[redacted] Rehge

FMI-1000-766

7th L.H. Gray Conf., Leeds, England, 13-15 April 1976. To be publ. in proc. by The Inst. of Physics, Bristol. no figs (only)

Quantitative Evaluation of Visual Detection Performance in Medicine: ROC Analysis and Determination of Diagnostic Benefit

CONF-760452--1

Charles E. Metz, Ph.D.

Stuart J. Starr, B.S.

Lee B. Lusted, M.D.

Center for Radiologic Image Research,

Department of Radiology,

The University of Chicago,

and

The Franklin McLean Memorial Research Institute,*

Chicago, Illinois, U.S.A.

*Operated by the University of Chicago for the U. S. Energy Research and Development Administration under Contract E(11-1)-69.

Presented at the 7th L. H. Gray Conference, "Medical Images: Formation, Perception and Measurement," held at Leeds, England, 13-15 April, 1976. To be published in the proceedings by The Institute of Physics, Bristol.

ABSTRACT

An ROC curve provides an empirical description of the trade-offs which are possible among the various types of correct and incorrect decisions as the human decision-maker varies one or more "confidence thresholds." Conventional ROC curves measured in simple decision-making situations can, in some cases, be used to predict human decision performance in more complex situations.

By considering both the consequences of the various types of diagnostic decisions and the overhead cost of a diagnostic study, one can use the ROC curve to evaluate the diagnostic usefulness of a study in any particular clinical context. Since the ROC curve describes the possible relationships among the probabilities of the various types of correct and incorrect decisions, it plays a central role in optimizing diagnostic strategies using the general techniques of decision analysis.

1. INTRODUCTION

With the introduction of new diagnostic imaging techniques in recent years, the problem of evaluating the relative usefulness of these new procedures in a meaningful way has become increasingly important. The question is often asked: Which imaging technique is best? This paper will attempt to describe an approach which can, in principle, provide an objective and quantitative answer to that kind of question.

The approach to be described here can be applied to both highly specific and also more general problems of evaluating diagnostic image quality. In order to answer a question regarding which of several imaging procedures is best, however, the question must be carefully phrased. What criteria are to be used to measure "best"? Best at doing what? Best for what human observer? Best in what situation or situations? The answer to a question regarding diagnostic image quality cannot be more precise than the question itself.

An important aspect of diagnostic medicine consists of making decisions regarding the actual health or disease state of the patient. On the basis of the information available to him, the diagnostic physician must decide whether the various attributes of a particular disease state are present or absent, must decide whether or not to request additional diagnostic tests, and must finally decide whether a particular disease state (or which one of several) is present in a particular patient.

The approach to diagnostic image evaluation described here is based upon measurement of the quality of the decisions which a physician is able

to make when using a diagnostic imaging technique in a given situation. The approach is divided into two parts: (1) measurement of the relationships among the relative frequencies of the various types of correct and incorrect decisions which the physician can achieve using the imaging technique, and (2) evaluation of the diagnostic benefit which can be gained from those possible combinations of decision frequencies. Principles of signal detection theory are used to guide the approach, to predict relationships among descriptions of decision performance in various situations, and to suggest optimal decision-making strategies. Essentially, however, the approach is empirical.

In the following sections we will review briefly the basic principles of Receiver Operating Characteristic (ROC) curve analysis; we will suggest ways in which the conventional ROC approach can be generalized to describe complex decision tasks; we will indicate how various measures of diagnostic quality can be derived from knowledge of the ROC curve; and we will point out applications of ROC analysis in a general decision-analytic approach to the diagnostic problem.

2. CONVENTIONAL ROC METHODS

To begin, consider the situation in which the diagnostic decision-maker must choose between two alternatives: for example, a particular disease may be present or absent, a lesion may be present or absent, or a disease may be of type A or type B. Two classes of "true states" exist which we can, for now, call "actually positive" and "actually negative"; and two decisions are possible, namely, a "positive" decision (i.e.,

the decision that the true state is "actually positive") and a "negative" decision (i.e., the decision that the true state is "actually negative").

Either decision may be correct or incorrect. Four types of decisions are therefore possible in this situation: a "true positive" decision (i.e., a "positive" decision for an "actually positive" true state); a "false positive" decision (i.e., a "positive" decision for an "actually negative" true state); a "true negative" decision (i.e., a "negative" decision for an "actually negative" true state); and a "false negative" decision (i.e., a "negative" decision for an "actually positive" true state). All four types of decisions will occur in almost any decision-making situation in the presence of uncertainty, and hence all four types of decisions must be taken into account when measuring (and attempting to specify) decision performance.

Suppose that a physician views a series of images (e.g., radiographs or scintigrams) which are similar except that some of the images are "actually negative" (e.g., are images of normal subjects), while the rest are "actually positive" (e.g., are images of subjects having a certain type of lesion or disease state). Because of statistical fluctuations in the image data (e.g., due to quantum statistics) or in normal anatomical structure or function (i.e., "structured noise" [Revesz, Kundel, and Graber, 1974]), because of neural and/or psychophysical fluctuations internal to the observer [Wickelgren, 1968; Tanner and Swets, 1954; Tanner and Birdsall, 1958; Green and Swets, 1966; Tanner and Norman, 1964; Tanner and Leshowitz, 1969; Goodenough and Metz, 1975], and perhaps because of statistical variations in the form of the lesion or manifestation of the disease, the presence or absence of the lesion or disease may not always be decided correctly.

2.a. The Concept of Confidence Threshold

It is a fundamental assumption of our approach that the observer decides whether or not a lesion (for example) is present by comparing his impression of the image with some "decision criterion" or "confidence threshold" and by stating that a lesion is present if and only if his confidence that the lesion exists in the image exceeds that threshold. The method by which the observer determines his confidence that the lesion is present is of no concern in the application of our approach, although the method used by the observer may influence his ability to detect the lesion. Training and experience would seem to be important factors in determining the skill with which the observer can utilize the image data to form an appropriate estimate of his confidence that a lesion or a disease state is present.

The relative frequencies of the possible correct ("true positive" and "true negative") and incorrect ("false positive" and "false negative") decisions will depend upon the confidence threshold adopted by the observer. We will discuss later the question of what confidence threshold the observer should adopt; let it suffice here to say that the degree of apparent "positiveness" which the observer requires before he is willing to make a "positive" decision should depend both upon the clinical consequences of the four types of correct and incorrect decisions for the case at hand and also upon the frequency of actually positive cases.

Since the underlying detectability of the lesion by the observer does not depend upon the confidence threshold chosen by the observer, however, any useful and complete description of the detectability of lesions (or disease

etc.) by a human observer in the presence of uncertainty must take into account the concept of a variable decision threshold.

As a case in point, if an observer is required to detect lesions in two series of images--of the same body part but produced by two different imaging systems, for example--then it may be found that the observer's decisions based upon one series of images yield both true-positive and false-positive decision frequencies which are greater than those elicited by the other series of images. In order to decide whether the differences in decision frequencies are due to different inherent lesion detectabilities provided by the two imaging systems, or whether they are due only to the (perhaps unconscious) adoption of different confidence thresholds by the observer, one needs a method of observer performance analysis which separates these two factors influencing the relative frequencies of the four types of decisions.

2.b. The Conventional ROC Curve

The combination of decision frequencies produced by an observer attempting to detect a given image feature or disease state using a given imaging system depends upon the confidence threshold adopted by the observer; the appropriate confidence threshold depends on the clinical context of the detection problem; and furthermore the confidence threshold which an observer may adopt is difficult to quantitate. For these reasons, it is useful to describe detection performance by the relationship among the various possible decision frequencies which are generated as different confidence thresholds are adopted by the observer, with the values of confidence threshold implicit, rather than explicit, in the analysis. Receiver Operating Characteristic (ROC) analysis does just that.

It is convenient to establish at this point a convention for notation which will facilitate our subsequent discussion. Let s and n represent the two true states "signal actually present" (i.e., "actually positive") and "noise only actually present" (i.e., "actually negative"), respectively. Let S and N represent the decisions "positive" and "negative," respectively. Then the conditional probability $P(S|s)$, for example, represents the relative frequency with which actually positive images are decided to be positive in a large number of trials, and $P(S|n)$ represents the relative frequency with which actually negative images are called positive. Similarly, $P(N|s)$ represents the relative frequency with which actually positive images are called negative, and $P(N|n)$ represents the relative frequency with which actually negative images are called negative. The four conditional probabilities described above are sometimes called the "conditional true positive probability," the "conditional false positive probability," the "conditional false negative probability," and the "conditional true negative probability," respectively. One should note that since all actually positive images must be called either positive or negative, $P(S|s) + P(N|s) = 1$, and similarly, $P(S|n) + P(N|n) = 1$. Thus specification of a combination of conditional true positive and false positive probabilities, for example, in fact specifies the conditional probabilities, and hence relative frequencies, of all four types of decisions.

A conventional ROC curve specifies the various trade-offs which are possible among the frequencies of the four types of correct and incorrect decisions as the confidence threshold is varied in a two-alternative detection

experiment by showing the various possible relationships between P(S|s), the conditional true positive probability, and P(S|n), the conditional false positive probability. Specifically, the conventional ROC curve is a graph of P(S|s) versus P(S|n).

A typical conventional ROC curve which might result from a two-alternative visual detection experiment is shown in figure 1. If the observer adopts some "moderate" confidence threshold in deciding whether to call each image "positive" or "negative," the resulting combination of conditional true and false positive decision probabilities might plot as point A. If the observer were to adopt a "strict" confidence threshold, on the other hand, in the sense that a suspected lesion must appear very convincing in order for the image to be called "positive," then false positive decisions would be less frequent, but true positive decisions would be less frequent also (i.e., false negative decisions would be more frequent, and more lesions would be "missed"). Hence, in this case, the resulting conditional frequencies might plot as point B. As a third possibility, if the observer were to adopt a rather "lax" confidence threshold, in the sense that any suggestion of a lesion-like structure in an image would cause the image to be called "positive," then true positive decisions would be more frequent than in the first case, but false positive decisions would be more frequent also. In this situation, the combination of conditional decision probabilities might plot as point C. As a continuum of various possible confidence thresholds is considered, the dotted curve is swept out. This is the conventional ROC curve, which describes the possible relationships between

conditional true and false postive decision probabilities and hence among the relative frequencies of the four types of correct and incorrect decisions as confidence threshold is varied.

ROC curve analysis was first developed for the evaluation of radar signal detectability and was later used to evaluate human detection performance in psychophysics [Green and Swets, 1966; Pastore and Scheirer, 1974]. More recently, ROC techniques have been applied to the problems of evaluating diagnostic medical decision-making [Lusted, 1968, 1971, 1975; Alcorn and O'Donnell, 1969; Morgan et al., 1973; McNeil, Varady, Burrows, and Adelstein, 1975; McNeil et al., 1976] and of evaluating the detectability provided by diagnostic medical imaging systems [Revesz and Kundel, 1971; Goodenough, 1972; Goodenough, Rossmann, and Lusted, 1972, 1973, 1974; Anderson et al., 1973; Andrus, Hunter, and Bird, 1975]. Introductory discussions of ROC analysis have been published by Swets [1973], Metz and Goodenough [1975], McNeil, Keeler, and Adelstein [1975], Metz, Starr, Lusted, and Rossmann [1975], and Swets [1976], for example.

2.c. Context-independent Measures of Decision Performance Related to the Conventional ROC Curve

Various measures of detection or decision performance have been proposed which are independent of the practical situation to which the decisions are applied and which can be related to the ROC curve.

Many investigators have used the index d' [Green and Swets, 1966, p. 60] to describe observer detection performance. This index is applicable only to detection situations which generate ROC curves of a particular form,

however, and can be misleading if applied incorrectly [Goodenough, Metz, and Lusted, 1973], especially in situations for which the conventional ROC curve itself is inappropriate [Metz and Goodenough, 1973].

The area under the conventional ROC curve can be shown to equal the fraction of correct responses that would result from an experiment in which the observer considers two images (or sets of data) simultaneously, one of which contains a signal (or is due to a disease state), and must choose the image contiaining the signal (or the data set associated with the disease state) [Green and Swets, 1966, p. 47]. Thus the area under an ROC curve provides a measure of discrimination ability.

It is an empirical fact that most experimentally-determined conventional ROC curves plot as almost straight lines on double-probability paper (i.e., on a graph with axes which are linear with respect to z, the inverse of the cumulative normal deviate probability distribution [Green and Swets, 1966, p. 61]). To the extent that this is approximately the case, the conventional ROC curve--and hence decision performance--can be described by a pair of numbers representing the slope and either one axis intercept or the distance from the origin to the ROC line along the negative diagonal on such a plot [Swets, 1976].

Other possible context-free measures of decision performance have been discussed by Swets [1976].

2.d. Practical Considerations

Although the ROC curve is usually regarded as describing the locus of combinations of conditional true and false positive decision frequencies which

are generated as the observer adopts various confidence thresholds, the curve is most often determined experimentally by requiring the observer to use, in effect, several thresholds simultaneously and to state into which one of several categories of confidence his impression of the image data (for example) falls. This so-called "rating method" [Green and Swets, 1966, p. 99; Goodenough, Rossmann, and Lusted, 1974; Metz and Goodenough, 1975] thus yields several points on the ROC curve from a single series of observations, through which a smooth curve can be drawn. It has been shown that ROC curves measured by the "yes-no" and "rating" methods are equivalent [Egan, Schulman and Greenberg, 1959].

A fundamental requirement for use of the ROC approach is that the true state must be known for each trial (e.g., each image). Although this requirement is sometimes tedious to satisfy in clinical applications of the approach, demanding careful long-term follow-up of cases, the requirement seems unavoidable in any objective approach to diagnostic evaluation. The quality of diagnostic decisions cannot be determined if the correct answers are not known.

In the design of observer performance experiments to measure the relative detectability provided by alternative imaging systems, due consideration must be given to the control of experimental variables. Aside from the perhaps obvious variables of display and viewing conditions, prior information, and observer experience, the characteristics of the population of images to be viewed, for example, must be taken into account, because the conclusions to be drawn from the experiment are applicable only to, and hence cannot be more specific than, the population of cases studied. To suggest a hypothetical and possibly extreme example, suppose that detection of liver carcinoma by two different scintigraphic imaging systems is to be evaluated, and suppose that in

fact one system detects diffuse metastases better while the other system detects solitary nodules better. If proven cases having diffuse metastases were pooled both with cases having solitary lesions and also with normal cases to form a series of cases for a detection experiment, the results might suggest that the detectability of liver carcinoma provided by the two systems is about equal. This would be a thoroughly valid conclusion for the population of cases used in the experiment, but would neglect the fact that the systems are better or worse for different types of disease. Thus this conclusion could be misleading if applied to populations heavily skewed to one type of disease or the other, and would fail to recognize that the imaging systems would not be of equivalent quality if one or the other form of disease could be ruled out by other tests.

Some parameters which should be considered in the design of an observer performance experiment have been listed by Metz and Goodenough [1975]. Other considerations of experimental design have been discussed by Green and Swets [1966, Appendix III] and by Kundel and Revesz [1974].

2.e. Conditions Under Which the Conventional ROC Approach is Valid

As we have mentioned, the conventional ROC approach can be applied to evaluate detection or decision performance in any situation for which true states can be divided into two classes and for which one of two decisions must be selected by the observer or decision-maker.

One should note carefully the distinction between, first, the use of the "rating method" to elicit confidence ratings in such "simple" detection or decision tasks and, second, more complex detection or decision tasks in which

multiple true states and decision alternatives are considered. In the former case, two true states are possible, "positive" or "negative," for example, and the decision-maker may be required to rate his confidence that one of the two true states exists as opposed to the other. Thus, although various categories of confidence may be elicited as responses from the decision-maker, these responses are directed toward selection of one or the other of two decisions: "positive" or "negative." If the observer is required to detect the presence of a lesion and to state in which quadrant of the image it is present, however, then five true states exist (no lesion, lesion in quadrant 1, lesion in quadrant 2, etc.) and five different decisions are possible, each of which may have associated with it some degree of confidence. Similarly, if the observer is required to detect a lesion and decide, if present, whether it is a lesion of type A or B, then three true states exist (no lesion, lesion type A, lesion type B) and three decisions are possible, each having some degree of confidence associated with it.

Although the conventional ROC approach cannot be applied directly to such complex detection or decision tasks, the fundamental principles of the approach can be generalized in a straightforward way to describe, and in some cases predict, decision performance in such situations.

3. GENERALIZED ROC METHODS

Fundamentally, the ROC approach describes the relationships or trade-offs which are possible among various types of correct and incorrect decisions as a confidence threshold (or perhaps several confidence thresholds) is varied. Viewing the ROC approach in this broad way, one can generalize the basic techniques

of the approach and extend them to analyze decision performance in complex detection or decision tasks relevant to clinical diagnostic situations.

3.a. Detection and Localization

Conventionally, ROC analysis is applicable to visual detection situations in which zero or one visual signal (e.g., lesion or disease) is present somewhere in the visual field and in which the observer is required only to state whether the signal is present or absent. If the signal is present and the observer states "Signal present," he is credited with a true positive response. In this situation the observer may, occasionally, receive credit for a correct positive response when, in fact, he misses the signal but falsely identifies a cluster of noise or normal background structure elsewhere in the image as due to the presence of the signal. Hence true positive decision frequency may be expected to decrease if the observer is required to determine correctly the position of the signal as well as its presence in order to receive credit for a correct positive decision.

If this more strictly defined "true positive, correct location" decision probability were plotted versus false positive decision probability, the resulting curve would be expected to lie below the conventional ROC curve. We call this new curve an "LROC" curve to indicate that it is analogous to an ROC curve, but with the additional requirement of correct localization. The position of the LROC curve depends upon the precision with which the observer is required to locate the signal.

We have developed a simple theoretical relationship which can be used to predict the LROC curve from knowledge of the conventional ROC curve under rather general conditions. Details of the derivation and experimental verification of

the predictions for the case of quadrant localization are reported elsewhere [Starr, Metz, Lusted, and Goodenough, 1975].

Since the sum of the conditional probabilities of a positive response with correct localization and of a positive response with incorrect localization when a signal is actually present in a detection and localization experiment must equal the conditional true positive decision probability in a simple detection experiment, the "generalized" ROC curve for a detection and localization experiment can be represented by a line in three-dimensional space, with the three axes taken to be the first two conditional probabilities mentioned above and also the conditional false positive decision probability. The LROC curve can be thought of as the projection of this generalized ROC curve onto the plane defined by the first and third axes.

3.b. Multiple Signal Detection

Some diagnostic medical imaging procedures, such as gallbladder radiography in search of gallstones and brain scintigraphy in search of metastases, may require determination of the number of visual signals or lesions present. Since conventional ROC curve analysis describes observer detection performance in situations where only two decision alternatives are available, and since conventional ROC curves are usually measured in situations where zero or one signal is known to be present, it would be both practically and conceptually useful to establish a relationship between the conventional ROC curve and a description of observer performance in the situation where more than one signal may be present.

We have derived theoretical relationships which can be used to predict human observer performance in a multiple signal detection experiment from knowledge of the conventional ROC curve. Experiments designed to test these preditions for the case in which zero, one, or two signals may be present appear to verify the theoretical predictions [Metz, Starr, and Lusted, 1976]. Since nine types of correct and incorrect decisions are possible for this particular case, since three constraint equations exist because probabilities conditional on the same true state must add to unity, and since determination of a single conditional decision probability determines the other eight according to our theory, the generalized ROC curve for this "0, 1, or 2" signal detection experiment can be represented by a line in six-dimensional space or, equivalently, by a set of five two-dimensional graphs.

As in the case of the detection and localization experiments, this demonstrated ability to predict observer performance in a complex detection task from knowledge of the conventional ROC curve determined in a simple detection task is reassuring and suggests not only that signal detection theory can be usefully applied to the problem of analyzing human observer behavior, but also that the conventional ROC curve provides a description of human observer detection performance which is applicable in situations clinically more relevant than those in which it is measured.

3.c. Other Types of Generalized ROC Curves

"Generalized ROC curves" describing the relationships among the relative frequencies of the various types of decisions in almost any diagnostic decison-making experiment can be imagined and can, in theory, be determined experimentally.

In some situations, however, the resulting relationships can be complex, difficult to measure to acceptable statistical accuracy without large numbers of trials, and difficult to interpret if an underlying theoretical structure is not available.

Consider, for example, analysis of the "detection and recognition" situation in which an observer must decide whether an image contains a lesion of type "a," a lesion of type "b," or no lesion at all. Let "a," "b," and "n" represent the tree true states and let "A," "B," and "N" represent the corresponding decisions. Nine types of correct and incorrect decisions are thus possible. Because of the fact that decision probabilities conditional on the same true state must add to unity, three constraint equations exist:

$$P(A|a) + P(B|a) + P(N|a) = 1;$$

$$P(A|b) + P(B|b) + P(N|b) = 1;$$

and

$$P(A|n) + P(B|n) + P(N|n) = 1.$$

Thus any combination of decision frequencies can be represented by a point in six-dimensional space, and observation of the relationships among six types of decision frequencies can always serve as an empirical description of observer performance in this situation.

Since more than one confidence threshold can be varied in this kind of experiment, however, the relationship among these decision frequencies is not necessarily represented by a line in the six-dimensional space. The number of degrees of freedom--and hence the theoretical and experimental complexity--actually necessary to describe human observer performance in a detection and

recognition situation is not yet established. Luce's "Choice Theory" [Luce, 1963] suggests that specification of two conditional decision frequencies should serve to determine the other eight, while a general signal detection theoretic approach [Swets and Birdsall, 1956; Whalen, 1971, p. 140] designed to maximize average decision utility by adoption of a theoretically optimum decision strategy suggests that five conditional decision frequencies must be specified in order to determine the other four.

The practical usefulness of the generalized ROC curve concept for analysis of decision performance in detection and recognition situations--and, indeed, in any situation requiring the use of multiple confidence thresholds--seems to depend upon our ability to discover theoretical points of view, similar to those described above for the detection-and-localization and multiple-signal-detection tasks, which can be used to illuminate fundamental relationships among the various decision frequencies.

4. MEASURES OF DIAGNOSTIC QUALITY DERIVED FROM ROC CURVES

A conventional or generalized ROC curve describes the trade-offs available among the frequencies of the various possible types of correct and incorrect decisions and thus provides a non-parametric description of discrimination ability in the situation used to determine the curve. If conventional ROC curves for detection of the same signal by the same observer are determined experimentally using two different imaging systems, for example and if the curve associated with one system is higher everywhere than the other curve, then one can justifiably conclude that the first system provides greater detectability of the given signal by the given observer.

Taken alone, however, the ROC curve does not answer several fundamentally important questions in applied decision-making situations, such as diagnostic medical imaging. "The ROC curve describes the trade-offs available among the various kinds of decisions, but what is the best trade-off of those available?" "How can the better imaging system be chosen if the system providing somewhat greater detectability is more costly in terms of money, radiation exposure, procedure complications and/or patient discomfort?" The following sections describe various attempts to come to grips with these questions. In each case, the ROC curve plays a key role.

4.a. Accuracy

Accuracy, or the fraction of all decisions which are correct, has been used frequently in the clinical literature as a measure of the "quality" of diagnostic decision-making.

For the simple detection situation in which one of two true states exists and one of two decisions is possible for each trial, the probability of a correct decision is equivalent to the probability of a true positive or a true negative decision. Hence, in the notation defined earlier,

$$\begin{aligned} P(\text{correct}) &= P(\text{true positive decision}) \\ &\quad + P(\text{true negative decision}) \\ &= P(s)P(S|s) + P(n)P(N|n) \\ &= P(s)P(S|s) + [1-P(s)][1-P(S|n)] \\ &= P(s)P(S|s) - [1-P(s)]P(S|n) + [1-P(s)]. \end{aligned}$$

By means of this expression, the probability of a correct decision can be computed for each point on a conventional ROC curve (i.e., for each pair of values

for P(S|s) and P(S|n)), given knowledge of the a priori probability of an actually positive true state, P(s)--that is, the relative frequency of actually positive cases.

Differentiating this expression and setting the result equal to zero, one finds that the probability of a correct decision is maximum when the operating point on the ROC curve is chosen such that the ROC curve slope satisfies the equation

$$\frac{dP(S|s)}{dP(S|n)} = \frac{1-P(s)}{P(s)} ,$$

in which the term on the right represents the inverse of the prior odds of an actually positive case. The maximum accuracy attainable in a given decision-making situation can then be used as a measure of diagnostic quality. One should note that both the optimal operating point on the ROC curve and also the accuracy which can be achieved depend upon the relative frequency of actually positive cases, P(s). By measuring the ROC curve, however, one can determine maximum attainable accuracy for any value of P(s).

The expression for accuracy shown above can be rather easily generalized to include more complex decision tasks. In such cases, the optimal operating point on the appropriate generalized ROC curve can be found, and the resulting maximum possible accuracy can be used as a measure of diagnostic quality.

The accuracy of decisions in a diagnostic situation appears to be of limited usefulness as a meaningful measure of diagnostic quality in most practical medical situations, however, for several reasons.

If the relative frequency of actually positive cases is small, as in diagnostic screening situations, for example, then diagnostic accuracy can be misleadingly high even when the diagnostic value of a test is poor. If only 5% of the cases studied are actually positive, for example, then one can achieve an accuracy of 95% simply by blindly calling all cases negative!

In addition, the accuracy measure ignores the medical fact that the consequences of different types of decisions can be quite different. The diagnostic impact of correctly finding an actually positive case is almost always considerably different from that of correctly diagnosing an actually normal case, for example. Similarly, the diagnostic consequences of a false negative decision are usually quite different from those of a false positive decision.

Finally, the accuracy measure does not account for the costs of a diagnostic study itself, such as financial cost, risk of complications, and radiation exposure.

4.b. Information

A popular definition of a good diagnostic test is that of one which provides a large amount of "information" regarding the true state of health or disease. Since the concept of "information" can be expressed as a reduction in uncertainty, and since uncertainty can be quantified in terms of entropy, the information corresponding to any combination of decision frequencies--and hence any point on a conventional or generalized ROC curve--can be computed using standard information theoretic techniques. Explicit applications of "information" as a measure of decision performance in a simple two-alternative decision task have been considered by Swets, Tanner, and Birdsall [1961] and by Metz, Goodenough, and

Rossmann [1973]. Generalization to more complex decision situations is straightforward. Measurement of the appropriate conventional or generalized ROC curve permits evaluation of the maximum information which can be gained from a given diagnostic test for any set of relative frequencies of the various possible true states.

The principal limitation of the information theoretic approach is that it does not appear to allow either the clinical consequences of the various types of decisions or the cost of the diagnostic test itself to be incorporated into the evaluation scheme.

4.c. Average Cost and Average Net Benefit

If the usefulness of a diagnostic study is to be evaluated in a generally meaningful way, it would seem necessary to consider both the possibility of different values for the consequences of the various types of correct and incorrect diagnostic decisions and also the cost, in the broad sense, of performing the diagnostic study itself. Only in this way does it seem possible to incorporate the factors which a physician must consider when choosing a diagnostic test or making a diagnostic decision and to allow for the fact that the medical information to be gained from a diagnostic study may not justify the expense and risk of performing the study.

A rather straightforward approach which satisfies these requirements can be developed by extending the concept of "expected value" or "average cost" [Green and Swets, 1966, p. 21; Whalen, 1971, p. 130]. The average cost per decision in a decision-making situation can be expressed by

$$\bar{C} = \begin{bmatrix}\text{"overhead"}\\ \text{cost per}\\ \text{decision}\end{bmatrix} + \sum_i \sum_j \begin{bmatrix}\text{cost of the consequences}\\ \text{of decision i when the}\\ \text{actual state is j.}\end{bmatrix}$$

$$\times \begin{bmatrix}\text{probability that decision i}\\ \text{is made and that the actual}\\ \text{state is j.}\end{bmatrix}$$

$$= \bar{C}_o + \sum_i \sum_j C_{ij}\, P(D_i|t_j)\, P(t_j)$$

in which $\bar{C}_o$ is the fixed "overhead" cost which must be invested in order to permit a decision to be made, $\{D_i\}$ is the set of all possible decisions, and $\{t_i\}$ is the set of all true states considered. Benefits of certain decisions can be interpreted as negative costs if desired.

In the simple two-alternative situation, $\{D_i\} = \{S,N\}$ and $\{t_i\} = \{s,n\}$, where S and N represent "positive" and "negative" decisions, respectively, and where s and n represent the corresponding true states. Then for this case the equation above becomes

$$\bar{C} = \bar{C}_o + C_{TP}\, P(S|s)\, P(s) + C_{FN}\, P(N|s)\, P(s)$$

$$+ C_{FP}\, P(S|n)\, P(n) + C_{TN}\, P(N|n)\, P(n)$$

$$= [\bar{C}_o + C_{FN}\, P(s) + C_{TN}\, P(n)] - [C_{FN} - C_{TP}]\, P(S|s)\, P(s)$$

$$+ [C_{FP} - C_{TN}]\, P(S|n)\, P(n)\ ,$$

where C_{TP} = cost of consequences of a true-positive decision; C_{FN} = cost of consequences of a false-negative decision; C_{FP} = cost of consequences of a false-positive decision; C_{TN} = cost of consequences of a true-negative decision; and $\bar{C}_o$ = overhead cost per decision.

This formulation assumes that decision-consequence and overhead costs can be combined linearly. Although this is often true, caution must be exercised in certain situations. In considering mortality risk, for example, one must realize that a person can die but once. The linear approximation is good, however, if such risks are small.

It should be clear that if values can be assigned to the various cost terms in this expression, then for any frequency of actually positive cases, P(s), the average cost per decision, $\bar{C}$, can be calculated for each possible operating point on the conventional ROC curve. Average cost is minimized when the operating point on the ROC curve is chosen such that

$$\frac{dP(S|s)}{dP(S|n)} = \frac{[1-P(s)]}{P(s)} \cdot \frac{[C_{FP}-C_{TN}]}{[C_{FN}-C_{TP}]} .$$

Since conventional ROC curves exhibit a monotonically decreasing slope as P(S|n) increases, this expression shows that the optimal operating point is on the lower left portion of the curve, for example, if actually positive cases are rare [P(s) << 1] or if the cost of the consequences of a false positive decision is relatively great.

The coordinates of the optimal operating point thus defined can be substituted into the previous expression to yield the minimum average cost per decision which is possible in the decision-making situation. This minimum attainable cost can then be used to characterize the "quality" of the diagnostic procedure for the set of costs and for the frequency of actually positive cases considered.

The "average net benefit" derived from a medical diagnostic study can be thought of as the reduction in average cost which is provided by performance of the study.

If no significant prior information is available regarding the state of health or disease in a particular case and if the study is not performed, then one may presume that in many situations no treatment is instituted and thus a "negative" decision is, in effect, made by default.* In this case, the average cost of not performing the diagnostic study is

$$\overline{C}\ (\text{no study}) = C_{FN}\ P(s) + C_{TN}\ P(n)$$

and the average net benefit, $\overline{NB}$, derived from performing the study is given by

$$\overline{NB} = \overline{C}\ (\text{no study}) - \overline{C}\ (\text{study})$$

$$= [C_{FN} - C_{TP}]\ P(s)\ P(S|s) - [C_{FP} - C_{TN}]\ P(n)\ P(S|n) - \overline{C}_o\ .$$

One can easily show that average net benefit is maximized when average cost is minimized.

If significant information regarding the state of health or disease in a particular case <u>is</u> available before the study in question is performed, however,

---

*Exceptions exist, of course. If the decision to be made is that of whether a person has been bitten by a rabid dog, the decision to be made in the absence of information is "positive." In this case, $\overline{C}(\text{no study}) = C_{TP}\ P(s) + C_{FP}\ P(n)$, and modifications required in the subsequent analysis should be obvious. Decision-making in the absence of information pertinent to the particular case in question is equivalent to guessing. One can easily show that the optimal guessing strategy is to choose "N" if $C_{FN}\ P(s) + C_{TN}\ P(n) < C_{TP}\ P(s) + C_{FP}\ P(n)$ and to choose "S" if the converse is true.

then the average net benefit to be gained from the additional test must be evaluated in a more general and somewhat more complicated way. Usually, prior information will yield decisions which are, on the average, more beneficial than those made when no information is available, and the net benefit of the additional test will, logically, depend upon the diagnostic improvement which it provides, less the overhead cost of the additional study. In this situation, one must measure two ROC curves: a curve describing decision performance based upon the prior information alone, and a curve describing decisions based upon both the prior data and also the data derived from the additional test. Then the average net benefit provided by the additional study is given by

$$\overline{NB} = \overline{C}(\text{prior data}) - \overline{C}(\text{study and prior data}),$$

where the two average costs are the smallest possible on the respective ROC curves.

The average net benefit approach seems intuitively meaningful and can be rather easily generalized to include complex decision-making situations. Perhaps the most fundamental problem in application of this approach to analyses of acutal medical detection and decision-making situations is that of choosing appropriate values for the required cost terms. We believe that these difficult value judgments confront the clinician in many decision-making tasks today, however, and the fact that decisions are made routinely implies that the associated value judgments are made routinely also. An important property of the average net benefit approach is that the usefulness of diagnostic performance, as specified by the ROC curve, can be evaluated for any system of cost

judgments. Assignment of values to the possible consequences of the various types of diagnostic decisions seems to be an inescapable responsibility of the physician, the patient, and society, and the fact that the analysis should include these values seems only proper. Despite the inevitable philosophical nature of certain cost assignments, thoughtful analysis of the nature and components of the costs inherent in the various types of decisions seems useful [Metz, Starr, Lusted, and Rossmann, 1975]. The concepts of "multi-attribute utility theory" [Edwards, Guttentag, and Snapper, 1975] should provide a useful approach to the problem of combining the various cost components, such as life expectancy, freedom from pain, risk of complications, and financial expense.

5. ROC CURVES AND DECISION ANALYSIS

In the preface to his introductory book on decision analysis, Raiffa [1968] states that, in his view, a proper approach to this subject presents neither a descriptive theory of actual behavior nor a positive theory of behavior for a superintelligent, fictitious being. Rather, decision analysis should prescribe "how an individual who is faced with a problem of choice under uncertainty should go about choosing a course of action that is consistent with his basic judgments and preferences. He must consciously police the consistency of his subjective inputs and calculate their implications for action. Such an approach is designed to help us reason and act a bit more systematically--when we choose to do so"!

Decision analysis provides a means for objectively optimizing decisions and actions while taking into account subjective judgments of various utilities

and likelihoods. The discipline of decision analysis has been applied to problems in medical diagnosis by Schwartz and colleagues [1973] and by Fryback [1974], and to problems in therapeutic medicine by McNeil and Adelstein [1975] and by Pauker and Kassirer [1975], for example.

ROC analysis is directly applicable in decision-analytic approaches to medical problems. In order to decide upon a course of action--such as what type of therapy to institute for a particular disease in a situation where the presence of disease is uncertain--the physician must consider the likelihood that a particular diagnostic decision is correct--that the disease is present, given a positive result for some test, for example. Since the odds that a disease is present, given a positive test result, are given by

$$\frac{P(s|S)}{P(n|S)} = \frac{P(S|s)}{P(S|n)} \cdot \frac{P(s)}{P(n)} ,$$

these odds are clearly dependent upon the operating point adopted on the ROC curve for the diagnostic study. Hence the optimal management of a clinical situation depends both upon an optimal selection of a cource of action, given the diagnostic decision, and also upon an optimal selection of the diagnostic decision itself, which is equivalent to selecting an optimal confidence threshold and a corresponding operating point on the ROC curve.

Given a certain amount of relevant information, the physician must decide whether to reach a diagnostic decision or, instead, to postpone his decision and to order additional tests which will be costly, in some sense. Elementary principles of decision analysis are directly applicable to such problems [Raiffa, 1968].

In the previous section, we discussed optimizing the operating point on the ROC curve for a diagnostic test by minimizing average cost or, equivalently, maximizing average net benefit. The same optimization would result from an explicitly decision-analytic approach in which the course of therapeutic action was determined solely by the diagnostic decision.

Also in the previous section, we considered the average net benefit of an additional diagnostic test when some diagnostic information is already available. In decision analytic terms, the additional test is worthwhile in that case only if its average net benefit is positive. Decision analysis shows that the average net benefit of the additional test can also be evaluated in a more specific way by considering two situations: one in which the decision based on previous information is "negative" and one in which it is "positive" [cf. Raiffa, 1968, Chapter 2]. The problem of deciding whether to order an additional test becomes one of evaluating whether or not the expected diagnostic gain is worth the cost; the answer depends upon the information available at the time when the test must be ordered. Again, analysis of this situation requires knowledge of ROC curves--that is, knowledge of the possible relationships among the relative frequencies of various types of correct and incorrect decisions.

We have discussed elsewhere the problem of evaluating combinations of diagnostic studies [Metz, Starr, Lusted, and Rossmann, 1975]. As one might expect, a highly definitive study with high overhead cost which is too "expensive" to be used routinely can, in some cases, be advantageous if it is applied to a "screened" population, and optimal decision thresholds for both the "screening study" and the "follow-up study" depend upon whether they are used alone or

in combination. Optimization of the diagnostic approach requires knowledge of the ROC curves for the various studies.

6. OVERVIEW

Most basically, an ROC curve provides an empirical description of the trade-offs which are possible among the various types of correct and incorrect decisions as the human decision-maker varies one or more "confidence thresholds." Conventional ROC curves measured in simple decision-making situations can, in some cases, be used to predict human decision performance in more complex decision situations.

By considering both the consequences of various types of diagnostic decisions and the overhead cost of a diagnostic study, one can use the ROC curve to evaluate the diagnostic usefulness of a study in any particular clinical context. Since the ROC curve describes the possible relationships among the probabilities of the various types of correct and incorrect decisions, it plays a central role in optimizing diagnostic strategies using the general techniques of decision analysis.

ACKNOWLEDGMENTS

The Center for Radiologic Image Research is supported by Grant GM-18940 from the U. S. Public Health Service. The Franklin McLean Memorial Research Institute is operated by The University of Chicago for the U. S. Energy Research and Development Administration under Contract E(11-1)-69.

We thank Stephen G. Pauker, M.D., of Tufts University for pointing out the approximate nature of linearly combining certain decision and overhead cost components, such as those involving mortality risk, in the average cost formulation.

REFERENCES

Alcorn, F.S. and O'Donnell, E., 1969, Cancer, 23:879.

Anderson, T.M. Jr., Mintzer, R.A., Hoffer, P.B., Lusted, P.B., Smith, V.C., and Pokorny, J., 1973, Invest. Radiol., 8:244.

Andrus, W.S., Hunter, C.H., and Bird, K.T., 1975, Chest, 64:463.

Edwards, W., Guttentag, M., and Snapper, K., 1975, "A Decision-Theoretic Approach to Evaluation Research." In: Handbook of Evaluation Research, Vol. 1, E.L. Struening and M. Guttentag, eds. (Beverly Hills, California: Sage Publications), Chapter 7.

Egan, J.P., Schulman, A.I., and Greenberg, G.Z., 1959, J. Acoust. Soc. Am., 31:768.

Fryback, D.G., 1974, The Use of Radiologists' Subjective Probability Estimates in a Medical Decision Making Problem (Ann Arbor, Michigan: Dept. of Psychology, Univ. of Michigan), Publication MMPP-74-14.

Goodenough, D.J., 1972, Radiographic Applications of Signal Detection Theory. Doctoral Thesis, University of Chicago.

Goodenough, D.J. and Metz, C.E., 1975, "Implications of a 'Noisy' Observer to Data Processing Techniques." In: Information Processing in Scintigraphy, C. Raynaud and A. Todd-Pokropek, eds. (Orsay, France: Commissariat à l'Énergie Atomique, Département de Biologie, Service Hospitalier Frédéric Joliot), p. 400.

Goodenough, D.J., Metz, C.E., and Lusted, L.B., 1973, Radiology, 106:565.

Goodenough, D.J., Rossmann, K., and Lusted, L.B., 1972, Radiology, 105:199.

Goodenough, D.J., Rossmann, K., and Lusted, L.B., 1973, Invest. Radiol., 8:339.

Goodenough, D.J., Rossmann, K., and Lusted, L.B., 1974, Radiology, 110:89.

Green, D.M. and Swets, J.A., 1966, Signal Detection Theory and Psychophysics (New York: John Wiley and Sons). Also in reprint with corrections, 1974 (Huntington, N.Y.: Robt. E. Krieger Publ. Co.).

Kundel, H.L. and Revesz, G., 1974, Invest. Radiol., 9:166.

Luce, R.D., 1963, "Detection and Recognition." In: Handbook of Mathematical Psychology, Vol. 1, R.D. Luce, R.R. Bush, and E. Galanter, eds. (New York: John Wiley and Sons), Chapter 3.

Lusted, L.B., 1968, Introduction to Medical Decision Making (Springfield, Ill.: Charles C Thomas Publisher).

Lusted, L.B., 1971, N. Engl. J. Med., 284:416.

Lusted, L.B., 1975, "Receiver Operating Characteristic Analysis and its Significance in the Interpretation of Radiological Images." In: Current Concepts in Radiology, Vol. 2, E.J. Potchen, ed. (St. Louis, Mo.: C. V. Mosby Co.), Chapter 6.

McNeil, B.J. and Adelstein, S.J., 1975, N. Engl. J. Med., 293:221.

McNeil, B.J., Hessel, S.J., Branch, N.T., Bjork, L., and Adelstein, S.J., 1976, J. Nucl. Med., 17:163.

McNeil, B.J., Keeler, E., and Adelstein, S.J., 1975, N. Engl. J. Med., 293:211.

McNeil, B.J., Varady, P.D., Burrows, B.A., and Adelstein, S.J., 1975, N. Engl. J. Med., 293:216.

Metz, C.E. and Goodenough, D.J., 1973, J. Nucl. Med., 14:873.

Metz, C.E. and Goodenough, D.J., 1975, "Quantitative Evaluation of Human Visual Detection Performance Using Empirical Receiver Operating Characteristic Curves." In: Information Processing in Scintigraphy, C.E. Metz, S.M. Pizer, and G.L. Brownell, eds. (Oak Ridge, Tenn.: USERDA-TIC), Publication CONF-730687.

Metz, C.E., Goodenough, D.J., and Rossmann, K., 1973, Radiology, 109:297.

Metz, C.E., Starr, S.J., and Lusted, L.B., 1976, "Observer Performance in Detecting Multiple Radiographic Signals: Prediction and Analysis Using a Generalized ROC Approach." Radiology, in press.

Metz, C.E., Starr, S.J., Lusted, L.B., and Rossmann, K., 1975. "Progress in Evaluation of Human Observer Visual Detection Performance Using the ROC Curve Approach." In: Information Processing in Scintigraphy, C. Raynaud and A. Todd-Pokropek, eds. (Orsay, France: Commissariat à l'Énergie Atomique, Département de Biologie, Service Hospitalier Frédéric Joliot), p. 420.

Morgan, R.H., Donner, M.W., Gayler, B.W., Margulies, S.J., Rao, P.S., and Wheeler, P.S., 1973, Am. J. Roentgenol., Rad. Therapy and Nuclear Med., 117:757.

Pauker, S.G. and Kassirer, J.P., 1975, N. Engl. J. Med., 293:229.

Pastore, R.E. and Scheirer, C.J., 1974, Psychological Bulletin, 81:945.

Raiffa, H., 1968, Decision Analysis: Introductory Lectures on Choices under Uncertainty (Reading, Mass.: Addison-Wesley).

Revesz, G. and Kundel, H.L., 1971, Invest. Radiol., 6:315.

Revesz, G. Kundel, H.L., and Graber, M.A., 1974, Invest. Radiol., 9:479.

Schwartz, W.B., Gorry, G.A., Kassirer, J.P., and Essig, A., 1973, Am. J. Med., 55:459.

Starr, S.J., Metz, C.E., Lusted, L.B., and Goodenough, D.J., 1975, Radiology, 116:533.

Swets, J.A., 1973, Science, 182:990.

Swets, J.A., 1976, "Signal Detection Theory Applied to the Evaluation of Imaging Techniques in Clinical Medicine." To be presented in a symposium on "New Trends in Psychophysics" at the Twenty-first International Congress of Psychology, 18-25 July, 1976, Paris.

Swets, J.A. and Birdsall, T.G., 1956, "The Human Use of Information, III. Decision-making in Signal Detection and Recognition Situations Involving Multiple Alternatives." In: Transactions of the Institute of Radio Engineers, Profession Group on Information Theory, IT-2:138.

Swets, J.A., Tanner, W.P., and Birdsall, T.G., 1961, Psychological Review, 68:301.

Treisman, M. and Leshowitz, B., 1969, Perception and Psychophysics, 6:281.

Whalen, A.D., 1971, Detection of Signals in Noise (New York: Academic Press).

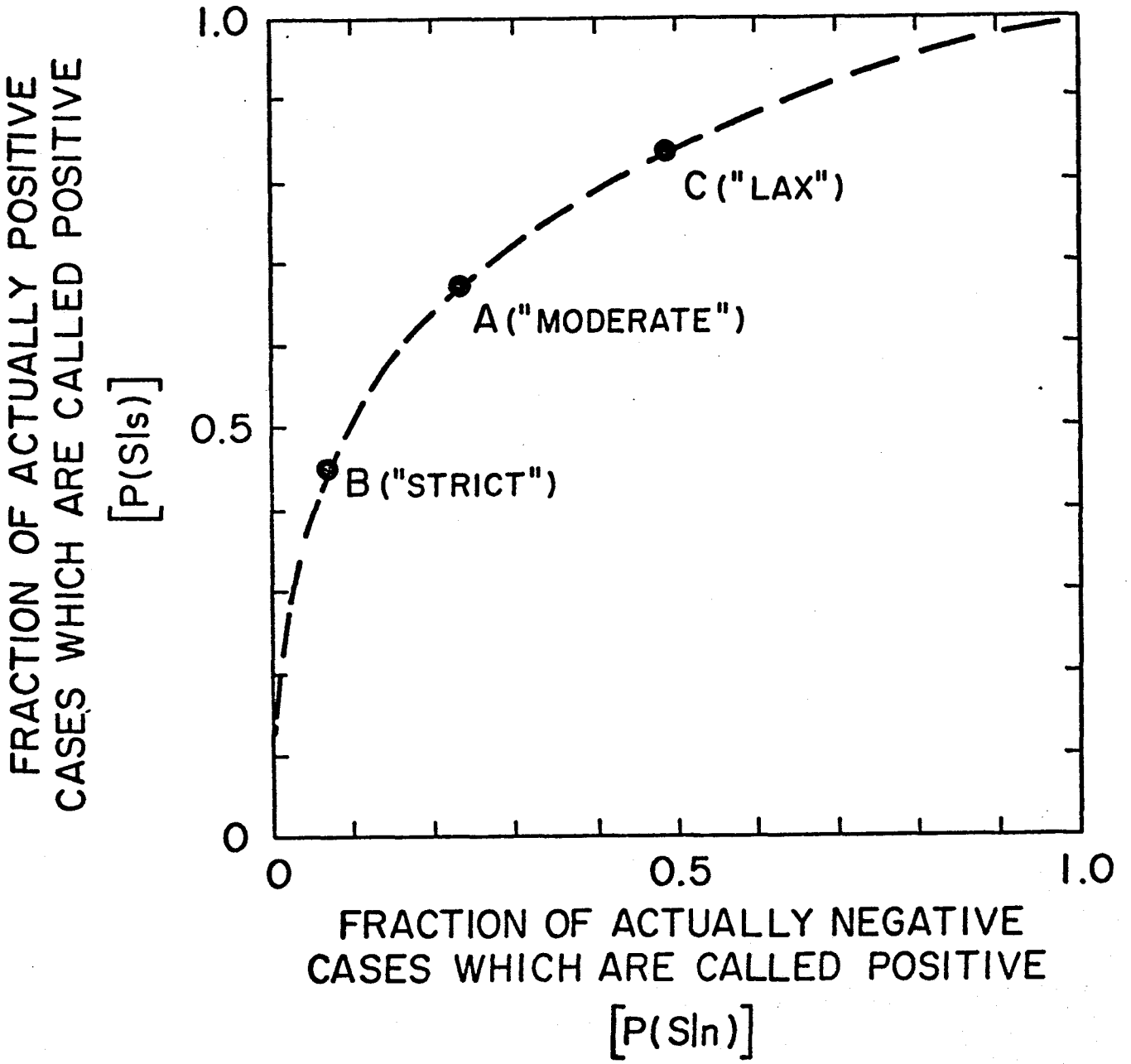
Wickelgren, W.A., 1968, J. Math. Psych., 5:102.

FIGURE LEGEND

Figure 1. A typical conventional ROC curve. The combination of relative decision frequencies represented by point "A" might result from use of a "moderate" confidence threshold, while combinations "B" and "C" would result from use of a more "strict" and a more "lax" confidence threshold, respectively. The ROC curve is swept out as a continuum of various possible confidence thresholds is considered.

SHORT TITLE

ROC Analysis and Determination of Diagnostic Benefit


1.0
0.5
0
0
0.5
1.0
FRACTION OF ACTUALLY POSITIVE CASES WHICH ARE CALLED POSITIVE
[P(S|s)]
FRACTION OF ACTUALLY NEGATIVE CASES WHICH ARE CALLED POSITIVE
[P(S|n)]
C ("LAX")
A ("MODERATE")
B ("STRICT")