Chapter 18: Diagnostic Tests

Toshi A. Furukawa; Sharon E. Strauss; Heiner C. Bucher; Agoritsas Thomas; Gordon Guyatt

Introduction

In the previous 2 chapters (Chapter 16, The Process of Diagnosis, and Chapter 17, 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 explain how to use an article that addresses 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 explain how to use articles that integrate a number of test results into a *clinical prediction rule* (Chapter 19.4, 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 saw a 70-year-old woman who lives alone and has been managing well. On this visit, she informed you of a long-standing problem, joint pain in her lower extremities. During the visit, you get 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 State Examination (MMSE), with which you are familiar, takes too long. You wonder if 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 to identify patients who need more extensive investigation for possible dementia?" To conduct a rapid and specific search, you access the PubMed Clinical Queries page (see Chapter 5, Finding Current Best Evidence). Typing in the search terms "identify dementia brief MMSE," you select "diagnosis" as the clinical study category and "narrow" as the scope of the filter. This search strategy yields 8 citations.

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 that reports results for an instrument named Six-Item Screener (SIS) meets both criteria.[1] 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.

How Serious Is the Risk of Bias?

Box 18-1 summarizes our *Users' Guides* for assessing the *risk of bias,* examining the results, and determining the applicability of a study reporting on the accuracy of a diagnostic test.

BOX 18-1

Users' Guide for an Article About Interpreting Diagnostic Test Results

How serious is the risk of bias?
- Did participating patients constitute a representative sample of those presenting with a diagnostic dilemma?
- Did investigators compare the test to an appropriate, independent reference standard?
- Were those interpreting the test and reference standard blind to the other result?
- Did all patients receive the same reference standard irrespective of the test results?

What are the results?
- What likelihood ratios were associated with the range of possible test results?

How can I apply the results to patient care?
- Will the reproducibility of the test results and their 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?

Did Participating Patients Constitute a Representative Sample of Those Presenting With a Diagnostic Dilemma?

A diagnostic test is useful only if it distinguishes among conditions or disorders that might otherwise be confused. 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.

We label studies with unrepresentative patient selection as suffering from *spectrum bias* (see Chapter 19.1, Spectrum Bias). There are 3 empirical studies that have systematically examined for various sources of *bias* in studies of diagnostic tests.[2-4] All 3 studies documented bias associated with unrepresentative patient selection.

The story of carcinoembryonic antigen (CEA) testing in patients with colorectal cancer reveals 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 healthy people, pregnant women, or patients with a variety of other conditions.[5] 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.

Enrolling *target-positive* patients (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 diagnostic test's power. This *case-control design* (where cases are known to be target positive and controls are known to be target negative) of a diagnostic test may be likened to a phase 2 efficacy trial: if it fails (ie, the test fails to discriminate target-positive from target- negative patients), the test is hopeless; if it succeeds, it cannot guarantee real-world effectiveness.

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.

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 independence as one criterion for a satisfactory reference standard.

For instance, consider a study that evaluated the utility of abdominojugular reflux for the diagnosis of congestive heart failure. Unfortunately, this study used clinical and radiographic criteria that included the abdominojugular reflex as the reference test.[6] Another example comes from a study evaluating screening instruments for depression in terminally ill people. The authors claimed perfect performance (*sensitivity* of 1.0 and *specificity* of 1.0) for a single question ("Are you depressed?") to detect depression. Their diagnostic criteria included 9 questions of which one was, "Are you depressed?"[7]

In reading articles about diagnostic tests, if you cannot accept the reference standard (within reason; after all, nothing is perfect), then the article is unlikely to provide trustworthy results.[3]

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

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

Consider how, once clinicians see a pulmonary nodule on a computed tomogram (CT), 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 independent 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 of Lijmer et al[2] found bias associated with unblinded assessments, although the magnitude was small.

Did Investigators Perform the Same Reference Standard in 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*[8,9] or *work-up bias).*[10,11] 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. This type of verification bias is known as *partial verification bias.*

Second, results of the index test may be verified by different reference standards. Use of different reference tests for positive and negative results is known as *differential verification bias.*

Verification bias proved a problem for the Prospective Investigation of Pulmonary Embolism 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 scans (92%). This 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-probability 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 or near-normal scans who did not undergo angiography (136 patients) or in 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 up 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 at the time 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 random sample of black persons 65 years and older in the general population; the other, a *consecutive sample* of unscreened patients referred by family, caregivers, or health care professionals 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 search for evidence). 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) and *International Statistical Classification of Diseases, 10th Revision* (*ICD-10*) criteria, based on an assessment by a geriatric psychiatrist or a neurologist that included history, physical and neurologic examination, a complete neuropsychological test battery that included the MMSE and 5 other tests, and an 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 result. To resolve the question, you email the first author and ask for clarification. A couple of emails 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, and social psychologist, a geriatrician, and a neuropsychologist" made the reference standard diagnoses. The author reports, "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 their ability to change our initial 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 back in the shoes 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 issue that her memory is not what it used to be.

The other is an 85-year-old woman, another long-standing 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 a mess. However, 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 (ie, their pretest probabilities) are quite different. In the first woman, the probability is relatively low, perhaps 20%; in the second, relatively high, perhaps 70%.

The results of a formal screening test (eg, the SIS) 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[1] to illustrate the usefulness of LRs. Table 18-1 presents the distribution of the SIS scores in the cohort of patients from the study by Callahan et al.

TABLE 18-1

Six-Item Screener (SIS) Scores in Patients With and Without Dementia and Corresponding Likelihood Ratios

SIS Score	Dementia	No Dementia	Likelihood Ratio
6	105	2	47
5	64	2	28
4	64	8	7.1
3	45	16	2.5
2	31	35	0.79
1	25	80	0.28
0	11	163	0.06
Total	345	306	

How likely is a test result of 6 among people who have dementia? Table 18-1 indicates that 105 of 345 people (30.4%) with the condition made 6 errors. We can also see that of 306 people without dementia, 2 (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 18-1 provides the LR for each possible SIS score.

How can we interpret LRs? Likelihood ratios indicate the extent to which a given diagnostic test result will raise or lower the pretest probability of the target disorder. An LR of 1 tells us that the posttest probability is exactly the same as the pretest probability. Likelihood ratios greater than 1.0 increase the probability that the target disorder is present; the higher the LR, the greater the 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.

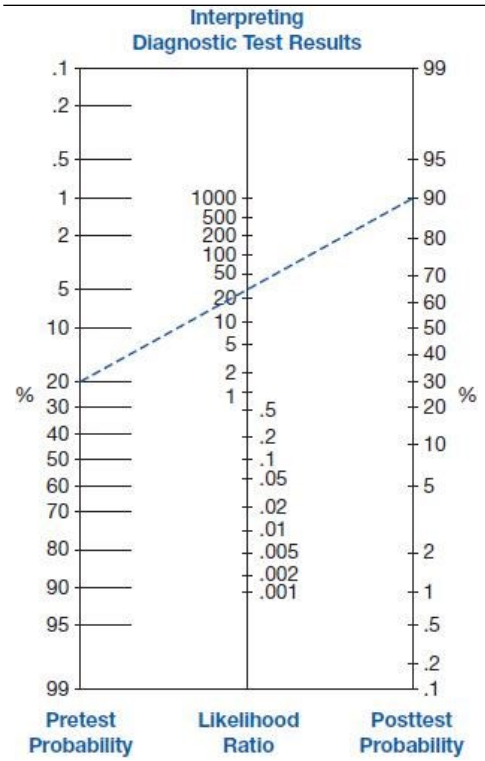
How big is a "big" LR, and how small is a "small" one? Use of LRs in your day-to-day practice will lead to your own sense of their interpretation, but consider the following a rough guide: LRs greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability, 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. If you wonder why the conversion to odds is necessary, consider the fact that LRs compare the likelihood of a test result between patients with and without a target disease (corresponding to the odds of that disease). The calculation is complicated, but there are now several Internet pages and smartphone applications that do this for you (http://meta.cche.net/clint/templates/calculators/lr_nomogram.asp and http://www.cebm.net/nomogram.asp or http://medcalc3000.com and https://itunes.apple.com/app/twobytwo/id436532323?mt=8).

When you do not have access to them, one strategy is to use the *nomogram* proposed by Fagan[13] (Figure 18-1), which does all of 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 represents 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.

FIGURE 18-1

Likelihood Ratio Nomogram



Recall the elderly woman from the opening scenario with suspected dementia. We have decided that the probability of this patient having the condition is approximately 20%. 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 get a posttest probability of approximately 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 17, 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 does not allow precise estimates of pretest probability, 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 get the posttest probability of approximately 80% and above 90%. Table 18-2 tabulates the posttest probabilities corresponding with each possible SIS score for the 65-year-old woman in the clinical scenario.

TABLE 18-2

Pretest Probabilities, Likelihood Ratios of the Six-Item Screener, and Posttest Probabilities in the 70-Year-Old Woman With Moderate Suspicion of Dementia

Pretest Probability, % (Range)[a]	SIS Score (LR)	Posttest Probability, % (Range)[a]
20 (10-30)	6 (47)	92 (84-95)
	5 (28)	88 (76-92)
	4 (7.1)	64 (44-75)
	3 (2.5)	38 (22-52)
	2 (0.79)	16 (8-25)
	1 (0.28)	7 (3-11)
	0 (0.06)	1 (1-3)

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

[a]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.

We can repeat this exercise for our second patient, the 85-year-old woman who had lost her way. You estimate that her history and presentation are compatible with a 70% probability of dementia. Using our nomogram (Figure 18-1), 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 of 60% to 80%), LRs, and posttest probabilities associated with each of these possible SIS scores are presented in Table 18-3.

TABLE 18-3

Pretest Probabilities, Likelihood Ratios of the Six-Item Screener, and Posttest Probabilities in the 85-Year-Old Woman With High Suspicion of Dementia

Pretest Probability, % (Range)[a]	SIS Score (LR)	Posttest Probability, % (Range)[a]
70 (60-80)	6 (47)	99 (99-99)
	5 (28)	98 (98-99)
	4 (7.1)	94 (91-97)
	3 (2.5)	85 (79-76)
	2 (0.79)	65 (54-76)
	1 (0.28)	40 (30-53)
	0 (0.06)	12 (8-19)

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

[a]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 60% to 80% would also be reasonable estimates.

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. The Rational Clinical Examination[14] is a series of systematic reviews of the diagnostic properties of the history and physical examination that have been published in *JAMA* (an updated database is available on the JAMAevidence homepage at jama evidence.com/resource/523). Chapter 19.2 lists a large number of examples of LRs. Further examples are accumulated on the JAMAevidence website (http://www.jamaevidence.com).

Dichotomizing Continuous Test Scores: Sensitivity, Specificity, and Likelihood Ratios

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 that address diagnostic tests report a 2 × 2 table and its associated sensitivity and specificity, as in Table 18-4, and to go along with it a figure that depicts the overall power of the diagnostic test (called a *receiver operating characteristic curve*).

TABLE 18-4

Comparison of the Results of a Diagnostic Test With the Results of Reference Standard Using a 2 × 2 Table

Test Results	Reference Standard	
	Disease Present	Disease Absent
Test result positive	True positive (TP)	False positive (FP)
Test result negative	False negative (FN)	True negative (TN)
$\text{Sensitivity} = \frac{TP}{TP + FN}$		
$\text{Specificity} = \frac{TN}{FP + TN}$		
$\text{Likelihood Ratio for Positive Test Result (LR+)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{\text{TP rate}}{\text{FP rate}} = \frac{TP/(TP + FN)}{FP/(FP + TN)}$		
$\text{Likelihood Ratio for Negative Test Result (LR−)} = \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{\text{FN rate}}{\text{TN rate}} = \frac{FN/(TP + FN)}{TN/(FP + TN)}$		

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.

The study by Callahan et al recommends a cutoff of 3 or more errors for the diagnosis of dementia. Table 18-5 provides the breakdown of the cohort of referred patients according to this cutoff.

TABLE 18-5

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

SIS Score	Dementia	No Dementia
≥3	278	28
<3	67	278
Total	345	306

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

When we set the cutoff of 3 or more, the SIS has a sensitivity of 0.81 (278/345) and a specificity of 0.91 (278/306). We can also calculate the LRs, exactly as we did in Table 18-1. The LR for an SIS score of 3 or greater is therefore (278/345)/(28/306) = 8.8, and the LR for an SIS score less than 3 is (67/345)/(278/306) = 0.21. The LR for a positive test result is often denoted as LR+ and that for a negative test result as LR−.

Let us now try to resolve our clinical scenario using this dichotomized 2 × 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 approximately 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 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 gentleman with a pretest probability of 50% of dementia who has surprised us by not making a single error on the SIS. With the dichotomous LR+/LR− approach (or, for that matter, with the sensitivity and 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 approximately 20%, very 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 18-1 (0.06) is only approximately 5%. With this posttest probability, you (and the patient and his family) can feel relieved and, at least for the time being, be spared further testing.

In summary, use of multiple cuts or thresholds (sometimes referred to as multilevel LRs or stratum-specific LRs) has 2 key advantages over the sensitivity and specificity approach. First, for a test that produces continuous scores or a number of categories (which many tests in medicine do, notably many laboratory tests), use of multiple thresholds retains as much information as possible. Second, knowing the LR of a particular test result, one can use a simple nomogram to move from the pretest to the posttest probability that is linked to your own patient.

USING THE GUIDE

Thus far, we have established that the results are likely true for the people who were included in the study, and we have calculated the multilevel LRs associated with each possible score of the test. We have indicated how the results could be applied to our patient (although we do not yet know the patient's score and have not decided how to proceed when we do).

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 (eg, variations in reagents in radioimmunoassay kits for determining hormone levels) or from its interpretation (eg, 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 one or more colleagues examine the same electrocardiogram, ultrasonogram, or CT (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 19.3, Measuring Agreement Beyond Chance), especially for issues that involve interpretation or judgment.

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 very useful. Under these circumstances, there is a good chance that the test can be readily applied to your clinical setting.

If reproducibility of a diagnostic test is very high, either the test is simple and unambiguous or those interpreting the results 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 (ie, sensitivity increases). If patients are all mildly affected, LRs move toward a value of 1.0 (ie, sensitivity decreases). If patients without the target disorder have competing conditions that mimic the test results seen in patients who have the target disorder, the LRs will move closer to 1.0, and the test will appear less useful (ie, 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 (ie, specificity increases). Differing prevalence in your setting may alert you to the possibility that the spectrum of target-positive and target-negative patients could differ in your practice.[15]

Investigators have reported the phenomenon of differing test properties in different subpopulations for exercise electrocardiography in the diagnosis of coronary artery disease. The more severe the coronary artery disease, the larger the LRs of abnormal exercise electrocardiograph results for angiographic narrowing of the coronary arteries.[16] Another example comes from the diagnosis of venous thromboembolism, where compression ultrasonography for proximal-vein thrombosis has proved more accurate in symptomatic outpatients than in asymptomatic postoperative patients.[17]

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 higher than 0.5 in those with low probability,[18] 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 of 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 here 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 a *systematic review* that summarizes the results of a number of studies.[19]

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 (ie, the treatment threshold). When the probability of the target disorder lies between the test and treatment thresholds, further testing is mandated (see Chapter 16, 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 18-1 allows us to determine the proportion of patients with extreme results (LR >10 or <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—a very impressive proportion and better than for most of our diagnostic tests.

A final comment has to do with the use of sequential tests. The LR approach fits in particularly well in thinking about the diagnostic pathway. Each item of history—or each finding on physical examination—represents a diagnostic test in itself. We can use one test to get 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 very closely related, however, application of the second test may provide little or no additional information, and the sequential application of LRs 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.[20] Once one has conducted an SIS, additional information from the MMSE is likely to be minimal.

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 19.4, Clinical Prediction Rules). For instance, in patients with suspected pulmonary embolism, one could use a rule that incorporates leg symptoms, heart rate, hemoptysis, and other aspects of the history and physical examination to accurately classify patients with suspected pulmonary embolism as being characterized by high, medium, and low probability.[21]

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.[22] 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 7, Therapy [Randomized Trials]). Establishing whether a test does more good than harm will involve (1) randomizing patients to a diagnostic strategy that includes the test under investigation and a management schedule linked to it, or to one in which the test is not available, and (2) following up patients in both groups forward in time 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 clinical 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 CT-angiogram for suspected pulmonary embolism. A high probability or normal or near-normal results of the CT-angiogram 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 reassurance or extensive investigation and ultimately planning for a tragic 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 4, 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 positron emission tomogram (PET)–CT and base further management on the results or use alternative diagnostic strategies. For this question, knowledge of the accuracy of CT is insufficient. A randomized trial of PET-CT–directed management or an alternative strategy for all patients is warranted. 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.

CLINICAL SCENARIO RESOLUTION

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. The SIS does not require any props or visual cues and is therefore unobtrusive, easy to administer, and takes only 1 to 2 minutes to complete (compared with 5 to 10 minutes for the MMSE). Although you note that trained research staff administered the SIS, the appendix of the 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

1. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. *Med Care*. 2002;40(9):771-–781. [PubMed: 12218768]

2. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061-–1066. [PubMed: 10493205]

3. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174(4):469-–476. [PubMed: 16477057]

4. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic

review. *Ann Intern Med*. 2004;140(3):189-–202. [PubMed: 14757617]

5. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci USA*. 1969;64(1): 161-–167. [PubMed: 5262998]

6. Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest*. 1990;97(4):776-–781. [PubMed: 2182296]

7. Chochinov HM, Wilson KG, Enns M, Lander S. "Are you depressed?" screening for depression in the terminally ill. *Am J Psychiatry*. 1997;154(5):674-–676. [PubMed: 9137124]

8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39(1):207-–215. [PubMed: 6871349]

9. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making*. 1984;4(2):151-–164. [PubMed: 6472063]

10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926-–930. [PubMed: 692598]

11. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45(6):581-–586. [PubMed: 1607897]

12. PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA*. 1990;263(20):2753-–2759. [PubMed: 2332918]

13. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med*. 1975;293(5):257.

14. Sackett DL, Rennie D. The science of the art of the clinical examination. *JAMA*. 1992;267(19):2650-–2652. [PubMed: 1573756]

15. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ*. 2013;185(11):E537-–E544. [PubMed: 23798453]

16. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med*. 1984;77(1):64-–71. [PubMed: 6741986]

17. Ginsberg JS, Caco CC, Brill-Edwards PA, et al. Venous thrombosis in patients who have undergone major hip or knee surgery: detection with compression US and impedance plethysmography. *Radiology*. 1991;181(3):651-–654. [PubMed: 1947076]

18. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117(2):135-–140. [PubMed: 1605428]

19. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*. 2008;149(12):889-–897. [PubMed: 19075208]

20. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med*. 1992;7(2):145-–153. [PubMed: 1487761]

21. van Belle A, Büller HR, Huisman MV, et al; Christopher Study Investigators. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *JAMA*. 2006;295(2):172-–179. [PubMed: 16403929]

22. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134(6):587--594. [PubMed: 3512062]