Why We Need Better Test Evaluations

Mark H. Zweig and E. Arthur Robertson

A laboratory test is clinically useful only if it successfully answers a question of consequence to patient management. Unfortunately, the results and conclusions of many published test evaluations are misleading or of uncertain validity because common-sense principles of study design are overlooked. This is illustrated by examples from recent literature. We suggest that tests should be evaluated with prospective studies of patients representative of the population for which the test will ultimately be used. The clinical question to be addressed by the test should be clearly stated, and then answered for each patient by means independent of the test being evaluated. When comparing tests with each other, decision levels should be chosen to give either the same sensitivity or specificity for each. The use of soundly designed protocols for the clinical evaluation of tests provides the information needed to select the most effective tests.

A test is clinically useful only if it successfully answers a clinical question of consequence to the patient’s management, a question that actually arises in the context of real-time patient care. While analytical limits of test performance may affect the test’s clinical usefulness (1), good analytical performance does not ensure that a new test is going to have real clinical value or any advantage over existing tests. Similarly, although surveys of results in healthy subjects and in various diseased patients may give clues to a test’s potential for application, the demonstration of higher results in diseased patients than in healthy controls does not establish clinical efficacy for the test. Scientific medicine admits new therapeutic agents into routine use only after appropriate prospective controlled clinical trials have demonstrated their efficacy. New diagnostic tests should be subjected to similar clinical trials to determine their diagnostic usefulness before being accepted into routine clinical use. The term “Phase I Trial” could be used to denote studies of the analytical precision, accuracy, sensitivity, and specificity of a laboratory test. “Phase II Trials” would refer to studies determining the usual range of results encountered in healthy subjects or comparing the results obtained in various disease states with this usual range. A prospective diagnostic trial, designed to establish the actual clinical usefulness of a test in a realistic clinical setting before it is adopted for general routine use, could be termed a “Phase III Trial.” In contrast to Phase II trials, which merely catalog results obtained in various clinical groups, the Phase III trial evaluates (in a blind, controlled, prospective, real-time study) the ability of the test to answer a specific clinical question in context.

The first step in a “Phase III Trial” is to define carefully the specific clinical question to be answered by the test(s) on trial. An explicit statement of (a) the clinical question, (b) the patient population about whom the question will be asked, and (c) the criteria for assigning the definitive answer or diagnosis should be constructed.

Patients are then admitted to the study prospectively, before either their test result or the answer to the clinical question is known. The test is performed on each patient without knowledge of the answer to the clinical question. Similarly, without knowledge of the test result, the clinical question is answered definitively by independent means, according to explicit diagnostic criteria. When tests are being compared, all are applied to every patient in the study group.

The design and conduct of the study must respect the rights of human subjects and protect them from inappropriate exposure to discomfort or risk because of additional testing, withholding of critical diagnostic information, or delaying therapy.

In analyzing the results of the diagnostic trial, the tests are compared on the basis of their true-positive rates, choosing appropriate decision levels for the tests so that all have the same false-positive rate. Alternatively, receiver operating characteristic curves can be used to assess the performance of the tests over all possible decision levels (2).

A well-designed and carefully conducted diagnostic trial entails substantial effort, but the reward is the confidence we can have in its conclusions. Unfortunately, the conclusions of many published studies are of uncertain validity, because these common-sense principles are often overlooked (3–5). Examples of such studies follow.

Selecting Appropriate Subjects

Most new tests are developed for application to a specific clinical situation. Accordingly, a test’s usefulness should be validated in that specific clinical setting. For example, radioimunoassays for serum prostatic acid phosphatase (PAP) have been advocated for detecting prostatic cancer. Griffiths (6) evaluated an RIA of PAP for the detection of prostatic cancer, using a “normal range” derived from serum PAP results for 223 men more than 20 years old who were clinically free of prostatic disease. But clinically, this test is applied to an elderly population, in whom there is a high incidence of clinically evident prostatic disorders. Therefore the effectiveness of the test should be evaluated in just such an elderly population. Although Griffiths did study patients with benign prostatic hypertrophy, he used a normal range based on patients free of prostatic disease. Because the healthy men apparently had lower serum PAP concentrations than men with benign prostatic hypertrophy, 9% of the latter group of men had PAP concentrations that exceeded the “normal range.” Applying the lower normal range obtained from the healthy men to the group with benign prostatic hypertrophy and carcinoma caused the test to have factiously high true-positive and false-positive rates. Use of an explicitly stated hypothesis for verification or rebuttal—such as, “In urology patients suspected of having prostatic cancer, PAP concen-

1 Nonstandard abbreviations: PAP, prostatic acid phosphatase (EC 3.1.3.2); AMI, acute myocardial infarction; L/S, lecithin/sphingomyelin; and CK, creatine kinase (EC 2.7.3.2).
trations cannot distinguish men who actually have cancer from those who do not"—could have led to the use of an appropriate study population. Unless the intent is to use the test to screen healthy men of all ages, it is not appropriate to use the hypothesis that was implicit (though not stated) in Griffiths' approach: "Prostatic acid phosphatase concentrations cannot distinguish men who have prostatic cancer from men clinically free of prostatic disease."

A study by DiMaggio et al. (7), comparing diagnostic tests for pancreatic cancer, illustrates the use of appropriate subjects. The subjects chosen were patients presenting with symptoms that were suggestive of pancreatic cancer and that could not be attributed to benign gastrointestinal disease on the basis of a routine workup. Seven diagnostic tests being evaluated were performed. Then the patients underwent liver biopsy or laparotomy to determine whether they actually had pancreatic cancer. This elegant prospective study illustrates the evaluation of tests in the specific clinical population for which they are intended.

Classifying Subjects Accurately

To assess accurately the ability of a test to classify patients into clinically important groups, the definitive diagnosis must be established for each subject. Often the "routine" diagnostic workup will be inadequate for evaluating a new test, and extraordinary means must then be used to establish the diagnoses. For example, if we wished to validate the usefulness of a test for pulmonary embolism, physical examinations and chest x-rays would be insufficient to establish the true definitive diagnosis. We might need to perform ventilation-perfusion lung scans on all subjects in the study and resolve equivocal cases with pulmonary angiography. Similarly, in evaluating a serum marker for diagnosing cancer of a particular organ, it might not be sufficient to compare the test results with needle biopsy results. If needle biopsy sometimes gave false-negative or false-positive results, it would not provide an accurate standard against which to compare the putative serum marker being studied. If the new test gave results identical to needle-biopsy results, it would appear perfect, even if it had the same inadequacies as needle biopsy. New tests in which results deviated from the needle-biopsy results would appear less perfect, even if they were better than needle biopsy. In fact, a test more nearly accurate than needle biopsy might appear identical in usefulness to a test that was less accurate, if they both deviated from needle biopsy results to the same degree. This might lead to the selection of a test that was not the most accurate one being considered.

Consider two recent papers addressing the assessment of fetal lung maturity. Brocklehurst and Wilde (8) studied alkaline phosphatase, heat-labile alkaline phosphatase, and the alkaline phosphatase/gamma-glutamyltransferase ratio in amniotic fluids, as potential indices of fetal lung maturity. Unfortunately, they judged fetal lung maturity by the lecithin/sphingomyelin (L/S) ratio. The L/S ratio is not 100% accurate in assessing lung maturity or in predicting respiratory distress syndrome and thus does not serve as a good standard for evaluating these putative markers of lung maturity or respiratory distress syndrome. The study simply describes the correlation of the new measurements with the L/S ratio; it does not tell us how well either the new tests or the L/S ratio correlates with the true diagnosis. A sounder approach is that of Torday et al. (9), who studied L/S ratios and saturated phosphatidylcholine concentrations in 322 amniotic fluid samples obtained within 72 h of birth. After birth, the definitive diagnosis (the presence or absence of respiratory distress syndrome) was determined for each infant on the basis of specific clinical, x-ray, and physiological criteria, without knowledge of the amniotic fluid results. The relative accuracy of the two tests in predicting respiratory distress syndrome was assessed by comparing results by each test with the definitive diagnosis. The authors concluded that the new test, saturated phosphatidylcholine concentration, was more nearly accurate than the older L/S ratio.

Classifying Subjects Independently of Tests Being Studied

In a diagnostic trial, the diagnoses must be established independently of the tests being evaluated. Including in the diagnostic criteria one of the tests under study biases the results in favor of that test. The correct approach is typified by the design of a study published by Galen et al. (10), who examined the relative efficiency of several enzyme tests in making the diagnosis of acute myocardial infarction (AMI). An appropriate study group was chosen: 100 consecutive patients admitted to a coronary care unit to rule out myocardial infarction. The 100 subjects were then divided into AMI and non-AMI patients "without any knowledge of any enzyme values."

Unfortunately, not all studies exclude the putative marker from the diagnostic criteria. In a more recent study by Ljungdahl et al. (11), 109 patients admitted to a coronary-care unit were classified as AMI or non-AMI based on electrocardiographic findings, total creatine kinase (CK) activity, CK-MB by electrophoresis, aminotransferase activity, and lactate dehydrogenase-H4 isoenzyme by electrophoresis. The objective of the authors was to compare CK-B by immuno- inhibition with results by the five above-mentioned tests. However, a table in that report presents the number of true positives, true negatives, false positives, and false negatives, and the predictive values of all the tests, even though all but CK-B were used in establishing the definitive diagnosis. Because they were part of the diagnostic criteria, the apparent diagnostic efficacy of each of these five tests would tend to be overestimated, potentially misleading readers examining the data in the table. Furthermore, although CK-B by inhibition was not included in the diagnostic criteria, its assessment was also biased somewhat by the inclusion of two very closely related tests—CK-MB by electrophoresis and total CK—in the diagnostic criteria.

In a recent study (12) serum CK-MB and other enzyme concentrations were examined in patients with chest pain and tachyarrhythmias who were admitted with the suspicion of AMI. Unfortunately, the diagnosis of AMI was based on whether the CK-MB concentration eventually became increased or not. It is not surprising, then, that in the figure showing the incidence of enzyme increases, CK-MB was abnormal in 100% of patients with infarcts and 0% of patients without infarction. This evaluation was clearly biased in favor of CK-MB. The conclusion of the paper that "elevated plasma MB CK activity appears to remain a good diagnostic marker of myocardial necrosis in patients with tachyarrhythmias" was essentially a circular conclusion, resulting from the design of the study.

Performing All Tests on All Subjects

When different assays are being compared, it is important that the subjects and specimens that are used be the same for all tests. Unfortunately, this is not always done. Minor deviations may cause negligible bias, but sometimes the deviations are large. For example, in a study by Guinan et al. (13), who examined 10 tests for the diagnosis of prostatic cancer, some tests were performed on all 300 subjects but some were performed on only 100 subjects. In fact, only two tests were done on all 300 subjects, while four were done on less than half the subjects. Another example occurs in the study by DiMaggio et al. (7),
cited earlier. Data for sensitivity, specificity, and predictive value were presented on four tests for the diagnosis of pancreatic disease and three tests for the diagnosis of pancreatic cancer. Even though 70 patients were used for the study, apparently no test was done on every patient, and each test was done on a different number of patients, ranging from 37 to 65.

Failure to use the identical subjects for evaluating each test can result in misleading conclusions based on sampling errors. Furthermore, subtle biases may affect the selection of subjects for the different groups. Thus apparent differences in test performance may simply be reflections of differences in the composition of the groups tested. In both of the examples, if some patients had more advanced and presumably more easily detectable disease, and were tested by only some of the tests, those tests could appear to have better sensitivity than the others. Conversely, inclusion of patients with minimal disease, which might be harder to detect, would tend to diminish the apparent sensitivity of tests performed on these patients, as compared with tests not done on these patients. Furthermore, performing all tests on all subjects ensures that differences in sensitivity and specificity are not simply due to inconsistent application of the diagnostic criteria. Similar considerations complicate comparisons between different studies.

Comparing Tests at Comparable True-Positive or False-Positive Rates

When tests are compared, the decision level selected for each test should be chosen so that all tests have the same false-positive rate (and thus the same specificity). Then the tests can be directly compared on the basis of their true-positive rates (or sensitivities). Alternatively, the reverse can be done: standardize the true-positive rates and then compare the false-positive rates. Most studies lack this approach. Rather, a conventional normal range is used to determine whether results are “positive” or “negative” in affected subjects and sometimes also in a control population with other conditions. This can result in misleading conclusions about test performance or give data that are not readily interpretable. For example, in a recent study we found that, with use of a conventionally obtained normal reference interval, tests A and B seemed to perform quite differently in classifying patients with chest pain into those who had an AMI and those who did not. The true-positive rates were 96% for A and 100% for B (Table 1). However, the false-positive rates were 0% for A and 38% for B. Thus, B appeared to be slightly more sensitive (higher true-positive rate) and considerably less specific (higher false-positive rate) than test A. Interpreting these data requires a somewhat arbitrary subjective judgment as to whether a false-positive rate of 0% instead of 38% is more important than the advantage of a true-positive rate of 100% instead of 96%. Fixing the true-positive rate for both tests permits an objective judgment based on the false-positive rate. When we chose the decision levels needed to give a 95% true-positive rate for both tests, both then had a 0% false-positive rate. The tests actually had nearly identical diagnostic abilities, but appeared to be different initially because the decision levels being used created “operating conditions” for the tests that were not comparable. In the case of test B, for example, the initial decision level was unnecessarily low.

This problem is illustrated in the recent study by Podolak et al. (14), who examined the relative diagnostic value of five serologic markers for the detection of pancreatic cancer. In this prospective study of 270 patients who had clinical indications of pancreatic disease, diagnoses of cancer (pancreatic or other) were based on examination of tissue (surgery, percutaneous biopsy, or autopsy) in all 103 cases. Diagnoses of benign disease (n = 167) were usually based on negative radiological and endoscopic examinations and the absence of new or further evidence of cancer on follow-up or, in 44 cases, tissue examination. Table 1 shows the true-positive and false-positive rates for the five markers when the patients are separated into those with pancreatic cancer and those with benign diseases. Apparently, galactosyltransferase is more effective than any of the others, but it is quite difficult to compare the tests directly. For example, ferritin has a higher true-positive rate than carcinoembryonic antigen, but also a higher false-positive rate. It is quite hard to judge which of these two would perform better clinically. If, however, we could change the decision level for each test to give a selected true-positive rate, such as 95%, we could then compare the markers on the basis of the corresponding false-positive rates.

A recent publication by Lindholm et al. (15) provides a good example of comparing two tests under comparable conditions. They compared PAP by RIA with acid phosphatase by enzymic assay for the detection of prostatic cancer. Decision levels were chosen to give specificities of 94% (6% false positives) and 97.5% (2.5% false positives), and then the tests were compared on the basis of the resulting true-positive rates (sensitivities). At both specificities, the RIA was more sensitive than the enzymic assay. Unfortunately, this approach has been taken only rarely.

A more comprehensive comparison of tests can be made by plotting all possible true-positive and false-positive rates for each test in the form of a receiver operating characteristic curve for each test—a plot of the true-positive rate as a function of the false-positive rate for all decision levels (see Figure 1). By using such curves, the tests may be compared without selecting any particular upper limit of normal or any particular true-positive or false-positive rate. Each decision level for a test has a corresponding pair of true-positive and false-positive rates, which are represented by one point on the receiver operating characteristic curve.

To illustrate the value of such analysis, consider a study in which several indirect methods for detecting lactase deficiency are compared, including measurements of blood glucose and galactose concentrations after oral lactose load (16). The definitive diagnosis of lactase deficiency was established or excluded by assaying lactase activity in peroral biopsies of the

---

**Table 1. Effect of Changing Decision Level on True- and False-Positive Rates**

<table>
<thead>
<tr>
<th>Test</th>
<th>Decision level</th>
<th>True-positive rate, %</th>
<th>False-positive rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial decision levels</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>10.0</td>
<td>96</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>6.0</td>
<td>100</td>
<td>38</td>
</tr>
<tr>
<td><strong>Modified decision levels</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>10.5</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>12.4</td>
<td>95</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Table 2. True- and False-Positive Rates for Five Markers of Pancreatic Cancer (14)**

<table>
<thead>
<tr>
<th>Marker</th>
<th>True-positive rate, %</th>
<th>False-positive rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-Fetoprotein</td>
<td>3.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Ferritin</td>
<td>50.0</td>
<td>22.1</td>
</tr>
<tr>
<td>RNase</td>
<td>29.7</td>
<td>13.6</td>
</tr>
<tr>
<td>Carcinoembryonic antigen</td>
<td>34.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Galactosyltransferase</td>
<td>67.2</td>
<td>1.7</td>
</tr>
</tbody>
</table>

---

1274 CLINICAL CHEMISTRY, Vol. 28, No. 6, 1982
small intestine. A receiver operating characteristic curve analysis of their raw data shows that, in fact, both tests can achieve high true-positive rates with low false-positive rates, and that neither test is clearly superior to the other. In the published report, however, the two tests are compared with use of a single decision level, taken from previous work or that of other investigators, for each test. At these decision levels, the authors concluded that the "plasma glucose test was clearly inferior to the plasma galactose." This erroneous conclusion resulted from an unfortunate choice of decision levels, which gave true-positive and false-positive rates of 76% and 4% for the plasma glucose test. This approach masked the ability of the glucose test to achieve true-positive and false-positive rates of 100% and 4% if another decision level was used. Receiver operating characteristic curves clearly show that the two tests in fact behaved very similarly in this study group. By showing the true-positive and false-positive rates for all possible decision levels, the receiver operating characteristic curves permit direct valid comparisons of tests' performances.

A Final Example

The examples described above illustrate subtle pitfalls in assessing the clinical efficacy of laboratory tests. These flaws in evaluating tests result in misleading data, mistaken conclusions, and wasteful practices. The recent activity about PAP as a marker for prostatic cancer serves as an example of the confusion, controversy, and waste that can occur.

In 1977 Foti et al. (17) described and evaluated an RIA for PAP. The data suggested to the authors that the RIA for PAP "has the potential to detect well over half the cases of intracapsular (and thus surgically curable) prostatic cancer (Stages I and II)." Needless to say, this led to a vigorous burst of activity by other investigators and kit manufacturers. Although subsequent studies by other workers demonstrated somewhat higher sensitivity for RIA than for enzymic assays, they did not confirm the markedly high sensitivity in occult and early stage disease reported by Foti et al. (6, 15, 18, 19). Nonetheless, RIA kits for the determination of PAP soon reached the marketplace. Advertising appeared in the scientific and lay press. PAP was promoted as a screening test for occult prostatic cancer in healthy men. The uncertainty about the real clinical advantages of the test, coupled with the availability of the kits and the publicity about the test, led to confusion and controversy among clinicians and laboratorians as to whether the test should be used or under what circumstances. Time and critical re-evaluation of the data have led to a more sober view (20–23).

If the original study of Foti et al. had been designed differently, the data might not have been so misleading and their conclusions might not have been so overly optimistic. The apparently high sensitivity in early-stage prostatic cancer may have been due to insufficiently rigorous diagnostic maneuvers, resulting in understaging. Others (24, 25) have shown that when the workup does not include procedures such as isotopic bone scan, lymphangiogram, or pelvic node dissection, patients often have advanced disease that is undetected, resulting in understaging. This suggests that, in the study of Foti et al., some of the patients who were thought to have only early minimal disease may actually have had more advanced disease (15, 26). Moreover, their use of "normal controls" to establish the upper limit of normal permitted an apparent high sensitivity, overshadowing the appreciable nonspecificity evidenced by the "abnormal" results they reported in other disorders.

Watson and Tang (20) recently showed that, even with the high sensitivity reported by Foti et al., the low prevalence of prostatic cancer leads to a low predictive value for PAP as a screening test. A better-designed clinical trial, with more thorough diagnoses, a more appropriate study group, consideration of predictive value, and the use of receiver operating characteristic analysis instead of normal ranges, might have produced more realistic conclusions.

Diagnostic tests proliferate, public demand and private enterprise plunge them into common use, and the need for careful trials to evaluate their worth becomes pressing. Expenditures for clinical laboratory testing consume about 0.6% of the U.S. gross national product (27). Mistaken endorsements of tests on the basis of flawed evaluations provoke enthusiastic but wasteful use. Conversely, critical evaluations occasion savings by forestalling inappropriate use of less-effective tests, while improving clinical care by encouraging use of the most efficacious.

References


