SPECIAL ARTICLE

DECISION-MAKING STUDIES IN PATIENT MANAGEMENT*

LEE B. LUSTED, M.D.

Abstract Decision analysis provides a method to help the physician choose a course of action consistent with his personal judgments, to relate his preferences to costs, and to act more systematically. Decision analysis uses personal probabilities and deals with the relation of values and costs of patient management procedures. The physician is able to introduce intuitive judgments directly into the decision problem by using a numerical scale to express his uncertainty about a symptom or a diagnosis. His preference for consequences of diagnoses and treatments can be numerically scaled as utility values.

Signal-detection theory has been used to develop performance criteria for radiologists' assistants and radiologic systems. The essential feature of the analysis, an operating characteristic curve, is a means for separating the detectability of a signal, a sensory process, from all other factors involved in the decision process.

MANY radiologists were disturbed by Knowles's articles,[1] "Radiology – a Case Study in Technology and Manpower." In his concluding paragraph, Knowles said in part,

> For example, one could and should ask how many renal arteriograms in patients with hypertension have resulted directly in the surgical or medical cure of the patients' hypertension. Or one could ask how many lung scans have altered the diagnosis, treatment or prognosis of the patient thought to have a pulmonary embolus. Such studies might conceivably put a discriminatory brake on the use of a technology seemingly run wild. As one looks at the cost of developing modern radiotherapeutic facilities, one must ask what the benefit of radiotherapy is as contrasted with the cost although in this instance even to ask the question, in an affluent, developed country will bring criticism raining down around the shoulders of the questioner.

What can a radiologist answer? Do the same questions about the relative values and costs of diagnostic tests and therapeutic procedures apply to all areas of medicine? Probably they do because the questions concern two larger problems – namely, the unmanaged proliferation of medical data and the distribution of health care.

As Kuhn[2] points out, two characteristics of a scientific revolution are professional insecurity and social pressure for reform. Professional insecurity is a result of a breakdown in the traditional way of doing business, and social pressure for reform builds up because existing institutions have ceased to meet adequately the problems posed by society. Under conditions of scientific revolution an environment is created that is favorable for seeking new solutions to old problems. It is in this context that I propose to consider decision-making studies as useful keys to the puzzles presented by Knowles.

I have chosen an example from diagnostic radiology as a focus of interest, and I will expand the discussion of decision making by the use of additional examples from other areas of medicine. The study from diagnostic radiology concerns the accuracy of interpretation of chest roentgenograms for the presence of active tuberculosis. This study will be analyzed by the use of signal-detection theory.

Accurate interpretation of a roentgenogram depends upon the visual perception by the radiologist of the images recorded on film. The images so recorded may be considered a signal, or a group of signals, and the radiologist observes these signals against a background of shadows that introduce confusion or noise. The radiologist, in his role as a decision maker, uses the signals to choose among alternative diagnoses, and, unfortunately, the decision must often be made under conditions of uncertainty.

COMMENTS ON SIGNAL DETECTION

The general theory of signal detection, based on probability theory and statistical decision theory, was developed most fully in the 1950's by mathematicians and engineers at the University of Michigan, at Harvard University and at the Massachusetts Institute of Technology.[3] Much of the material resulted from an analysis of radar and information-transmission systems in situations where it was necessary for an observer to differentiate a signal plus noise from noise alone.

Signal-detection theory was also applied by psychophysicists[4,5] in experiments on vision and audition, with a heavy emphasis on the decision aspects of detection. In a visual experiment an observer is asked to observe an illuminated screen upon which a signal of pulsed light is projected. The observer is to state during each observation interval whether a signal is or is not present. This type of study and other related experiments have demonstrated that a host of factors that affect the observer's attitude are compressed into a single variable called the decision criterion. The approach clearly isolates the inherent detectability of a signal, a sensory process, from the observer's decision process. This procedure has been useful in the study of a number of sensory problems – for instance, visual and auditory thresholds. The theory will now be applied to the interpretation of chest roentgenograms.

*Address reprint requests to Dr. Lusted at the Department of Radiology, University of Chicago, 950 East 59th St., Chicago, Ill. 60637.

Chest X-ray Interpretation as a Signal-Detection Study

We are going to perform an experiment similar to one designed by Yerushalmy[6,7] to study the accuracy of chest-film interpretation for the presence of active tuberculosis. You are one of 10 physicians participating in the study, and you will be asked to make a diagnosis about the presence or absence of active disease on a series of 70-mm chest photofluorograms selected from a large population (over 14,000) of college students. You will be asked to provide an average of 3000 film readings according to a randomized scheme so that you will provide 10 independent interpretations of each film.

Since the photofluorograms are all "proved cases," we will record from your interpretations the number of cases of tuberculosis that you diagnose correctly (true positive), the number of cases actually negative that you call positive (false positive), the number of negative cases that you diagnose correctly (true negative) and, finally, the number of cases truly positive for tuberculosis that you call negative (false negative).

We can now find the accuracy and error rates of interpretation. We will call these your attitude about the diagnostic criteria for tuberculosis.

Your true-positive (T.P.) percentage is

$$\text{T.P. }(\%) = \frac{\text{No. true positives you read}}{\text{No. true positives you read} + \text{no. false negatives you read}} \times 100.$$

Your false-positive (F.P.) percentage is

$$\text{F.P. }(\%) = \frac{\text{No. false positives you read}}{\text{No. true negatives you read} + \text{no. false positives you read}} \times 100.$$

True-negative and false-negative percentages are calculated in the same way.

The four possible outcomes for your diagnoses may be displayed in a matrix as shown below. This is called a decision matrix in decision-theory studies.

		Your Diagnosis: Positive	Your Diagnosis: Negative	
Disease Category	Positive	% true positive	% false negative	= 100% diseased patients tested
	Negative	% false positive	% true negative	= 100% negative patients tested

Following the convention in psychophysics, data will be plotted only in terms of two independent quantities — namely, the percentage of true-positive and the percentage of false-positive diagnoses. Your diagnostic performance for each series of film interpretations is indicated by a "box score" of percentage of true-positive and percentage of false-positive diagnoses. I participated in one of Yerushalmy's studies, and my performance was 80 per cent true positive and 4 per cent false positive.

Figure 1 shows the results of a large number of studies by Yerushalmy,[6,7] by Garland[8] and by other investigators.[9] Note the reciprocal relation between percentage of true-positive and percentage of false-positive diagnoses. The curve is called a receiver-operating-characteristic (ROC) curve in detection theory.

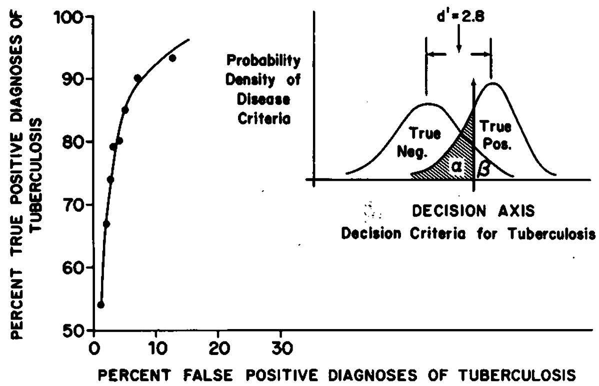

Figure 1. Receiver Operating Characteristic (ROC) Curve for the Interpretation of Chest Photofluorograms for the Presence of Pulmonary Tuberculosis.

A reciprocal relation is demonstrated between the percentage of true-positive and the percentage of false-positive diagnoses. Hypothetical population density curves that generated the ROC curve are shown in the upper right diagram. The α and β areas represent false-negative and false-positive diagnoses. d' is an index of detectability and is defined as the mean separation of the two distributions divided by the standard deviation of one distribution. This is written

$$d' = \frac{{}^{m}\text{T.P.} - {}^{m}\text{T.N.}}{\sigma},$$

where ^{m}T.P. = mean of true-positive population and where ^{m}T.N. = mean of true-negative population.

It is convenient to assume that data-generating processes that underlie the ROC curve are as shown in the right-hand diagram of Figure 1. The area under the true-positive distribution curve is divided by a vertical line into two parts — namely, a true-positive area (T.P.%) and a false-negative (α) area (F.N.%). Likewise, the true-negative distribution curve is divided into a true-negative area (T.N.%) and a false-positive (β) area (F.P.%). For each position of the vertical line there is a corresponding set of values in the four cells of the decision matrix and a corresponding point on an ROC curve. The vertical line represents the critical decision criteria that the decision maker has selected on the decision axis. The decision maker divides the chest x-ray-film observations into two mutually exclusive regions. If a set of observations for tuberculosis is noted, he says that tuberculosis is present. If the observations fail to meet his criteria, he says that no tuberculosis is present. The vertical line represents the critical set of radiographic criteria of a decision maker that he uses to separate the films into two groups.

Why an ROC Curve Is a Useful Device

An ROC curve provides a natural distinction between the inherent detectability of a signal (radiographic image) and the judgment of the observer (the physician). The unique feature of this presentation is that the results are independent of any assumptions that one might make about the statistical distribution of the sensory events produced by the signal plus noise or by noise alone.

An ROC curve may be generated by holding constant the physical characteristics of the stimulus situation while the observer is asked to change his decision attitude. Garland[8] attempted to assess the effect of a film reader's attitude by asking the radiologist to read a group of films twice: first with a lax or liberal attitude and then with a strict or conservative attitude. With a liberal attitude the radiologist would be more likely to adopt a policy such as "better safe than sorry" or "when in doubt call the shadow positive." This policy increases the number of true-positive diagnoses, but it concomitantly increases the false-positive diagnoses. In Figure 1 a more liberal decision attitude would generate points on the upper portion of the curve, whereas increasingly conservative attitudes would generate points progressing downward.

Garland[8] found that a radiologist could change his point of operation on the ROC curve by changing his decision attitude. It seems probable that a physician cannot maintain a consistent decision attitude for an extended time. This may help to explain the results of observer-error studies,[8] which disclosed that a radiologist will disagree with his own film interpretations about one out of five times on a second reading of the same films, and he will disagree with his colleague's interpretation about one out of three times.

Unfortunately, the ROC curve does not seem to help sort out the factors that contribute to the radiologist's error rate. Errors caused by inadequate eye scanning of the film, misinterpretation of images, lapses of memory and so forth are lumped together, although future research may show how some types of errors may be isolated for study.

Would it be useful to the radiologist to know his ROC curve for each type of roentgenographic examination? Can he hold his false-positive errors within certain limits? What percentage of false-positive and false-negative diagnoses is he willing to accept? These questions lead to a consideration of decision rules and likelihood ratios.

Decision Rules and Likelihood Ratios

Without specifying a decision objective there is no way to evaluate a decision procedure. Your decision objective might be to make as many positive diagnoses as possible or it might be to hold false-positive diagnoses within a fixed limit, perhaps 10 per cent.

Decision theory says that the goals of a decision maker can be as varied and different as the practical situations in which decisions are made. Is it possible to formulate a single decision procedure that applies to more than one goal? Although an unqualified "yes" is not possible, it is possible to show that a likelihood ratio provides a convenient and simple means of stating optimal decision rules applicable to several different decision goals. Examples of such goals are to maximize the percentage of correct diagnoses, to maximize a weighted combination of diagnoses, or to maximize the expected value of diagnosis and treatment.

A likelihood is a conditional probability and a likelihood ratio is the ratio of two conditional probabilities. It is appropriate to speak of a likelihood as a conditional probability because the probability is influenced by information already known. In Figure 1 the percentage of true-positive diagnoses represents a conditional probability or likelihood of a true-positive diagnosis of tuberculosis given a group of patients some of whom are known to have active tuberculosis.

A likelihood ratio in Figure 1 is the ratio of percentage of true-positive interpretations to percentage of false-positive interpretations. At my point of operation on the ROC surve the likelihood ratio is $\frac{0.80\ \text{T.P.}}{0.04\ \text{F.P.}} = 20.$

Estimation of Values and Costs of Decisions from an ROC Curve

The expected values for the four possible outcomes of a decision can be used to study the relative weight that the radiologist attaches to the values of correct decisions compared with the costs of errors. Expected value of a decision strategy is the sum of four terms, each of which represents the value or cost associated with an outcome weighted by the probabilitythat outcome will occur under the strategy in question.[10]

Maximizing the expected value is equivalent to maximizing an expression of the form

$$L_c = \frac{p\,(TN)}{p\,(TP)} \cdot \frac{V_{TN} + C_{FP}}{V_{TP} + C_{FN}},$$

where L_c = cut off or critical value of criteria on the basis of which the physician makes his decision,
p (TN) = probability of a true-negative diagnosis,
p (TP) = probability of a true-positive diagnosis,
V_{TN} = value (for the physician) of a true-negative diagnosis,
V_{TP} = value (for the physician) of a true-positive diagnosis,
C_{FP} = cost (for the physician) of a false-positive diagnosis and
C_{FN} = cost (for the physician) of a false-negative diagnosis.

Green and Swets[11] show that the slope of an ROC curve at any point is equal to the likelihood ratio criterion that generates that point. Developing an ROC curve makes it possible to investigate the values of an observer without asking for an explicit expression of the values.

At my operating point of 80 per cent true-positive and 4 per cent false-positive diagnoses in Figure 1, the slope of the ROC curve is 5. The preceding equation may now be rewritten with the use of the data from Yerushalmy[6] for $\frac{p\ (TN)}{p\ (TP)} = 2000$:

$$5 = 2000 \cdot \frac{V_{TN} + C_{FP}}{V_{TP} + C_{FN}}.$$

This equation may be rewritten:

$$\frac{1}{400} = \frac{V_{TN} + C_{FP}}{V_{TP} + C_{FN}}.$$

With the help of this expression I wish to explore my preferences for costs of false-positive and false-negative errors in the diagnosis of active tuberculosis. I begin by asking some questions about the relative values of true-positive and true-negative diagnoses. Since the patient does not participate in this exercise I must try to interpret values for this patient.

Questions about Values	My Attitude about Relative Value of True-Positive vs True-Negative Diagnoses
SUBJECTIVE VALUES	$\frac{V_{TP}}{V_{TN}}$
1. Value to my self-esteem of correct diagnosis	1
2. Value to referring physicians of correct diagnosis	1
3. Value to my reputation with referring physicians	1
4. Value to patient of correct diagnosis	<1
5. Value to my reputation with patients	1
6. Value to society of correct diagnosis	>1
7. et cetera . . .	
OBJECTIVE VALUES	
1. My professional fee	1
2. et cetera . . .	

Following this line of argument, I conclude that for me for this diagnosis $V_{TP} = V_{TN}$, and I will thus arbitrarily set $V_{TP} = V_{TN} = 1$ unit.

Then the previous expression is rewritten,

$$\frac{1}{400} = \frac{1 + C_{FP}}{1 + C_{FN}}.$$

I proceed to ak questions about the relative cost of diagnostic errors compared with the value of correct diagnoses. Dollar costs need to be considered in this argument. I conclude that the cost of false-positive and false-negative errors is much greater than the 1 unit assigned to values for correct diagnoses. This is indicated as $C_{FN} >> 1$; $C_{FP} >> 1$.

The preceding expression is now written

$$\frac{1}{400} = \frac{C_{FP}}{C_{FN}} \text{ or } C_{FN} = 400\ C_{FP}.$$

In decision-analysis terms this expression can be interpreted to mean that for me for the diagnosis of active tuberculosis, the cost of a false-negative diagnosis must be 400 times the cost of a false-positive diagnosis to warrant a positive diagnosis of tuberculosis (that is, the consequences of ignoring tuberculosis when it is present must be at least 400 times as serious as the consequences of further testing or treatment of tuberculosis when the patient does not in fact have tuberculosis). I am willing to trade the cost of one false-negative for 400 false-positive diagnoses.

Values and money are closely related in considerations of patient management, and more studies are needed.[12,13] Adelstein, Parker and Wagner[14] have followed an argument using likelihood ratios and weighting functions to obtain an objective evaluation of new diagnostic tests. They show a specific application to the uptake of ^{131}I by the thyroid gland. The results indicate that although radioiodine uptake can be predicted with some accuracy from history, signs and symptoms, the prediction was incorrect in 21 per cent of cases when computer evaluated and in 40 per cent when evaluated by physicians. The authors raise questions whether measurements of radioiodine uptake alone should direct therapy or whether other tests of thyroid function can serve more efficiently to guide patient management. These questions are similar to those of Knowles. Diagnostic tests now in use and new diagnostic tests need objective evaluation to determine whether they optimize the number of patients correctly managed.[15]

Ginsburg[16] has developed a sequential decision model to be used as an aid in patient management. Starting with the problem of pleural effusion he developed a model composed of diagnostic-decision trees, probability estimates for the tree branchings and a utility model with dollar tradeoffs for such consequences as days in bed with no pain, short-term severe pain, days of restricted activity and death. This is the first study that I have seen investigating in detail a relation of dollar value to other patient values such as pain and blindness. It appears that a knowledgeable physician teamed with a person versed in decision analysis can develop useful components of probability and utility models to assist the physician to patient management.

Application of ROC Curves to the Training of Mammogram Screeners and to the Evaluation of Radiographic Imaging Systems

Two challenging problems, related to questions raised by Knowles, concern the use of assistants to screen radiographic examinations and the evaluation of radiographic equipment for diagnostic usefulness. In both these areas decision-making studies and ROC curves appear to be helpful tools.

Radiologists' Assistants in Mammography Screening

Physicians' assistants are being trained for several medical specialties, one of which is radiology. Technologists are doing gastrointestinal examinations,[17] and technologists are screening chest roentgenograms,[18] barium-enema films[19] and mammograms.[20-22] ROC curves can be used for training and assessing the performance of nonradiologic personnel to screen mammograms or other types of radiographic examinations.

Alcorn and O'Donnell[20,21] worked with two medical secretaries, two x-ray technologists experienced in mammography and two senior x-ray technologists not experienced in mammography. The objective was to train the screeners to separate the abnormal breast films from the normal films, and special emphasis was placed on finding breast cancer. Films of proved cases were used for training, and a different set of proved-case films were used to test the screeners. The proved cases were obtained from Egan.[23] Curve A in Figure 2 was constructed from

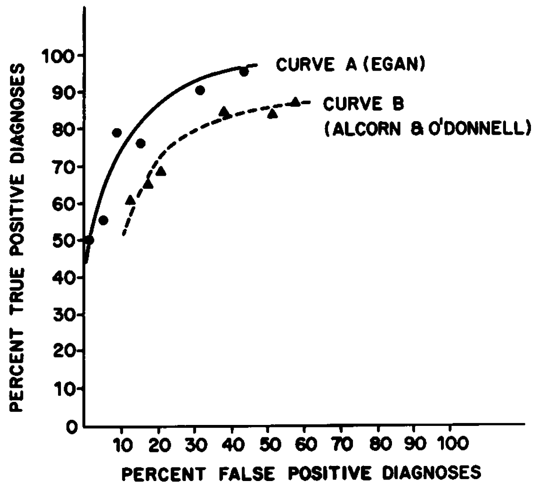

Figure 2. ROC Curves for the Interpretation of Mammograms.

Curve A has been constructed from the data of Egan,[23] and Curve B from the data of Alcorn and O'Donnell.[20,21] Curve A represents interpretation by radiologists, and Curve B interpretation by nonradiologic personnel. Curve A, with a larger value for d' than Curve B, indicates greater sensitivity of the observers.

the data of Egan,[23] and evaluation of the radiologist's performance is based on the pathologist's diagnosis that we accept as "the truth." In decision-making studies it is important to recognize that a reference is accepted as "the truth."[24] This reference is frequently the pathologist's opinion, but in other situations, it may be the considered judgment of a group of experts – for instance, a tumor board. Curve B shows the performance of the screeners, and a comparison can be made with the performance of the radiologists in Curve A. Is this performance acceptable, or should the screeners have more training? A value judgment is necessary since there is no "right" answer to the question. Would 90 per cent true-positive and 10 per cent false-positive diagnoses be acceptable? To reach this level, the screeners would have to surpass the radiologist's performance. Another method whereby the radiologist can improve his own performance as well as the screener's is the development of flow charts and decision trees.

Use of Flow Charts and Decision Trees

A flow chart consists of a hierarchy or sequential system of progressively more specific questions leading ultimately to the most specific diagnostic output justified by the evidence at hand. Tuddenham[19,25] has developed a series of logical flow charts as an aid in teaching x-ray diagnosis. A portion of a flow chart in Figure 3 shows a programmed search strategy used in examining x-ray films of the barium-filled colon.

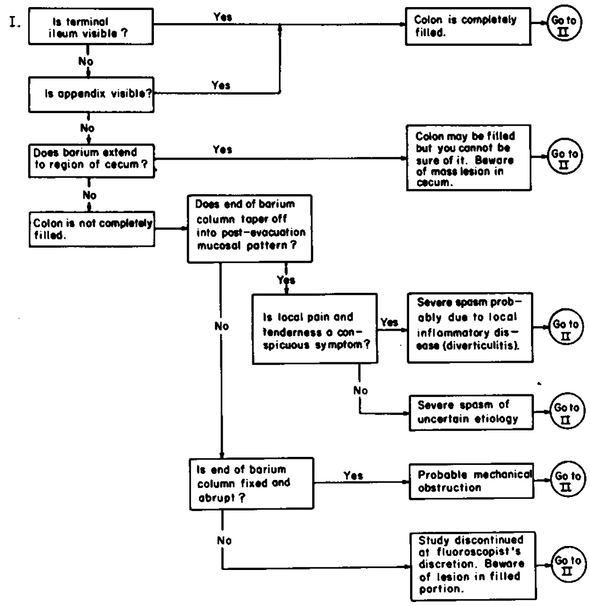

Figure 3. Flow Chart Showing a Programmed Search Strategy Used in Examination of X-Ray Films of the Colon (Adapted from the Work of Dr. William J. Tuddenham, Pennsylvania School of Medicine, Who Has Devised Many More Flow Charts).[25]

Only the first steps are shown.

One of the most difficult and time-consuming tasks in the construction of a patient-management decision model is synthesizing a decision tree that relates the set of alternative acts and outcomes at each stage of decision. A short tree structure of a neurologist's diagnostic process[26] is shown in Figure 4. Unless the pathologic states and diagnostic tests are few, the number of branches in the tree becomes too large to manage. This is when the tree becomes a bushy mess, as Raiffa[27] notes, and it is necessary to prune the tree, a difficult task because it requires concentrated thought by the physician about the step-by-step details of patient management that he usually does not explicitly assess. It requires that he assign probabilities to branches of the tree and utilities to alternative treatment consequences. Ginsberg[16] shows how a tree of decision

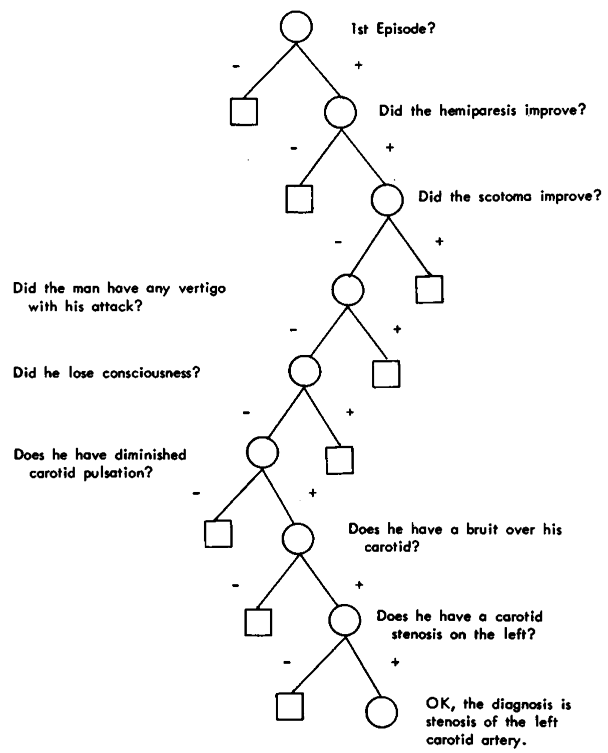

Figure 4. Tree Structure of a Neurologist's Diagnostic Game in Which the Information Given Was "Sudden Left Central Scotoma and Right Hemiparesis in a 55-Year-Old Man" (Reproduced from Kleinmuntz[26] with the Permission of the Publishers).

The circles represent nodes at which the physician may give a differential diagnosis, or he may elect to proceed by asking for more information. The plus and minus branches represent the presence or absence of the preceding symptom.

trees may be pruned in the management of pleural effusions.

A tree structure of a physician's diagnostic process can be visualized as the skeleton of a computer program that is used to study a sequence of steps in medical diagnosis and treatment, a process called sequential decision making. Gorry and Barnett[28-30] have developed a computer program that employs sequential decision making to balance the cost of making a diagnosis against the cost of further testing and the value of evidence that can be obtained. With the use of data on congenital heart disease the diagnostic accuracy achieved by the sequential program and a nonsequential or complete type of program was compared. The sequential program used seven tests, and the nonsequential program used 31 tests to achieve the same degree of diagnostic accuracy. The complete type of program is comparable in performance to the expert human diagnostician.

The capability to select a few critical tests could affect favorably the costs of patient management, and the sequential approach to computer-aided diagnosis appears to be a fertile area of study.

One reason that the physician has difficulty assigning probabilities to branches of a decision tree is that he may be uncertain about the relative importance of a symptom or a group of symptoms for a given diagnosis. This is so because there are few studies on the quantitative relations among signs, symptoms, laboratory tests and diagnoses. Medical data have not been examined for redundancies to determine which combinations of symptoms and tests give optimal differentiation of diseases. To illustrate data reduction and identification of optimal disease descriptors, an example was selected from the work of Pipberger[31] on the differential diagnosis of chest pain.

Data Reduction of Clinical Information in the Differential Diagnosis of Chest Pain

Pipberger[31] studied 1238 patients admitted to five Veterans Administration hospitals with the chief complaint of chest pain. The aim of the investigation was to define the relative value of signs and symptoms and to identify redundant and irrelevant information.

A comprehensive literature review was undertaken to identify all signs, symptoms and laboratory tests that were related to entities associated with chest pain. A list was constructed of 429 questions of the yes-no type and 69 items of numerical data from laboratory tests, blood pressure and so forth. Eighteen entities were found, but only four groups that contained 200 or more patients were selected for further study: acute myocardial infarction; old myocardial infarction, with coronary-artery insufficiency; angina pectoris; and pneumonia (with or without pleural involvement).

The results of the study will be illustrated by comparison of the patient group with old myocardial infarction with that of pneumonia. Only signs and symptoms that exceeded a frequency of 25 per cent were retained. However, lists based purely on frequency did not indicate which signs and symptoms were most useful in a differential diagnosis between old myocardial infarction and pneumonia.

The discriminative power of signs and symptoms exceeding 25 per cent prevalence was studied. The authors, using contingency table analysis and chi-square tests, concluded that chi-square values of 40* and greater were necessary to obtain efficient separation of entities. Table 1 shows the signs and symptoms arranged in order of ability to discriminate between old myocardial infarction and pneumonia. Some items of high frequency were found to lack discriminative power. A history of smoking occurred in 63 per cent of patients with old myocardial infarction and in 72 per cent of those with

*Chi-square test was used to indicate the importance of a sign or symptom in ability to separate two entities. The larger the value of chi-square, the more certain you are that the sign favors one of the diseases. The authors made a value judgment about "efficient" separation of diseases, and then a minimum chi-square of 40 was selected.

pneumonia. Because the prevalence rates are so similar the item lacks discriminative power and does not appear in Table 1.

Table 1. Signs and Symptoms Differentiating Old Myocardial Infarction from Pneumonia, Listed in Order of Importance.*

Discriminating Signs & Symptoms	Chi-Square Value†	Incidence Rates (%)		Indicating
		Old Myocardial Infarction	Pneumonia	
1. Pain ↑, physical exertion, present	144	71	8	Old myocardial infarction
2. Pain, stabbing, knifelike, present	127	9	71	Pneumonia
3. Pain ↑, physical exertion, past	126	63	6	Old myocardial infarction
4. Cough & expectoration	115	21	80	Pneumonia
5. Pain, retrosternal, past	115	59	6	Old myocardial infarction
6. Pain, retrosternal, present	109	63	9	Old myocardial infarction
7. Signs of consolidation on physical examination	100	1	51	Pneumonia
8. Pain radiation, left arm, hand, fingers, present	93	41	0	Old myocardial infarction
9. Pain for days, present	93	5	55	Pneumonia
10. Temperature ≧101°F	91	6	56	Pneumonia
11. Pain, lateral chest, present	88	1	47	Pneumonia
12. Pain, several hours, past	82	45	4	Old myocardial infarction
13. Pain radiation, left arm, hand, fingers, past	80	41	3	Old myocardial infarction
14. Respiratory rate ≧20	75	14	61	Pneumonia
15. Moist basilar rales	63	17	61	Pneumonia
16. Pain, constriction, tightness, present	63	38	4	Old myocardial infarction
17. Pain, numerous seizures daily, present	61	31	74	Pneumonia
18. Lagging hemithorax with respiration	59	2	37	Pneumonia
19. Pain at irregular intervals, past	56	34	3	Old myocardial infarction
20. Pain, constriction, tightness, past	53	35	5	Old myocardial infarction
21. Pain at irregular intervals, present	52	32	3	Old myocardial infarction
22. Pain, pressure-like, present	48	29	3	Old myocardial infarction
23. Pain, pressure-like, past	45	24	1	Old myocardial infarction
24. Pain, right anterior part of chest, present	42	1	27	Pneumonia
25. Temperature elevation >2 days	42	1	27	Pneumonia
26. Heart rate >100	40	6	35	Pneumonia

*Adapted from Table 10 of Pipberger et al.[31]

†Chi-square discrimination ≧40.

From the original list of 498 information items only 46 exceeded the limits of 25 per cent frequency and 40 for chi-square discrimination. This reduction in the number of items indicates that about 90 per cent of the total information was either redundant or irrelevant for the description of certain entities or for separation of these two diseases.

The authors proceeded to test the separating powers of discriminators by a variety of classification procedures. The 27 items that yielded the largest chi-square values in the discrimination studies were selected, and discriminant function analysis was applied. The list could be reduced further. Six items could classify patients with coronary-artery disease versus those with pneumonia with 95 per cent accuracy.

This example illustrates methods and objectives that could be applied to all areas of clinical medicine. The objectives are, first of all, to develop a frequency list of signs and symptoms for an entity, secondly, to develop a list of discriminating signs, symptoms and laboratory tests to be used in differential diagnosis and, thirdly, to test the classification accuracy of the discriminators to correlate the power of groups of discriminators with degrees of diagnostic accuracy. Most textbooks note certain signs and symptoms as pathognomonic. Their presence or absence is essential to diagnosis. Pipberger[31] found few items that could be called pathognomonic.

Evaluation of the Relative Usefulness of a Complex Radiographic Imaging System

A second problem for the radiologist is to evaluate the diagnostic usefulness of complex and costly radiographic imaging systems. This problem is at the heart of Knowles's[1] comment concerning "a technology seemingly run wild."

Rossmann[32] has studied the problem of radiographic image quality, and recently wrote as follows:

> . . . the central problem in a study of radiographic image quality is to gain knowledge regarding the effect of physical image quality on the diagnosis and not necessarily to design "high fidelity" imaging systems. For example, image intensifier-television outputs used in fluoroscopy are poor from a physical standpoint because of high quantum noise and low resolution. Nevertheless, a diagnosis can be made from these images in many cases before spot-films of better physical quality are viewed. The implication is that image intensifier-television images are often of sufficient diagnostic quality.

Sufficient diagnostic quality is the key phrase because it implies that the system should be evaluated in terms of the observer, the radiologist, instead of a physical property such as the modulation transfer function.[33] Evaluation in terms of the observer makes use of ROC curves.

Suppose that the installation of a closed-loop television system is to be made between the hospital radiology department and a remote station in the hospital. It will be convenient for the radiologists and other physicians to view the television monitor image transmitted from the radiology department, but the radiologist is concerned that the transmission system will degrade the diagnostic quality of the x-ray image.

A way to evaluate the performance of the television system is to develop ROC curves. A series of proved cases — for example, cholecystograms — is interpreted by the radiologists with direct viewing, and an ROC curve is developed. The same series of proved cases is transmitted via the television system and interpreted by the same radiologists, who now view the films on a television monitor. A second ROC curve is developed and compared with the first ROC curve.

Kundel, Revesz and Stauffer[34] have used this method of evaluation in their laboratory. Chest roentgenograms containing simulated nodules in the lung were used to test a television chain. The results are shown in Table 2.

Table 2. Comparison of the Error Rates of Observers for Detecting Nodules Using Different Modes of Viewing the Films.*

Viewing Mode	No. of Observers	False-Negative Diagnosis	False-Positive Diagnosis	Index of Detectability (d')
Direct viewing	21	0.14	0.24	1.91
Unprocessed television	3	0.34	0.30	0.98
Television contrast enhancement	10	0.24	0.26	1.46

*Adapted from Table 1 of Kundel et al.[34]

It is clear that television viewing alone results in an increased error rate and that image processing improves the performance. Television with contrast enhancement might be pictured as ROC Curve B in Figure 2, whereas direct viewing performance would be closer to ROC Curve A in Figure 2.

ROC curves and decision analysis are useful tools for the study of imaging systems.[35,36] Evaluation of radiographic image systems in terms of observer performance rather than in terms of physical constants of the system represents a noteworthy change from the present point of view, but it is essential that whenever possible, all new radiographic equipment and technics be so evaluated.

References

1. Knowles JH: Radiology — a case study in technology and manpower. New Eng J Med 280:1271-1278, 1323-1329, 1969
2. Kuhn TS: The Structure of Scientific Revolutions. Second edition. Chicago, University of Chicago Press, 1970, p 92
3. Detection theory and applications. Proc IEEE 58:607-852, May, 1970
4. Signal Detection and Recognition by Human Observers: Contemporary readings. Edited by JA Swets. New York, John Wiley and Sons, 1964
5. Green DM, Swets JA: Signal Detection Theory and Psychophysics. New York, John Wiley and Sons, 1966
6. Yerushalmy J: Reliability of chest radiography in the diagnosis of pulmonary lesions. Amer J Surg 89:231-240, 1955
7. *Idem:* The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. Radiol Clin N Amer 7:381-392, 1969
8. Garland LH: Studies on the accuracy of diagnostic procedures. Amer J Roentgen 82:25-38, 1959
9. Lusted LB: Logical analysis in roentgen diagnosis. Radiology 74: 178-193, 1960
10. Green DM, Swets JA:[5] p 22
11. *Idem:*[5] p 36
12. Collen MF, Feldman R, Siegelaub AB, et al: Dollar cost per positive test for automated multiphasic screening. New Eng J Med 283:459-463, 1970
13. Bell RS, Loop JW: The utility and futility of radiographic skull examination for trauma. New Eng J Med 284:236-239, 1971
14. Adelstein SJ, Parker R, Wagner HN Jr: First phase in objective evaluation of new diagnostic tests: application to thyroid uptake of ^{131}I. Invest Radiol 5:153-163, 1970
15. Taylor TR: Computer-guided diagnosis. J Roy Coll Physicians Lond 4:188-195, 1970
16. Ginsberg AS: Decision Analysis in Clinical Patient Management with an Application to the Pleural Effusion Problem. Santa Monica, California, Rand Corporation, 1969 (Document D-19753-RC)
17. Campbell JA, Lieberman M, Miller RE, et al: Experience with technician performance of gastrointestinal examinations. Radiology 92:65-73, 1969
18. Sheft DJ, Jones MD, Brown RF, et al: Screening of chest roentgenograms by advanced roentgen technologists. Radiology 94:427-429, 1970
19. Tuddenham WJ, Houser LM, Tuddenham PS, et al: Preliminary evaluation of effectiveness of logical flow charts in teaching roentgen diagnosis. Radiology 93:17-24, 1969
20. Alcorn FS, O'Donnell E: Mammogram screeners: modified program learning for nonradiologic personnel. Radiology 90:336-338, 1968
21. *Idem:* The training of nonphysician personnel for use in a mammography program. Cancer 23:879-883, 1969
22. Dowdy AH, Lagasse LD, Roach P, et al: Lay screeners in mammographic survey programs. Radiology 95:619-621, 1970
23. Egan RL: Mammography. Springfield, Illinois, Charles C Thomas, 1964
24. Coppleson LW, Factor RM, Strum SB, et al: Observer disagreement in the classification and histology of Hodgkin's disease. J Nat Cancer Inst 45:731-740, 1970
25. Tuddenham WJ: The use of logical flow charts as an aid in teaching roentgen diagnosis. Amer J Roentgen 102:797-803, 1968
26. Kleinmuntz B: Diagnostic problem solving by computer. Jap Psychol Res 7:189-194, 1965
27. Raiffa H: Decision Analysis: Introductory lectures on choices under uncertainty. Reading, Massachusetts, Addison-Wesley Publishing Company, 1968, p 240

28. Gorry GA, Barnett GO: Sequential diagnosis by computer. JAMA 205:849-854, 1968
29. Gorry GA: Strategies for computer-aided diagnosis. Math Biosci 2:293-318, 1968
30. *Idem:* Modeling the diagnostic process. J Med Educ 45:293-302, 1970
31. Pipberger HV, Klingeman JD, Cosma J: Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain. Meth Inform Med 7:79-92, 1968
32. Rossmann K, Wiley BE: The central problem in the study of radiographic image quality. Radiology 96:113-118, 1970
33. Rossmann K: Point spread-function, line spread-function, and modulation transfer function: tools for the study of imaging systems. Radiology 93:257-272, 1969
34. Kundel HL, Revesz G, Stauffer HM: The electro-optical processing of radiographic images. Radiol Clin N Amer 7:447-460, 1969
35. Lusted LB: Introduction to Medical Decision Making. Springfield, Illinois, Charles C Thomas, 1968, p 112
36. *Idem:* Perception of the roentgen image: applications of signal detectability theory. Radiol Clin N Amer 7:435-445, 1969

MEDICAL PROGRESS

ABDOMINAL SURGERY (First of Three Parts)*

Claude E. Welch, M.D.

THE last progress report on this subject was made two years ago. In conformity with the others in this series, this will serve as a guide to important papers published during this period. Articles have been chosen primarily because of their value for the clinical surgeon. Though the most important conclusions of the listed references are cited, the reader should refer to the original articles for additional details.

Lower Esophagus

Achalasia

The treatment of achalasia (cardiospasm) depends upon the severity of symptoms and may be medical, by dilatation or by surgery. Since no method guarantees success, surgery usually is reserved for the most serious cases. The procedure of choice, it is generally agreed, is the Heller operation, which involves a long myomotomy of the lower esophagus and cardia.

Rees et al.[1] followed 84 patients who were treated surgically. There were no postoperative deaths. Of the 18 patients who had various cardioplastic procedures or esophagogastrectomy, 12 had poor results; eight required reoperation for peptic esophagitis. Forty-four patients had a myomotomy alone; eight had poor results chiefly because of esophagitis. The authors believe the Heller operation gives good results in 64 to 87 per cent of patients. Since the main problem after operation is esophagitis, if this is noted before operation additional operative measures to reduce acid are desirable.

Nemir et al.[2] also studied the causes of failure of esophagomyomotomy. They had 74 operations, with one death. There were 18 unsatisfactory results (24 per cent), and 11 of these patients required a further operation. Factors leading to a poor result include long standing disease and severe scarring about the esophagogastric junction at operation. A pyloroplasty and hiatus-hernia repair is necessary in addition to the Heller operation in patients who have both achalasia and esophagitis.

Esophageal Replacement

Portions of the esophagus may be replaced by stomach, jejunum or colon. In recent years the colon has become the favored organ, since adequate segments can nearly always be obtained. As a conduit to bypass extensive lye strictures or cancer, the descending colon, with a pedicle based on the left colic artery, has been used most often.

El-Domeiri et al.[3] have noted that 75 colon interposition operations were done in a 13-year period in the Memorial Hospital, New York City. The descending colon was used in 53 cases and has been the segment of choice since 1960. Esophageal resection and radiation therapy were added in favorable cases. Fifty-three patients survived the postoperative period. Five survived to 10 years.

Abbreviations Used	
CEA:	carcinoembryonic antigen
GR:	gastric resection
VA:	vagotomy and antrectomy
VGE:	vagotomy and gastroenterostomy
VP:	vagotomy and pyloroplasty
VPL:	vagotomy, pyloroplasty and ligature
VR:	vagotomy and gastric resection

When the stomach is used, the fundus usually has been drawn up into the chest to replace the excised esophagus. Tubes made from the greater curvature have usually been antiperistaltic and are not satisfactory. Yamagishi[4] has developed a tube, 25 to 30 cm, that is isoperistaltic. It is made from the greater curvature and based on the right gastroepiploic artery. He has used it in 17 patients without a death; there was no regurgitation after operation.

Jejunal interposition had been introduced by Merendino for benign lesions of the lower esophagus, such as esophagitis, achalasia or stenosis. This

*Address reprint requests to Dr. Welch at Suite 1104, Warren Bldg., 275 Charles St., Boston, Mass. 02114.