1C2

DIAGNOSTIC TESTS

Roman Jaeschke, Gordon Guyatt, and Jeroen Lijmer

The following EBM Working Group members also made substantive contributions to this section: Peter Wyer, Virginia Moyer, Deborah Cook, Jonathan Craig, Luz Maria Letelier, John Williams, C. David Naylor, W. Scott Richardson, Mark Wilson, and James Nishikawa

IN THIS SECTION

CLINICAL SCENARIO

How Accurate is CT Scanning in Suspected Appendicitis?

A 32-year-old woman enters the emergency department presenting with right lower quadrant pain. She is single and is employed by a company that sells Internet-related products. She is sexually active, having had three sexual partners during the past year, and her last menstrual period ended 3 weeks ago. Yesterday, she began to feel unwell and lost her appetite. During the past few hours the pain became much worse and she felt febrile, but she did not take her temperature. She has not experienced any vaginal discharge. She came to the emergency department when the pain became so severe that she started to worry whether something serious might be wrong.

On examination, you see a moderately ill woman with a temperature of 38.2° C and otherwise normal vital signs, who displays tenderness and guarding in the right lower quadrant and questionable rebound tenderness. You find no cervical motion tenderness, nor do you see cervical discharge. Laboratory examination findings include a white blood cell count of 11,000/mm^3. Your differential diagnosis includes appendicitis, pelvic inflammatory disease, and ectopic pregancy; as you are debating whether to refer directly to surgery or to begin by obtaining a gynecologist's opinion, your colleague, an interventional radiologist, stops by on his way back from performing an emergency pulmonary angiogram. You describe the patient you are attending to and he mentions that up to 15% of needless laparotomies and up to 20% of admissions can be avoided if a computed tomographic (CT) scan is performed in patients like this one. He mentions "a very good paper that you must read, since it was published in the *New England Journal of Medicine*" although the citation and the details of the investigators' methods and study results currently escape him.

The patient is stable and currently comfortable, and the emergency department has quieted down since the morning rush. A colleague is ready to allow you a break and you decide you can afford to invest 30 minutes to look for and examine the paper recommended by the radiologist.

FINDING THE EVIDENCE

Upstairs in the library, you use the computer to search the PubMed database. You select "diagnosis" and "specificity" from the clinical queries page (www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html) to have a preformatted search for diagnostic test studies. With the key words "CT" and "appendicitis," the search yields 39 citations. When you limit the search to English-language papers with abstracts that were published during the past 5 years, you find that 18 recent articles remain. The 18 abstracts include two narrative reviews, four retrospective studies, two studies focusing on specific imaging signs, and two studies focusing on a selected group of patients. Two of the abstracts provide no quantitative information about the test's performance and one is from a journal your library does not carry. The remaining five abstracts report a high level of accuracy of the test. The title of the most recent article best fits the patient with right lower quadrant tenderness in that it refers to the value of helical CT scanning for differentiating between appendicitis and acute gynecologic conditions.[1] Furthermore, the *New England Journal of Medicine* article is older and seems less relevant in that it analyzes issues related to cost and patient impact, rather than focusing on diagnostic test accuracy. You decide to retrieve the more recent paper.

In the ensuing discussion of the validity, results, and applicability of studies examining the properties of diagnostic tests, we will focus both on the scenario that included the diagnosis of pulmonary embolism using ventilation-perfusion scanning (see Part 1C, "The Process of Diagnosis") and on the article about the value of CT scanning in the diagnosis of appendicitis. Table 1C-2 summarizes our Users' Guide for a study of interpreting test results.

TABLE 1C–2

Users' Guide for an Article About Interpreting Diagnostic Test Results

Are the results valid?

- Did clinicians face diagnostic uncertainty?
- Was there a blind comparison with an independent gold standard applied similarly to the treatment group and to the control group?
- Did the results of the test being evaluated influence the decision to perform the gold standard?

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 result and its interpretation be satisfactory in my clinical setting?
- Are the results applicable to the patient in my practice?
- Will the results change my management strategy?
- Will patients be better off as a result of the test?

ARE THE RESULTS VALID?

Did Clinicians Face Diagnostic Uncertainty?

A diagnostic test is useful only to the extent that it distinguishes between conditions or disorders that might otherwise be confused. Almost any test can differentiate healthy persons from severely affected ones; this ability, however, tells us nothing about the clinical utility of a test. The true, pragmatic value of a test is therefore established only in a study that closely resembles clinical practice. Another way to understand this point is to refer back to Figure 1C-1 in Part 1C, "The Process of Diagnosis." Note that the population of interest comprises patients whose predicament falls between the test and treatment thresholds.

A vivid example of how choosing the right population can dash the hopes raised with the introduction of a diagnostic test comes from the story of carcinoembryonic antigen (CEA) testing in patients with colorectal cancer. When measured in 36 people with known advanced cancer of the colon or rectum, CEA was elevated in 35 of them. At the same time, much lower levels were found in people without cancer who suffered from 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 and clinicians abandoned CEA measurement for cancer diagnosis and screening. Carcinoembryonic antigen testing has proved useful only as one element in the follow-up of patients with known colorectal cancer.[3]

In an empiric study of design-related bias in studies of diagnostic tests, Lijmer and colleagues related features of the design to the power of tests.[4] Their findings included a large overestimate of the power of the test to distinguish between target-positive and target-negative patients when the investigators enrolled separate test and normal control populations (relative diagnostic odds ratio, [OR] 3.0; 95% confidence interval [CI], 2.0-4.5).

This example contrasts with the PIOPED study that demonstrated the utility of ventilation-perfusion scanning in the diagnosis of pulmonary embolism.[5] Here, investigators recruited the whole spectrum of patients suspected of having pulmonary embolism, including those who entered the study with high, medium, and low clinical suspicion of the condition. The patient sample in the helical CT study from the scenario related to scanning and appendicitis mentioned earlier in this section was appropriate because it comprised consecutive, nonpregnant women presenting to the emergency department of a large general hospital—ones in whom acute appendicitis or an acute gynecologic condition was suspected.

Was There a Blind Comparison With an Independent Gold Standard Applied Similarly to the Treatment Group and the Control Group?

The accuracy of a diagnostic test is best determined by comparing it to the "truth." Accordingly, readers must assure themselves that an appropriate *reference standard* (such as biopsy, surgery, autopsy, or long-term follow-up) has been applied to every patient, along with the test under investigation.[6] In the PIOPED study, the investigators used the pulmonary angiogram as the reference standard, and this was as "gold" as could be achieved without sacrificing the patients.

One way a *gold standard* can go wrong is if the test is part of the gold standard. For instance, one study evaluated the utility of measuring both serum and urinary amylase in making the diagnosis of pancreatitis.[7] The investigators constructed a gold standard that relied on a number of tests, including ones for serum and urinary amylase. This incorporation of the test into the gold standard is likely to inflate the estimate of the test's diagnostic power. Thus, clinicians should insist on the independence of the test and gold standard.

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 for your purposes. If you do accept the reference standard, the next question to ask is whether the test results and the reference standard were assessed blindly (that is, by interpreters who were unaware of the results of the other investigation). Clinical experience demonstrates the importance of this independence or *blinding*. Once clinicians see a pulmonary nodule on a CT scan, they can see the previously undetected lesion on the chest radiograph; once they learn the results of an echocardiogram, they hear the previously inaudible cardiac murmur. The Lijmer et al empiric study of diagnostic test bias to which we have referred demonstrated the bias associated with unblinding even though the magnitude was small (relative diagnostic OR, 1.3; 95% CI, 1.0-1.9).[4]

The more likely that knowledge of the reference standard result could influence the interpretation of a new test, the greater is the importance of the blinded interpretation. Similarly, the more susceptible the gold standard is to changes in interpretation as a result of knowledge of the test, the more important is the blinding of the gold standard interpreter. In their study, the PIOPED investigators did not state explicitly that the tests were interpreted blindly. However, one could deduce from the effort they put into ensuring reproducible, independent readings that the interpreters were, in fact, blind; through correspondence with one of the authors, we have confirmed that this was indeed the case.

In the study of the use of CT in the diagnosis of suspected appendicitis, the investigators used surgical and pathologic findings as the reference standard for patients who went to surgery. For patients who did not go to surgery, the findings at clinical follow-up—including outpatient clinic visits and telephone calls during at least a 2-month period after the CT scan—provided the gold standard. The researchers did not report blinding of the physicians for the results of the helical CT scan. Particularly for patients in whom the diagnosis was made by long-term follow-up, knowledge of the CT result could have created a bias toward making the test look better than it really was.

Did the Results of the Test Being Evaluated Influence the Decision to Perform the Reference Standard?

The properties of a diagnostic test will be distorted if its results influence whether patients undergo confirmation by the reference standard. This situation, sometimes called *verification bias*[8,9] or *workup bias,*[10,11] applies when, for example, patients with suspected coronary artery disease whose exercise test results are positive are more likely to undergo coronary angiography (the gold standard) than those whose exercise test results are negative. The Lijmer et al study showed a large magnitude of bias associated with use of different reference tests for positive and negative results.[4]

Verification bias proved a problem for the PIOPED study as well. Patients whose ventilation-perfusion scans 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, since clinicians might be reluctant to subject patients with a low probability of pulmonary embolism to the risks of angiography.

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. However, the PIOPED investigators applied a second reference standard to the 150 patients with low-probability or normal/near normal scans who failed to undergo angiography (136 patients) or in whom angiogram interpretation was uncertain (14 patients): they would be judged to be free of pulmonary embolism if they did well without treatment. Accordingly, the PIOPED investigators followed each of these patients for 1 year without treating them with anticoagulant drugs. Clinically evident pulmonary embolism developed in none of these patients during this time, from which we can conclude that clinically important pulmonary embolism (if we define clinically important pulmonary embolism as requiring anticoagulation therapy to prevent subsequent adverse events) was not present at the time they underwent ventilation-perfusion scanning.

In the helical CT study, the investigators established the reference standard in all patients. However, the test results probably influenced which reference standard—surgery or follow-up—was chosen. As we have mentioned previously, to the extent that CT results influenced the decision regarding the final diagnosis, the study provides an excessively optimistic picture of the test properties.

WHAT ARE THE RESULTS?

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

The starting point of any diagnostic process is the patient presenting with a constellation of symptoms and signs. Consider two patients with nonspecific chest pain and shortness of breath without findings suggesting diagnoses such as pneumonia, airflow obstruction, or heart failure, in whom the clinician suspects pulmonary embolism. One is a 78-year-old woman 10 days after surgery and the other is a 28-year-old man experiencing a high level of anxiety. Our clinical hunches about the probability of pulmonary embolism as the explanation for these two patients' complaints—that is, their pretest probabilities—are very different. In the older woman, the probability is high; in the young man, it is low. As a result, even if both patients have intermediate-probability ventilation-perfusion scans, subsequent management is likely to differ in each. One might well treat the elderly woman immediately with heparin but order additional investigations in the young man.

Two conclusions emerge from this line of reasoning. First, regardless of the results of the ventilation-perfusion scan, they do not tell us whether pulmonary embolism is present. What they do accomplish is to 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 likelihood ratio.

As depicted in Table 1C-3, constructed from the results of the PIOPED study, there were 251 people with angiographically proven pulmonary embolism and 630 people whose angiograms or follow-up excluded that diagnosis. For all patients, ventilation-perfusion scans were classified into four levels: high probability, intermediate probability, low probability, and normal or near-normal. How likely is a high-probability scan among people who do have pulmonary embolism? Table 1C-3 shows that 102 of 251 (or 0.406) people with the condition had high-probability scans. How often is the same test result, a high-probability scan, found among people in whom pulmonary embolism was suspected but has been ruled out? The answer is 14 of 630 (or 0.022) of them. The ratio of these two likelihoods is called the *likelihood ratio* (LR); for a high probability scan, it equals 0.406 ÷ 0.022 (or 18.3). In other words, a high-probability ventilation-perfusion scan is 18.3 times as likely to occur in a patient with—as opposed to without—a pulmonary embolism.

TABLE 1C–3

Test Properties of Ventilation Perfusion (V/Q) Scanning

			Pulmonary Embolism		
	Present		Absent		
Scan Results	Number	Proportion	Likelihood Number	Proportion	Ratio
High probability	102	102/251 = 0.406	14	14/630 = 0.022	18.3
Intermediate probability	105	105/251 = 0.418	217	217/630 = 0.344	1.20
Low probability	39	39/251 = 0.155	273	273/630 = 0.433	0.36
Normal/near normal	5	5/251 = 0.020	126	126/630 = 0.200	0.10
Total	251		630		

In a similar fashion, we can calculate the likelihood ratio for each level of the diagnostic test results. Each calculation involves answering two questions: First, how likely it is to obtain a given test result (say, a low-probability ventilation-perfusion scan) among people with the target disorder (pulmonary embolism)? Second, how likely it is to obtain the same test result (again, a low-probability scan) among people without the target disorder? For a low-probability ventilation-perfusion scan, these likelihoods are 39/251 (0.155) and 273/630 (0.433), respectively, and their ratio (the likelihood ratio for low-probability scan) is 0.36. Table 1C-3 provides the results of the calculations for the other scan results.

What do all these numbers mean? The Likelihood ratios indicate by how much a given diagnostic test result will raise or lower the pretest probability of the target disorder. A likelihood ratio of 1.0 means that the posttest probability is exactly the same as the pretest probability. Likelihood ratios >1.0 increase the probability that the target disorder is present, and the higher the likelihood ratio, the greater is this increase. Conversely, likelihood ratios <1.0 decrease the probability of the target disorder, and the smaller the likelihood ratio, the greater is the decrease in probability and the smaller is its final value.

How big is a "big" likelihood ratio, and how small is a "small" one? Using likelihood ratios in your day-to-day practice will lead to your own sense of their interpretation, but consider the following a rough guide:

- Likelihood ratios of >10 or < 0.1 generate large and often conclusive changes from pre- to posttest probability;

- Likelihood ratios of 5–10 and 0.1–0.2 generate moderate shifts in pre- to posttest probability;

- Likelihood ratios of 2–5 and 0.5–0.2 generate small (but sometimes important) changes in probability; and

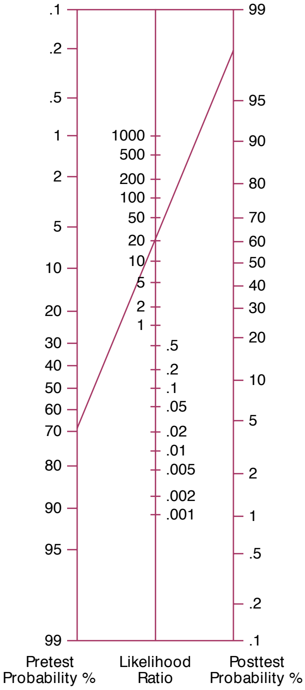
- Likelihood ratios of 1–2 and 0.5–1 alter probability to a small (and rarely important) degree.

Having determined the magnitude and significance of the likelihood ratios, how do we use them to go from pretest to posttest probability? We cannot combine likelihoods directly, the way we can combine probabilities or percentages; their formal use requires converting pretest probability to odds, multiplying the result by the Likelihood ratio, and converting the consequent posttest odds into a posttest probability. Although it is not too difficult (see Part 2B2, "Therapy and Understanding the Results, Measures of Association"), this calculation can be tedious and off-putting; fortunately, there is an easier way.

A *nomogram* proposed by Fagan[12] (Figure 1C-2) 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 likelihood ratio, 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 likelihood ratio for the observed test result.

FIGURE 1C–2

Likelihood Ratio Nomogram



Copyright © 1975 Massachusetts Medical Society. All rights reserved. Reproduced with permission from the Massachusetts Medical Society.

Recall the elderly woman mentioned earlier with suspected pulmonary embolism after abdominal surgery. Most clinicians would agree that the probability of this patient having the condition is quite high—about 70%. This value then represents the pretest probability. Suppose that her ventilation-perfusion scan was reported as being within the realm of high probability. Figure 1C-2 shows how you can anchor a ruler at her pretest probability of 70% and align it with the Likelihood ratio of 18.3 associated with a high-probability scan. The results: her posttest probability is >97%. If, by contrast, her ventilation-perfusion scan result is reported as intermediate (Likelihood ratio, 1.2), the probability of pulmonary embolism hardly changes (it increases to 74%), whereas a near-normal result yields a posttest probability of 19%.

The pretest probability is an estimate. We have already pointed out that the literature dealing with differential diagnosis can help us in establishing the pretest probability (see Part 1C, "The Process of Diagnosis"). Clinicians can deal with residual uncertainty by examining the implications of a plausible range of pretest probabilities. Let us assume the pretest probability in this case is as low as 60%, or as high as 80%. The posttest probabilities that would follow from these different pretest probabilities appear in Table 1C-4.

TABLE 1C–4

Pretest Probabilities, Likelihood Ratios of Ventilation-Perfusion Scan Results, and Posttest Probabilities in Two Patients With Suspected Pulmonary Embolism

Pretest Probability %/(Range)*	Scan Result (LR)	Posttest Probability %/(Range)*
78-Year-Old Woman With Sudden Onset of Dyspnea Following Abdominal Surgery		
70 (60-80)	High Probability (18.3)	97 (96-99)
70 (60-80)	Intermediate Probability (1.2)	74 (64-83)
70 (60-80)	Low Probability (0.36)	46 (35-59)
70 (60-80)	Normal/Near Normal (0.1)	19 (13-29)
28-Year-Old Man With Dyspnea and Atypical Chest Pain		
20 (10-30)	High Probability (18.3)	82 (67-89)
20 (10-30)	Intermediate Probability (1.2)	23 (12-34)
20 (10-30)	Low Probability (0.36)	8 (4-6)
20 (10-30)	Normal/Near Normal (0.1)	2 (1-4)

* 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.

LR indicates Likelihood ratio.

We can repeat this exercise for our second patient, the 28-year-old man. Let us consider that his presentation is compatible with a 20% probability of pulmonary embolism. Using our nomogram (see Figure 1C-2), the posttest probability with a high-probability scan result is 82%; with an intermediate-probability result, it is 23%; and with a near-normal result, it is 2%. The pretest probability (with a range of possible pretest probabilities from 10% to 30%), likelihood ratios, and posttest probabilities associated with each of the four possible scan results also appear in Table 1C-4.

The investigation of women with possible appendicitis showed that the CT scan was positive in all 32 in whom that diagnosis was ultimately confirmed. Of the 68 who did not have appendicitis, 66 had negative scan results. These data translate into a Likelihood ratio of 0 associated with a negative test and a Likelihood ratio of 34 for a positive test. These numbers effectively mean that the test is extremely powerful. A negative result excludes appendicitis, and a positive test makes appendicitis highly likely.

Having learned to use likelihood ratios, you may be curious about where to find easy access to the Likelihood ratios of the tests you use regularly in your own practice. The Rational Clinical Examination[13] is a series of systematic reviews of the diagnostic properties of the history and physical examination that have been published in *JAMA*. Black and colleagues have summarized much of the available information about diagnostic test properties in the form of a medical text.[14] In addition, we provide our own summary of the likelihood ratios of some common tests in another section of this book (see Part 2C, "Diagnosis, Examples of Likelihood Ratios").

Sensitivity and Specificity.

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 two other terms in the lexicon of diagnostic testing: sensitivity and specificity.

You may have noted that our discussion of likelihood ratios omitted any talk of "normal" and "abnormal" tests. Instead, we presented four different ventilation-perfusion scan interpretations, each with its own Likelihood ratio. However, this is not the way the PIOPED investigators presented their results. They relied on the older (but less useful) concepts of sensitivity and specificity.

Sensitivity is the proportion of people with the target disorder in whom a test result is positive and *specificity* is the proportion of people without the target disorder in whom a test result is negative. To use these concepts, we have to divide test results into normal and abnormal categories; in other words, we must create a two-column x two-column table. Table 1C-5 presents the general form of a 2 x 2 table that we use to understand sensitivity and specificity. Look again at Table 1C-5 and observe that we could transform a 4 x 2 table such as Table 1C-4 into any of three such 2 x 2 tables, depending on what we call normal or abnormal (or depending on what we call negative and positive test results). Let us assume that we call only high-probability scans abnormal (or positive).

TABLE 1C-5

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

	Reference Standard	
Test Results	Disease Present	Disease Absent
Disease present	True Positive (*a*)	False Positive (*b*)
Disease absent	False Negative (*c*)	True Negative (*d*)

* Sensitivity (Sens) $= \dfrac{a}{a + c}$

Specificity (Spec) $= \dfrac{d}{b + d}$

Likelihood ratio for positive test (LR+) $= \dfrac{sens}{1 - spec} = \dfrac{a/(a + c)}{b/(b + d)}$

Likelihood ratio for negative test (LR–) $= \dfrac{1 - sens}{spec} = \dfrac{c/(a + c)}{d/(b + d)}$

Table 1C-6 presents a 2 x 2 table comparing the results of a ventilation perfusion scan with the results of pulmonary angiogram as a reference standard.

TABLE 1C–6

Comparison of the Results of Diagnostic Test (Ventilation-Perfusion Scan) With the Results of Reference Standard (Pulmonary Angiogram) Assuming Only High-Probability Scans Are Positive (Truly Abnormal)*

	Angiogram	
Scan Category	Pulmonary Embolism Present	Pulmonary Embolism Absent
High probability	102	14
Others	149	616
Total	**251**	**630**

* Sensitivity, 41%; specificity, 98%; Likelihood ratio of a high-probability test result, 18.3; Likelihood ratio of other results, 0.61.

To calculate sensitivity from the data in Table 1C-6, we look at the number of people with proven pulmonary embolism (251) who were diagnosed as having the target disorder on ventilation-perfusion scan (102) characterized by a sensitivity of 102/251, or approximately 41% (a/a+c). To calculate specificity, we look at the number of people without the target disorder (630) whose ventilation-perfusion scan results were classified as normal (616), yielding a specificity of 616/630, or 98% (d/b+d). We can also calculate likelihood ratios for the positive and negative test results using this cutpoint: 18.3 and 0.61, respectively.

Let us see how the test performs if we decide to put the threshold of positive vs negative in a different place in the table. For example, let us call only the normal/near-normal ventilation perfusion scan result negative. As shown in the 2 x 2 table depicted in Table 1C-7, the sensitivity is now 246/251, or 98% (among 251 people with pulmonary embolism, 246 are diagnosed on ventilation-perfusion scan), but what has happened to specificity? Among 630 people without pulmonary embolism, test results in only 126 are negative (specificity, 20%). The corresponding likelihood ratios are 1.23 and 0.1. Note that with this cut we not only lose the diagnostic information associated with the high-probability scan result, but we also interpret intermediate- and low-probability results as if they increase the likelihood of pulmonary embolism, when in fact they decrease the likelihood. You can generate the third 2 x 2 table by setting the cutpoint in the middle. If your sensitivity and specificity values are 82% and 63%, respectively, and associated Likelihood ratios of a positive and a negative test are 2.25 and 0.28, you have it right.

TABLE 1C-7

Comparison of the Results of Diagnostic Test (Ventilation-Perfusion Scan) With the Results of Reference Standard (Pulmonary Angiogram) Assuming Only Normal/Near-Normal Scans Are Negative (Truly Normal)*

	Angiogram	
Scan Category	**Pulmonary Embolism Present**	**Pulmonary Embolism Absent**
High, intermediate, and low probability	246	504
Near normal/normal	5	126
Total	**251**	**630**

* Sensitivity, 98%; specificity, 20%; Likelihood ratio of high, intermediate, and low probability, 1.23; Likelihood ratio of near normal/normal, 0.1.

In using sensitivity and specificity you must either discard important information or recalculate sensitivity and specificity for every cutpoint. We recommend the Likelihood ratio approach because it is much simpler and much more efficient.

USING THE GUIDE

Thus far, we have established that the results are likely true for the people who were included in the PIOPED study, and we have ascertained the Likelihood ratio associated with different results of the test. We have concluded that the helical CT scanning study may have overestimated the power of the test, but not so seriously as to completely invalidate the results. How useful are the tests likely to be in our clinical practice?

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). A second cause of different test results in stable patients arises whenever a test requires interpretation (eg, the extent of ST-segment elevation on an electrocardiogram). 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 Part 2C, "Diagnosis, Measuring Agreement Beyond Chance"). This is especially important when expertise is required in performing or interpreting the test. You can confirm this by recalling the clinical disagreements that arise when you and one or more colleagues examine the same ECG, ultrasound, or CT scan, even when all of you are experts.

If the 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, it is very useful. Under these circumstances, the likelihood is good that the test can be readily applied to your clinical setting. If reproducibility of a diagnostic test is very high and observer variation is very low, 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.

Using the Guide

The helical CT study made no reference to reproducibility, other than to say that the residents initially interpreted the scans and the consultants agreed in all but one case. The authors did not describe the degree of experience of the radiologists, but the residents' involvement suggests that unusual expertise is not mandatory for accurate interpretation of the images.

Are the Results Applicable to the Patient 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, likelihood ratios will move away from a value of 1.0 (sensitivity increases). If patients are all mildly affected, likelihood ratios move toward a value of 1.0 (sensitivity decreases). If patients without the target disorder have competing conditions that mimic the test results seen in patients who do have the target disorder, the likelihood ratios will move closer to 1.0 and the test will appear less useful (specificity decreases). In a different clinical setting in which fewer of the disease-free patients have these competing conditions, the likelihood ratios will move away from 1.0 and the test will appear more useful (sensitivity increases).

The phenomenon of differing test properties in different subpopulations has been demonstrated most strikingly for exercise electrocardiography in the diagnosis of coronary artery disease. For instance, the more extensive the severity of coronary artery disease, the larger are the likelihood ratios of abnormal exercise electrocardiography for angiographic narrowing of the coronary arteries.[15] Another example comes from the diagnosis of venous thromboembolism, where compression ultrasound for proximal-vein thrombosis has proved more accurate in symptomatic outpatients than in asymptomatic postoperative patients.[16]

Sometimes, a test fails in just the patients one hopes it will best serve. The likelihood ratio of a negative dipstick test 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 over 0.5 in those with low probability,[17] rendering it of little help in ruling out infection in the latter. If you practice in a setting similar to that of the investigation and if the patient under consideration meets all the study inclusion criteria and does not violate any of the exclusion criteria, you can be confident that the results are applicable. If not, a judgment is required. 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. The issue of generalizability may be resolved if you can find an overview that pools the results of a number of studies.[18]

USING THE GUIDE

The participants in the PIOPED study were a representative sample of patients with suspected pulmonary embolism from a number of large general hospitals. Therefore, the results are readily applicable to most clinical practices in North America. There are groups such as critically ill patients to whom we might be reluctant to generalize the results; such patients were excluded from the study and are likely to have had a different spectrum of competing conditions than other patients.

The patients enrolled in the study of CT scanning in acute appendicitis constitute a representative sample of women presenting to the emergency department with right lower quadrant pain. The patient before you, in whom the differential diagnosis includes appendicitis and pelvic inflammatory disease, meets study eligibility criteria. Thus, you can be confident that the results will apply in her case.

Will the Results Change My Management Strategy?

It is useful, when making, learning, teaching, and communicating management decisions, to link them explicitly to the probability of the target disorder. As we have described, 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[19] (see Part 1C, "The Process of Diagnosis").

Once we decide what our test and treatment thresholds are, posttest probabilities have direct treatment implications. Let us suppose that we are willing to treat those patients with a probability of pulmonary embolism of 80% or higher (knowing that we will be treating 20% of them unnecessarily). Furthermore, let us suppose we are willing to dismiss the diagnosis of pulmonary embolism in those with a posttest probability of 10% or less. You may wish to apply different numbers here; the treatment and test thresholds are a matter of judgment and they differ for different conditions depending on the risks of therapy (if risky, you want to be more certain of your diagnosis) and the danger of the disease if left untreated (if the danger of missing the disease is high—such as in pulmonary embolism—you want your posttest probability to be very low before abandoning the diagnostic search). In the 28-year-old man discussed earlier in this section, a high-probability scan results in a posttest probability of 82% and may dictate treatment (or, at least, further investigation) and an intermediate probability scan (23% posttest probability) will dictate further testing (perhaps bilateral leg venography, ultrasound, or pulmonary angiography), whereas a low-probability or normal scan (probabilities of less than 10%) will exclude the diagnosis of pulmonary embolism. In the elderly woman, a high-probability scan dictates treatment (97% posttest probability of pulmonary embolism) and an intermediate result (74% posttest probability) may be compatible with either treatment or further testing (likely a pulmonary angiogram), whereas any other result mandates further testing.

If most patients have test results with Likelihood ratios near 1.0, the test will not be very useful. 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 Likelihood ratios. In the patients suspected of having pulmonary embolism in our ventilation-perfusion scan example, a review of Table 1C-3 allows us to determine the proportion of patients with extreme results (either high probability with an Likelihood ratio of over 10, or normal/near-normal scans with an Likelihood ratio of 0.1). The proportion can be calculated as (102+14+5+126)/881, or 247/881 = 28%. Clinicians who have been frustrated by frequent intermediate- or low-probability results in patients with suspected pulmonary embolism will already know that this proportion (28%) is far from optimal. Thus, despite the high Likelihood ratio associated with a high-probability scan and the low Likelihood ratio associated with a normal/near-normal result, ventilation perfusion scanning is of limited usefulness in patients with suspected pulmonary embolism.

A final comment has to do with the use of sequential tests. We have demonstrated how each item of history—or each finding on physical examination—represents a diagnostic test. We generate pretest probabilities that we modify with each new finding. In general, we can also use laboratory tests or imaging procedures in the same way. However, if two tests are very closely related, application of the second test may provide little or no information, and the sequential application of likelihood ratios 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] *Clinical prediction rules* deal with the lack of independence of a series of tests that can be applied to a diagnostic dilemma and provide the clinician with a way of combining their results (see Part 2C, "Diagnosis, Clinical Prediction Rules"). For instance, the clinician in the scenario that opened this section could have used a rule that incorporates respiratory symptoms, heart rate, leg symptoms, oxygen saturation, electrocardiographic findings, and other aspects of history and physical examination to accurately classify patients with suspected pulmonary embolism as being characterized by high, medium, and low probability.[21]

USING THE GUIDE

Given the extreme likelihood ratios of helical CT scanning in women with abdominal pain, CT results are very likely to change management. For any patient with an intermediate likelihood of appendicitis, a positive scan will suggest immediate surgery, and a negative scan will mandate continued observation with treatment of alternative diagnostic possibilities (in this case, pelvic inflammatory disease).

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 Part 1B1, "Therapy"). Establishing whether a test does more good than harm will involve (1) randomizing patients to a diagnostic strategy that includes the test under investigation or to one in which the test is not available and (2) following patients in both groups forward in time to determine the frequency of patient-important target 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 both of the tests we have considered in detail in this section. A high probability or normal/near-normal results 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).

The researchers who conducted the investigation of helical CT scanning in women with abdominal pain asked clinicians to formulate management plans before CT results were available and compared the plan to the one that clinicians followed after receiving the CT result. Of 100 patients, clinicians sent home 43 patients whom they would otherwise have admitted for observation, and they sent 13 others, whom they would otherwise have observed, to the operating room for immediate appendectomy. The evident benefits for patients and for the health care system—patients prefer to be at home than in a hospital, along with the fact that delayed appendectomy risks additional complications—eliminate the need for a randomized trial of CT scanning vs standard diagnostic approaches in women presenting to the emergency department with abdominal pain.

In other clinical situations, tests may be accurate and management may even change as a result of their application, but their impact on patient outcome may be far less certain. Consider one of the issues we raised in our discussion of framing clinical questions (see Part 1A1, "Finding the Evidence"). We presented 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, bronchoalveolar lavage for critically ill patients with possible pulmonary sepsis, bronchial provocation testing for patients with asthma, and the incremental value of magnetic resonance imaging over CT for a wide variety of problems. For these and many other tests, confidence in the right management strategy must await the conduct of well-designed and adequately powered randomized trials.

CLINICAL RESOLUTION

You are sufficiently impressed by the information in the article about helical CT scanning that you decide to bypass the gynecologic consultation. Your radiologist colleague facilitates an emergent scan and soon calls you back, triumphantly announcing that the results are characteristic of appendicitis. The surgeons are soon having the patient whisked to the operating room, and you later hear that the patient is recovering uneventfully after the removal of her inflamed appendix.

References

1. Rao PM, Feltmate CM, Rhea JT, Schulick AH, Novelline RA. Helical computed tomography in differentiating appendicitis and acute gynecologic conditions. *Obstet Gynecol*. 1999;93:417-421.
2. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci U S A*. 1969;64:161-167.
3. Bates SE. Clinical applications of serum tumor markers. *Ann Intern Med*. 1991;115:623-638.
4. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diangostic tests. *JAMA*. 1999;282:1061-1066.
5. The PIOPED investigators. Value of ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA*. 1990;263:2753-2759.
6. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology, A Basic Science for Clinical Medicine*. 2nd ed. Boston: Little, Brown and Company; 1991:53-57.
7. Kemppainen EA, Hedstrom JI, Puolakkainen PA, et al. Rapid measurement of urinary trypsinogen-2 as a screening test for acute pancreatitis. *N Engl J Med.* 1997;336:1788-1793.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207-215.
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;4:151-164.
10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-930.
11. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45:581-586.
12. Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med*. 1975;293:257.
13. Sackett DL, Rennie D. The science and art of the clinical examination. *JAMA*. 1992;267:2650-2652.
14. Black ER, Bordley DR, Tape TG, Panzer RJ. *Diagnostic Strategies for Common Medical Problems*. 2nd ed. Philadelphia: American College of Physicians; 1999.
15. Hlatky MA, Pryor DB, Harrell FE. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med*. 1984;77:64-71.
16. 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;181:651-654.

17. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the repid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117:135-140.
18. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120:667-676.
19. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology, a Basic Science for Clinical Medicine*. 2nd ed. Boston: Little, Brown and Company; 1991:145-148.
20. Guyatt GH, Oxman A, Ali M, et al. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med*. 1992;7:145-153.
21. Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med*. 1998;129:997-1005.
22. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134:587-594.
23. Canadian Lung Oncology Group. Investigation for mediastinal disease in patients with apparently operable lung cancer. *Ann Thorac Surg*. 1995;60:1382-1389.