jnm/ ADJUNCTIVE MEDICAL KNOWLEDGE

Determining the Value of Diagnostic and Screening Tests

Barbara J. McNeil and S. James Adelstein

Peter Bent Brigham Hospital and Harvard Medical School, Boston, Massachusetts

Rapid advances in medical technology frequently lead to the development of new diagnostic procedures whose value should be determined before they are used widely. These values can be measured in terms of health and money. Health values relate to the accurate identification and successful treatment of disease; financial values relate to the husbanding of monetary resources expended for health services. In this discussion we shall review briefly the fundamentals of decision analysis as applied to medical diagnosis. We shall then show how these lead to a general approach to the evaluation of new tests using examples from our own work and that of others.

FUNDAMENTALS OF DECISION MAKING

A number of methods have been used to evaluate diagnostic procedures (*1–3*). For purposes of this discussion it is necessary to summarize the two used most frequently in analyses of radiologic and scintigraphic data: the decision matrix and the receiver operating characteristic (ROC) curve.

The decision matrix. The decision matrix (Table 1) relates results of a diagnostic test with a binary outcome (normal, abnormal) to clinical or pathologic findings, also with a binary outcome (disease, no disease). Five ratios can be derived from this table and are used to characterize such binary tests:

TABLE 1. A GENERAL DECISION MATRIX

Test results	Disease present (D+)	Disease absent (D−)	Total
Abnormal (T+)	a	c	a + c
Normal (T−)	b	d	b + d
Total	a + b	c + d	a + b + c + d

True-positive ratio (sensitivity) $= \frac{a}{a+b}$

True-negative ratio (specificity) $= \frac{d}{c+d}$

Accuracy $= \frac{a+d}{a+b+c+d}$

1. The true-positive (TP) ratio is the proportion of positive tests in all patients with disease, $a/(a + b)$, and is the *sensitivity* of the test.
2. The false-positive (FP) ratio is the proportion of positive tests in all patients without disease, $c/(c + d)$.
3. The true-negative (TN) ratio is the proportion of negative tests in all patients without disease, $d/(c + d)$, and is the *specificity* of the test.
4. The false-negative (FN) ratio is the proportion of negative tests in all patients with disease, $b/(a + b)$.
5. The likelihood ratio (L) of a test is the ratio of the TP ratio to the FP ratio:

$$\frac{a}{(a+b)} \times \frac{(c+d)}{c}.$$

(Another proportion $(a + d)/(a + b + c + d)$ is frequently called the *accuracy* of the test: it is the ratio of correct outcomes to all outcomes.)

It is important to emphasize that these ratios describe the sensitivity and specificity of the test; they cannot be used alone to determine the significance of a positive or negative test. An extended analysis is required to determine the probability that a patient does or does not have disease, given the test result. This calculation is frequently made using Bayes' theorem, in which the prevalence of disease in the patient population under study (the prior probability) is combined with the true-positive and true-negative values to form the posterior probabilities, i.e., the probability that the patient does or does not have disease, given the test result. In the special case where the patient population used to form the decision matrix has the same prevalence of disease as the larger general population, the probability of

Received Feb. 25, 1976; original accepted Feb. 26, 1976.

For reprints contact: Barbara J. McNeil, Dept. of Radiology, Harvard Medical School, 25 Shattuck St., Boston, Mass. 02115.

disease becomes a/(a + c), and the probability of no disease becomes d/(b + d).*

An example of the application of the decision matrix and Bayes' theorem can be taken from an analysis of liver scintigraphy. Scintigraphic examinations of the liver were performed on 650 patients referred to our nuclear medicine clinic, and the outcomes were categorized as normal or abnormal (*4*). Histologic examination by closed biopsy, surgery, or autopsy was obtained in 344 of the 650, and the results were categorized as pathologic or nonpathologic. The decision matrix for this subgroup of 344 patients is shown in Table 2; the TP ratio is 0.90, the FP ratio is 0.37, and the accuracy is 0.83. In the entire group of 650 patients, the prevalence of liver disease, or its prior probability [designated P(D+)], was determined by other means to be 0.66. For these 650 patients the posterior probability of having liver disease in the presence of an abnormal scintigram, designated P(D+|T+), is calculated from Bayes' theorem to be 0.83, i.e., a positive test changes the probability of liver disease from 0.66 to 0.83. On the other hand, the posterior probability of having liver disease if the scintigram is normal, designated as P(D+|T−), is calculated to be 0.24, i.e., a negative test changes the probability from 0.66 to 0.24, a considerably greater difference.

The better the test (TP ratio near 1.00), the greater the proportion of diseased patients identified and the smaller the proportion of patients without disease who have positive results (FP ratio near 0.00). A good test, therefore, provides high posterior probabilities of either having or not having disease after positive or negative test results, respectively, have been obtained. As these TP and FP ratios vary from 1.0 and 0.0, respectively, the posterior probabilities decrease: the reliability with which we can estimate the state of health of a patient decreases. Similarly, as the prevalence of disease decreases for a given pair of TP and FP ratios, the reliability of the estimates also decreases. These relationships can be expressed graphically using Bayes' theorem (Fig. 1). This presentation assumes a perfectly sensitive test (TP ratio = 1.00) with varying false-positive ratios; the latter have been used to generate a series of hyperbolic curves (*5*). These curves show that high posterior probabilities of disease (>90%) can be achieved for false-positive ratios greater than 0.10 only if the prevalence of disease is high (>45%). For lower disease prevalences, much lower false-positive ratios are necessary. In other words, diagnostic tests performed on patients coming from a population with a low prevalence of disease must be much more accurate than tests performed on patients coming from a group with a higher prevalence in order to achieve the same level of posterior probabilities.

The receiver operating characteristic curve. When tests do not have binary outcomes, but rather have a continuum of values (any one of which can be selected as the boundary between normal and abnormal), the true and false positive ratios vary with the value selected as the cutoff point. Routine chemistry examinations and radioimmunoassays are examples of such tests. We can graphically visualize the effect of changes in the cutoff point on test sensitivity by using a receiver operating characteristic (ROC) curve, a plot of the true-positive ratio against the false-positive ratio for varying cutoff points (Fig. 2).

Similar ROC curves can also be constructed for some situations not involving a single test with a continuous scale of outcomes. For example, this plot can be used for visual detection situations (e.g., radiographic or scintigraphic interpretations) in which the observer is required to make his interpretation at varying thresholds (i.e., strict or lax); each threshold criterion corresponds to a different pair of true-positive-to-false-positive ratios (*1*). This technique has been applied extensively by Metz et al in their determination of optimum radiographic film–screen combinations (*6*).

When a series of consecutive diagnostic modalities (i.e., history, physical examination, varying levels of tests) is used for disease detection, then an ROC

* The use of disease prevalence rates for prior probabilities can only be applied to the population as a whole. For any individual patient the constellation of clinical and laboratory findings may change the prior probabilities above or below the average values.

TABLE 2. DECISION MATRIX FOR HEPATIC SCINTIGRAPHY IN THOSE WITH PATHOLOGIC CORRELATION (344 PATIENTS) (4)

	Histologic findings	
Scan results	Disease present (D+)	Disease absent (D−)
Abnormal (T+)	231 (TP = 0.90)	32 (FP = 0.37)
Normal (T−)	27 (FN = 0.10)	54 (TN = 0.63)

Calculation of posterior probabilities for entire group of 650 patients having hepatic scintigraphy.
Prior probabilities: P(D+) = 0.66; P(D−) = 0.34
Posterior probabilities as calculated from Bayes' theorem:

$$P(D+|T+) = \frac{TP \times P(D+)}{TP \times P(D+) + FP \times P(D-)} = 0.83,$$

$$P(D+|T-) = \frac{FN \times P(D+)}{FN \times P(D+) + TN \times P(D-)} = 0.24.$$

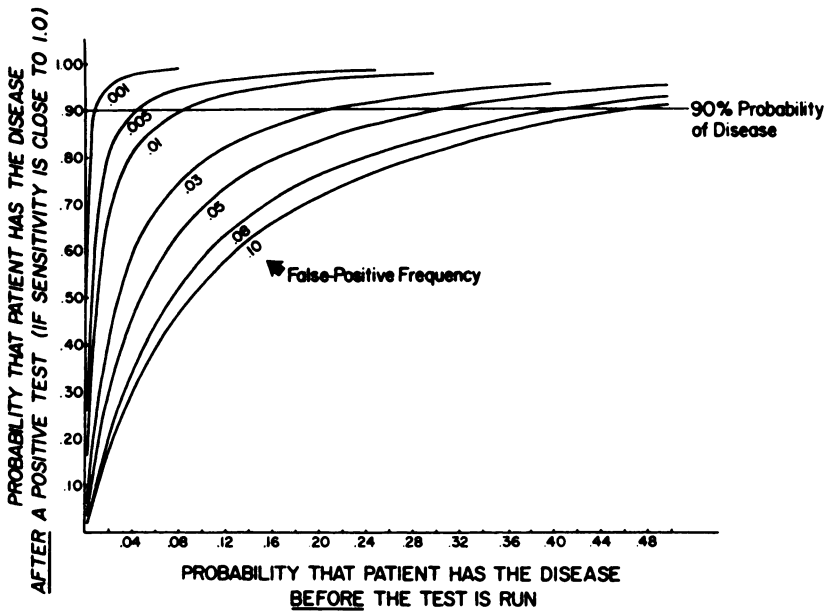

FIG. 1. Probability graph to help physician determine probability that patient with positive test has disease if test is extremely sensitive (TP ratio = 1). Each curve represents different false-positive ratio. (Reprinted with permission from *N Engl J Med* 291: 116, 1974.)

curve can be plotted to show the gains achieved by introducing each modality. In the studies using this approach, an initial ROC curve is obtained after performing a disjunctive analysis of the varying modalities as they are introduced (*7,8*): First, is A_1 present? Then, is A_1 or A_2 present ($A_1 \cup A_2$)? Finally, is A_1 or A_2 or $A_3 \cdots$ present ($A_1 \cup A_2 \cup A_3 \cdots$)? If an additional test is performed only on selected subsets of the initial group, then its effect can be visualized by conjoining A and B ($A_1 \cap B$; $A_1 \cup A_2 \cap B$; $A_1 \cup A_2 \cup A_3 \cdots \cap B$) at each stage of the diagnostic process.

GENERAL CONSIDERATIONS

Health values. "Health values" associated with diagnostic tests are best understood through a simplified model of the diagnostic and therapeutic process (Fig. 3). In this model, a patient with a symptom complex or a syndrome enters the diagnostic process. At the first decision node (Square 1), a diagnostic test is either performed or not performed. In the former case, the first chance node (Circle 1) depicts the results of this test in terms of the amount of information achieved. The test can provide new information not previously available to the diagnostician (+), it can provide no additional information (0), or lastly it can provide misleading and therefore negative information (—). This stage of the diagnostic process provides the first point at which we can measure the value of a diagnostic test and satisfies those who would claim that the ultimate test of a diagnostic procedure is its ability to sort patients with regard to specific diseases.

After diagnostic testing, treatment is instituted (Square 2a) and the results of treatment are depicted at the second and final chance node (Circle 2a). These outcomes represent a continuum of states ranging from perfect health (a cure) to death. This stage of the diagnostic process provides the second point at which we can measure the value of a diagnostic test and satisfies more operationally oriented physicians who claim that the ultimate test of diag-

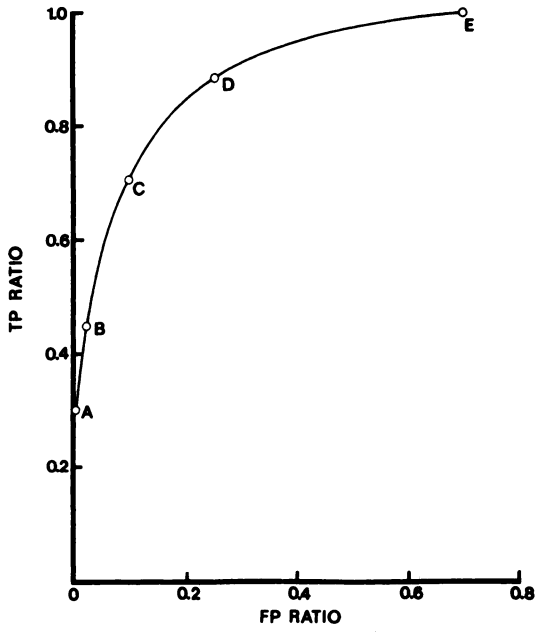

FIG. 2. Hypothetical ROC curve. Vertical scale is TP ratio, and horizontal scale is FP ratio. At extreme point A, test has poor sensitivity (TP ratio = 0.30) but good specificity (TN ratio = 1.00, FP ratio = 0.00). At other extreme, E, test has high sensitivity (TP ratio = 1) but poor specificity (TN ratio = 0.30, FP ratio = 0.70). (Reprinted with permission from *N Engl J Med* 293: 212, 1975.)

DIAGNOSTIC AND THERAPEUTIC PROCESS: HEALTH MEASURES OF DIAGNOSTIC VALUE

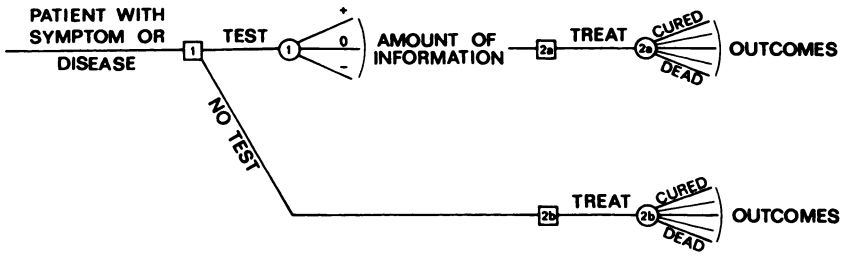

FIG. 3. Model of diagnostic process. At first decision node (square node 1), test is either performed or not performed. In former case it can provide additional information (+), no additional information (0), or misleading information (—). After testing, treatment is instituted with continuum of therapeutic outcomes.

nosis is the extent to which it can save lives, restore health, or alleviate suffering. If the test is not performed, treatment is instituted on the basis of available information (Square 2b) with the same continuum of outcomes.

Financial values. The financial aspects of the diagnostic and therapeutic process can also be considered. In broad terms, the financial value of a test lies in its ability, if truly negative, to eliminate costs associated with unnecessary diagnostic procedures and therapeutic regimens and, if truly positive, to eliminate financial costs caused by the progression of untreated disease. These benefits are difficult to measure directly. Therefore, three other financial measures are frequently used in evaluating diagnostic tests: (A) the total cost of diagnosis and therapy once the test is introduced; (B) the average cost of achieving a given unit of health (e.g., case finding, life saving) by use of the test; and (C) the marginal cost of achieving one additional unit of health by one procedure over another.

SPECIFIC ILLUSTRATIONS

A number of diagnostic procedures have been evaluated by one or more of these approaches. The remainder of this review will discuss several of these in the context of the schema proposed in Fig. 3.

The first measure used to evaluate diagnostic tests is concerned with the amount of information provided. In this discussion "information" means knowledge obtained from an investigation or study and is not restricted to the meaning usually used in information theory, i.e., a reduction in uncertainty. Thus, there are three levels of information available from a diagnostic test. First, it can be used for research purposes and can lead to a greater understanding of underlying pathophysiologic mechanisms. Measurement of glucocerebroside-splitting enzymes in patients with Gaucher's disease falls into this category. Second, a diagnostic test can be of prognostic value to the physician or patient, regardless of subsequent therapeutic measures. A positive test identifying patients with pancreatic carcinoma and many negative tests fall into this category. Third, a test can have an effect on subsequent therapy. Thus, the first question to ask of a new diagnostic test in relationship to its diagnostic value is "Does it provide additional information?" or can its results be predicted frequently and accurately from other data, for example, from the history of the patient, from his physical examination, or from the results of other laboratory tests.

Example 1. *Results of the test cannot be predicted perfectly from the clinical manifestations of disease: the uptake of radioactive iodine.*

In this study, radioactive iodine (RAI) uptakes were made on 90 patients with suspected thyroid abnormalities, and these uptakes were categorized as high, low, or normal (9). In addition, and independently, the presence or absence of 21 symptoms and 18 signs was scored. By a modification of Bayes' theorem and with use of computer analysis, the likelihood of a high, low, or normal value in each of these patients was determined by a quantitative analysis of these clinical signs and symptoms. The measured results obtained in a given patient were compared with the results predicted from the Bayesian computer analysis (Table 3): 77% (17/22) of the high RAI values could be predicted from clinical manifestations, 87% of the low values could be predicted, and 74% of the normal values could be predicted. Thus, in this situation, on the average, 21% of patients had test results with information which could not be predicted from clinical signs and symptoms. In these patients the RAI uptake test provided additional information not obtainable from the clinical examination alone.

If the outcome of the test is not predictable, then one must determine if the test improves diagnostic accuracy and if other diagnostic tests are superior in terms of sensitivity or specificity.

Example 2. *Introduction of the test improves diagnostic accuracy (sensitivity and/or specificity): the lung scan in the definition of pleuritic chest pain; serum folate assays in the detection of alcoholism.*

The hospital records of 97 patients under 40 years of age who presented with acute pleuritic chest pain and who were suspected of having pulmonary emboli were reviewed in an attempt to define the role of pulmonary scintigraphy in this clinical setting (7). The results were plotted in the form of a ROC curve with the true-positive ratio corresponding to abnormal finding(s) in patients with pulmonary embolism and a false-positive ratio corresponding to abnormal finding(s) in patients without pulmonary embolism. A disjunctive analysis of seven symptoms or historical findings, 14 clinical signs, and 11 other laboratory tests showed that maximum information could be obtained from two points of history, one physical finding, and the chest x-ray (Fig. 4, triangles). Using the two pieces of historical information (i.e., is the patient postoperative and does he have old venous disease) allows detection of 65% of the patients with pulmonary embolism and falsely categorizes 18% of those without disease. Introducing data from the physical examination increases the detection rate to 80% but also increases the false-positive rate to 23%. If radiographic findings are also introduced and the patient has any of the preceding signs or a pleural effusion, then 95% of the patients with disease are detected, but at the same time 40% of the patients without pulmonary embolism also have positive findings.

When a lung scan read with certain strict criteria for abnormalities is combined with each point in the preceding disjunctive process, the ROC curve (Fig. 4, circles) becomes more concave. For example, before the lung scan was introduced, a chest radiograph or several pieces of historical information identified 95% of the patients with disease. With the lung scan, 95% are still detected but the FP ratio drops from 40% to 5%. Introduction of the lung scan has markedly shifted the ROC curve to the left, thus enhancing specificity at each stage of the diagnostic workup. Performance of a lung scan in all patients (irrespective of symptoms and x-ray findings) increases sensitivity slightly, from 95% to 100%, but nearly doubles the cost (7). Thus, for young patients with pleuritic pain, the lung scan is better than other diagnostic tools in terms of increased specificity and slightly better in terms of increased sensitivity.

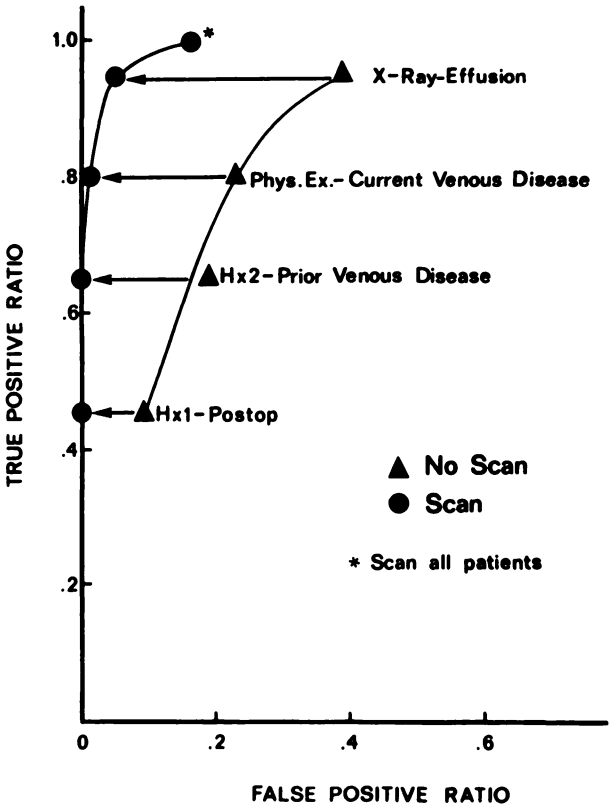

FIG. 4. Experimental ROC curve for diagnosis of pulmonary embolism in young patients presenting with pleuritic chest pain. Triangles represent new variables created by disjoining individual variables. Circles represent subsequent combinational analysis. Arrows indicate that lung scans were performed only on patients belonging to groups created by above disjunctive process, and asterisk indicates that lung scans were performed on all patients irrespective of other symptoms, signs, or laboratory tests.

TABLE 3. MEASURED AND PREDICTED RAI UPTAKES: MEASURE OF ADDITIONAL INFORMATION (9)

	No. predicted			
Measured	High	Low	Normal	% Predicted
High	17	2	3	72
Low	0	26	4	87
Normal	2	8	28	74
			Average	79%

This approach is also exemplified through another diagnostic problem, the diagnosis of alcoholism in a series of random hospital admissions (8). Hepatomegaly was found in 31% of alcoholic patients and the alcoholics were detected by using a diagnostic criterion of either hepatomegaly, an abnormal mean corpuscular red cell volume (MCV) increased the detection rate to 77% and simultaneously increased the false-positive ratio to 14%. Over 90% of the alcoholics were detected by using a diagnostic criterion of either hepatomegaly, an abnormal MCV value, or an abnormal SGOT level. At the same time, however, nearly 30% of nonalcoholics

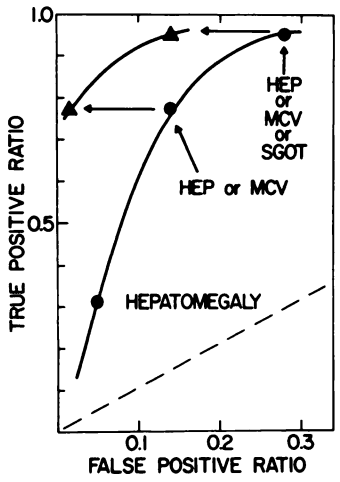

FIG. 5. Experimental ROC curve for diagnosis of alcoholism in random hospital admissions. Circles indicate disjoined variables and triangles indicate effect of abnormal serum folate level on groups isolated from disjunctive process. (Courtesy of Dr. David E. Drum.)

TABLE 4. IMPACT OF PELVIMETRY ON NEONATAL MORBIDITY AND MORTALITY (*10*)

	n	Cesarean sections (%)	Neonatal mortality (%)	Neonatal morbidity (%)
Pelvimetry	151	43	3	13
No pelvimetry	154	32	5	20

TABLE 5. CORRELATION OF SCREENING RESULTS (CLINICAL EXAMINATION AND MAMMOGRAPHY) WITH DISEASE STATE IN STUDY GROUP (*11*)

	Disease		
Test	D+	D−	No. examinations
T+	127	960	1,087
T−	63	63,660	63,723
Total	190	64,620	64,810

Sensitivity: 127/190 = 67%
Specificity: 63,660/64,620 = 98%

were falsely included by this diagnostic criterion. As in the case of pulmonary embolism, addition of a specific diagnostic test significantly reduced the false-positive ratio (Fig. 5, triangles), the analog of the lung scan being a serum folate measurement. For example, in patients with hepatomegaly, an abnormal MCV value, or an abnormal SGOT as well as a serum folate level below 3 ng/ml, the true-positive ratio remained above 90%, but the false-positive ratio dropped dramatically from 28% to 14%. For the detection of alcoholism in this small random patient population, the folate radioassay proved highly effective in improving the specificity of the preliminary screening tests.

A second measure of value for diagnostic or screening procedures relates diagnosis and therapy and is concerned with short- or long-term health outcomes evaluated from either randomized or nonrandomized studies. Optimally, randomized series are traditionally preferred for this purpose. Frequently, however, difficulties arise in attempts to institute such studies when diagnostic rather than screening or therapeutic information is to be obtained. Nonetheless, at least two studies in the radiologic literature have used double-blind randomized trials, one for a diagnostic test and one for a screening program.

Example 3. *Introduction of the test leads to a reduction in morbidity and/or mortality: evaluation through randomized series of a diagnostic test (pelvimetry) or a screening program (mammography).*

In 1962 Derk Crichton started a randomized trial on the use of pelvimetry in patients referred by practicing obstetricians for this procedure (*10*). Of the 305 patients referred, pelvimetry was performed on 154 and not performed on the remaining 151. Differences in the two groups were observed (Table 4). For the mothers, there was a larger percentage of cesarean sections in those having pelvimetry than in those refused the procedure: 43% versus 32%. More striking and more important, however, was the marked difference in the short-term infant mortality and morbidity. There were 2% fewer infant deaths in the group having pelvimetry and 7% fewer morbid events. The value of pelvimetry in this clinical situation appears to be clear: it markedly reduces short-term morbidity and mortality in infants.

A second example in the radiologic literature involves a randomized trial of a screening test and uses long-term health outcomes as a measure of value. The Health Insurance Plan's evaluation of mammography in the detection of early breast cancer involved over 61,000 women (*11–13*). Thirty thousand were selected to have an annual screening consisting of a breast examination and mammography, and a similarly sized control group was selected to have routine health care only. Study patients were examined for 3 years and complete followup information was available on all patients 5 years later. In the study group 190 patients developed cancer; 127 of these were detected on screening, leading to a sensitivity of 67% (Table 5). Of the 127 carcinomas detected, 33% were detected by mammography alone, 44%

were detected by clinical examination alone, and 23% were detected by both modalities. Of the 64,620 patients without disease, 63,660 had negative examinations, giving a specificity of 98% for the screening regimen (Table 5).

It is important to consider the health benefits from this screening program because they show that an intermediate stage in the diagnostic process, case finding, may not be equivalent to the desired endpoint, life saving. Over the entire time span, fewer people died in the study group than in the control group. At 5 years, the respective case fatality rates (when corrected for a 1-year lead time in detection due to screening) were 27.9% and 42.1%. However, most of this benefit arose in the 50–59-year age group where half as many patients died in the study group as in the control group. Below the age of 50 years approximately the same number of patients died in both groups, and in patients over 60 a slightly smaller number of patients died in the study group.

Evaluation of health outcomes from randomized series of patients studied with and without a diagnostic test is not always possible. Therefore, such evaluations must frequently be achieved using health outcomes from nonrandomized groups of patients believed comparable in certain critical ways but differing in other ways, for example, in their geographic location, their time of study, or the specificity of their treatment.

Example 4. *Two similar but geographically separated patient populations studied at the same time show that a new test can reduce morbidity: the radiometric assay for digoxin.*

The radiometric assay for digoxin was developed by Haber and Smith at Massachusetts General Hospital (MGH) and its effects subsequently studied by their collaborators (*14*). During the time in which the assay was available on a daily basis at MGH and only intermittently available at another Harvard teaching hospital (Peter Bent Brigham Hospital), the incidence of digitalis toxicity was compared. At the Massachusetts General Hospital the incidence was 4%, while at the Peter Bent Brigham Hospital it was 10%. The hazards of such comparative analyses are plain, but they do represent a method of determining the potential impact of a test.

Example 5. *The course of disease is unchanged in the same hospital before, during, and after the introduction of a test: the impact of brain scanning on the surgical treatment of brain tumors.*

A study on the comparison of health outcomes in comparable patient populations before and after the introduction of a diagnostic test is exemplified in a study at Johns Hopkins Hospital in patients with primary and secondary brain tumors (*15*). In 1962 approximately 300 brain scans were done at Johns Hopkins Hospital, and in 1972 3,000 were performed. During this period the sensitivity of the brain scan was constant at approximately 77%. Differing factors in this decade were the number of patients operated on and the duration of symptoms prior to operation. For example, the number of patients operated on in 1962 was about 60, and in 1972, approximately 70. The duration of symptoms prior to operation was nearly 4 years in the former case and less than 1 year in the latter case. Despite these two favorable indices, the long-term health outcomes of patients with primary and secondary brain malignancies did not change over the decade 1962–1972. The percentage of patients surviving 50 months after surgery during the time period of this study was relatively constant for all tumor types. These data, therefore, indicate that while the number of brain scans performed at Johns Hopkins increased by a factor of over 10 in the past decade, there was no associated increase in survival rates in patients with brain tumors. From the survival point of view, then, the value of brain scans is not great. However, other issues must be considered in this case and perhaps the most obvious one relates to the importance of the prognostic information made available as a result of a normal test in a patient without disease.

Example 6. *The test provides new information in a well-studied representative patient population, but its ultimate value as determined by relating the results to the natural history of disease is not great: diagnostic examinations searching for patients with renovascular hypertension.*

Long-term health outcomes were compared in two theoretical groups of patients with renovascular disease (*16,17*). One group was evaluated with the intravenous pyelogram (IVP) and/or renogram and then treated surgically when indicated; another group of patients was not studied by either of these tests but instead was treated empirically with antihypertensive medication. This study was designed to answer the question "What is the value of evaluating all hypertensive patients for renovascular disease?" If the results of the specific surgical and the nonspecific medical treatment were the same, then there would be no health value associated with the routine performance of either of these tests. The technique used to answer this question can be expressed explicitly in the form of decision flow diagrams (Figs.

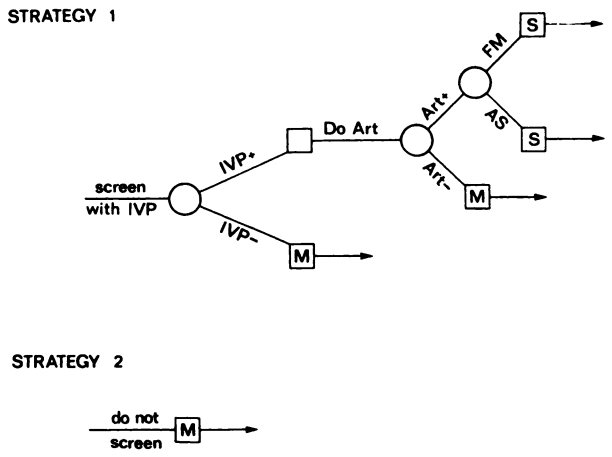

FIG. 6. Decision flow diagrams for treatment of patients with hypertensive disease. Squares denote decision nodes and circles denote chance nodes. Strategy 1: Testing is performed on all patients with hypertension. If intravenous pyelogram is abnormal (IVP+), arteriography (Art) is always performed. When IVP or arteriogram is negative, medical therapy (M) is followed. When arteriogram is positive (Art+) and patients are found to have either fibromuscular (FM) or atherosclerotic (AS) disease, surgical treatment is chosen (S). Outcomes of operation and medicine are shown in Fig. 7. Strategy 2: No testing is done and all patients are treated medically (M).

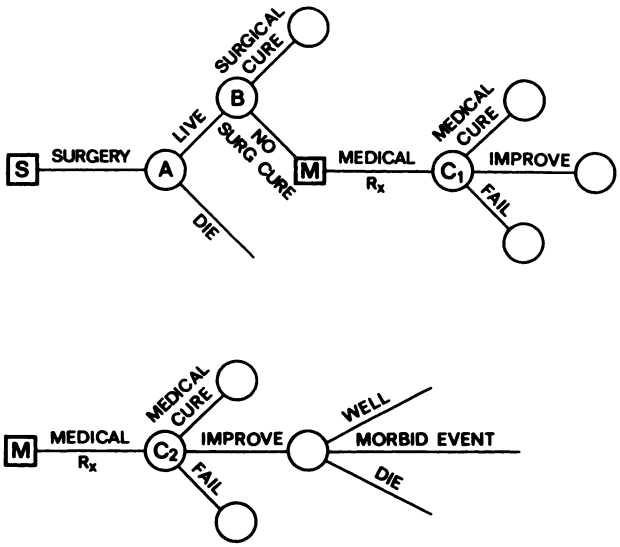

FIG. 7. Outcomes associated with surgical and medical regimens. Results of surgical management of patients with renovascular disease are detailed at chance nodes A and B. Results of primary or supplemental medical therapy are expressed at nodes C_1 and C_2. All terminal chance nodes are associated with three possible outcomes: cure, morbid event, or death. (Reprinted with permission from *N Engl J Med* 293: 222, 1975.)

6 and 7). In the diagnostic testing situation (Strategy 1), all patients with an abnormal test, here an IVP, undergo arteriography. All patients with either fibromuscular or atherosclerotic renovascular disease are treated surgically. Patients with normal urograms and arteriograms are treated medically. In the no-test situation (Strategy 2), all patients are treated medically. The results of surgery can be followed in Fig. 7 (top). Some patients do not survive the operation; those who do are either cured or not cured as a result. Those not cured by surgery receive medical therapy (node C_1) which results in medical cure, improvement (incomplete reduction in blood pressure), or failure (no change in blood pressure). If the initial treatment is medical, either because screening was not performed or because the patients screened were not suitable surgical candidates, the results can be categorized in a similar manner (node C_2). For both strategies the chance node at the end of each branch of the decision diagram indicates the patient's state, i.e., he is well, he has suffered a nonfatal morbid event (coronary heart disease, cerebrovascular accident), or he is dead (postoperatively or as a consequence of his hypertension).

The difference between the numbers of well patients, patients suffering a morbid event, or dead patients in the two strategies is a measure of the value of the diagnostic testing strategy. These numbers were calculated using data from the Cooperative Study on Renovascular Disease on the sensitivity and specificity of the IVP and renogram and from the Framingham Study on the probabilities of fatal and nonfatal morbid events (*18–21*). Several initial diastolic blood pressures (90–135 mm Hg) and two compliances (50% and 84%)* were used for this purpose.

These calculations indicate little difference in outcomes for patients with renovascular disease regardless of the endpoint measured (Fig. 8). In all cases the maximum difference is about 10% and, assuming an incidence of renovascular disease of 10%, this figure extrapolates to a difference of only 1% for the total hypertensive population. When deaths are enumerated within these small differences, medical therapy is better than surgical therapy except at diastolic blood pressures higher than 135 mm Hg. When nonfatal morbid events are enumerated, surgery is always better, and when the numbers of well patients are considered, medical therapy is better at high compliance rates and surgical therapy at low compliance rates. These results indicate that while there is some potential benefit to the identification and surgical treatment of patients with reno-

* In this discussion we have used compliance rates to reflect not only the degree to which patients follow advice, but also the vigor with which physicians prescribe antihypertensive medication. The 50% compliance rate is probably the most common one in a typical American population and indicates that 50% of known hypertensive patients are lost to followup, 25% are cured, and 25% are improved. The 84% compliance rate is an unusual one found only in specially designed programs where 84% of the patients are cured and 16% are lost to followup.

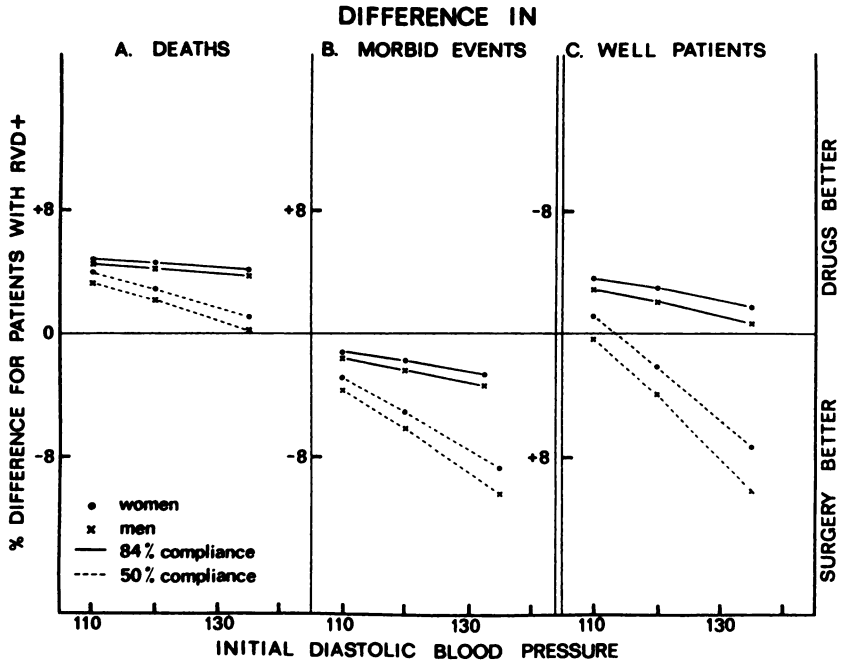

FIG. 8. Difference in number of deaths (Panel A), nonfatal morbid events (Panel B), and well patients (Panel C) as function of sex, compliance, and initial diastolic blood pressure. Vertical scale represents difference between results of surgical and medical therapy in 100 patients with renovascular disease of both atherosclerotic (AS) and fibromuscular (FM) types. Points above horizontal line indicate superiority of medical therapy, and points below indicate superiority of surgical therapy.

vascular disease, the mode of treatment, and hence diagnosis, contributes much less to ultimate prognosis than does initial blood pressure and compliance.

The financial costs of the diagnostic testing strategy can also be calculated. The average cost of a case finding using either the IVP or renogram is approximately $2,000, while the average cost of curing a patient with surgery is nearly ten times more. Extrapolation of these average cost calculations to the total American hypertensive population shows that the costs of diagnosis and surgery for the entire hypertensive American population would cost billions of dollars, nearly 60% more than the medical regimen.

The elementary principles and the clinical examples reviewed in this article have been presented in order to provide a systematic approach to the measurement of the health and financial values of diagnostic and therapeutic intervention. It is clear that measurement of these values is becoming increasingly important as new and untested procedures and instruments are introduced for both diagnosis and therapy. Hopefully, with knowledge of these values, the resources allocated for medical care can be optimally utilized.

ACKNOWLEDGMENTS

This work was supported in part by USPHS grants GM-02201 and GM-18674.

REFERENCES

1. LUSTED L: *Introduction to Medical Decision Making.* Springfield, Ill., CC Thomas, 1968

2. BARNOON S, WOLFE H: *Measuring the Effectiveness of Medical Decisions.* Springfield, Ill., CC Thomas, 1972

3. MCNEIL BJ, KEELER E, ADELSTEIN SJ: Primer on certain elements of medical decision making. *N Engl J Med* 293: 211–215, 1975

4. DRUM DE, CHRISTACOPOULOS JS: Hepatic scintigraphy in clinical decision making. *J Nucl Med* 13: 908–915, 1972

5. KATZ MA: A probability graph describing the predictive value of a highly sensitive diagnostic test. *N Engl J Med* 291: 115–116, 1974

6. METZ CE, GOODENOUGH DJ, ROSSMAN K: Evaluation of receiver operating characteristic curve data in terms of information theory with applications to radiography. *Radiology* 109: 297–304, 1973

7. MCNEIL BJ, HESSEL SJ, BRANCH WT, et al: Measures of clinical efficacy. III. The value of the lung scan in the evaluation of young patients with pleuritic chest pain. *J Nucl Med* 17: 163–169, 1976

8. DRUM DE: Detection of alcoholism in general hospital patients: In preparation

9. ADELSTEIN SJ: PARKER R, WAGNER HN: First phase in the objective evaluation of new diagnostic tests. *Invest Radiol* 5: 153–163, 1970

10. CRICHTON D: The accuracy and value of cephalopelvimetry. *J Obstet Gynaecol Br Commonw* 69: 366–378, 1962

11. SHAPIRO S, STRAX P, VENET L: Periodic breast cancer screening in reducing mortality from breast cancer. *JAMA* 215: 1777–1785, 1971

12. SHAPIRO S: Screening for early detection of cancer and heart disease. *Bull NY Acad Med* 51: 80–95, 1975

13. HUTCHINSON G: Personal communication, 1976

14. DUHME DW, GREENBLATT DJ, KOCH-WESER J: Reduction of digoxin toxicity associated with measurement of serum levels. *Ann Intern Med* 80: 516–519, 1974

15. GEORGE RA, WAGNER HN: Ten years of brain tumor scanning at Johns Hopkins: 1962–1972. In *Noninvasive Brain Imaging*, DeBlanc HJ, Sorenson JA, eds. New York, Society of Nuclear Medicine, 1975, pp 3–16

16. MCNEIL BJ, VARADY PD, BURROWS BA, et al: Measures of clinical efficacy. Cost-effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. *N Engl J Med* 293: 216–221, 1975

17. MCNEIL BJ, ADELSTEIN SJ: Measures of clinical efficacy. The value of case finding in hypertensive renovascular disease. *N Engl J Med* 293: 221–226, 1975

18. BOOKSTEIN JJ, ABRAMS HL, BUENGER RE, et al: Radiologic aspects of renovascular hypertension. I. Aims and methods of the radiology study group. *JAMA* 220: 1218–1224, 1972

19. BOOKSTEIN JJ, ABRAMS HL, BUENGER RE, et al: Radiologic aspects of renovascular hypertension. II. The role of urography in unilateral renovascular disease. *JAMA* 220: 1225–1230, 1972

20. KANNEL WB, GORDON T: Some characteristics related to the incidence of cardiovascular disease and death: Framingham study, 16-year follow-up. In *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*. Washington, D.C., Government Printing Office, 1970

21. KANNEL WB, GORDON T: Survival following certain cardiovascular events. In *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*. Washington, D.C., Government Printing Office, 1970

New MIRD Committee Publications

Pamphlet #1, Revised—A Revised Schema for Calculating the Absorbed Dose from Biologically Distributed Radionuclides—12 pp.

Describes how to calculate the radiation dose and establishes a mathematical formalism for simplifying dose calibrations. This number is a revision of Pamphlet #1, which was first published February 1968 as part of MIRD Supplement #1. It introduces the term "S," the absorbed dose per unit cumulated activity, and offers more information on the requirements of a kinetic model.

$6.75 with binder; $4.50 without binder.

Pamphlet #10—Radionuclide Decay Schemes and Nuclear Parameters for Use in Radiation-Dose Estimation—Approx. 125 pp.

Provides essential radioactive decay scheme information in convenient form on more than 120 medically important radionuclides. This publication updates and supersedes Pamphlets 4 and 6 which provided data for 54 radionuclides. In loose-leaf binder format for ease of updating and adding additional radionuclides.

$8.75 with binder; $6.50 without binder.

Pamphlet #11—"S" Absorbed Dose per Unit Cumulated Activity for Selected Radionuclides and Organs—Approx. 255 pp.

The tabulated values of "S" in this publication simplify dose calculations. Instead of requiring separate consideration of each radiation of the decay scheme and its associated absorbed fraction, the "S" tabulation permits dose calculations by simply referring to a single table entry for each organ combination. This pamphlet provides "S" values for 117 radionuclides plus 6 parent and short-lived daughter combinations as a uniformly distributed source in 20 source organs irradiating 20 target organs which include ovaries, red bone marrow, testes, and total body. In loose-leaf binder format for ease of updating and adding additional radionuclides and source and target organs.

$10.20 with binder: $7.95 without binder.

Extra binders available at $3.75 each.

Please address all orders to:

MIRD Pamphlets
Society of Nuclear Medicine
475 Park Avenue South
New York, N.Y. 10016

CHECKS MADE PAYABLE TO THE "SOCIETY OF NUCLEAR MEDICINE" OR A PURCHASE ORDER MUST ACCOMPANY ALL ORDERS.