IN THIS CHAPTER:

Introduction

Clinical Scenario

How Can We Identify Dementia Quickly and Accurately?

Finding the Evidence

Are the Results Valid?

Did Participating Patients Present a Diagnostic Dilemma?

Did the Investigators Compare the Test to an Appropriate, Independent Reference Standard?

Were Those Interpreting the Test and Reference Standard Blind to the Other Results?

Did Investigators Perform the Same Reference Standard to All Patients Regardless of the Results of the Test Under Investigation?

What Are the Results?

What Likelihood Ratios Were Associated With the Range of Possible Test Results?

Dichotomizing Continuous Test Scores, Sensitivity and Specificity, and LR+ and LR–

How Can I Apply the Results to Patient Care?

Will the Reproducibility of the Test Result and Its Interpretation Be Satisfactory in My Clinical Setting?

Are the Study Results Applicable to the Patients in My Practice?

Will the Test Results Change My Management Strategy?

Will Patients Be Better Off as a Result of the Test?
INTRODUCTION

In the previous 2 chapters (Chapter 14, The Process of Diagnosis, and Chapter 15, Differential Diagnosis), we explained the process of diagnosis, the way diagnostic test results move clinicians across the test threshold and the therapeutic threshold, and how to use studies to help obtain an accurate pretest probability. In this chapter, we show you how to use an article addressing the ability of a diagnostic test to move clinicians toward the extremely high (ruling in) and extremely low (ruling out) posttest probabilities they seek. Later in this book, we will show you how to use articles that integrate a number of test results into a clinical prediction rule (Chapter 17.4, Clinical Prediction Rules).

CLINICAL SCENARIO

How Can We Identify Dementia Quickly and Accurately?

You are a busy primary care practitioner with a large proportion of elderly patients in your practice. Earlier in the day, you treated a 70-year-old woman who lives alone and has been managing well. On this visit, she complained about a longstanding problem, joint pain in her lower extremities. During the visit, you have the impression that, as you put it to yourself, “she isn’t quite all there,” although you find it hard to specify further. On specific questioning about memory and function, she acknowledges that her memory is not what it used to be but otherwise denies problems. Pressed for time, you deal with the osteoarthritis and move on to the next patient.

That evening, you ponder the problem of making a quick assessment of your elderly patients when the possibility of cognitive impairment occurs to you. The Mini-Mental Status Examination (MMSE), with which you are familiar, takes too long. You wonder whether there are any brief instruments that allow a reasonably accurate rapid diagnosis of cognitive impairment to help you identify patients who need more extensive investigation.

FINDING THE EVIDENCE

You formulate the clinical question: In older patients with suspected cognitive impairment, what is the accuracy of a brief screening tool for diagnosing dementia (or for identifying those who need more extensive investigation)? You select “diagnosis” and “narrow, specific search” from the PubMed Clinical Queries page. Using search terms “dementia AND screen* AND brief,” the search yields 48 citations. Limiting to English-language studies of humans in the last 5 years cuts the list to 21.
You survey the abstracts, looking for articles that focus on patients with suspected dementia and report accuracy similar to your previous standard, the MMSE. An article reporting results for an instrument named Six-Item Screener (SIS) meets both criteria. You retrieve the full-text article electronically and start to read it, hoping its methods and results will justify using the instrument in your office.

**ARE THE RESULTS VALID?**

Table 16-1 summarizes our Users' Guides for assessing the validity, examining the results, and determining the applicability of a study reporting on the accuracy of a diagnostic test.

**Did Participating Patients Present a Diagnostic Dilemma?**

A diagnostic test is useful only if it distinguishes between conditions and disorders that might otherwise be confusing. Although most tests can differentiate healthy persons from severely affected ones, this ability will not help us in clinical practice. Studies that confine themselves to florid cases vs asymptomatic healthy volunteers are unhelpful because, when the diagnosis is obvious, we do not need a diagnostic test. Only a study that closely resembles clinical practice and includes patients with mild, early manifestations of the target condition can establish a test’s true value.

**TABLE 16-1**

Users' Guide for an Article About Interpreting Diagnostic Test Results

<table>
<thead>
<tr>
<th>Are the results valid?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Did participating patients present a diagnostic dilemma?</td>
</tr>
<tr>
<td>• Did investigators compare the test to an appropriate, independent reference standard?</td>
</tr>
<tr>
<td>• Were those interpreting the test and reference standard blind to the other results?</td>
</tr>
<tr>
<td>• Did investigators perform the same reference standard to all patients regardless of</td>
</tr>
<tr>
<td>the results of the test under investigation?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What are the results?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• What likelihood ratios were associated with the range of possible test results?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How can I apply the results to patient care?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Will the reproducibility of the test result and its interpretation be satisfactory in</td>
</tr>
<tr>
<td>my clinical setting?</td>
</tr>
<tr>
<td>• Are the study results applicable to the patients in my practice?</td>
</tr>
<tr>
<td>• Will the test results change my management strategy?</td>
</tr>
<tr>
<td>• Will patients be better off as a result of the test?</td>
</tr>
</tbody>
</table>
The story of carcinoembryonic antigen (CEA) testing in patients with colorectal cancer shows how choosing the wrong spectrum of patients can dash the hopes raised with the introduction of a diagnostic test. A study found that CEA was elevated in 35 of 36 people with known advanced cancer of the colon or rectum. The investigators found much lower levels in normal people, pregnant women, or in patients with a variety of other conditions.\(^2\) The results suggested that CEA might be useful in diagnosing colorectal cancer or even in screening for the disease. In subsequent studies of patients with less advanced stages of colorectal cancer (and therefore lower disease severity) and patients with other cancers or other gastrointestinal disorders (and therefore different but potentially confused disorders), the accuracy of CEA testing as a diagnostic tool plummeted. Clinicians appropriately abandoned CEA measurement for new cancer diagnosis and screening.

There have been 3 systematic, empirical examinations of design-related bias in studies of diagnostic tests. Lijmer et al\(^3\) and Rutjes et al\(^4\) collected meta-analyses of diagnostic tests and examined what aspects of study design influenced the apparent diagnostic power of the tests. Whiting et al\(^5\) systematically collected and reviewed primary studies that investigated the effects of bias on estimates of diagnostic test performances.

All 3 studies documented substantial bias associated with unrepresentative patient selection. Enrolling target-positive (those with the underlying condition of interest—in our scenario, people with dementia) and target-negative patients (those without the target condition) from separate populations results in overestimates of the test’s power (relative diagnostic odds ratio [RDOR], 3.0; 95% confidence interval [CI], 2.0-4.5; and RDOR, 4.9; 95% CI, 0.6-37.3).\(^3,4\) Even if investigators enroll target-positive and target-negative patients from the same population, nonconsecutive patient sampling and retrospective data collection may inflate estimates of diagnostic test performances (RDOR, 1.5; 95% CI, 1.0-2.1; and RDOR, 1.6; 95% CI, 1.1-2.2, respectively).\(^2,3\) We label studies with unrepresentative patient selection as having spectrum bias (see Chapter 17.1, Spectrum Bias).

Table 16-2 summarizes the empirically supported sources of bias in studies of diagnostic tests.

**Did the Investigators Compare the Test to an Appropriate, Independent Reference Standard?**

The accuracy of a diagnostic test is best determined by comparing it to the “truth.” Readers must assure themselves that investigators have applied an appropriate reference, criterion, or gold standard (such as biopsy, surgery, autopsy, or long-term follow-up without treatment) to every patient who undergoes the test under investigation.

One way a study can go wrong is if the test that is being evaluated is part of the reference standard. The incorporation of the test into the reference standard is likely to inflate the estimate of the test’s diagnostic power. Thus, clinicians should insist on the independence as one criterion for a satisfactory reference standard.
### TABLE 16-2

**Empirical Evidence of Sources of Bias in Diagnostic Accuracy Studies**

<table>
<thead>
<tr>
<th>Source of Bias</th>
<th>Lijmer et al&lt;sup&gt;3&lt;/sup&gt; (RDOR; 95% CI)</th>
<th>Whiting et al&lt;sup&gt;5&lt;/sup&gt;</th>
<th>Rutjes et al&lt;sup&gt;4&lt;/sup&gt; (RDOR; 95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did participating patients present a diagnostic dilemma?</td>
<td>Case-control design (3.0; 2.0-4.5)</td>
<td>Distorted selection of participants (some empirical support)</td>
<td>Case-control design (4.9; 0.6-37.3)</td>
</tr>
<tr>
<td>Nonconsecutive patient selection</td>
<td>Nonconsecutive sampling (0.9; 0.7-1.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrospective data collection</td>
<td>Retrospective data collection (1.0; 0.7-1.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did investigators compare the test to an appropriate, independent reference standard?</td>
<td>Inappropriate reference standard (some empirical support)</td>
<td></td>
<td>Incorporation bias (using test as part of reference standard) (no empirical support)</td>
</tr>
<tr>
<td>Were those interpreting the test and reference standard blind to the other result?</td>
<td>Not blinded (1.3; 1.0-1.9)</td>
<td>Review bias (some empirical support)</td>
<td>Single or non-blinded reading (1.1; 0.8-1.6)</td>
</tr>
<tr>
<td>Did investigators perform the same reference standard to all patients regardless of the results of the test under investigation?</td>
<td>Different reference tests (2.2; 1.5-3.3)</td>
<td>Differential verification bias (some empirical support)</td>
<td>Differential verification (1.6; 0.9-2.9)</td>
</tr>
<tr>
<td>Partial verification</td>
<td>Partial verification bias (strong empirical support)</td>
<td></td>
<td>Partial verification (1.1; 0.7-1.7)</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; RDOR, relative diagnostic odds ratio.

*RDOR, point estimates, and 95% CIs are shown.*
For instance, consider a study that evaluated the utility of abdominoujugal reflux for the diagnosis of congestive heart failure. This study used, however, clinical and radiographic criteria, including abdominoujugal reflex, as the reference test. Another example comes from a study evaluating screening instruments for depression in terminally ill people. The authors claimed perfect performance (sensitivity = 1.0, specificity = 1.0) for a single question (Are you depressed?) to detect depression. Their diagnostic criteria included 9 questions, of which 1 was “Are you depressed?”

In reading articles about diagnostic tests, if you cannot accept the reference standard (within reason, that is; after all, nothing is perfect), then the article is unlikely to provide valid results (Table 16-2).

**Were Those Interpreting the Test and Reference Standard Blind to the Other Results?**

If you accept the reference standard, the next question is whether the interpreters of the test and reference standard were aware of the results of the other investigation (blind assessment).

Consider how, once clinicians see a pulmonary nodule on a computed tomographic (CT) scan, they can see the previously undetected lesion on the chest radiograph, or, once they learn the results of an echocardiogram, they hear a previously inaudible cardiac murmur.

The more likely that knowledge of the reference standard result can influence the interpretation of a test, the greater the importance of the blinded interpretation. Similarly, the more susceptible the reference standard is to changes in interpretation as a result of knowledge of the test being evaluated, the more important the blinding of the reference standard interpreter. The empirical study by Lijmer et al demonstrated bias associated with unblinding, although the magnitude was small (RDOR, 1.3; 95% CI, 1.0-1.9), whereas Rutjes et al found a compatible although statistically nonsignificant RDOR (RDOR, 1.1; 95% CI, 0.8-1.6) (Table 16-2).

**Did Investigators Perform the Same Reference Standard to All Patients Regardless of the Results of the Test Under Investigation?**

The properties of a diagnostic test will be distorted if its results influence whether patients undergo confirmation by the reference standard (verification or work-up bias). This can occur in 2 ways. First, only a selected sample of patients who underwent the index test may be verified by the reference standard. For example, patients with suspected coronary artery disease whose exercise test results are positive may be more likely to undergo coronary angiography (the reference standard) than those whose exercise test results are negative. Whiting et al reviewed several documented instances of this type of verification bias, known as partial verification bias.

Second, results of the index test may be verified by different reference standards. Lijmer et al and Rutjes et al found a large magnitude of bias associated with the
use of different reference tests for positive and negative results. The RDOR for this type of bias, also known as differential verification bias, was 2.2; 95% CI, 1.5-3.3\(^3\) and 1.6; 95% CI, 0.9-2.9,\(^4\) respectively, in these 2 systematic reviews (Table 16-2).

Verification bias proved a problem for the Prospective Investigation of Pulmonary Embolus Diagnosis (PIOPED) study that evaluated the utility of ventilation perfusion scanning in the diagnosis of pulmonary embolism. Patients whose ventilation perfusion scan results were interpreted as “normal/near normal” and “low probability” were less likely to undergo pulmonary angiography (69%) than those with more positive ventilation perfusion scan results (92%), which is not surprising because clinicians might be reluctant to subject patients with a low probability of pulmonary embolism to the risks of angiography.\(^12\)

Most articles would stop here, and readers would have to conclude that the magnitude of the bias resulting from different proportions of patients with high- and low-probability ventilation perfusion scans undergoing adequate angiography is uncertain but perhaps large. The PIOPED investigators, however, applied a second reference standard to the 150 patients with low-probability or normal/near-normal scan results who failed to undergo angiography (136 patients) or for whom angiogram interpretation was uncertain (14 patients). They judged such patients to be free of pulmonary embolism if they did well without treatment. Accordingly, they followed all such patients for 1 year without treating them with anticoagulant drugs. No patient developed clinically evident pulmonary embolism during follow-up, allowing us to conclude that patient-important pulmonary embolism (if we define patient-important pulmonary embolism as requiring anticoagulation therapy to prevent subsequent adverse events) was not present when they underwent ventilation perfusion scanning. Thus, the PIOPED study achieved the goal of applying a reference standard assessment to all patients but failed to apply the same standard to all.

**USING THE GUIDE**

The study of a brief diagnostic test for cognitive impairment included 2 cohorts. One was a stratified random sample of community-dwelling black persons aged 65 years and older; the other was a consecutive sample of nonselected nonscreened patients referred by family, caregivers, or providers for cognitive evaluation at the Alzheimer Disease Center. In the former group, the authors included all patients with a high suspicion of dementia on a detailed screening test and a random sample of those with moderate and low suspicion. The investigators faced diagnostic uncertainty in both populations. The populations are not perfect: the former included individuals without any suspicion of dementia, and the latter had already passed an initial screen at the primary care level (indeed, whether to refer for full geriatric assessment is one of the questions you are trying to resolve for the patient who triggered your literature search). Fortunately, test properties proved similar in the 2 populations, considerably lessening your concern.
All patients received the SIS, which asks the patient to remember 3 words (apple, table, penny); then to say the day of the week, month, and year; and finally to recall the 3 words without prompts. The number of errors provides a result with a range of 0 to 6.

For the reference standard diagnosis of dementia, patients had to satisfy both Diagnostic and Statistical Manual of Mental Disorders (Third Edition Revised) (DSM-III-R) and the International Classification of Diseases, Tenth Revision (ICD-10) criteria, according to an assessment by a geriatric psychiatrist or a neurologist that included medical history and physical and neurologic examination; a complete neuropsychological test battery, including MMSE and 5 other tests; and interview with a relative of the participant.

Although you are satisfied with this reference standard, the published article leaves you unsure whether those making the SIS and the reference diagnosis were blind to the other results. To resolve the question, you e-mail the first author and ask for clarification. A couple of e-mails later, you have learned that “research assistants who had been trained and tested” administered the neuropsychological battery. On the other hand, “a consensus team composed of a geriatric psychiatrist, social psychologist, a geriatrician, and a neuropsychologist” made the reference standard diagnoses. The author reports that “there were open discussions of the case, and they had access to the entire medical record, including results of neuropsychological testing, at their disposal.” The 6 items included in the SIS are derived from the MMSE but “were not pulled out as a separate instrument in the consensus team conference.”

Thus, although there was no blinding, you suspect that this did not create important bias and are therefore ready to consider its results.

**What Are the Results?**

**What Likelihood Ratios Were Associated With the Range of Possible Test Results?**

In deciding how to interpret diagnostic tests results, we will consider its ability to change our estimate of the likelihood the patient has the target condition (we call this the pretest probability) to a more accurate estimate (we call this the posttest probability of the target disorder). The likelihood ratio (LR) for a particular test result moves us from the pretest probability to a posttest probability.

Put yourself in the place of the primary care physician in the scenario and consider 2 patients with suspected cognitive impairment with clear consciousness. The first is the 70-year-old woman in the clinical scenario who seems to be managing rather well but has a specific complaint that her memory is not what it used to be. The other is an 85-year-old woman, another longstanding patient, who arrives accompanied, for the first time, by her son. The concerned son tells you that she has, on one of her usual morning walks, lost her way. A neighbor happened to catch her a few miles away...
from home and notified him of the incident. On visiting his mother’s house, he was surprised to find her room in a mess. Yet in your office, she greets you politely and protests that she was just having a bad day and does not think the incident warrants any fuss (at which point the son looks to the ceiling in frustrated disbelief). Your clinical hunches about the probability of dementia for these 2 people—that is, their pretest probabilities—are different. For the first woman, the probability is relatively low, perhaps 20%; for the second, it is relatively high, perhaps 70%.

The results of a formal screening test, the SIS in our example, will not tell us definitively whether dementia is present; rather, the results modify the pretest probability of that condition, yielding a new posttest probability. The direction and magnitude of this change from pretest to posttest probability are determined by the test’s properties, and the property of most value is the LR.

We will use the results of the study by Callahan et al to illustrate LRs. Table 16-3 presents the distribution of SIS scores in cohort of patients in the study by Callahan et al.1

How likely is a test result of 6 among people who do have dementia? Table 16-3 shows that 105 of 345 (or 30.4%) people with the condition made 6 errors. We can also see that of 306 people without dementia, 2 (or 0.65%) made 6 errors. How likely is this test result (ie, making 6 errors) in someone with dementia as opposed to someone without? Determining this requires us to look at the ratio of the 2 likelihoods that we have just calculated (30.4/0.65) and equals 47. In other words, the test result of 6 is 47 times as likely to occur in a patient with, as opposed to without, dementia.

In a similar fashion, we can calculate the LR associated with a test result of each score. For example, the LR for the test score of 5 is (64/345)/(2/306) = 28. Table 16-3 provides the LR for each possible SIS score.

How can we interpret LRs? LRs indicate the extent to which a given diagnostic test result will increase or decrease the pretest probability of the target disorder. An LR of

<table>
<thead>
<tr>
<th>TABLE 16-3</th>
<th>Six-Item Screener Scores in Patients With and Without Dementia, and Corresponding Likelihood Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dementia (+)</td>
</tr>
<tr>
<td>SIS = 6</td>
<td>105</td>
</tr>
<tr>
<td>SIS = 5</td>
<td>64</td>
</tr>
<tr>
<td>SIS = 4</td>
<td>64</td>
</tr>
<tr>
<td>SIS = 3</td>
<td>45</td>
</tr>
<tr>
<td>SIS = 2</td>
<td>31</td>
</tr>
<tr>
<td>SIS = 1</td>
<td>25</td>
</tr>
<tr>
<td>SIS = 0</td>
<td>11</td>
</tr>
<tr>
<td>Sum</td>
<td>345</td>
</tr>
</tbody>
</table>

Abbreviation: SIS, Six-Item Screener.

Data from Callahan et al.1
1 tells us that the posttest probability is exactly the same as the pretest probability. LRs greater than 1.0 increase the probability that the target disorder is present; the higher the LR, the greater this increase. Conversely, LRs less than 1.0 decrease the probability of the target disorder, and the smaller the LR, the greater the decrease in probability.

How big is a “big” LR, and how small is a “small” one? Using LRs in your day-to-day practice will lead to your own sense of their interpretation, but consider the following a rough guide:

- LRs of greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability;
- LRs of 5 to 10 and 0.1 to 0.2 generate moderate shifts in pretest to posttest probability;
- LRs of 2 to 5 and 0.5 to 0.2 generate small (but sometimes important) changes in probability; and
- LRs of 1 to 2 and 0.5 to 1 alter probability to a small (and rarely important) degree.

Having determined the magnitude and significance of LRs, how do we use them to go from pretest to posttest probability? One way is to convert pretest probability to odds, multiply the result by the LR, and convert the consequent posttest odds into a posttest probability. A much easier strategy uses a nomogram proposed by Fagan13 (Figure 16-1) that does all the conversions and allows an easy transition from pretest to posttest probability.

The left-hand column of this nomogram represents the pretest probability, the middle column represents the LR, and the right-hand column shows the posttest probability. You obtain the posttest probability by anchoring a ruler at the pretest probability and rotating it until it lines up with the LR for the observed test result. There is also a Web-based interactive program (http://www.JAMAevidence.com) that will do this for you. You can enter exact numbers for a pretest probability and an LR to obtain the exact posttest probability.

Recall the elderly woman from the opening scenario, who has suspected dementia. We have decided that the probability of this patient’s having the condition is about 20%. Let us suppose that the patient made 5 errors on the SIS. Anchoring a ruler at her pretest probability of 20% and aligning it with the LR of 28 associated with the test result of 5, you can obtain her posttest probability, around 90%.

The pretest probability is an estimate. Although the literature dealing with differential diagnosis can sometimes help us in establishing the pretest probability (see Chapter 15, Differential Diagnosis), we know of no such study that will complement our intuition in arriving at a pretest probability when the suspicion of dementia arises. Although our intuition makes precise estimates of pretest probability difficult, we can deal with residual uncertainty by examining the implications of a plausible range of pretest probabilities.

For example, if the pretest probability in this case is as low as 10% or as high as 30%, using the nomogram, we will obtain the posttest probability of about 80% and above 90%. Table 16-4 tabulates the posttest probabilities corresponding with each possible SIS score for the 70-year-old woman in the clinical scenario.
We can repeat this exercise for our second patient, the 85-year-old woman who had lost her way. You estimate that her medical history and presentation are compatible with a 70% probability of dementia. With our nomogram (Figure 16-1), the posttest probability with an SIS score of 6 or 5 is almost 100%; with an SIS score of 4, it is 94%; with an SIS score of 3, it is 85% and so on. The pretest probability (with a range of possible pretest probabilities from 60% to 80%), LRs, and posttest probabilities associated with each of these possible SIS scores are presented in Table 16-5.

Having learned to use LRs, you may be curious about where to find easy access to the LRs of the tests you use regularly in your own practice.
Rational Clinical Examination is a series of systematic reviews of the diagnostic properties of the medical history and physical examination that have been published in JAMA. Chapter 17.2, Examples of Likelihood Ratios, lists some examples of LRs. Further examples are accumulated on the Users’ Guides Web site (http://www.JAMAevidence.com).

### TABLE 16-4

<table>
<thead>
<tr>
<th>Pretest Probability, % (Range)(^a)</th>
<th>SIS Result (LR)</th>
<th>Posttest Probability, % (Range)(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 (10-30)</td>
<td>SIS = 6 (47)</td>
<td>92 (84-95)</td>
</tr>
<tr>
<td></td>
<td>SIS = 5 (28)</td>
<td>88 (76-92)</td>
</tr>
<tr>
<td></td>
<td>SIS = 4 (7.1)</td>
<td>64 (44-75)</td>
</tr>
<tr>
<td></td>
<td>SIS = 3 (2.5)</td>
<td>38 (22-52)</td>
</tr>
<tr>
<td></td>
<td>SIS = 2 (0.79)</td>
<td>16 (8-25)</td>
</tr>
<tr>
<td></td>
<td>SIS = 1 (0.28)</td>
<td>7 (3-11)</td>
</tr>
<tr>
<td></td>
<td>SIS = 0 (0.06)</td>
<td>1 (1-3)</td>
</tr>
</tbody>
</table>

Abbreviations: LR, likelihood ratio; SIS, Six-Item Screener.

*The values in parentheses represent a plausible range of pretest probabilities; that is, although the best guess as to the pretest probability is 20%, values of 10% to 30% would also be reasonable estimates.

### TABLE 16-5

<table>
<thead>
<tr>
<th>Pretest Probability, % (Range)(^a)</th>
<th>SIS Result (LR)</th>
<th>Posttest Probability, % (Range)(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70 (60-80)</td>
<td>SIS = 6 (47)</td>
<td>99 (99-99)</td>
</tr>
<tr>
<td></td>
<td>SIS = 5 (28)</td>
<td>98 (98-99)</td>
</tr>
<tr>
<td></td>
<td>SIS = 4 (7.1)</td>
<td>94 (91-97)</td>
</tr>
<tr>
<td></td>
<td>SIS = 3 (2.5)</td>
<td>85 (79-76)</td>
</tr>
<tr>
<td></td>
<td>SIS = 2 (0.79)</td>
<td>65 (54-76)</td>
</tr>
<tr>
<td></td>
<td>SIS = 1 (0.28)</td>
<td>40 (30-53)</td>
</tr>
<tr>
<td></td>
<td>SIS = 0 (0.06)</td>
<td>12 (8-19)</td>
</tr>
</tbody>
</table>

Abbreviations: LR, likelihood ratio; SIS, Six-Item Screener.

*The values in parentheses represent a plausible range of pretest probabilities. That is, although the best guess as to the pretest probability is 70%, values of 60% to 80% would also be reasonable estimates.
Dichotomizing Continuous Test Scores, Sensitivity and Specificity, and LR+ and LR–

Readers who have followed the discussion to this point will understand the essentials of interpretation of diagnostic tests. In part because they remain in wide use, it is also helpful to understand 2 other terms in the lexicon of diagnostic testing: sensitivity and specificity. Many articles on diagnostic tests report a $2 \times 2$ table and its associated sensitivity and specificity, as in Table 16-6, and to go along with it a figure that depicts the overall power of the diagnostic test (called a receiver operating characteristic [ROC] curve).

The study by Callahan et al recommends a cutoff of 3 or more errors for the diagnosis of dementia. Table 16-7 provides the breakdown of the cohort of referred patients according to this cutoff.

When we set the cutoff of 3 or more, SIS has a sensitivity of 0.81 (278/345) and a specificity of 0.91 (278/306). We can also calculate the LR, exactly as we did in Table 16-3. The LR for SIS greater than or equal to 3 is therefore $(278/345)/(28/306) = 8.8$, and the LR for SIS less than 3 is $(67/345)/(278/306) = 0.21$. LR for a positive test result is often denoted as LR+, and that for a negative test result is denoted as LR–.

Let us now try to resolve our clinical scenario using this dichotomized $2 \times 2$ table. We had supposed that the pretest probability for the woman in the opening scenario was 20%, and she had made 5 errors. Because the SIS score of 5 is associated here with an LR+ of 8.8, using Fagan’s nomogram, we arrive at the posttest probability of around 70%, a figure considerably lower than the 90% that we had arrived at when we had a specific LR for 5 errors. This is because the

<table>
<thead>
<tr>
<th>TABLE 16-6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison of the Results of a Diagnostic Test With the Results of Reference Standard Using a $2 \times 2$ Table</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Results</th>
<th>Disease Present</th>
<th>Disease Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Test negative</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Sensitivity (Sens) = \[ \frac{TP}{TP + FN} \]

Specificity (Spec) = \[ \frac{TN}{FP + TN} \]

Likelihood ratio for positive test (LR+) = \[ \frac{Sens}{1 - Spec} \] True positive rate = \[ \frac{TP}{TP + FN} \]

False positive rate = \[ \frac{FP}{FP + TN} \] Likelihood ratio for negative test (LR–) = \[ \frac{1 - Sens}{Spec} \] False negative rate = \[ \frac{FN}{TP + FN} \]

True negative rate = \[ \frac{TN}{FP + TN} \]

Abbreviations: FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Sensitivity is the proportion of people with a positive test result among those with the target condition. Specificity is the proportion of people with a negative test result among those without the target condition.
TABLE 16-7

Comparison of the Results of a Diagnostic Test (Six-Item Screener) With the Results of Reference Standard (Consensus DSM-IV and ICD-10 Diagnosis) Using the Recommended Cutoff

<table>
<thead>
<tr>
<th></th>
<th>Dementia (+)</th>
<th>Dementia (–)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS ≥ 3</td>
<td>278</td>
<td>28</td>
</tr>
<tr>
<td>SIS &lt; 3</td>
<td>67</td>
<td>278</td>
</tr>
<tr>
<td>Sum</td>
<td>345</td>
<td>306</td>
</tr>
</tbody>
</table>

Abbreviations: DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, (Fourth Edition); ICD-10, International Classification of Diseases, Tenth Revision; SIS, Six-Item Screener.

dichotomized LR+ for SIS scores of 3 or more pooled strata for SIS scores of 3, 4, 5, and 6, and the resultant LR is thus diluted by the adjacent strata.

Although the difference between 70% and 90% may not dictate change in management strategies for the case in the clinical scenario, this will not always be the case. Consider a third patient, an elderly man with a pretest probability of 50% of dementia who has surprised us by making not a single error on the SIS. With the dichotomous LR+/LR– approach (or, for that matter, with the sensitivity/specificity approach, because these are mathematically equivalent and interchangeable), you combine the pretest probability of 50% with the LR– of 0.21 and arrive at the posttest probability of about 20%, likely necessitating further neuropsychological and other examinations. The true posttest probability for this man when we apply the LR associated with a score of 0 from Table 16-3 (0.06) is only about 5%. With this posttest probability, you (and the patient and his family) can feel relieved and be spared of further testing and further distress.

In summary, using multiple cuts or thresholds (sometimes referred to as multilevel LRs or stratum-specific LRs) has 2 key advantages over the sensitivity/specificity approach. First, for a test that produces continuous scores or a number of categories (which many tests in medicine do), using multiple thresholds retains as much information as possible. Second, knowing the LR of a particular test result, you can use a simple nomogram to move from the pretest to the posttest probability that is linked to your patient.
**How Can I Apply the Results to Patient Care?**

**Will the Reproducibility of the Test Result and Its Interpretation Be Satisfactory in My Clinical Setting?**

The value of any test depends on its ability to yield the same result when reapplied to stable patients. Poor reproducibility can result from problems with the test itself (e.g., variations in reagents in radioimmunoassay kits for determining hormone levels) or from its interpretation (e.g., the extent of ST-segment elevation on an electrocardiogram). You can easily confirm this when you recall the clinical disagreements that arise when you and 1 or more colleagues examine the same electrocardiogram, ultrasonograph, or CT scan (even when all of you are experts).

Ideally, an article about a diagnostic test will address the reproducibility of the test results using a measure that corrects for agreement by chance (see Chapter 17.3, Measuring Agreement Beyond Chance), especially for issues of interpretation.

If the reported reproducibility of a test in the study setting is mediocre and disagreement between observers is common, and yet the test still discriminates well between those with and without the target condition, the test is likely to be useful. Under these circumstances, the likelihood is good that the test can be readily applied to your clinical setting.

If, on the other hand, reproducibility of a diagnostic test is high, either the test is simple and unambiguous or those interpreting it are highly skilled. If the latter applies, less skilled interpreters in your own clinical setting may not do as well. You will either need to obtain appropriate training (or ensure that those interpreting the test in your setting have that training) or look for an easier and more robust test.

**Are the Study Results Applicable to the Patients in My Practice?**

Test properties may change with a different mix of disease severity or with a different distribution of competing conditions. When patients with the target disorder all have severe disease, LRs will move away from a value of 1.0 (i.e., sensitivity increases). If patients are all mildly affected, LRs move toward a value of 1.0 (i.e., sensitivity decreases). If patients without the target disorder have competing conditions that mimic the test results observed for patients who do have the target disorder, the LRs will move closer to 1.0 and the test will appear less useful (i.e., specificity decreases). In a different clinical setting in which fewer of the disease-free patients have these competing conditions, the LRs will move away from 1.0 and the test will appear more useful (i.e., specificity increases).

Investigators have demonstrated the phenomenon of differing test properties in different subpopulations for exercise electrocardiography in the diagnosis of coronary artery disease. The more extensive the severity of coronary artery disease, the larger the LRs of abnormal exercise electrocardiography for angiographic narrowing of the coronary arteries. Another example comes from the diagnosis of venous thromboembolism, in which compression ultrasonography for proximal-vein thrombosis has proved more accurate in symptomatic outpatients than in asymptomatic postoperative patients.
Sometimes, a test fails in just the patients one hopes it will best serve. The LR of a negative dipstick test result for the rapid diagnosis of urinary tract infection is approximately 0.2 in patients with clear symptoms and thus a high probability of urinary tract infection but is more than 0.5 in those with low probability, rendering it of little help in ruling out infection in the latter situation.

If you practice in a setting similar to that of the study, and if the patient under consideration meets all the study eligibility criteria, you can be confident that the results are applicable. If not, you must make a judgment. As with therapeutic interventions, you should ask whether there are compelling reasons why the results should not be applied to the patients in your practice, either because of the severity of disease in those patients or because the mix of competing conditions is so different that generalization is unwarranted. You may resolve the issue of generalizability if you can find an overview that summarizes the results of a number of studies.

**Will the Test Results Change My Management Strategy?**

It is useful, when making and communicating management decisions, to link them explicitly to the probability of the target disorder. For any target disorder, there are probabilities below which a clinician would dismiss a diagnosis and order no further tests—the test threshold. Similarly, there are probabilities above which a clinician would consider the diagnosis confirmed and would stop testing and initiate treatment—the treatment threshold. When the probability of the target disorder lies between the test and treatment thresholds, further testing is mandated (see Chapter 14, The Process of Diagnosis).

If most patients have test results with LRs near 1.0, test results will seldom move us across the test or treatment threshold. Thus, the usefulness of a diagnostic test is strongly influenced by the proportion of patients suspected of having the target disorder whose test results have very high or very low LRs. Among the patients suspected of having dementia, a review of Table 16-3 allows us to determine the proportion of patients with extreme results (either LR > 10 or LR < 0.1). The proportion can be calculated as \((105 + 2 + 64 + 2 + 11 + 163)/(345 + 306)\) or 347/651 = 53%. The SIS is likely to move the posttest probability in a decisive manner in half of the patients suspected of having dementia and examined, an impressive proportion and better than for most of our diagnostic tests.

A final comment has to do with the use of sequential tests. A new test can be integrated into the existing diagnostic pathway in 3 main ways—as replacement, triage, or add-on (Figure 16-2). That is, a new test can replace an existing test in the existing diagnostic pathway, can be performed before the old test so that only patients with particular results on this triage test continue the testing pathway, or can be placed after the old test so that only patients with a particular result on the old test may need this add-on new test.

The LR approach fits in particularly well in thinking about the diagnostic pathway. Each item of the medical history, or each finding on physical examination, represents a diagnostic test. We can use one test to obtain a certain posttest probability that can be further increased or decreased by using another subsequent test. In general, we can also use laboratory tests or imaging procedures in the same way. If 2 tests are closely related, however, application of the second test may provide little or no information, and the
sequential application of LR is will yield misleading results. For example, once one has the results of the most powerful laboratory test for iron deficiency, serum ferritin, additional tests such as serum iron or transferrin saturation add no further useful information.

Clinical prediction rules deal with the lack of independence of a series of tests and provide the clinician with a way of combining their results (see Chapter 17.4, Clinical Prediction Rules). For instance, for patients with suspected pulmonary embolism, one could use a rule that incorporates respiratory symptoms, heart rate, leg symptoms, oxygen saturation, electrocardiographic findings, and other aspects of medical history and physical examination to accurately classify patients with suspected pulmonary embolism as being characterized by high, medium, and low probability.

Will Patients Be Better Off as a Result of the Test?

The ultimate criterion for the usefulness of a diagnostic test is whether the benefits that accrue to patients are greater than the associated risks. How can we establish the benefits and risks of applying a diagnostic test? The answer lies in thinking of a diagnostic test as a therapeutic maneuver (see Chapter 6, Therapy). Establishing whether a test does more good than harm will involve randomizing patients to a diagnostic strategy that includes the test under investigation, creating a management schedule linked to the diagnostic strategy or to one in which the test is not available, and following up patients in both groups to determine the frequency of patient-important outcomes.
When is demonstrating accuracy sufficient to mandate the use of a test, and when does one require a randomized controlled trial? The value of an accurate test will be undisputed when the target disorder is dangerous if left undiagnosed, if the test has acceptable risks, and if effective treatment exists. This is the case for the ventilation perfusion scan for suspected pulmonary embolism. A high-probability or normal/near-normal result of a ventilation perfusion scan may well eliminate the need for further investigation and may result in anticoagulant agents being appropriately given or appropriately withheld (with either course of action having a substantial positive influence on patient outcome).

Sometimes, a test may be completely benign, represent a low resource investment, be evidently accurate, and clearly lead to useful changes in management. Such is the case for use of the SIS in patients with suspected dementia, when test results may dictate reinsurance or extensive investigation and ultimately planning for a deteriorating course.

In other clinical situations, tests may be accurate, and management may even change as a result of their application, but their effect on patient outcome may be far less certain. Consider one of the issues we raised in our discussion of framing clinical questions (see Chapter 3, What Is the Question). There, we considered a patient with apparently resectable non–small-cell carcinoma of the lung and wondered whether the clinician should order a CT scan and base further management on the results or whether an immediate mediastinoscopy should be undertaken. For this question, knowledge of the accuracy of CT scanning is insufficient. A randomized trial of CT-directed management or mediastinoscopy for all patients is warranted, and indeed, investigators have conducted such a trial.²³ Other examples include catheterization of the right side of the heart for critically ill patients with uncertain hemodynamic status and bronchoalveolar lavage for critically ill patients with possible pulmonary infection. For these tests, randomized trials have helped elucidate optimal management strategies.

**USING THE GUIDE**

Although the study itself does not report reproducibility, its scoring is simple and straightforward because you need only count the number of errors made to 6 questions. It does not require any props or visual cues and is therefore unobtrusive and easy to administer. The SIS takes only 1 to 2 minutes to complete (compared with 5 to 10 minutes for the MMSE). The appendix of the published article gives a detailed word-by-word instruction on how to administer the SIS. You believe that you too can administer this scale reliably.

The patient in the clinical scenario is an older woman who was able to come to your clinic by herself but appeared no longer as lucid as she used to be. The Alzheimer Disease Center cohort in the study we had been examining in this chapter consists of people suspected of having dementia by their caregivers and brought to a tertiary care center directly. Their test characteristics were reported to be similar to those observed in the general population cohort, that is, in a sample with less severe presentations. You decide that there is no compelling reason that the study results would not apply to your patient.
You invite your patient back to the office for a follow-up visit and administer the SIS. The result is a score of 4, which, given your pretest probability of 20%, increases the probability to more than 60%. After hearing that you are concerned about her memory and possibly about her function, she agrees to a referral to a geriatrician for more extensive investigation.

References


