPRs have been developed. We take two of them to illustrate this actuarial turn in the field of violence risk assessment.

Violence Risk Appraisal Guide. The most studied SPR for risk appraisal among criminal patients is the Violence Risk Appraisal Guide (VRAG) (Harris et al., 1993; Quinsey et al., 1998; Rice & Harris, 1995). A sample of over 600 men from a maximum-security hospital in Canada served as subjects. All had been charged with a serious criminal offense. Approximately 50 predictor variables were coded from institutional files. The criterion variable to be predicted was any new criminal charge for a violent offense, or return to the institution for a similar act, over a time at risk in the community that averaged approximately 7 years after discharge. A series of stepwise regression models identified 12 variables for inclusion in the final SPR, including the Hare Psychopathy Checklist-Revised, elementary school maladjustment, and a diagnosis of schizophrenia (which had a negative weight). When the scores on this SPR were dichotomized into "high" and "low," the results were that 55% of the group scoring high committed a new violent offense (115/209), compared with 19% of the group scoring low (76/409). Using a wide range of decision thresholds to calculate a ROC gave a *A* index of .76, well above chance.

Iterative Classification Tree. More recently, a SPR for assessing risk of violence among persons being discharged from acute psychiatric facilities has been developed by a group sponsored by the MacArthur Foundation (Steadman et al., 2000), in a project that included one of this article's authors (JM). A sample of over 900 men and women from three civil hospitals in the United States served as subjects. None had a criminal charge pending. Based on a review of the patients' files as well as interviews with the patients, 134 risk factors were coded. The criterion variable of violence was measured by arrest records, hospitalization records, patient self-report, or the report of a collateral informant, over a time at risk in the community of 20 weeks after hospital discharge.

A variant of a "classification-tree" approach, which the MacArthur group called an Iterative Classification Tree (ICT), was used to construct their SPR. A classification tree reflects an interactive and contingent model of violence, in that dichotomous (or trichotomous) classifications are made on individual predictive variables in a conditional sequence, with each classification determining the variable considered next. This procedure serves to tailor the scoring to the case at hand: it allows many different combinations of risk factors to classify

a person as high or low risk, unlike a "main effects" linear regression analysis, which applies the same risk factors to all persons being assessed.

Risk factors identified for the ICT for given groups of patients included a screening version of the Hare Psychopathy Checklist, serious abuse as a child, and whether the patient was suicidal (which had a negative weight). Of the patients scoring in the low-risk category on this SPR, 4% committed a violent act during the follow-up, whereas of the patients scoring in the high-risk category, 44% committed a violent act. ROC analysis gave a *A* index of .82. Figure 4 shows the paired values of P(TP) and P(FP) that may be attained with $A = .82$.

Such an accuracy level is relatively high for predicting behavior and it suggests that any information loss that might have resulted from the classification-tree approach of adopting a decision threshold (or two) for each individual predictive variable, rather than just for a final, combined variable, is not very large. This suggestion of small loss is consistent with another study that compared classification-tree and logistic-regression techniques in the emergency-room diagnosis of myocardial infarction; both methods gave $A = .94$ (Tsien et al., 1998).

Clinical vs. actuarial prediction. The question of whether or not a clinician's making adjustments in the SPR's probability estimate (or categorization) helps or hurts the accuracy of prognosis has been debated actively in the violence field. The VRAG developers once thought it might help (Webster et al., 1994) and now believe it hurts (Quinsey et al., 1998). Others believe adjustment by the clinician is desirable (Hanson, 1998). Two factors are adduced to support the clinician's option of making an adjustment. One is "questionable validity generalization," an issue that arises when using a SPR based on one population to predict for another—for example, using the VRAG, which is based on male offenders who were predominantly white Canadians, to predict for the MacArthur sample of male and female white, African-American, and Hispanic patients not referred by a court, who consented to participating in the research—or vice versa. Although some evidence indicates that risk factors found in both the VRAG and the MacArthur ICT are predictive of violence in diverse groups (see a review by Hemphill et al., 1998), attempts to generalize the validity of some other SPRs for violence have not found success (Klassen and O'Connor, 1990).

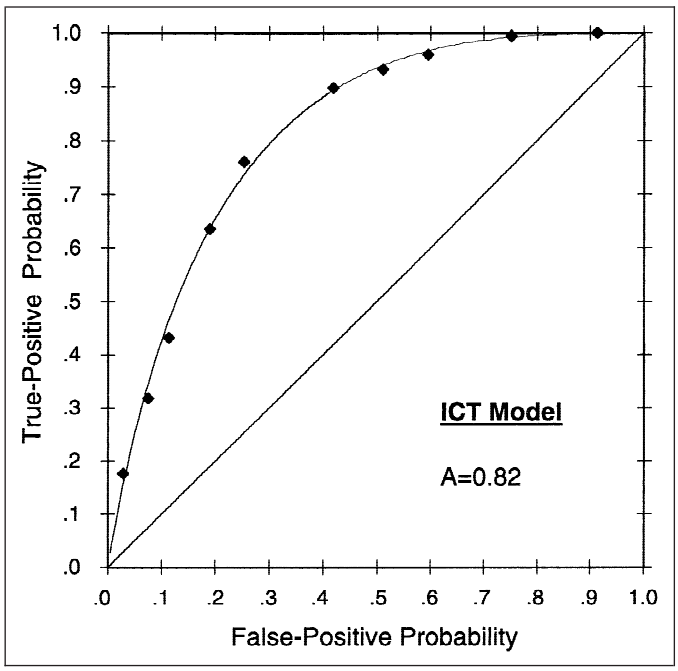

Fig. 4. Empirical ROC (receiver operating characteristic) for the SPR (statistical prediction rule) of the Iterative Classification Tree (ICT) for predicting violence. A computer program for fitting ROC data sorted the nearly continuous output of the rule into categories to yield 10 decision thresholds and their corresponding ROC data points.

The second factor used to support a clinician's option to adjust the actuarial prediction has been termed "broken leg countervailings" (Grove and Meehl, 1996, following Meehl, 1954). The story is simple: a SPR predicts with great accuracy when people will go to the movies and yields an estimate of probability .84 that Professor X will go to the movies tomorrow. But the clinician has learned that Professor X has just broken his leg and is immobilized in a cast. The story could be taken to be an analogue, for example, of the situation where a direct threat of violence by a patient to a named victim occurs, although such threats do not occur frequently enough to appear as a variable in the VRAG or the MacArthur ICT. As to the aptness of the analogy, interested parties have disagreed.

Diagnosing cancer

The potentially beneficial use of SPRs in the diagnosis of breast and prostate cancer has been shown during the past 15 years in studies by a research team in which an author of this article (JAS) participates, as well as by several other investigators. These studies focused on image-based diagnoses: on mammography for the detection and/or the malignant-benign classification of breast abnormalities and on magnetic resonance (MR) imaging for staging the extent of prostate cancer. In each instance the relevant perceptual features of the image were determined for inclusion as variables in the SPR and, in some instances, demographic and laboratory data were also utilized in the same SPR. A general discussion of approach is followed here by some specific results of the studies. Some earlier work on SPRs in medical diagnosis and a few other contemporary examples in medicine are cited briefly at the conclusion of this section.

General approach. The initial step in constructing a SPR for an image-based diagnosis is to obtain an exhaustive list of possibly relevant perceptual features in the image. This step is accomplished mainly by literature review and through close observation of, and interviews with, radiologists who specialize in interpreting the kind of image in question. Secondly,

perceptual tests analyzed by a *multidimensional scaling* (MDS) technique (e.g., Shiffman et al., 1981; Young and Hamer, 1987) may supply other features that are not verbalized by the radiologists, but are nonetheless used or usable in diagnosis.

In brief, in such perceptual tests, radiologists are invited to rate (on a ten-point scale) the degree of similarity between members of pairs of images; various representative images are presented successively in all possible pair-wise combinations. MDS analysis converts those similarity ratings into distances between the various pairs of images in a conceptual geometric space of several dimensions, wherein ratings of greater dissimilarity correspond to greater inter-image distances. The dimensions of the space (or the axes of its representation) are calculated in a manner to rationalize the total set of ratings and thereby to reveal the various aspects of appearance along which the images vary, or the perceptual dimensions inherent in the structure of the images. (To imagine how this analysis is accomplished, think of rating the straight-line distances between all pairs of state capitals to solve for a map of the U.S. showing its two dimensions, or measuring the distances between selected pairs of stars to give a space of three dimensions.)

The dimensions calculated by MDS are candidate features for the SPR. To determine which ones might actually be relevant features, the investigator successively arrays all the images along each dimension according to their respective coordinate values on that axis and asks the experts each time what perceptual feature is varying along that array. For some arrays, expert radiologists will give consistent feature names and express the belief that the feature is indeed diagnostic. The candidate features for a particular type of image and disease, as determined by interview alone or by interview plus MDS, have numbered between 30 and 70 in the studies discussed here.

Some paring of the set of candidate features may take place in discussion among radiologists and investigators; for example, it may be evident that a given feature is present twice because different radiologists gave it different names, or that two distinct features are highly correlated in occurrence. Rating scales are designed for the features remaining, on which a rating may signify the observer's confidence that the feature is present, a measurement of the size or extent of the feature, or a judgment of grade or degree or clarity. A consensus group of radiologists gives names to the features and selects particular images to represent points along the scale (particularly anchors at the endpoints).

In the next step, several radiologists view a set of a hundred or more "known" images, whose truth (presence or absence of cancer) has been established by pathology examinations (and possibly long-term, negative follow-up), and they rate each candidate feature for each case. At this point, one or another multivariate statistical analysis or pattern recognition technique (as described above in the section on SPRs) is applied to determine quantitatively how diagnostic or predictive each feature is in combination with others, and the features meeting some criterion of diagnosticity or predictive power are selected. The result is converted to a SPR that takes ratings of all features for any given case and issues an estimate of the probability that cancer is present in that case. A typical number of perceptual features contained in a SPR is about a dozen; this set is deemed necessary and sufficient for the required diagnosis.

Breast cancer. The first SPR for breast cancer in our series of studies (Getty et al., 1988) was developed with six mammography specialists of a university hospital and used to augment the diagnostic performance of six general radiologists in community hospitals. The task was to determine whether cases with evident focal abnormalities were malignant or benign. Specifically, the specialists helped with the choice of a master list of features, and rated a set of (100) training cases (half malignant, half benign) on those features to provide the basis for a SPR. The generalists first read (interpreted) a set of (118) test cases in their usual, unaided manner and months later read those cases with the checklist of the features that were incorporated in the SPR. In that augmented reading, they rated the SPR features for each case, and were given the SPR's probability estimate for each case before making their own judgment, on a five-category scale, of the likelihood that cancer was present. The second reading by the generalists provided a cross-validation of the SPR based on the specialists.

The ROCs for the generalists' baseline and augmented readings are shown in Figure 5. Each curve is based on four thresholds (and hence four points) corresponding to the internal boundaries of the five categories of the rating scale. The curve for the augmented readings is uniformly higher, having a *A* index of .87, compared to .81 for the baseline reading. The specialists, aided by the master checklist of features, and with a different set of cases, produced a ROC (not shown) with the same *A* index as the augmented generalists, .87. The SPR, by itself, with the generalists' feature ratings, yielded $A = .85$. All of the differences are statistically significant, so the results are that the generalists given the SPR probability estimate were more accurate than the SPR alone (which used their feature ratings), and the SPR enabled the generalists to reach the level of the specialists.

To see the clinical significance of these results, a decision threshold for the baseline ROC was chosen that approximated the thresholds obtained in four clinical studies of mammography accuracy that were comparable and available at the time (as described by Getty et al., 1988). This threshold point has ROC coordinates $P(FP) = .20$ and $P(TP) = .67$. The threshold point for the augmented ROC at the same P(FP) has a $P(TP) = .80$, i.e., .13 higher than the baseline performance (see vertical dashed line in Figure 5). So, if one chose to take the accuracy gain in additional TPs, there would be 13 more cancers found in 100 cases of malignancy. If it were desired to utilize the accuracy gain to reduce P(FP) rather than to increase P(TP), one can read the graph of Figure 5 in the horizontal direction: The augmented curve has a false-positive probability

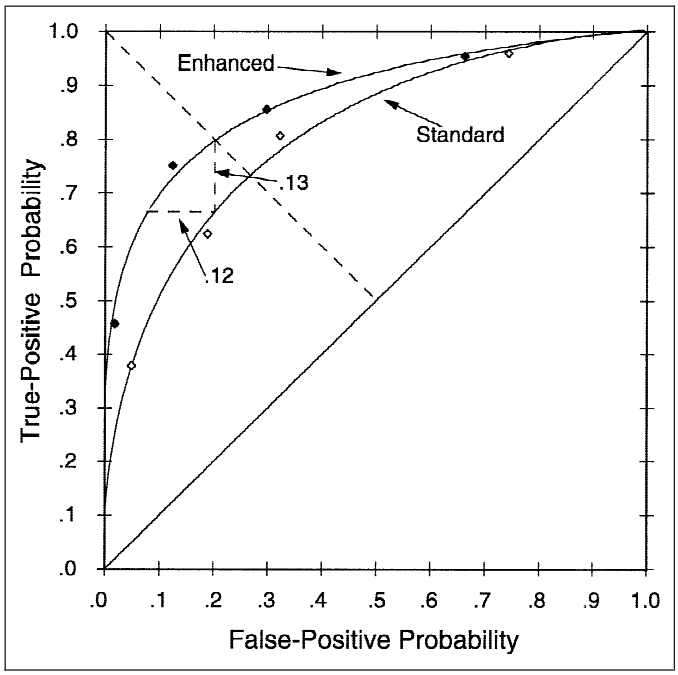

Fig. 5. Empirical ROCs (receiver operating characteristics) for general radiologists reading mammograms to distinguish malignant from benign lesions. The lower curve represents a baseline accuracy, for readings in the usual manner. The upper curve shows the accuracy obtained when the radiologists gave feature ratings and received the probability estimate of the statistical prediction rule (SPR). Curves are based on the pooled ratings of five radiologists who used a five-category rating scale for likelihood of malignancy. Two possible realizations of the gain in accuracy are indicated: an increase of .13 in the true-positive probability, P(TP), and a decrease of .12 in the false-positive probability, P(FP).

.12 less than the baseline curve, with P(FP) dropping from .20 to .08 (see horizontal dashed line in Figure 5). These probabilities can be translated into numbers of (additional) cases correctly diagnosed. Assuming a succession of 1000 cases in a referral hospital with a base rate of cancer of .32, the SPR procedure could find an additional 42 malignancies present or make 82 fewer false-positive decisions. Other ways to distribute the accuracy gain between increased TPs and decreased FPs could be achieved by adjusting the decision threshold.

We note, however, that although these calculations were likely valid when they were made, they may overestimate gains available now. Recent development of a less invasive (needle) biopsy technique, with relatively low morbidity, has served to reduce the cost of a false-positive diagnosis, with the result that the decision threshold in practice may have shifted to a more lenient setting (to the right along the ROC). We do not have data to enable estimating just where this current setting may be and hence can not estimate the size of currently available gains, but we point out that the gains may be smaller than those previously attainable: The relevant comparison of aided and unaided decision making may be at a point where their ROCs are closer together, especially on the TP axis. On the other hand if the SPR is used only for difficult cases, which may be a practical way to use it, then the typical threshold may again be near the middle of the graph, as in Figure 5, and the potential gains shown there may continue to be a good estimate.

The second study in this series (Swets et al., 1991) showed that the amount of increased accuracy provided by a SPR depends on the difficulty of the cases in the test set, with larger accuracy improvement for more difficult cases. Whereas for the full set (of 146 cases) the increased true-positive or decreased false-positive proportions were about .05, for the most difficult (56) cases, the changes in these proportions were on the order of .16. Their ROCs are seen in Figure 6. The difference between the top two curves (all cases) in the *A* index is .02 and for the bottom two curves (difficult cases) the difference is .12, from .60 to .72. Note particularly that the SPR had a beneficial effect even when the baseline performance was close to the chance level (the dashed diagonal running from lower left to upper right).

Another study showed the potential for determining relevant perceptual features not verbalized by image experts by means of multidimensional scaling (MDS) analyses of perceptual tests, as described above. Working with the experimental, untried image modality of diaphanography (light scanning), 9 of

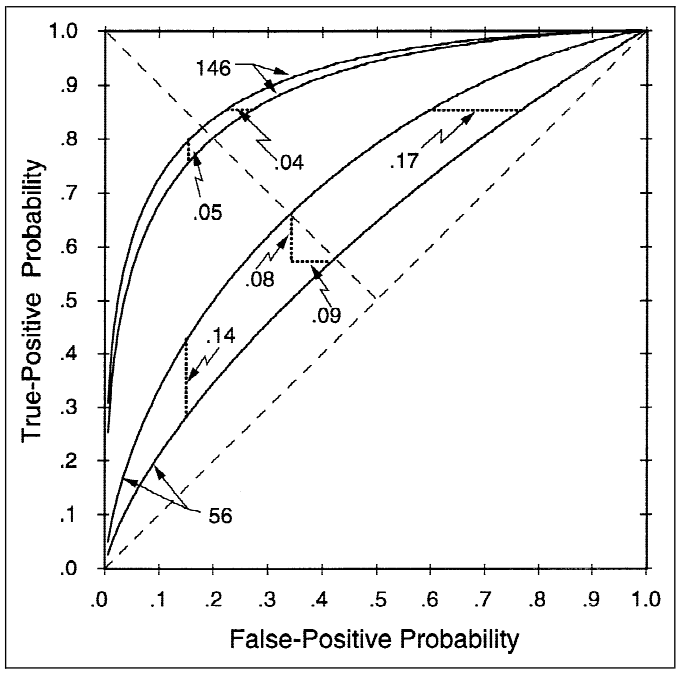

Fig. 6. Empirical ROCs (receiver operating characteristics) showing relative enhancement effects of a SPR (statistical prediction rule) applied to an easy and a difficult case set, with a larger gain for difficult cases. The curves are based on the pooled data of six radiologists. Dotted lines indicate illustrative gains in true-positive probability, P(TP), at a false-positive probability, P(FP), of .15, or, alternatively, decreases in P(FP) at P(TP) = .85. These differences are .14 and .17, respectively, for the difficult cases. Dotted lines near the center of the graph indicate the possibility of a simultaneous increase in P(TP) and decrease in P(FP), of about .08 for difficult cases.

a total of 17 features supplied by multivariate analysis for the SPR were provided by the MDS analysis (Getty & Swets, 1991). The diaphanography study also showed the importance of enhancing a new imaging modality by feature analysis and SPR methods before conducting an evaluation study to estimate its potential. Whereas unaided readers yielded a *A* index near .60, a SPR based on their ratings of the full feature set gave a *A* near .80. The implications of such a difference can range from likely rejection to possible further consideration of a new technique.

Prostate cancer. Magnetic resonance (MR) imaging is used to determine the extent of biopsy-proven prostate cancer, primarily to determine whether the cancer is contained within the gland and is therefore likely to be curable, or has spread beyond the gland and hence can be treated only palliatively. Our first study employed five radiologists who specialized in prostate MR and four radiologists who typically read MR of the body generally, each group reading one of two sets of 100 cases (Seltzer et al., 1997). Figure 7 shows ROCs obtained from the specialists, the lower curve when they gave ratings to the master set of features for each case to provide data for construction of a SPR, as well as giving their own estimates of the probability of extended cancer, and the higher curve based on the SPR calculated from their ratings. The *A* indexes are .79 for the lower curve and .87 for the higher curve. Even the specialists could be improved.

Figure 8 shows the ROCs from the generalists, the lower curve from a standard, baseline reading, and the upper curve from the SPR developed in their second reading with feature ratings of the same case set. The *A* indexes are .66 and .79, respectively. As in the breast study, the SPR brought generalists to the level of (feature-aided) specialists. As in the diaphanography study, it could bring a decision about potential usage of a technique from rejection to acceptance.

A second study showed the improvement of accuracy in prostate staging that could be achieved by constructing successively more inclusive SPRs (Getty et al., 1997). The objective variables of patient age, PSA (prostate specific antigen) test value, and the biopsy Gleason score (based on a pathologist's evaluation of tissue specimens) were considered along with a SPR based just on the perceptual features of the MR image. SPRs were constructed based on age only, on age plus PSA, on those two variables plus Gleason score, and on those three variables plus MR features. Figure 9 shows the ROCs of the four prediction rules, with *A* indexes progressing from .58 to .74 to .81 to .86. In a subset of difficult cases for which the PSA value and Gleason score were in an intermediate, incon-

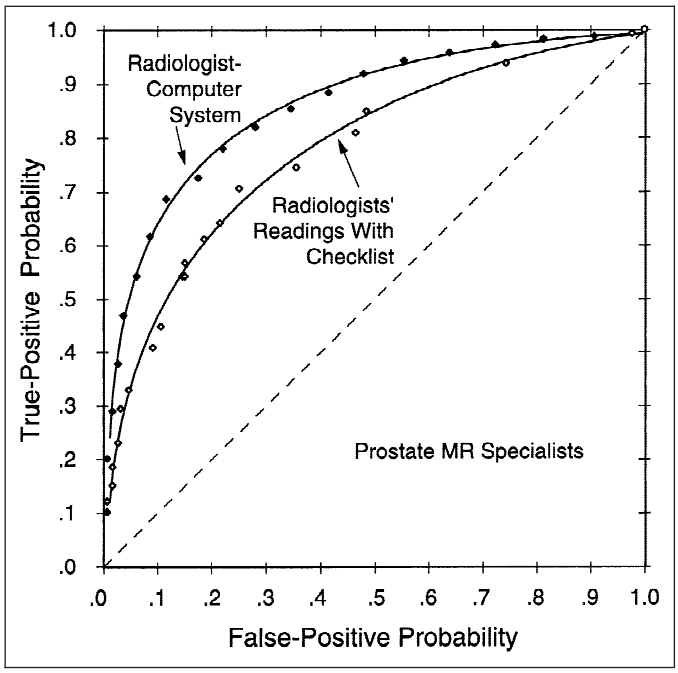

Fig. 7. Empirical ROCs (receiver operating characteristics) for specialists' readings of magnetic resonance (MR) images to determine the extent of prostate cancer. Pooled data from five radiologists. The lower curve was obtained when the readers were making feature ratings as well as an estimate of the probability of advanced cancer. The upper curve shows the performance of a SPR (statistical prediction rule) based on those feature ratings. For both curves, a computer curve-fitting program placed the probability estimates in categories to yield 19 decision thresholds and data points.

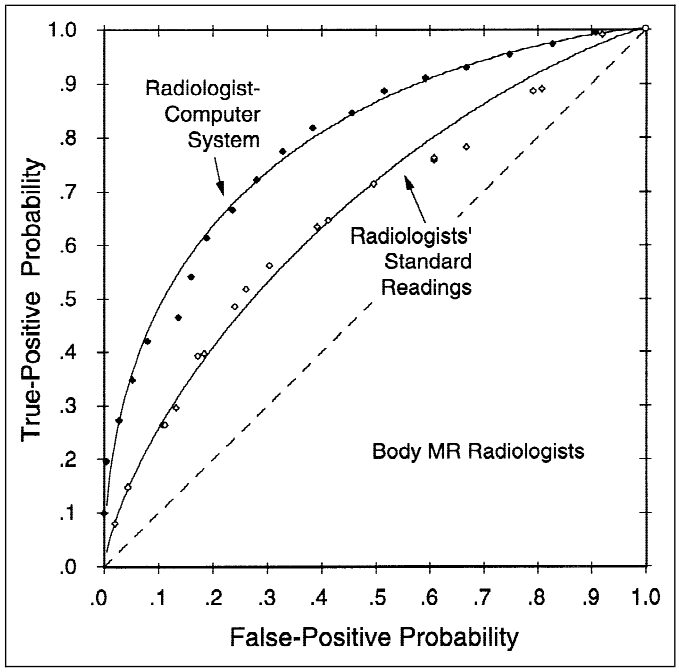

Fig. 8. Empirical ROCs (receiver operating characteristics) obtained from general body radiologists reading magnetic resonance (MR) images to determine the extent of prostate cancer. The curves are based on the pooled data of four radiologists. The lower curve represents a baseline reading and the upper curve represents the performance of a SPR (statistical prediction rule) developed from the radiologists' feature ratings in a second reading. A computer curve-fitting program placed the probability estimates, of both radiologists and the SPR, in categories to yield 19 decision thresholds and data points.

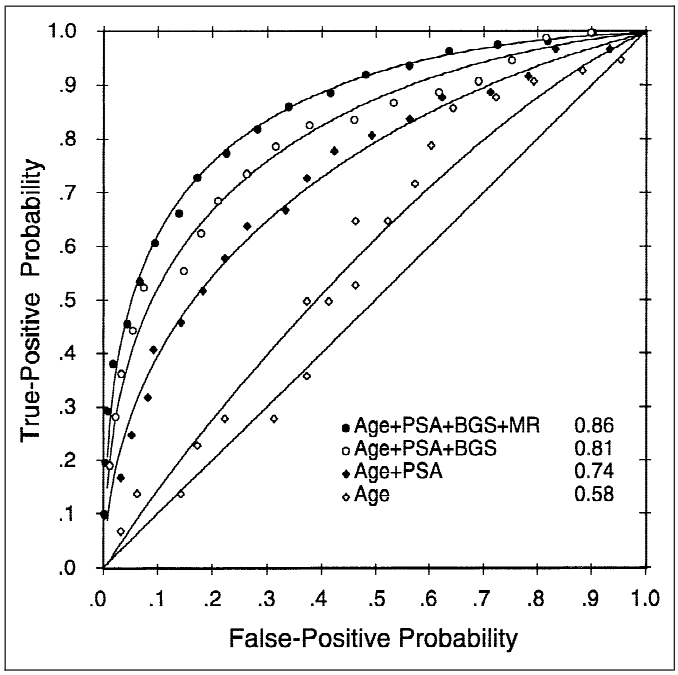

Fig. 9. Empirical ROCs (receiver operating characteristics) for determining the extent of prostate cancer, based on SPRs (statistical prediction rules) using one, two, three, or four predictor variables. Additional variables were added in the order in which they become available in the clinic. PSA is the test value of prostate specific antigen; BGS is the biopsy Gleason score; MR represents the SPR developed for readings of magnetic resonance images. For each rule, probability estimates of advanced cancer were categorized to yield 19 decision thresholds and data points. The accuracy measures *A* in the inset show the more inclusive SPRs to be increasingly more accurate.

clusive range, the *A* index for the SPR based on age, PSA, and Gleason score was .69; adding MR data to that SPR gave $A = .79$. In terms of the conditional probabilities: at P(FP) = .10, the value of P(TP) was increased by enhanced MR from .29 to .47.

Successful application of a SPR for prostate cancer. The staging of the extent of prostate cancer has provided an outstanding example of a highly useful SPR widely used in clinical practice (Partin et al., 1997). Data from three major medical centers were combined to predict the pathological stage of cancer, for men with clinically localized cancer, from the variables of PSA, clinical stage, and biopsy Gleason score. Four pathological stages considered were organ-confined disease and three levels of invasion beyond the organ. Charts were constructed so that for any combination of the predictive variables one can read the probability of cancer at the various pathological stages (along with 95% confidence intervals). So, for example, a patient having cancer of clinical grade T1c (cancer discovered only by biopsy following a PSA test), a PSA value between 4 and 10, and a Gleason score of 6, has a probability $p = .67$ of organ-confined disease, $p = .30$ of capsular penetration, $p = .02$ of seminal-vesicle involvement, and $p = .01$ of pelvic lymph-node involvement. The charts are used productively to counsel patients having a choice to make among alternative therapies. The authors give references to a dozen other studies providing confirmation of their results. Such data are the basis for the decision trees of the Treatment Guidelines for Patients published by the American Cancer Society.

Other work. Work on SPRs in medicine from the early 1960s, including some of his own on bone diseases, was reviewed by Lodwick (1986). Current studies of breast cancer include some using an artificial neural network as the basis for a SPR (e.g., Jiang et al., 1996; Lo et al., 1997). Another recent study used automated computer analysis of the mammogram image without human viewing, along with a linear discriminant SPR, and found $A = .87$ without the SPR and $A = .91$ with it (Chan et al., 1999). Other MR prostate studies include that of Yu et al. (1997).

Increased Utility: Setting the Best Decision Threshold

Concern for setting an appropriate decision threshold emerged early in medicine. The cost-benefit formula presented earlier in this article was promoted, for example, in influential books by Lusted (1968) and Weinstein et al. (1980), who were among the founders of the Society for Medical Decision Making. The following examples show how diagnostic decision making can be enhanced by optimizing decision thresholds.

Screening for the HIV of AIDS

Prominent screening tests for the virus (HIV) of AIDS consist of blood analyses that yield a continuous scale of a physical quantity (optical density). The selection of a decision threshold for any of the several tests available, as approved by the Federal Drug Administration, was made by its manufacturer. There is some suggestion that these thresholds were chosen to best discriminate between positive and negative cases (maximize the percent correct decisions of either kind), but there seem to be no published rationales for the particular thresholds chosen. Moreover, they vary considerably from one manufacturer's test to another. Informal and published recommendations that some formula for setting an optimal threshold be used for such medical tests (e.g., Lusted, 1968, and Weinstein et al., 1980) have not been heeded. An offer made to a drug company, of software that physicians might use to define an appropriate decision threshold for the company's test in any particular situation, was not accepted (A. G. Mulley, personal communication, 1990).

ROC data for the HIV. Three widely used HIV tests were evaluated by Nishanian et al. (1987) and the data were subjected to ROC analysis by Schwartz et al. (1988), as shown in Figure 10. They are seen to yield values of the threshold mea-

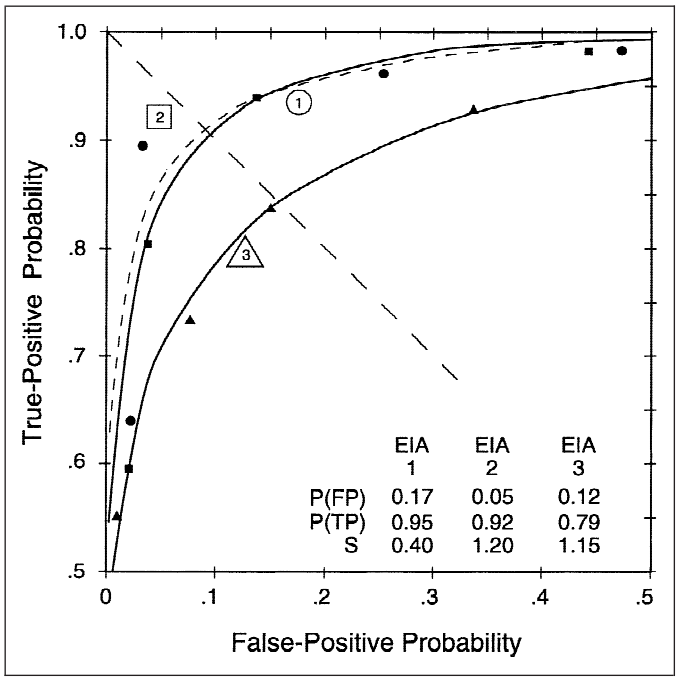

Fig. 10. Empirical ROCs (receiver operating characteristics) for three screening tests for the human immunodeficiency virus (HIV). The tests are called enzyme-linked immunoassays, abbreviated EIA; the three tests are numbered. The curve for each test is based on the points of five decision thresholds; curves for a given test are symbolized by circles, squares or triangles. The open data points with the tests' identifying numbers indicate the threshold used for each test in practice. The inset gives the measure *S* for each of these thresholds and the corresponding values of the false-positive probability P(FP) and the true-positive probability P(TP). (Note that this graph includes only the upper left quadrant of the usual ROC graph.)

sure *S* of .40, 1.20, and 1.15. That is, one threshold is on the lenient side of neutral and two are on the strict side. Their pairs of P(FP) and P(TP) values were (.17, .95), (.05, .92), and (.12, .79) , respectively. Consider the first and second tests listed; the first test picks up a few more TPs than the second (from .92 to .95) at the expense of issuing three times as many FPs—17 per hundred versus 5 per hundred. It is difficult to imagine a good reason for both tests to be approved for use in the same settings with their diverse thresholds. Incidentally, those two tests were substantially more accurate (*A =.97)* than the other ($A = .92$) for the case sample tested.

Fixed vs. changing threshold. Of further concern is the fact that the thresholds for these tests were originally chosen when the tests were used to screen donated blood and then left unchanged when the tests became used to diagnose people. The difference between the costs of an FP decision for discarding a pint of uncontaminated blood, on the one hand, and for pursuing further tests for an uninfected person, on the other, would seem large enough to call for some shift in threshold. Similarly, thresholds were not reconsidered when the tests were applied to different populations characterized by very different rates, or prior probabilities, of the disease. Thresholds remained fixed across low-risk blood donors, high-risk blood donors, military recruits, and methadone-clinic visitors, for which the numbers of infected individuals per 100,000 were estimated to range from 30, through 95, through 150, to 45,000, respectively (Schwartz et al., 1998). Assuming for the moment constant benefits and costs of decisions, that amount of variation in the base rates would move the optimal threshold value of *S* over a large range, from 3,000 to near 1, i.e., from a point very near the lower left corner of the ROC graph to a point near the center. The corresponding variation in P(FP) and in P(TP) would be very large, on the order of .50.

Other cost-benefit factors that might affect the placement of the decision threshold include whether the test is voluntary or mandatory, and mandatory for what group of persons, for whose good. For example, testing is done in connection with life-insurance and health-insurance examinations, where false-positive decisions can have significant lifetime costs to individuals. There are other mandatory tests, such as those for certain international travelers, for which the benefits of detection are small. Still other factors that might affect the threshold setting are whether the results are confidential or anonymous and how effective the therapy may be (Meyer & Pauker, 1987; Weiss & Thier, 1988).

Screening low-risk populations. Consider another instance of screening low-probability populations, namely a company's employees, for whom the prior probability of HIV is about .003 (Bloom & Glied, 1991). Ordinarily in such settings, a positive outcome on a typical screening test is followed by a more conclusive (and expensive) confirmatory test to reduce the number of false-positives (Schwartz et al, 1988). The College of American Pathologists' estimates of P(TP) and P(FP) for the best screening and confirmatory tests lead to the result that after a positive result on both tests, the probability of HIV is .13 (Bloom & Glied, 1991). Hence, six of seven individuals diagnosed as positive in this manner would be told they have the HIV when in fact they do not (Swets, 1992).

Detecting cracks in airplane wings

The principal techniques for nondestructive testing of metal structures provide a visual pattern for interpretation by technicians, for example, ultrasound and eddy current. In both cases, the basis for a decision is the judged weight of the perceptual evidence and so the observer acquires a degree of confidence, or probability estimate, that a flaw is present. The two-valued diagnosis of flaw present or not, usually a crack caused by metal fatigue, requires that a decision threshold be set along the scale.

In looking for cracks in airplane wings, the costs of incorrect decisions are large and obvious. A false-negative decision, missing a crack actually there, can jeopardize the lives of many passengers. On the other hand, a false-positive decision takes a plane out of service unnecessarily, possibly at great inconve-

nience and large dollar cost. On the face of it, the benefits and costs point toward a lenient threshold for declaring a flaw—lives versus dollars. Still, the prior probability of a flaw is very low and such low prior probabilities, even with moderate to strict thresholds, tend to produce an unworkable number of false-positive decisions. Setting the best decision threshold, again, must involve probabilities as well as benefits and costs. There are a few tentative references to the problem and the solution in the nondestructive-testing literature, e.g., Rau (1979), Rummel (1988), and Sweeting (1995), but no instances of experimental, let alone systematic, use of reasoned thresholds seem to have been made.

Test data in this field are hard to come by: A reliable determination of "truth" for flaw present or not in each specimen requires destructive testing of the entire test set of specimens which then, of course, are not available for comparative evaluation of the next diagnostic technique or the next group of technicians to come along. A classic, atypically ambitious study was mounted by the U. S. Air Force in the 1970s. It was characterized as "Have Cracks, Will Travel" because it brought 149 metal specimens to 17 bases where they were inspected by 121 technicians using ultrasound and 133 technicians using eddy-current displays.

ROC data. As reviewed in earlier publications (Swets, 1983, 1992), the study asked the technicians for only a binary response and hence obtained just a single ROC point from each. The data, however, are highly revealing. Figures 11 and 12 show that the ROC points obtained cover almost the entire

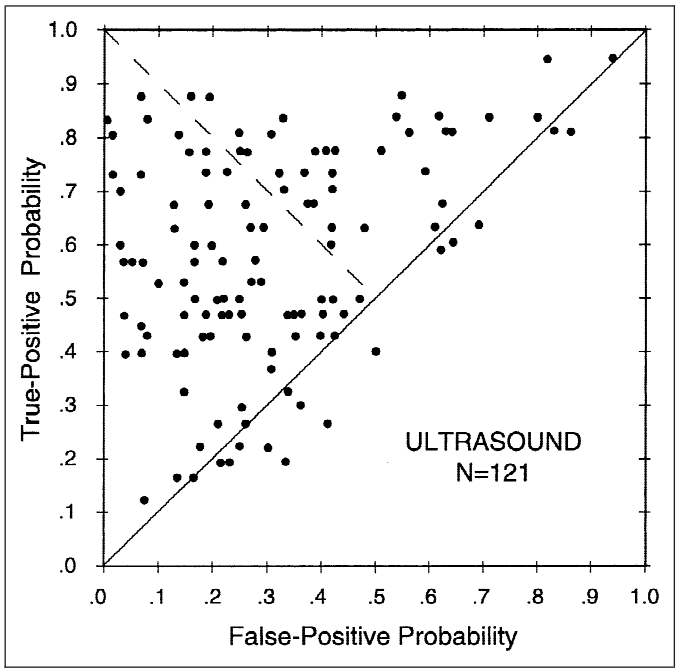

Fig. 11. Single, empirical ROC (receiver operating characteristic) points for each of 121 technicians inspecting 149 metal specimens for cracks with an ultrasound image.

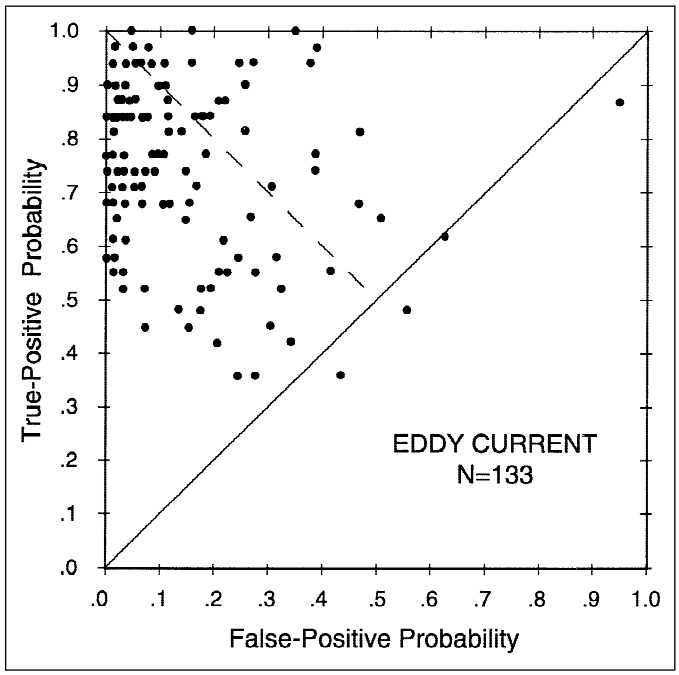

Fig. 12. Single, empirical ROC (receiver operating characteristic) points for each of 133 technicians inspecting 149 metal specimens for cracks with an eddy-current image.

usable ROC space, for both imaging techniques. The spread of data points from P(FP) = 0 to almost 1.0 demonstrates total inconsistency among, and no control over, the technicians' decision thresholds. No publication, either in the report or open literature on materials testing, has appeared to us to suggest that there has been an adaptive response to this state of affairs.

A note on accuracy. A break down by air-force base of the data points in Figures 11 and 12 (not shown) indicates that accuracy varied extensively from one base to another, with technicians at a given base being quite consistent, and with the average accuracies of the bases varying uniformly across the full range of possible accuracies. Roughly, bases had an average ranging from $A = .95$ to .55 (see Figure 3). The strong suggestion is that the perceptual features and weights used by technicians at the highly accurate bases could be analyzed in the manner used for mammography experts as described above and the result carried to the under-performing bases. Thus, Figures 11 and 12 point up the potential of an SPR to increase accuracy as well as the potential for threshold analysis to increase utility.

A confirming study. A study of a commonly used eddy-current display for examining steam generators showed wide variation among technicians to persist in non-destructive materials testing (Harris, 1991). For one representative fault, for example, the observed variation in P(TP) across technicians was nearly .40. The (plus and minus) one-standard-deviation

range was from .69 to .93, indicating that one-third of the technicians fell outside of that range.

Other Examples in Brief

Weather forecasting

The National Weather Service estimates the risk of certain hazards, such as tornadoes, hurricanes, and heavy rains, which pose a threat to life or property. To assist in assessing risk, information is routinely collected on variables (e.g., barometric pressure, wind speeds, cloud formation) known to be predictors of one or another of these hazards. This information is analyzed by regression-based computer programs that incorporate models of the association between patterns of these predictors and the occurrence of given hazards in the past. These programs yield objective predictions of various weather events. These objective predictions are given at regular periods to meteorologists in local areas. These local meteorologists may then modify the objective predictions in light of predictors that they believe were not adequately accounted for in the computer model, or in response to new information that has become available since the objective prediction was formulated. The objective predictions, the SPR's, are often referred to as "guidance, not gospel" by the local meteorologists. A subjective prediction is then publicly issued, and this risk message is referred to as the forecast. Weather forecasting is one area in which the "clinical" adjustment of a SPR's output actually increases, rather than decreases, predictive accuracy. The subjectively adjusted SPR predictions of temperature and precipitation are consistently more valid than the unadjusted objective SPR predictions (Carter & Polger, 1986).

Weather forecasting for commercial interests adopted the practice of setting optimal decision thresholds, indeed using the formula presented earlier in this article, more than 25 years ago (Miller & Thompson, 1975).

Law school admissions

Decisions about whom to admit to universities and to graduate and professional schools have for many years been made with the help of a SPR. In the case of law schools, for example, the admissions committee is typically presented with an "Admissions Index," which is the applicant's score on a SPR that predicts first-year grades at that particular law school. Two variables usually go into the SPR: undergraduate grade point average (GPA) and Law School Admissions Test (LSAT) score. If an applicant scores above a certain decision threshold on the Admissions Index, he or she is presumed to be an "admit." That is, it would take a flaw elsewhere in the application (e.g., the impressive GPA was achieved by enrolling in very weak courses) to deny the applicant admission. Likewise, schools set a decision threshold for "presumed reject," whereby any applicant with an Admissions Index below this score will be rejected absent unusual factors' (e.g., graduate work of extraordinary quality) being found elsewhere on the application.

Each law school sets its own decision thresholds for presumed admit and presumed reject, with the more selective schools setting the thresholds higher than the less selective schools. Applicants scoring between these two thresholds have their applications reviewed more intensively by members of the admissions committee who can, in effect, adjust the Admissions Index by taking into account additional variables not in the SPR, such as the quality of the undergraduate institution attended and the stringency of its grading standards (and the farther apart the two decision thresholds are set, the larger this middle group will be). These adjustments are made "clinically," by members of the admissions committee.

It is interesting to note, in light of issues raised earlier in this article, that at least some of these "subjective" variables can be quantified and incorporated into the SPR. For example, at the University of Virginia School of Law, where one of us (JM) teaches, a new and expanded Admissions Index is being used as a tool in selecting the class of 2003. This index includes two additional variables: the mean LSAT score achieved by all students from the applicant's college who took the LSAT (a proxy for the quality of the undergraduate institution) and the mean GPA achieved by students from the applicant's college who applied to law school (a proxy for the extent of grade inflation, and having a negative weight in the SPR). This new four-variable SPR predicts first-year law school grades (correlation $r = .48$) significantly better than the old two-variable SPR (correlation $r = .41$) (P. Mahoney, personal communication, 1999). Note that the results of this expanded SPR are still adjusted by the admissions committee to take into account other, harder-to-quantify variables, such as unusual burdens borne or achievements experienced during college, to produce the final decision to admit or reject. The degree of adjustment is less than it was previously, however, because two formerly "subjective" variables have become "objective" and now contribute to the SPR itself.

For the related problem of making personnel decisions based on aptitude tests, an approach akin to the formula given above for setting the optimal decision threshold has been in use for many years (Cronbach & Gleser, 1965).

Aircraft cockpit warnings

Based on specialized sensing devices, warnings are given to airborne pilots that another plane is too close or is threatening to be, that they are getting too close to the ground, there is engine failure, or wind shear is present in the landing area. A problem gaining recognition, for example, by the National Aeronautics and Space Administration, is how the various decision thresholds should be set to avoid missing a dangerous condition while not crying wolf so often that the pilot comes to ignore or respond slowly to the warning signal. Unfortunately, the moderate diagnostic accuracies of the sensing devices along with the low prior probabilities of the dangers conspire

to produce many false alarms (Getty et al., 1995)—so many that officials have asked whether just one true alarm among 20 total alarms is enough to maintain the pilot's rapid response. Indeed, a few years ago the Federal Aviation Administration ordered a shutdown of collision-warning devices on commercial airliners because of the serious distractions they presented both to pilots and air traffic controllers. As a particular example, just one aircraft responding to a wind-shear alarm by circling the field before landing can put air-traffic control at a busy field under additional strain for several hours. It is still common practice, however, for purchasers of cockpit warning systems to set specifications for the product that require only a high P(TP), without mentioning the P(FP), and manufacturers have been willing to comply.

That the typically low prior probabilities in some settings can lead to extreme results is exemplified by the calculated performance of a detector of plastic explosives in luggage as considered by the Federal Aviation Administration. With an apparently passable accuracy of P(TP) = .95 and P(FP) = .05, it was estimated to produce in standard operation 5 million false-positives for each true-positive result (Speer, 1989).

Disability determination

Applicants for disability status under the Social Security Administration presently undergo a five-step sequential evaluation. A binary (positive-negative) determination is made for these variables in turn: (1) whether the applicant is engaging in substantial gainful activity; (2) whether the impairment is severe; (3) whether the impairment is on a list of qualifying impairments; (4) whether the applicant is able to do work done previously by the applicant; and (5) whether the applicant is able to do other work. At each step, the application is denied, accepted, or forwarded to the next step.

Four of these variables (excepting number 3) are essentially continuous and hence require a "judgment call" for a binary decision: substantiality of gainful activity, severity of impairment, residual functional capacity for past work, or for any work. The assessment of each variable could be made on a rating scale, and so the question arises if accuracy of disability determination might be increased by rating them for each case and entering them as continuous variables in a SPR, which would give them proper weights and then issue what is essentially a "disability score." In principle, accuracy might be enhanced because then the information loss that may come with dichotomizing continuous variables would not be at risk four times, but would be confined to the final decision variable, the score. (There would probably be, under this scheme, different SPRs for mental and physical disability.)

Perhaps more important, given a disability score, the placement of a decision threshold for allowance could be discussed precisely, and given "sensitivity" testing for its best location. At present, the effective threshold changes dramatically from the initial level of a claim evaluation to an appeals level: roughly two-thirds of cases first denied and then appealed are allowed at the second level; award rates at the appeals level are more than twice those at the initial level (General Accounting Office, 1997; Social Security Advisory Board, 1998). Indeed, a class-action suit by a Florida legal association was brought to remedy the plight of initially denied applicants who are not aware of the potential of making a formal appeal.

Quality of sound in opera houses

To take some respite from our *Sturm und Drang* (and crashes and diseases), consider the objective and subjective evaluation of 23 opera houses in Europe, Japan, and the Americas (Hidaka & Beranek, 2000). Twenty-two conductors rated the several opera houses for acoustical quality on a five-category scale, and the average ratings of the respective houses were related to several physically measured acoustical variables. The purposes of the evaluation were to establish, in effect, a SPR as a framework for evaluating existing opera houses and for suggesting guidelines for use in the acoustical design of new opera houses.

Five important, independent, objective acoustical variables measured in the audience areas were: reverberation times; the time difference and loss of loudness as sound transverses the head from one ear to the other (related to an impression of "spaciousness"); the time delay between the direct sound from the stage and its first wall reflection (related to "intimacy"); the strength of the sound; and the bass ratio. Two additional variables thought to be important, but difficult to measure physically, are "texture," having to do with the number and quality of early, lateral reflections, and "diffusion," resulting from large irregularities on the walls and ceiling where reverberant sound is formed (e.g., niches and coffers) and small irregularities on the lower side walls and balcony fronts that give "patina" to the early sound reflections.

The four opera houses that received average ratings of 4 or higher on the conductors' 5-point scale, and were highly evaluated objectively, are those in Buenos Aires, Dresden, Milan, and Tokyo. The Tokyo opera house rates high despite just two years of service; it is the one that was explicitly designed by Hidaka and Beranek with the above-mentioned variables centrally in mind.

"It's laptop vs. nose"

This section heading is quoted from a *New York Times* article under the byline of Peter Passell (1990), entitled "Wine equation puts some noses out of joint," and introduces a second topic in our small foray into aesthetics.

Among the most interesting uses of a SPR is Ashenfelter, Ashmore, and Lalonde's (1995) successful attempt to predict the quality of the vintage for red Bordeaux wines. Taking the market price at auction of mature Bordeaux wines as their index of "quality," Ashenfelter et al. show how the vintage of a wine strongly determines its quality. Some vintages are very good, some very bad, and most in between. By the time a Bordeaux wine is mature and drinkable, there is usually con-

siderable agreement among wine drinkers as to the quality of its vintage. The trick is how to predict this quality (i.e., auction price) decades in advance, when the wine is young and undrinkable. The typical way this is done is "clinically," by having wine experts swirl, smell, and taste the young wine. Ashenfelter et al., however, observed that the weather during the growing season is a key determinant of the quality of any fruit, including grapes. More specifically, "great vintages for Bordeaux wines correspond to the years in which August and September are dry, the growing season is warm, and the previous winter has been wet." They developed a multiple-regression SPR, which they refer to as the "Bordeaux equation," that consists of the age of the vintage and indices that reflect the above-mentioned temperature and precipitation variables (e.g., millimeters of rain in the Bordeaux region during certain months of the vintage year). This SPR accounts for fully 83% of the variance in the price that mature Bordeaux red wine commands at auction. Moreover, the SPR produces its predictions as soon as the growing season is complete and the grapes are picked—before any "expert" has even sipped the young wine. As might be imagined, the reaction of the wine-tasting industry to the Ashenfelter et al. SPR has been "somewhere between violent and hysterical" (Passell, 1990). But drinkers of, and investors in, red Bordeaux wine have reason to be grateful to the developers of this SPR.

CONCLUSIONS AND DISCUSSION

It seems fair to conclude from the examples provided above that SPRs can bring substantial improvements in the accuracy of repetitive diagnostic decisions and that decision analysis can improve the utility of such decisions. We mention a few other benefits that may accrue from these methods, just from having the right set of features or variables specified.

Additional Benefits of a Systematic Approach to Predictor Variables

Speeding the specification of diagnostic features

Consider the manner in which visual features of medical images of a new modality are identified in typical practice. Ordinarily, depending on their own initiative or that of the equipment manufacturer, individual radiologists propose some features depending on their own (and possibly close colleagues') experience, and present them in seminars, teaching files, or journal articles. The accumulation of data is slow, in part because radiologists do not always have systematic feedback on the pathology results of their own patients, let alone of other patients. Then, perhaps a few years later, a synthesis of features may appear in a manual or textbook. Such a laissez faire approach is unnecessarily slow and quite out of synchrony with the pace of modern development in equipment for medical imaging. In contrast, application of multivariate feature analysis and development of a SPR can motivate fast and wide data collection of a proven case set and reveal in a matter of months a fairly complete set of features and their relative weights as well. (The follow-up of normal cases needs to be pursued longer as a refinement.) The distribution channel may be print media, but it could also be interactive over whatever computer network is appropriate.

Facilitating communication among diagnosticians

Even when the radiological SPR is not put into wide use, the radiologists who have become acquainted with the technique generally agree that the sets of features and weights it has identified can be very useful in facilitating communication. Features have clear advantages over a holistic approach to image interpretation in this regard. Mammogram features identified in the work described above contributed to a standardized reporting system, for reports from the radiologist to the referring physician and surgeon, developed by the American College of Radiology (Kopans & D'Orsi, 1992, 1993). The possibility that the radiologist's quantitative ratings of the mammogram's features can be translated automatically, by computer-based linguistic techniques, into a useful report of findings has received support (Swets, 1998).

Another result of SPR-based feature analysis could be to facilitate discussion between radiologists holding different opinions about a given medical image. Still another use of a well-defined feature set would be for teaching purposes, even for highly experienced radiologists in continuing education. Moreover, as shown above, a feature analysis has the potential to promote general radiologists to the level of specialists for a given organ or type of image. It may bring novices more quickly to the approximate level of experts. Because the perceptual-feature approach does not depend on knowledge of underlying anatomy or pathology, we consider the possibility that it may help to teach paramedics to read images, which might have special value in countries or regions where radiologists are in short supply.

Why Are These Methods Relatively Little Used?

There are several hindrances, if not roadblocks, that impede progress on the decision-support front. Grove and Meehl (1996) list 17 reasons that clinicians have given for not adopting SPRs for psychiatric prognoses. They focus on attitudinal factors in a climate in which the SPR is viewed as replacing or degrading the clinician. Other fields have more benign issues. We imagine that the weather forecaster has no problem receiving a SPR contribution to the forecast (for example, that there is a 31% chance that Hurricane Floyd will hit the Cape and a 33% chance that it will hit the Islands). We consider below some purely attitudinal factors, but treat mainly logistic and other practical matters.

Maybe the main attitudinal factor is that diagnosticians,

perhaps especially those with patients awaiting the result, naturally want to feel that they "understand" their diagnoses and recommendations; they want to know why they arrived at some point and they want to be able to give a narrative account of their thought process. Such a plot line may be difficult to find in a SPR's results, perhaps with some of its statistically important predictor variables' not being self evident to anyone, and with other seemingly obvious variables not present in the SPR ("your father's drug use matters, your mother's does not"). Producing the largest area under the ROC is technical, may seem like dust-bowl empiricism at best, and is simply not satisfying to diagnostician or client.

These methods are little known

To be sure these methods are part of our culture; literate people know they exist, perhaps in weather models or insurance models of life expectancy. Still, "SPR" (or any synonym) is not a household word. The concept is not very clear. Many people know of isolated examples, but have not integrated over enough of them to "have the concept," to see SPRs as forming a significant class, much less as a phenomenon that can be studied and exploited. Similarly, making decisions based on odds and the costs/benefits of outcomes is something every human does. Yet, the idea that diagnostic decision thresholds can be set deliberately and perhaps optimally is often not there. Indeed, the existence of a very broad class of problems that may be called "diagnostic" is not a common idea, much less that there may be a science of diagnostics. These are hurdles for decision-support enthusiasts when trying to persuade administrators that their agencies need a science of diagnostics, for examples, the Federal Aviation Agency and the Food and Drug Administration.

The need for adaptive SPRs

In the context of our violence example we raised the issue of whether an SPR based on one population of cases will generalize well enough to another. That question arises in many, perhaps most, diagnostic fields. In medicine, for example, the characteristics of patients undergoing mammography will vary from university to community hospitals and across regions of the country. Again, optimal weather models will vary with locale. SPRs for cracks in airplane wings may differ from large commercial planes to small private planes. Hence, it is desirable to build SPRs that can adapt their variables and variables' weights automatically to case samples for which different ones will be optimal. As the sample of (proven) cases grows in any particular setting, the SPR in use there should change, at whatever convenient interval, so that it becomes tuned or tailored to that setting.

Not only does the world vary from one location to another, but it changes dynamically and thereby creates a larger problem for some fixed SPRs. For example, a fixed SPR in medical imaging may not be current for long enough to make its implementation worthwhile. One can assemble a useful SPR for MR imaging of some disease/organ only to have a modification of MR technique come along in months, and by virtue of rearranging the physics and perception of the image, call for a reanalysis of the SPR's image features as well as their weights. In this case, automatic adaptation is not possible. Creative, human intervention is required to ascertain what new perceptual features are necessarily added to the existing SPR. On the other hand, as assumed in the preceding paragraph, the problem is less severe in fields in which the predictor variables are largely objective. Thus, discovery of a new risk factor related to violence can be accommodated by an SPR geared to adapt so that a new predictor variable is as easily handled as are new weights for a given set of variables.

Several methods for updating SPRs have recently been reviewed along with a successful application to law school admissions (Dawes et al., 1992). However, there are apparently not many such adaptive SPRs in routine, practical use at present. Fortunately, the rate at which computer databases are being assembled and shared over networks suggests that common use of self-tuning SPRs need not be far off. All of us concerned with diagnostics should be anticipating the day when handling data is not a problem.

Although flexibility is desirable in many settings, its desirability should not be used as a reason for abandoning a SPR in favor of human intuition. The latter is indeed flexible, but often in an unsystematic way not open to scrutiny. An adaptive SPR can be flexible as well, but in a systematic manner whose validity can consequently be evaluated.

Accountability

When a human contributes subjective estimates of the values of a SPR's predictor variables (e.g., ratings of perceptual features of an image), then each small piece of the diagnosis is objectified and placed indelibly on the record. Whether liable or not in a legal sense, the diagnostician may well feel likely to be called on to be responsible (by an employer or patient) for the ultimate validity of that entire, detailed, quantitative record. A more comfortable position would be to be responsible merely for written notes of one's impressions.

Inconvenience

The "inconvenience" of using SPRs covers a multitude of sins, beginning perhaps with computer issues: the sheer need to face a computer, and perhaps to have a computer selected, purchased, installed, maintained, and upgraded. Other issues have to do with efficiency and workflow; for example, must a radiologist lay aside the microphone used for dictation to enter data via a keyboard? Such questions may have answers; for example, speech-recognition systems will allow the radiologist full control and data entry through the microphone. Will the data entry take more time? Perhaps, but it may also produce automatically the "report of findings" from radiologist to re-

ferring physician and end up saving time. In short, the "human-factors" problems may be soluble.

The ideas are technical

Probabilities. We claim in the Appendix that probabilities are useful and straightforward. But that does not help people who are put off immediately and completely by their impression that a major effort will be involved in gaining an understanding. On the other hand, having read this far, we hope that the reader will find it difficult to imagine diagnoses undertaken competently without some fundamental acquaintance with probabilities.

Forging agreement on threshold probabilities will often be difficult. Either a lack of understanding of probabilities, or a lack of consensus if such understanding exists, can bedevil the process. Illustrative data on threshold probabilities obtained in a survey of medical directors of AIDS counseling and testing centers revealed that whereas 25% of the respondents would initiate a discussion of decreased life expectancy with patients having a probability of infection greater than .15, fully 50% of them would require a probability of infection of .95 or higher before having that discussion. Further, whereas 43% of the directors would advise against pregnancy for patients with a probability of infection above .15, another 30% would require a probability of .95 to do so (Mulley & Barry, 1986).

The tools are cumbersome

A recent proposal is that complex SPRs and decision analysis be replaced by "fast and frugal" versions (Gigerenzer et al., 1999). In such simplifications, a SPR may use just one predictor variable, for example, or treat all variables as having the same weight (following Dawes and Corrigan, 1974; Dawes, 1979). We think that such simple heuristics bear study for the day-to-day ad lib decisions of individuals. However, we do not expect them to help generally in repetitive problems of the same form, largely for professionals, as we have considered in this article. For our type of problem, possibly excepting such as the hospital emergency room, speed and simplicity are not at issue. So the law school admissions office can set up an SPR and decision threshold(s) and apply them cost-effectively to its thousands of applicants in a given year; using one variable rather than four saves nothing of consequence. Likewise, the weather forecaster is not tempted to discard the fifth or tenth highest rated predictor variable if it contributes to accuracy. The radiologist should be led to rate the dozen or so perceptual features that make a difference and the SPR might as well use all available ratings. Predicting violence may need to be fast in outpatient treatment, but not in a forensic facility where nobody is going anywhere soon. Also in these cases, the selection of a decision threshold is usually an important societal matter, warranting a good deal of time and effort. The speed-accuracy tradeoff is a cost-benefit question and, generally, even small increments in accuracy or utility are to be preferred to savings in time or effort.

Defining benefits and costs

Assessing benefits and costs can be problematic; publicizing them can leave the decision maker vulnerable to criticism. How many safe people should be hospitalized as "dangerous" to prevent discharging one patient who turns out to be violent? No court has ever answered that question with a number. Judges are notoriously reluctant to set decision thresholds that depend on overt cost-benefit consideration, as are many other professionals and officials. The decision analyst's position is that making consistent decisions requires a stable threshold; that any threshold implies some cost-benefit structure; and that an explicit consideration of benefits and costs may be preferable, for some purposes, to sweeping them under the rug. The decision maker who is explicit does indeed invite criticism, but such vulnerability to criticism in itself may be a positive source of improvement. In contrast, an appeal to ineffable intuition ends with the appeal itself (because there is no way of disputing it); hence it precludes critical evaluation and consequently precludes productive modification of the way in which the decision was made. There are certainly fields where, realistically, benefits and costs will continue to be left vague. To some they may suggest boundary conditions for the sort of decision analysis advanced here; to others they will be a challenge.

More complex computer-based systems have not done well

It may be that computer-based systems for two-alternative diagnoses suffer by inappropriate generalization from experience with medical systems built to contend with more complex diagnoses, in which the diagnostician describes a patient's symptoms and looks for a listing of all diseases that should be considered. Such systems have been based on artificial intelligence (expert systems), probabilistic reasoning, a combination of the two, or on other methods. Performance deficiencies of four prominent examples were reviewed by Berner et al. (1994). An accompanying editorial in *The New England Journal of Medicine* gave these examples a grade of "C" (Kassirer, 1994).

What do these hindrances add up to?

It may well be that one particular hindrance just mentioned is sufficient to preclude use of a SPR or an analysis of decision utility in a given situation. Together, they may seem overwhelming. Yet, SPRs and decision analysis have been regarded as practical and have been in routine use in some settings for decades, weather forecasting perhaps being the main "existence proof" among the examples of this article.

Making sense of the historical pattern of use of the decision-support methods is not easy. On the one hand, one might understand why radiologists, who are technically oriented, would become acquainted with the methods before administrators of broad national health programs, such as HIV screening, and of the regulatory agencies that approve the screening

tests. On the other hand, it is not evident (to us) why meteorologists have shown for decades a sophistication not mirrored by materials scientists who provide the science for nondestructive testing programs.

The Importance of Public Awareness of Decision-Support Methods

This article is published in *Psychological Science in the Public Interest*, featured in an accompanying press conference, and rewritten in shorter form for *Scientific American*, precisely because these mechanisms were created to bring such information to the public and its decision makers. An earlier effort to reach decision makers was a *Science and Public Policy Seminar* for government officials and congressional staff, sponsored by the Federation of Behavioral, Psychological and Cognitive Sciences (Swets, 1991). That presentation led to this line of work being selected as the illustration of practical benefits of basic research in the behavioral sciences in a White House science policy report (co-signed by Clinton and Gore, 1994). In the longer run, a national awareness may help to make inroads in the procedures and regulations of policy makers.

Acknowledgements—We appreciate helpful comments for exposition from the editor's reviewers: Klaus Fiedler, Elliot Hirshman, and John Wixted. Steven Seltzer also made helpful suggestions in this regard. We acknowledge the essential contributions of our colleagues in this line of work over several years. With Swets: David J. Getty and Ronald M. Pickett at BBN Technologies; Carl J. D'Orsi at the University of Massachusetts Medical Center; Charles E. Metz at the University of Chicago Medical School; and Robert A. Greenes, Barbara J. McNeil, Steven E. Seltzer, and Clare M. C. Tempany at the Brigham and Women's Hospital, Harvard Medical School. With Dawes: David Faust at the University of Rhode Island; Paul E. Meehl at the University of Minnesota; and William Chaplin, Lewis R. Goldberg, Paul J. Hoffman, John W. Howard, and Leonard G. Rorer at Oregon Research Institute and the University of Oregon. With Monahan: Paul Appelbaum, Steven Banks, Thomas Grisso, Edward Mulvey, Pamela Robbins, Loren Roth, Eric Silver, and Henry Steadman, all on the MacArthur Violence Risk Assessment Study.

REFERENCES

Ashenfelter, O., Ashmore, D., & Lalonde, R. (1995). Bordeaux wine vintage quality and the weather. *Chance, 8*, 7–14.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society, 53*, 370–418.

Berner, E.S., Webster, G.D., et al. (1994). Performance of four computer-based diagnostic systems. *The New England Journal of Medicine, 330*, 1792–1796.

Bloom, D.E. & Glied, S. (1991). Benefits and costs of HIV testing. *Science, 252*, 1798–1804.

Carter, G. & Polger, P. (1986). A 20-year summary of National Weather Service verification results for temperature and precipitation. *Technical Memorandum NWS FCST 31*. Washington DC: National Oceanic and Atmospheric Administration.

Chan, H-P., Sahiner, B., Helvie, M.A., Petrick, N., et al. (1999). Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study. *Radiology, 21*, 817–827.

Clinton, W.J. & Gore, A., Jr. (1994). *Science in the national interest.* White House Office of Science and Technology Policy, Washington D. C.

Cronbach, L.J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana IL: University of Illinois Press.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Dawes, R.M.. Faust, D., & Meehl. P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Dawes, R.M., Faust, D., & Meehl, P.E. (1992). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351–367). Mahwah, NJ: Erlbaum.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia PA: Society for Industrial and Applied Mathematics.

General Accounting Office (1997). Social security disability: SSA actions to reduce backlogs and achieve more consistent decisions deserve high priority. *T−HEHS-97-118.* Washington DC.

Getty, D.M., Pickett, R.M., D'Orsi, C.J., & Swets, J.A. (1988). Enhanced interpretation of diagnostic images. *Investigative Radiology, 23*, 240–252

Getty, D.J., Seltzer. S.E., Tempany, C.M.C., Pickett, R.M., Swets, J.A., & McNeil, B.J. (1997). Prostate cancer: Relative effects of demographic, clinical, histologic, and MR imaging variables on the accuracy of staging. *Radiology, 204*, 471–479.

Getty, D.M. & Swets, J.A. (1991). Determining the diagnostic features worthy of attention in a new imaging modality. *Fechner Day 91*, International Society of Psychophysics, Durham, NC, 45–47.

Getty, D.J., Swets. J.A., Pickett, R.M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1*, 19–33.

Gigerenzer, G., Todd, P.M. et al. (1999). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Gish, H. (1990). A probabilistic approach to the understanding and training of neural network classifiers. IEEE International Conference on Acoustics, Speech, and Signal Processing. Albuquerque, NM, April 3–6, 1361–1368.

Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association, 81*, 108–113.

Gottfredson, D., Wilkins, L., & Hoffman, P. (1978). Guidelines for parole and sentencing: A policy control method. Lexington, MA: Lexington Books.

Grove, W. & Meehl, P. (1996). Comparative efficacy of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.

Hanley, J.A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making, 8*, 197–203.

Hanson, R. (1998). What do we know about sex offender risk assessment? *Psychology, Public Policy, and Law, 4*, 50–72.

Harris, D.H. (1991, October). *Eddy current steam generator data analysis performance.* Paper presented at the ASME International Joint Power Generation Conference, San Diego, CA.

Harris, G.T., Rice, M.E., & Quinsey, V.L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20*, 315–335.

Hemphill, J., Hare, R., & Wong, S. (1998). Psychopathy and recidivism: A review. *Legal and Criminological Psychology, 3*, 139–170.

Hertz, J., Krogh, A., & Palmer, R.G. (1991). *Introduction to the theory of neural computation.* Reading, MA: Addison-Wesley.

Hidaka, T. & Beranek, L. L. (2000). Objective and subjective evaluations of twenty-three opera houses in Europe, Japan, and the Americas. *Journal of the Acoustical Society of America, 107*, 368–383.

Hosmer, D.W. & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Jiang, Y., Nishikawa, R.M., Wolverton, D.E., Metz, C.E., Giger, M.L., Schmidt, R.A., Vyborny, C.J., & Doi, K. (1996). Malignant and benign clustered microcalcifications: Automated feature analysis and classification. *Radiology, 198*, 671–678.

Kassirer, J.P. (1994). A report card on computer-assisted diagnosis. *The New England Journal of Medicine, 330*, 1824–1825.

Kopans, D. & D'Orsi, C.J. (1992). ACR system enhances mammogram reporting. *Diagnostic Imaging, Sept.*, 125–132.

Kopans, D. & D'Orsi, C.J. (1993). Mammographic feature analysis. *Seminars in Roentgenology, 28*, 204–230.

Klassen, D. & O'Connor, W. (1990). Assessing the risk of violence in released mental patients: A cross validation study. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*, 75 –81.

Lachenbruch, P. (1975). *Discriminant analysis.* New York:Hafner.

Lidz, C., Mulvey, E., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of the American Medical Association, 269*, 1007–1011.

Lo, J.Y., Baker, J.A., Kornguth, P.J., Iglehart, J.D., & Floyd, C.E., Jr., (1997). Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features. *Radiology, 203*, 159–163.

Lodwick, G.S. (1986). Computers in radiologic diagnosis. *Applied Radiology, Jan/Feb*, 61–65.

Lusted, L.B. (1968). *Introduction to medical decision making.* Springfield IL: Charles C. Thomas.

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota.

Meyer, K.B. & Pauker, S.G. (1987). Screening for HIV: Can we afford the false-positive rate? *The New England Journal of Medicine, 317*, 238–241.

Miller, A. & Thompson, J.C. (1975). *Elements of meteorology*. Columbus, OH: Merrill.

Mulley, A.G. & Barry, M.J. (1986). *Clinical utility of tests for HTLV-III/LAV infection.* (Grant application to the Public Health Service.) Boston, MA: Massachusetts General Hospital.

Nishanian, P., Taylor, J.M.G., Korns, E., Detels, R., Saah, A., & Fahey, J.L. (1987). Significance of quantitative enzyme-linked immunosorbent assay (ELIZA) results in evaluation of three ELIZAs and Western blot tests for detection of antibodies to human immunodeficiency virus in a high-risk population. *Journal of Clinical Microbiology, 25*, 395–400.

Partin, A.W., Kattan, M.W., Subong, E.N.P., Walsh, P.C., Wojno, K.J., Oesterling, J.E., Scardino, P.T., & Pearson, J.D. (1997). Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. *Journal of the American Medical Association, 277*, 1445–1451.

Passell, P. (1990). Wine equation puts some noses out of joint. *New York Times, March 4*, p. 1.

Peterson, W.W., Birdsall, T.G., & Fox, W.C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory PGIT−4*, 171–212.

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Rau, C.A., Jr. (1979). Proceedings of the ARPA/AFMI, *Review of Progress in Quantitative NDE, 150–161.*

Rice, M.E. & Harris, G.T. (1995). Violent recidivism: assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737–748.

Richard, M.D. & Lippmann, R.P. (1991). Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Computation, 3*, 461–483.

Rummel, W.D., Christner, B.K., & Long, D.L. (1988). Methodology for analysis and characterization of nondestructive inspection capability. *Review of Progress in Quantitative NDE, 7*, 1769–1776.

Schwartz, J.S., Dans, P.E., & Kinosian, B.P. (1988). Human immunodeficiency virus test evaluation, performance, and use: Proposals to make good tests better. *Journal of the American Medical Association, 259*, 2574–2579.

Seltzer, S.E., Getty, D.J., Tempany, C.M.C., Pickett, R.M., Schnall, M.D., McNeil, B.J., & Swets, J.A. (1997). Staging prostate cancer with MR imaging: A combined radiologist−computer system. *Radiology, 202*, 219–226.

Shiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). *Introduction to multidimensional scaling.* New York: Academic Press..

Social Security Advisory Board (1998). How SSA's disability programs can be improved. Washington DC.

Speer, J.R. (1989). Detection of plastic explosives. *Science, 243*, 1651.

Steadman, H., Silver, E., Monahan, J., Appelbaum, P., Robbins, P., Mulvey, E., Grisso, T., Roth, L., & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*, 83–100.

Sweeting, T. (1995). Statistical models for nondestructive testing. *International Statistical Review, 63*, 199–214.

Swets, J.A. (1983). Assessment of NDT systems: Part 1. The relationship of true and false detections; Part 2. Indices of performance. *Materials Evaluation, 41*, 1294–1303.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.

Swets, J.A. (1991). The science of high stakes decision making in an uncertain world. *Science and public policy seminar*, Federation of Behavioral, Psychological, and Cognitive Sciences, Rayburn House Office Building, Washington, DC (September 6).

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.

Swets, J.A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Mahwah, NJ: Erlbaum.

Swets, J.A., Getty, D.J., Pickett, R.M., D'Orsi, C.J., Seltzer, S.E., & McNeil, B.J. (1991). Enhancing and evaluating diagnostic accuracy. *Medical Decision Making, 11*, 9–18.

Swets, J.A. (1998). Increasing the accuracy of mammogram interpretation. *Final report under Contract DAMD17-94-C-4082.* U.S. Army Medical Research and Materiel Command, Fort Detrick, Frederick, MD.

Tsien, C.L., Fraser, H.S., Long, W.J., & Kennedy, R.L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infection. *Medinfo, 9*, 493–497.

Webster, C., Harris, G., Rice, M., Cormier, C., & Quinsey, V. (1994). *The violence prediction scheme; Assessing dangerousness in high risk men.* Centre of Criminology, University of Toronto, Canada.

Weinstein, M.C., Fineberg, H.V., Elstein, A.S., Frazier, H.S., Neuhauser, D., Neutra, R.R., & McNeil, B.J. (1980). *Clinical decision analysis.* Philadelphia PA: W.B. Saunders.

Weiss, R. & Their, S.O. (1988). HIV testing is the answer—What's the question? *The New England Journal of Medicine, 319*, 1010–1012.

Young, F.W.; R.M. Hamer (Ed.) (1987). *Multidimensional scaling: History, theory, and applications.* Mahwah, NJ: Erlbaum.

Yu, K.K., Hricak, H., Alagappan, R., Chernoff, D.M., Bachetti, P., & Zaloudeck, C.J.(1997). Detection of extracapsular extension of prostate carcinoma with endorectal and phased−array coil MR imaging: Multivariate feature analysis. *Radiology, 202*, 697–702.

APPENDIX: SOME CONCEPTS OF PROBABILITY

Though they are not all used formally in the body of this article, certain concepts of probability are fundamental to the ideas and analyses presented there. These concepts are not necessary to appreciate the gist of the article, but a brief review of them may promote a more sophisticated view. As it happens, the same concepts are quite generally useful in human affairs. Seemingly intricate or technical at first glance, and indeed they are often misunderstood and confused, they can be seen to be straightforward and likely to repay some attention.

Recall our focus on the two elements of a diagnostic task: (1) the presence or absence of a condition of interest, and (2) a decision that the condition is or is not present. We spoke of the actual presence or not of the condition as the "truth" about the condition and designated the two truth states as T+ (condition present) and T− (condition absent). Similarly, we designated the positive and negative decision as D+ and D−.

Joint probabilities

We wish to make probability statements about two ways in which T and D values may combine. One is the co-occurrence, or joint occurrence, of a T value and a D value (say, the values T− and D+, which together represent a false-positive outcome). An expression such as P(T− & D+) denotes a joint probability; the other three possible coincidences of T and D values also have associated joint probabilities. In words, one speaks, say, of the probability of a cancer being absent and the diagnosis being positive.

Conditional probabilities

The second way in which T and D values may combine is in a conditional relationship. We may ask about the probability of D+ occurring conditional on, or given, the occurrence of T−. Here, the notation is P(D+ | T−). In this example, note that we are conditioning D on T. That is, the quantity of interest in this example is the probability of a positive cancer diagnosis given the actual absence of cancer. Another possibility is to condition in the other direction, on the decision D. For example, P(T+ | D+) expresses the probability of there being cancer in truth given that the decision made is positive for cancer. It may be noted that the direction from decision to truth gives the probability that usually interests the patient and the doctor; it represents what is termed the "predictive value" of the diagnosis. For some purposes, it interests evaluators of diagnostic performance. However, probabilities proceeding from truth to diagnosis are of principal utility in the present context: As seen in our discussion of ROC analysis, they are the basis for valid measures of diagnostic accuracy and the diagnostic decision threshold.

Strictly, one should be careful to qualify any probability referring to combinations of T and D as either a joint or conditional probability. In the present context, we focus predominantly on conditional probabilities and have not carried the qualifier along when no confusion is likely. Also, it is particularly important to be clear about the direction of a conditional probability: from truth to decision or the reverse. Confusion in this respect is widespread and plagues communication about probabilities in diagnostics. Still, here, where we are primarily concerned about probabilities of decisions given the truth, we also drop that qualifier when permissible. So here, a "false-positive probability," for example, is a conditional probability and it is P(D+ | T−), or conditioned on the truth.

Prior probabilities

The third and final kind of probability required here is the prior probability of the condition or event of interest, "prior" to a decision, which is denoted either P(T+) or P(T−). One must know, for example, whether the probability of breast cancer in a given diagnostic setting is relatively low, .03, as it might be in broad-based mammography screening, or relatively high, .33, as it might be in a referral hospital for symptomatic cases. As discussed, this variable affects the decision threshold that is selected for making a positive diagnosis. So there is a need to know the prior probability of a condition or event: of cancer, violence, or severe weather, say, in any population or locale under study.

Relation among the three probabilities

The joint, conditional, and prior probabilities are simply related; the joint probability is the product of the other two. Considering just positive quantities, for example: $P(T+ \& D+) = P(D+ \mid T+) \times P(T+)$.

Calculation of probabilities from a frequency table

The computation of the three types of probabilities is based on frequency data in a two-by-two contingency table. Such a table has two columns, headed by T+ and T−, and two rows, labeled D+ and D−, as shown in Table A-I. The frequencies of cases that fall in each of the four cells are denoted *a*, *b*, *c*, and *d*, respectively. Thus, *a* is the number of cases for which T+ and D+ co-occur, and so forth. The margins give cell totals, e.g., $a+c$ is the total number of times T+ occurs, whether associated with a positive decision (*a*) or a negative decision (*c*). Likewise, $a+b$ is the number of times that the positive decision D+ is made, whether to T+ or T−. The total sample size is $N = a+b+c+d$. Various proportions calculated from a sample's table are taken as estimates of corresponding probabilities in the population from which the sample is drawn.

Note that dividing the column sums by the sample size gives the proportions of times that T+ and T− occur–which are taken

Table A-I. *Two-by-two table of truth and decision:* a, b, c, d *are the frequencies of the four possible decision outcomes. The important proportions, or probabilities, are defined.*

		Truth		
		Positive	*Negative*	
Decision	*Positive*	*a* True positive	*b* False positive	*a + b*
	Negative	*c* False negative	*d* True negative	*c + d*
		a + c	*b + d*	*a + b + c + d =N*

Prior probabilities: *Joint probabilities:*
$(a + c)/N = P(T+)$ $a/N = P(T+ \& D+)$ $c/N = P(T+ \& D-)$
$(b + d)/N = P(T-)$ $b/N = P(T- \& D+)$ $d/N = P(T- \& D-)$

Conditional probabilities (of decision conditional on truth):
$a/(a + c) = P(D+ \mid T+)$ $c/(a + c) = P(D- \mid T+)$
$b/(b + d) = P(D- \mid T-)$ $d/(b + d) = P(D- \mid T-)$

as the *prior* probabilities, respectively, of T+ and T−. Specifically, the proportion *(a+c) / (a+b+c+d)* is equal to P(T+) and the proportion *(b+d) / (a+b+c+d)* is equal to P(T−).

Dividing a cell frequency by a column or row total gives a *conditional* probability. For example, the proportion *(a) / (a+c)* is the conditional probability of a true-positive decision conditioned on T+, namely, P(D+ | T+). The proportion *(a) / (a+b)* is the true-positive probability conditioned on D+, namely, P(T+ | D+).

Lastly, dividing a cell frequency by the sample size gives a *joint* probability. So, e.g., *(d) / (a+b+c+d)* is the joint probability of a true-negative outcome, P(T− & D−).

Two basic probabilities

Two conditional probabilities based on the frequencies in Table A-1 suffice to provide all of the information contained in the four conditional probabilities just described. Provided they are truth-conditional probabilities, two will do, because the other two are their complements. That is, *a / (a+c)* and *c / (a+c)*—namely, the two proportions derived from the left column—add to 1.0 (when T+ holds, the decision is either D+ or D−). Similarly, their probabilities, P(D+ | T+) and P(D− | T+), add to one. Also, the two conditional probabilities of the right column are complements. The two probabilities often used to summarize the data are the true-positive and false-positive probabilities: *a / (a+c)* or P(D+ | T+) and *b / (b+d)* or P(D+ | T−). These are the two conditional probabilities of a positive decision, given T+ or T−. Their notation may be simplified as P(TP) and P(FP).

These two probabilities are independent of the prior probabilities (by virtue of using the priors in the denominators of their defining ratios). The significance of this fact is that ROC measures do not depend on the proportions of positive and negative instances in any test sample, and hence, generalize across samples made up of different proportions. All other existing measures of accuracy vary with the test sample's proportions and are specific to the proportions of the sample from which they are taken.

Inverse probabilities and Bayes' theorem

We mentioned in passing conditional probabilities that are conditioned on the decision rather than the truth, so called "inverse" probabilities, and relegated them to secondary interest. However, the concept of inverse probabilities is centrally important to our developments when applied not to the decision, but to the data or evidence that underlie the decision. Whereas the construction of a SPR is based on probabilities of items of information (data, symptoms, pieces of evidence) that are dependent (conditional) upon known positive and negative instances of truth, the use of the SPR as a decision aid is based on the inverse probability: the probability of the positive truth state given the (collective) data. It is this latter probability that the SPR supplies for diagnosis and forms the continuum of evidence along which a decision threshold is set to permit a binary, positive or negative, decision.

Inverse probabilities are often called "Bayesian" probabilities because they may be calculated by means of the clergyman Thomas Bayes' (1763) theorem from the truth-conditional probabilities along with the prior probabilities. Specifically, using the symbol "e" to denote the evidence for a decision, the theorem (stated here for just the positive alternative) is:

$$P(T+ \mid e) = \frac{P(e \mid T+) \times P(T+)}{P(e)},$$

where $P(e) = [P(e \mid T+) \times P(T+)] + [P(e \mid T-) \times P(T-)]$, that is, the sum of the values of P(e) under the two possible alternatives.

The theorem illustrates that the quantity produced for the decision maker by a SPR incorporates the prior probability. Though that fact is sometimes forgotten, the decision maker should be consistently aware of it and resist the tendency to make a further adjustment for the prior probability, or base rate, that characterizes the situation at hand.