2.
Evaluation Model

2

Evaluation Models

To understand the strengths and weaknesses of evaluation, one must keep in mind its fundamental purpose: to inform those who make decisions. The inferences drawn from an evaluative study of a diagnostic procedure are important, because they may influence decisions about its use. Therefore, one must carefully examine the assumptions underlying any approach to evaluation to determine the extent to which they provide direction or misdirection for decisions about the use of a procedure.

Ideally, one would want an evaluative study of a diagnostic procedure to provide precise measurement of the full array of medical and nonmedical benefits, risks, and costs resulting from alternative diagnostic strategies applied to a patient with specific signs, symptoms, and risk factors. Additionally, the ideal study would provide some method for comparing one kind of benefit or cost with another. Then, the decisionmaker could apply the findings in a relatively straightforward fashion to choose among alternative strategies.

Reality is not so accommodating. No studies of medical procedures deal with all possible benefits, risks, and costs. Nor do they often provide a method for collapsing multiple dimensions of benefit and cost into a single composite measure. Often, they aggregate findings over patients with diverse signs, symptoms, and risk factors, thereby limiting their usefulness for diagnostic decisions. Research costs, unfeasibility of measurement, and inability to make value tradeoffs dictate that all evaluative studies stop far short of the ideal. Yet, even partial approaches provide evidence that is useful. For example, most studies of the four X-ray procedures discussed here do not explicitly consider issues of cost, but some do reach important findings about the magnitude of the clinical benefits to be derived from the procedure. Indeed, the most important feature of diagnostic-procedure evaluation is not whether it has considered all possible dimensions of benefit, risk, and cost but what it uses as the "endpoint" of the evaluation (79). A study of the cost or accuracy of diagnosis may carry a message for medical decisionmakers that is quite different from a study that chooses as its endpoint the patient's ultimate health status or costs.

In our review of the literature on the four X-ray procedures, we were able to classify all studies into one of five broad categories based largely on their evaluative endpoints:

- studies of diagnostic "efficiency,"
- studies of diagnostic yield,
- Ž studies of high-yield criteria,
- studies of diagnostic information, and
- . studies of outcomes.

Each of these approaches is discussed in detail below. Each has its advantages and disadvantages. One must consider an evaluative model in terms of the cost and feasibility of anaIysis and the reasonableness of the assumptions about factors outside its scope. In short, no single category of evaluation is best in all cases; each has a place in the evaluator's bag of techniques.

STUDIES OF DIAGNOSTIC EFFICIENCY

Diagnostic efficiency is used here to refer to the capacity of a test to meet its immediate objective: correct diagnosis, A perfectly efficient test is one in which either of two kinds of error would never occur: 1) the detection of disease when in truth none is there (false positives), and 2) the failure to detect disease when it exists (false negatives). A perfectly inefficient test

would always be in error. Between these two extremes lie **an** infinite number of combinations of frequency of the two kinds of error.

The classic measures of diagnostic efficiency corresponding to each of the two errors are sensitivity and specificity. Sensitivity is the proportion of individuals with disease whose test results are positive (true-positive rate); specificity is the proportion of normal individuals whose test results are negative (true-negative rate).

Knowledge of these error rates is essential to the appropriate interpretation of test results by clinicians. Only by knowing the test's sensitivity and specificity and the prevalence of the suspected disease in the population can the clinician accurately interpret a positive or negative finding on a test (55). Though probably no clinician is prepared to perform sophisticated analyses of pretest and posttest disease probabilities for every test he or she orders, implicit processing of such information does take place, and accurate assessments of these two components of diagnostic efficiency in different situations are useful in this regard.

The use of diagnostic efficiency to assist in the choice among alternative diagnostic strategies is another matter altogether. Here, we say that if one test is both more sensitive and more specific than another, it is more efficient and may be the procedure of choice, although differences in the cost of each test should also be considered before such a decision can be made.

The comparison among tests is frequently not so straightforward. The more sensitive test is often less specific than its competitor. Then, it is impossible to avoid considering the implications of the two kinds of errors. If the consequences of a false positive are major—perhaps the performance of expensive or risky followup tests or even application of inappropriate and dangerous therapy—high specificity is desirable. Conversely, if the implication of a false negative is dire—the deterioration of the patient's condition to an unredeemable state—then a highly sensitive test may be preferred. Indeed, whether either test is worth performing at all cannot be determined by examination of its sensitivity and specificity alone. One must know the implications of each kind of test result (true positive, false positive, true negative, false negative) for the health and well-being of the patient and for the cost of medical care (54). Nevertheless, these components of diagnostic efficiency must be known if such analysis of the test's implications for outcomes is to take place. The problem with studies using these efficiency measures is that they often draw inferences about the usefulness of one diagnostic strategy versus another without further analysis.

Summary indexes of diagnostic efficiency have been developed by collapsing sensitivity and specificity into composite measures. The most common is diagnostic accuracy, * defined as the proportion of all test results that are correct, and measured as the sum of true-positive and true-negative results divided by the total number of tests. This index of efficiency is attractive because it is comprehensible, but it is more dangerous than the separate use of sensitivity and specificity because it assumes that the two kinds of testing error (false positive and false negatives) are equally important. This assumption is arbitrary and, for the vast majority of diagnostic tests, invalid. By using it, the investigator has bought into the equal valuation of the two kinds of error. Inferences, usually incorrect, are almost inevitable.

Diagnostic accuracy has also been criticized because, unlike sensitivity and specificity, it systematically varies with the prevalence of disease in the sample of patients under study. If the sensitivity of a test exceeds its specificity, then the higher the prevalence of disease in the study sample, the higher will be the measured diagnostic accuracy. This property is viewed as dangerous by some, for it implies that by manipulating the selection of patients under study, the investigator has the power to predetermine measured accuracy (41). However, the source of this problem is not the measure itself but the inappropriate generalization of study findings to populations not represented by the study sample. If one test proves more accurate than another in a random sample of men over 65 who

*Semantic problems abound in the literature. Although accuracy is generally defined as above, the same measure is referred to by at least one author as "validity" (41).

are hospitalized for suspected lung disease, and if there are no suspected sampling biases, * then the test will also be more accurate in all such patients. In any event, this problem is minor compared to the assumptions of equal importance of all kinds of error.

Another index of diagnostic efficiency is the likelihood ratio, L, defined as the ratio of the true-positive rate (sensitivity) to the false-positive rate (1 – specificity), or,

$$L = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

Since sensitivity and specificity take on values between O and 1, L must lie between O and 8. L can also be interpreted as the rate at which the odds in favor of disease prior to having test results are translated into odds in favor of disease after the test results are known. * * Thus, it is a measure of the information content of the test, Very high or very low values of L imply high information content, while a value of L near 1 implies that a test adds little information to assist in diagnosis.

Unlike diagnostic accuracy, the likelihood ratio does not vary with disease prevalence, but it, too, makes implicit assumptions about the relative importance of the different kinds of testing errors. L is essentially a measure of the ability of a test to remove uncertainty, but uncertainty of different kinds may be more or less important to the patient. By collapsing sensitivity and specificity into a single composite index, one loses the ingredients necessary for such an analysis.

Other measures of diagnostic efficiency such as the predictive value of a positive test (55) are also available. They are essentially variations of the indexes described here, and they also suffer from the general limitations described above.

To summarize, measurement of the basic constituents of diagnostic efficiency—sensitivity and specificity—is a critical first step in obtaining information necessary for decisions about the use of a diagnostic test, but it is generally an insufficient guide for decisionmaking unless much is known or can be assumed about the importance of each kind of testing error. In any case, the use of summary indexes of diagnostic accuracy in the absence of reporting on sensitivity and specificity is inadequate.

*Most studies are based on patients at a single hospital or institution, The sample may be representative of those presenting at the hospital, but there is always a problem in generalizing beyond the particular institutions, for patient populations vary widely among hospitals.

* *The mathematical derivation of this interpretation is given by McNeil and Mellins (81),

STUDIES OF DIAGNOSTIC YIELD

Diagnostic yield is defined as the proportion of all test results found to be positive or abnormal. This measure of effectiveness is employed in evaluating the usefulness of a diagnostic X-ray procedure in a group of individuals with a specified set of signs, symptoms, or risk factors. Investigators typically compare yields of two or more alternative diagnostic strategies, one involving the X-ray procedure, the other(s) not. If the diagnostic yield of the strategy using the X-ray procedure is low relative to competing strategies, the inference is that the X-ray procedure is unjustified.

The concept of diagnostic yield is employed most frequently in evaluating the usefulness of an X-ray procedure as a screening tool. Here, the study population is defined by demographic and behavioral risk factors such as age, sex, occupation, history of smoking, or admission to a hospital. Symptoms or physical signs causing suspicion of a disease detectable by the X-ray are absent. The X-ray procedure has potential for detecting radiographic signs of occult (symptomatic) disease, presumably in early and more manageable stages than would appear with symptoms. The diagnostic strategies in contention are generally straightforward: screening v. no screening. The diagnoslic yield of the screening procedure thus represents the net difference in the number of cases detected between the screening and no screening options.

The cost of the screening program is often introduced as an element of the analysis. The case-finding cost (i.e., cost per abnormal) is estimated as the unit cost of the procedure divided by diagnostic yield. High case-finding costs are generally interpreted as evidence against the use of an X-ray procedure on the class of individuals in question.

Just as measures of diagnostic efficiency give inadequate consideration to the implications of different kinds of errors, so too does diagnostic yield. The negative finding, whether correct or incorrect, appears to have no value. Yet, negative results may point the way to other possible diagnosis and often reassure the patient, a function of considerable value in some situations (51).

Diagnostic yield is also insensitive to the potential significance of positive findings. When abnormalities detected by the X-ray are already known or could have been detected by simpler diagnostic approaches not considered, or when knowledge of an abnormality does not affect subsequent management of the case because it is either clinically insignificant or not amenable to treatment, then the use of diagnostic yield as an evaluative criterion overestimates the clinical value of a test. Consequently, many researchers use a modified definition of diagnostic yield, counting as abnormal only those cases whose test results are "important," "clinically significant," or unknown prior to the test. This approach attempts to incorporate into the definition of "abnormal" some consideration of the health and economic implications of the finding.

When is the diagnostic yield a useful evaluative endpoint? Three conditions must hold: 1) it must be reasonable to ignore the reassurance value of a negative finding, 2) a false-positive result should not imply costly or risky followup procedures, and 3) a true-positive result should be likely to significantly influence therapy and outcome. Under these conditions, comparison of the diagnostic yield of alternative tests or testing strategies will generally result in appropriate medical decisions. Few diagnostic testing alternatives meet all of these conditions. To the extent that they do not, the results of such analyses should be interpreted cautiously.

STUDIES OF HIGH-YIELD CRITERIA

The "high-yield criteria" approach is an interesting variation on diagnostic yield. Whereas studies of diagnostic yield attempt to identify the best diagnostic strategy for individuals or patients with prespecified presenting conditions, studies of high-yield criteria have as their objective the identification of the presenting conditions (signs, symptoms, risk factors, etc.) that would justify a diagnostic strategy (namely, the use of the X-ray test). A diverse set of individuals is partitioned into two groups: those for whom the diagnostic X-ray strategy is preferred to other strategies, and those for whom the X-ray strategy is inferior to other strategies. The high-yield criteria are the factors used to partition the universe of patients.

This approach is most frequent in studies of the use of X-ray procedures in symptomatic patients, where there are many combinations of presenting signs and symptoms that are potential indications for the procedure. The analysis proceeds on the assumption that if presenting conditions can be identified which account for the vast majority of positive findings, then the diagnostic yield can be improved by limiting the procedure only to patients with these conditions. The high-yield conditions so identified become X-ray referral criteria.

The performance of a set of high-yield criteria must be assessed, but this cannot be accomplished by applying the new criteria to the sample of patients used to create them. Testing on independent patient samples taken from the same population is necessary, and performance, as measured by the number of positives who were missed and the number of X-rays saved, will always deteriorate from the original data set (86). Moreover, if the criteria are applied in pa-

tient populations that are inherently different from the sample used to create them, performance may increase or decrease markedly.

Virtually all high-yield criteria studies analyze data on a sample of patients referred for the X-ray procedure under study. The sample is biased in that patients with the same presenting conditions who are not referred for X-ray are not included in the study. The result is probably to overestimate diagnostic yield in patients with moderate signs and symptoms.

The greatest weakness of studies of high-yield criteria stems from their use of diagnostic yield as the relevant evaluative endpoint, The problems that plague diagnostic yield as a valid measure are equally germane to this discussion. How can one logically choose among signs and symptoms to maximize the diagnostic yield of a test when the implications of an abnormal finding (or of a normal finding) are unknown or their benefits doubtful? These limitations should be kept in mind in assessing whether the decision situation is appropriate to the application of this technique.

STUDIES OF DIAGNOSTIC INFORMATION

An alternative to the use of objectively measurable endpoints, such as sensitivity, specificity, or diagnostic yield, is to measure the impact of the test on the probabilities that physicians subjectively assign to possible diagnoses. If a test finding, positive or negative, has very little impact on such subjective probabilities, then it may be assumed that it has little impact on therapy and outcomes. The information value of the test, then, is measured as the degree of change in physicians' subjective probabilities brought about by the performance of the test.

The information value of the test is attractive as an evaluative measure because it makes a connection, admittedly loose, with therapy and outcomes, and because it considers both positive and negative test findings as important. But the use of physicians' subjective probabilities as the basis for measuring information content is troublesome. Since the likelihood ratio, discussed above, is a completely analogous measure of information content based on objective measures (sensitivity and specificity), it is questionable why one would wish to introduce into an evaluation of a diagnostic strategy the inaccuracy inherent in subjective probability assessments by physicians.

The design of a study to assess the information content of a test is difficult at best. Contrived clinical situations would be needed to protect against biased posttest probability assessments by physicians who had committed themselves to ordering the test. Though the design of studies of test sensitivity and specificity is also difficult for interpretive tests such as X-rays, the introduction of respondent bias is an unnecessary complication.

This approach assumes that the alternative strategy to the test is a state of zero information. The information content Of the test is measured at two points in time: once before and once after its performance. The change in probabilities between these times is attributed wholly to the test results. The assumption is that probabilities would not have changed over time in the absence of the X-ray, or that if more time were allowed to elapse, more information would not be forthcoming as the clinical course progressed. The more appropriate comparison would be between the information available immediately from the X-ray and the information that would be obtained in time as further clinical information emerges. This tradeoff between early and late information can only be compared by looking at the implications of a delay in therapy for patient outcomes and costs.

It is difficult to imagine how the information content of a test could be used to influence medical decisions. Supppose one found that over half of the procedures performed added little information. Would that imply that the procedure is being wasted? Not necessarily, for if the cost

of missing a rare disease is high (that is, if lives are lost), then one might be willing to use a test to rule out the disease, depending on the dollar costs and medical risks of the test. But using a test in this way would insure that a substantial proportion of cases would have similar pretest and posttest probabilities and, hence, index values close to zero.

STUDIES OF OUTCOMES

The inadequacies of diagnostic endpoints can be overcome only by research into the implications of test findings for patient outcomes, including mortality, morbidity, and quality of life. A few studies have attempted to relate the findings from diagnostic X-rays to ultimate outcomes of the medical process. Outcome measures such as 1- or 5-year survival rates and years of life saved have been applied in a few studies. Alternative diagnostic strategies are compared for their ultimate impact on health outcomes and sometimes on the cost of achieving those outcomes.

Studies of this kind require that either a great deal of evidence be available in the literature on the outcome of alternative therapies or expensive prospective and well-controlled studies will be performed to obtain outcome data. Given these requirements, it is understandable that few studies fall into this category.