How to Use an Article About a Diagnostic Test

Roman Jaeschke, Gordon H. Guyatt, David L. Sackett, and the Evidence Based Medicine Working Group

Based on the "Users' Guides to the Medical Literature" and reproduced with permission from JAMA. (1994;271(5):389-391) and (1994;271(9):703-707).

Clinical Scenario

You are a medical consultant asked by a surgical colleague to see a 78 year old woman, now 10 days after abdominal surgery, who has become increasingly short of breath over the last 24 hours. She has also been experiencing what she describes as chest discomfort which is sometimes made worse by taking a deep breath (but sometimes not). Abnormal findings on physical examination are restricted to residual tenderness in the abdomen and scattered crackles at both lung bases. Chest radiograph reveals a small right pleural effusion, but this is the first radiograph since the operation. Arterial blood gases show a PO2 of 70, with a saturation of 92%. The electrocardiogram shows only non-specific changes.

You suspect that the patient, despite receiving 5000 U of heparin twice a day, may have had a pulmonary embolus (PE). You request a ventilation-perfusion scan (V/Q scan) and the result reported to the nurse over the phone is "intermediate probability" for PE. Though still somewhat uncertain about the diagnosis, you order full anticoagulation. Although you have used this test frequently in the past and think you have a fairly good notion of how to use the results, you realize that your understanding is based on intuition and the local practice rather than the properties of V/Q scanning from the original literature. Consequently, on your way to the nuclear medicine department to review the scan, you stop off in the library.

The Search

Your plan is to find a study that will tell you about the properties of V/Q scanning as it applies to your clinical practice in general, and this patient in particular. You are familiar with using the software program "Grateful Med" and utilize this for your search. The program provides a listing of Medical subject (MeSH) headings, and your first choice is "pulmonary embolism". Since there are 1749 articles with that MeSH heading published between 1989 and 1992 (the range of your search) you are going to have to pare down your search. You choose two strategies: you will pick only articles that have "radionuclide imaging" as a subheading, and also have the associated MeSH heading "comparative study" (since you will need a study comparing V/Q scanning to some reference standard). This search yields 31 papers, of which you exclude 11 which evaluate new diagnostic techniques, 9 which relate to the diagnosis and treatment of deep venous thrombosis, and one which examines the natural history of PE. The remaining 11 address V/Q scanning in PE. One, however, is an editorial; four are limited in their scope (dealing with perfusion scans only, with situations in which the diagnostic workup should begin with pulmonary angiography, or with a single perfusion defect). Of the remainder, the "PIOPED study" catches your eye both because it is in a widely read journal with which you are familiar, and

because it is referred to in the titles of several of the other papers [1]. You print the abstract of this article, and find it includes the following piece of information: among people with an intermediate result of the V/Q scan, 33% had PE. You conclude you have made a good choice, and retrieve the article from the library shelves.

Introduction

Clinicians regularly confront dilemmas when ordering and interpreting diagnostic tests. The continuing proliferation of medical technology renders the clinician's ability to assess diagnostic test articles ever more important. Accordingly, this article will present the principles of efficiently assessing articles about diagnostic tests and optimally using the information they provide. Once you decide, as was illustrated in the clinical scenario with the PIOPED paper, that an article is potentially relevant (that is, the title and abstract suggest the information is directly relevant to the patient problem you are addressing) you can invoke the same three questions that we suggested in the introduction and the guides about therapy (Table 1).

Table 1. Evaluating and applying the results of studies of diagnostic tests.

I. Are the results in the study valid?

- Primary Guides
  - Was there an independent, blind comparison with a reference standard?
  - Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
- Secondary Guides
  - Did the results of the test being evaluated influence the decision to perform the reference standard?
  - Were the methods for performing the test described in sufficient detail to permit replication?

II. What are the results?

- Are likelihood ratios for the test results presented or data necessary for their calculation provided?

III. Will the results help me in caring for my patients?

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

- **Are the results of the study valid?**
  Whether one can believe the results of a study is determined by the methods used to carry it out. To say that the results are valid implies that the accuracy of the diagnostic test, as reported, is close enough to the truth to render the further examination of the study worthwhile. First, you must determine if you can believe the results of the study by considering how the authors assembled their patients and how they applied the test and an appropriate reference (or "gold" or "criterion") standard to the patients.
- **What are the results of the study?**
  If you decide that the study results are valid, the next step is to determine the diagnostic test's

accuracy. This is done by examining (or calculating for yourself) the test's likelihood ratios (often referred to as the test's "properties").

- **Will the results help me in caring for my patients?**
  The third step is to decide how to use the test, both for the individual patient and for your practice in general. Are the results of the study generalizable -- i.e. can you apply them to this particular patient and to the kind of patients you see most often? How often are the test results likely to yield valuable information? Does the test provide additional information above and beyond the history and physical examination? Is it less expensive or more easily available than other diagnostic tests for the same target disorder? Ultimately, are patients better off if the test is used?

In this article we deal with the first question in detail, while in the next article in the series we address the second and third questions. We use the PIOPED article to illustrate the process.

In the PIOPED study 731 consenting patients suspected of having PE underwent both V/Q scanning and pulmonary angiography. The pulmonary angiogram was considered to be the best way to prove whether a patient really had a PE, and therefore was the reference standard. Each angiogram was interpreted as showing one of three results: PE present, PE uncertain, or PE absent. The accuracy of the V/Q scan was compared with the angiogram, and its results were reported in one of four categories: "high probability" (for PE), "intermediate probability", "low probability", or "near normal or normal". The comparisons of the V/Q scans and angiograms are shown in Tables 2A and 2B. We'll get to the differences between these tables later; for now, let's apply the first of the three questions to this report.

I. Are the results in this article valid?

A. Primary guides

1. Was there an independent, blind comparison with a reference standard?

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 [2]. In the PIOPED study the pulmonary angiogram was employed as the reference standard and this was as "gold" as could be achieved without sacrificing the patients. In reading articles about diagnostic tests, if you can't accept the reference standard (within reason, that is - 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 is whether the test results and the reference standard were assessed independently of each other (that is, by interpreters who were unaware of the results of the other investigation). Our own clinical experience shows us why this is important. Once we have been shown a pulmonary nodule on a CT scan, we see the previously undetected lesion on the chest radiograph; once we learn the results of the echocardiogram, we hear the previously inaudible cardiac murmur. The more likely that the interpretation of a new test could be influenced by knowledge of the reference standard result (or vice versa), the greater the importance of the independent interpretation of both. The PIOPED investigators did not state explicitly that the tests were interpreted blindly in the paper. However, one could deduce from the effort they put into ensuring reproducible, independent readings that the interpreters were in fact blind, and we have confirmed through correspondence with one of the authors that this was so. When such matters are in doubt, most authors are happy to clarify if directly contacted.

2. Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?

A diagnostic test is really useful only to the extent it distinguishes between target disorders or states that might otherwise be confused. Almost any test can distinguish the healthy from the severely affected; this ability 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.

A vivid example of how the hopes raised with the introduction of a diagnostic test can be dashed by subsequent investigations comes from the story of carcino-embryonic-antigen (CEA) in colorectal cancer. CEA, when measured in 36 people with known advanced cancer of the colon or rectum, was elevated in 35 of them. At the same time, much lower levels were found in normal people and in a variety of other conditions [3]. 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 colo-rectal 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 plummeted and CEA for cancer diagnosis and screening was abandoned. CEA is now recommended only as one element in the follow-up of patients with known colorectal cancer [4].

In the PIOPED study the whole spectrum of patients suspected of having PE were eligible and recruited, including those who entered the study with high, medium, and low clinical suspicion of PE. We thus may conclude that the appropriate patient sample was chosen.

B. Secondary guides

Once you are convinced that the article is describing an appropriate spectrum of patients who underwent the independent, blind comparison of a diagnostic test and a reference standard, most likely its results represent an unbiased estimate of the real accuracy of the test -- that is, an estimate that doesn't systematically distort the truth. However, you can further reduce your chances of being misled by considering a number of other issues.

3. 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 result influences whether patients undergo confirmation by the reference standard. This situation, sometimes called "verification bias" [5] [6] or "work-up bias" [7] [8] would apply, for example, when patients with suspected coronary artery disease and positive exercise tests were more likely to undergo coronary angiography (the reference standard) than those with negative exercise tests.
Verification bias was a problem for the PIOPED study; patients whose V/Q scans were interpreted as "normal/near normal" and "low probability" were less likely to undergo pulmonary angiography (69%) than those with more positive V/Q scans (92%). This is not surprising, since clinicians might be reluctant to subject patients with a low probability of PE to the risks of angiography. PIOPED results restricted to those patients with successful angiography are presented in Table 2A.

Table 2A. The relationship between the results of pulmonary angiograms and V/Q scan results (only patients with successful angiograms).

Scan category	Angiogram	
	PE present	PE absent
High probability	102	14
Intermediate probability	105	217
Low probability	39	199
Near normal / normal	5	50

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 V/Q 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 PE if they

did well without treatment. Accordingly, they followed every one of them for one year without treating them with anticoagulants. Not one of these patients developed clinically evident PE during this time, from which we can conclude that clinically important PE (if we define clinically important PE as requiring anticoagulation to prevent subsequent adverse events) was not present at the time they underwent V/Q scanning. When these 150 patients, judged free of PE by this second reference standard of a good prognosis without anticoagulant therapy, are added to the 480 patients with negative angiograms in Table 2A, the result is Table 2B. We hope you agree with us that the better estimate of the accuracy of V/Q scanning comes from Table 2B, which includes the 150 patients who, from follow up, did not have clinically important PE. Accordingly, we will use these data in subsequent calculations.

Table 2B. The relationship between the results of pulmonary angiograms and V/Q scan results (including 150 patients with low probability/near normal/normal V/Q scans, no (136) or uninterpretable (14) angiograms, and no clinically important thromboembolism on follow up).

Scan Category	Angiogram	
	PE present	PE absent
High probability	102	14
Intermediate probability	105	217
Low probability	39	273
Near normal / normal	5	126
Total	251	630

There were still another 50 patients with either high or intermediate probability scans who either did not undergo angiography, or whose angiograms were uninterpretable. It is possible that these individuals could bias the results. However, they are a relatively small proportion of the population, and if their clinical characteristics are not clearly different from those who underwent angiography, it is unlikely that the test properties would differ systematically in this sub-population. Therefore, we can proceed with relative confidence in the PIOPED results.

4. Were the methods for performing the test described in sufficient detail to permit replication?

If the authors have concluded that you should use a diagnostic test, they must tell you how to use it. This description should cover all issues that are important in the preparation of the patient (diet, drugs to be avoided, precautions after the test), the performance of the test (technique, possibility of pain), and the analysis and interpretation of its results.

Once the reader is confident that the article's results constitute an unbiased estimate of the test properties, she can determine exactly what (and how helpful) those test properties are. While not pristine (studies almost never are) we can strongly infer that the results are a valid estimate of the properties of the V/Q scan. We will describe how to interpret and apply the results in the next article of this series.

II. What are the Results?

Clinical Scenario
You are back where we put you: in the library studying an article that will guide you in interpreting Ventilation/Perfusion (V/Q) lung scans. Using the criteria in Table 1, you have decided that the PIOPED study [1] will provide you with valid information. Just then, another physician comes looking for an article to help with the interpretation of V/Q scanning. Her patient is 28 year old man whose acute onset of shortness of breath and

vague chest pain began shortly after completing a 10 hour auto trip. He experienced several episodes of similar discomfort in the past, but none this severe, and is very apprehensive about his symptoms. After a normal physical examination, electrocardiogram, and chest radiograph, and blood gases that showed a PCO2 of 32 and a PO2 of 82 mm Hg, your colleague ordered a V/Q scan. The results were reported as an "intermediate probability" scan.

You tell your colleague how you used Grateful Med to find an excellent paper addressing the accuracy of V/Q scanning. She is pleased that you found the article valid, and you agree to combine forces in applying it to both your patients.

In the previous article on diagnostic tests, we presented an approach to deciding whether a study was valid, and the results therefore worth considering. In this instalment, we explore the next steps which involve understanding and using the results of valid studies of diagnostic tests.

1. Are likelihood ratios for the test results presented or data necessary for their calculation included?

Pre-test Probability

The starting point of any diagnostic process is the patient, presenting with a constellation of symptoms and signs. Consider the two patients who opened to this exercise - the 78 year old woman 10 days after surgery and the 28 year old anxious man, both with shortness of breath and non-specific chest pain. Our clinical hunches about the probability of pulmonary embolus (PE) as the explanation for these two patients' complaints, that is, their pre-test probabilities, are very different: the probability in the older woman is high, and in the young man is low. As a result, even if both have intermediate probability V/Q scans, subsequent management is likely to differ. One might well treat the elderly woman but order additional investigations in the young man.

Two conclusions emerge from this line of reasoning. First, whatever the results of the V/Q scan, they do not tell us whether PE is present. What they do accomplish is to modify the pre-test probability of PE, yielding a new, post-test probability. The direction and magnitude of this change from pre-test to post-test probability is determined by the test's properties, and the property that we shall focus on in this series is the likelihood ratio.

The second conclusion we can draw from our two, contrasting patients is that the pre-test probability exerts a major influence on the diagnostic process. Each item of the history and physical examination is a diagnostic test that either increases or decreases the probability of a target disorder. Consider the young man who presented to your colleague. The fact that he presents with shortness of breath raises the possibility of pulmonary embolism. The fact that he has been immobile for ten hours increases this probability, but his age, lack of antecedent disease, and normal physical examination, chest radiograph, and arterial blood gases all decrease this probability. If we knew the properties of each of these pieces of information (and for some of them, we do [9] [10]) we could move sequentially through them, incorporating each piece of information as we go and continuously recalculating the probability of the target disorder. Clinicians do proceed in this fashion, but because the properties of the individual items of history and physical examination usually are not available, they often must rely on clinical experience and intuition to arrive at the pre-test probability that precedes ordering a diagnostic test. For some clinical problems, including the diagnosis of pulmonary embolism, their intuition has proved surprisingly accurate [1].
Nevertheless, the limited information about the properties of items of history and physical examination often results in clinicians varying widely in their estimates of pre-test probabilities. There are a number of solutions to this problem. First, clinical investigators should study the history and physical examination to learn more about the properties of these diagnostic tests. Fortunately, such investigations are becoming common. Panzer and colleagues have summarized much of the available information in the form of a medical text [11], and overviews on the accuracy and precision of the history and physical examination is being published concurrently with the Users' Guides in the JAMA series on The Rational Clinical Examination [12]. In addition, for some target disorders such as myocardial ischemia, multivariable analyses can provide physicians with ways of combining information to generate very precise pre-test probabilities [13]. Second, when we don't know the properties of history and physical examination we can consult colleagues about their probability estimates; the consensus view is likely to be more accurate than our individual intuition. Finally, when we remain uncertain about the pre-test probability, we can assume the highest plausible pre-test probability, and the lowest possible pre-test probability, and see if this changes our clinical course of action. We will illustrate how one might do this later in

this discussion.

Likelihood Ratios

The clinical usefulness of a diagnostic test is largely determined by the accuracy with which it identifies its target disorder, and the accuracy measure we shall focus on is the likelihood ratio. Let's now look at Table 3, constructed from the results of the PIOPED study. There were 251 people with angiographically-proven PE and 630 people whose angiograms or followup excluded PE. For all patients, V/Q scans were classified into four levels, from high probability to normal or near normal. How likely is a high probability scan among people who do have PE? Table 3 shows that 102 out of 251 (or 0.406) people with PE had high probability scans. How often is the same test result, a high probability scan, found among people who, although suspected of it, do not have PE? The answer is 14 out of 630 or 0.022 of them. The ratio of these two likelihoods is called the likelihood ratio (LR) and for a high probability scan equals 0.406 ¸ 0.022 or 18.3. In other words, a high probability lung scan is 18.3 times as likely to occur in a patient with, as opposed to a patient without, a pulmonary embolus. In a similar fashion, the LR can be calculated for each level of the diagnostic test result. Each calculation involves answering two questions: first, how likely it is to get a given test result (say low probability V/Q scan) among people with the target disorder (PE) and second, how likely it is to get the same test result (again, low probability scan) among people without the target disorder (no PE). For low probability V/Q scan these likelihoods are 39/251 (0.16) and 273/630 (0.43), and their ratio (the LR for low probability scan) is 0.36. As shown in Table 3, we can repeat these calculations for the other scan results.

Table 3. Test Properties of Ventilation/Perfusion (V/Q) Scanning

		Pulmonary Embolism				
		Present		Absent		
		Number	Proportion	Number	Proportion	Likelihood Ratio
V/Q Scan Result	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	1.2
	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		

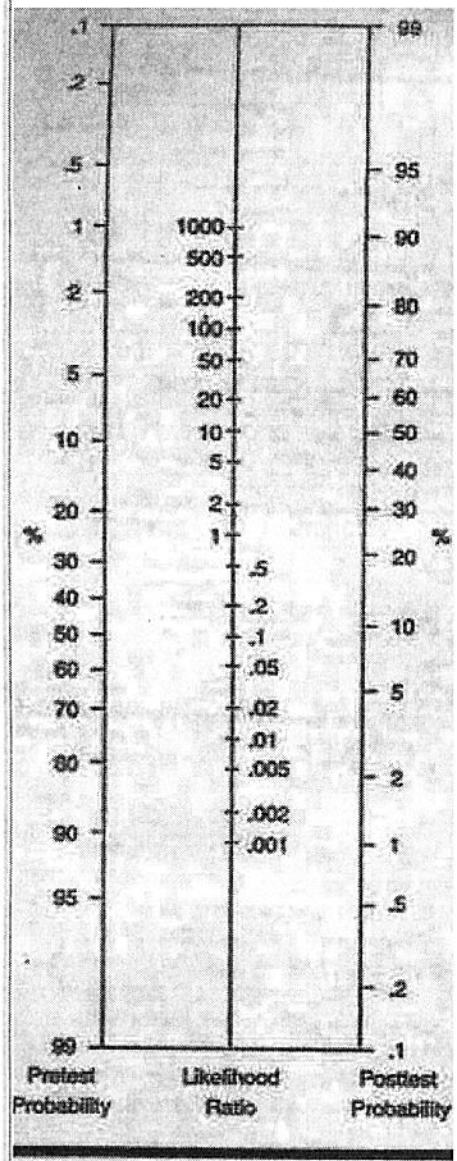
What do all these numbers mean? The LRs indicate by how much a given diagnostic test result will raise or lower the pre-test probability of the target disorder. A LR of 1 means that the post-test probability is exactly the same as the pre-test probability. LRs greater than 1 increase the probability that the target disorder is present, and the higher the LR the greater this increase. Conversely, LRs less than 1 decrease the probability of the target disorder, and the smaller the LR, the greater the decrease in probability and the smaller its final value.

How big is a big LR, and how small is a small one? Using LRs in your day-to-day practice will lead to your own sense of their interpretation, but as a rough Content:

- LRs >10 or < 0.1 generate large, and often conclusive changes from pre- to post-test probability;
- LRs of 5-10 and 0.1-0.2 generate moderate shifts in pre- to post-test probability;
- LRs of 2-5 and 0.5-0.2 generate small (but sometimes important) changes in probability; and
- LRs of 1-2 and 0.5-1 alter probability to a small (and rarely important) degree.

Having determined the magnitude and significance of the LRs, how do we use them to go from pre- to post-test probability? We can't combine likelihoods directly, the way we can combine probabilities or percentages; their formal use requires converting pre-test probability to odds, multiplying the result by the LR, and converting the consequent post-test odds into a post-test probability. While not too difficult [F1], this calculation can be tedious and off-putting; fortunately, there is an easier way. A nomogram proposed by Fagan [14] (Figure 1) does all the conversions for us and allows us to go very simply from pre- to post-test probability. The first column of this nomogram represents the pre-test probability, the second column represents the LR, the third shows the post-test probability. You obtain the post-test probability by anchoring a ruler at the pre-test probability and rotating it until it lines up with the LR for the observed test result.

Figure 1: Likelihood ratio nomogram



Nomogram for interpreting diagnostic test results. Adapted from Fagan.*

Recall our elderly woman with suspected PE after abdominal surgery. Most clinicians would agree that the probability of this patient having PE is quite high, at about 70%. This value then represents the pre-test probability. Suppose that her V/Q scan was reported as high probability. Anchoring a ruler at her pre-test probability of 70% and aligning it with the LR of 18.3 associated with a high-probability scan, her post-test

probability is over 97%. What if her V/Q scan yielded a different result? If her V/Q scan result is reported as intermediate (LR 1.2) the probability of PE hardly changes (to 74%), while a near normal result yields a post-test probability of 19%.

We have pointed out that the pre-test probability is an estimate, and that one way of dealing with the uncertainty is to examine the implications of a plausible range of pre-test probabilities. Let us assume the pre-test probability in this case is as low as 60%, or as high as 80%. The post-test probabilities that would follow from these different pre-test probabilities appear in Table 4.

Table 4. Pre-test probabilities, likelihood ratios of ventilation/perfusion scan results, and post-test probabilities in two patients with pulmonary embolus

Pre-test probability, % (Range)*	Scan result and likelihood ratio	Post-test probability, % (Range)*
78 year old woman with sudden onset of dyspnea following abdominal surgery		
(60)* 70% (80)	High probability: LR = 18.3	(96%) 97% (99%)
(60) 70% (80)	Intermediate probability: LR = 1.2	(64%) 74% (83%)
(60) 70% (80)	Low probability: LR = 0.36	(35%) 46% (59%)
(60) 70% (80)	Normal/near normal: LR = 0.1	(13%) 19% (29%)
28 year-old man with dyspnea and atypical chest pain		
(10) 20% (30)	High probability: LR = 18.3	(67%) 82% (89%)
(10) 20% (30)	Intermediate probability: LR = 1.2	(12%) 23% (34%)
(10) 20% (30)	Low probability: LR = 0.36	(4%) 8% (6%)
(10) 20% (30)	Normal/near normal: LR = 0.1	(1%) 2% (4%)

* The values in brackets represent a plausible range of pre-test probabilities. That is, while the best guess as to the pre-test probability is 70%, values of 60 to 80% would also be reasonable estimates.

The same exercise may be repeated for our second patient, the young man with non-specific chest pain and hyperventilation. If you consider that his presentation is compatible with a 20% probability of PE, using our nomogram the post-test probability with a high-probability scan result is 82%, an intermediate probability 23%, and a near normal is 2%. The pre-test probability (with a range of possible pre-test probabilites from 10 to 30%), LRs and post-test probabilities associated with each of the four possible scan results also appear in Table 4.

Readers who have followed the discussion to this point will understand the essentials of interpretation of diagnostic tests, and can stop here! They should consider the next section, which deals with sensitivity and specificity, optional. We include it largely because clinicians still will encounter studies that present their results in terms of sensitivity and specificity, and may wish to understand this alternative framework for summarizing the properties of diagnostic tests.

Sensitivity and Specificity

You may have noted that our discussion of likelihood ratios ignored any talk of "normal" and "abnormal" tests. Instead, we presented four different V/Q scan interpretations, each with their own LR. This is not, however, the way the PIOPED investigators presented their results. They fell back on the older (but less useful) concepts of sensitivity and specificity.

Sensitivity is the proportion of people with the target disorder in whom the test result is positive and specificity is the proportion of people without the target disorder in whom test result is negative. To use these concepts we have to divide test results into normal and abnormal; in other words, create a 2x2 table. The general form of a 2X2 table which we use to understand sensitivity and specificity is presented in Table 5A. Look again at Table 3 and observe that we could transform our 4x2 table into any of three such 2x2 tables, depending on what we call normal or abnormal (or what we call negative and positive test results). Let's assume that we call only high probability scan abnormal (or positive). The resulting 2X2 table is presented in Table 5B.

Table 5A. Comparison of the results of diagnostic test with the result of reference standard.

Test Result	Reference Standard	
	Disease Present	**Disease Absent**
Disease Present	True positive *(a)*	False positive *(b)*
Disease Absent	False negative *(c)*	True negative *(d)*

Sensitivity = a/(a+c)

Specificity = d/(b+d)

LR for positive test result = (a/(a+c)) / (b/(b+d))

LR for negative test result = (c/(a+c)) / (d/(b+d))

Table 5B. Comparison of the results of diagnostic test (V/Q scan) with the result of reference standard (pulmonary angiogram) assuming only high probability scans are positive (truly abnormal).

Scan Category	Angiogram	
	Pulmonary Embolus Present	**Pulmonary Embolus 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 3 we look at the number of people with proven PE (251) who were diagnosed as having the target disorder on V/Q scan (102) - 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) was classified on V/Q scan as normals (616) - specificity of 616/630 or 98% (d/b+d). We can also calculate likelihood ratios for the "positive" and "negative" test results using this cut-point, 18.3 and 0.6 respectively.
Let's see how the test performs if we decide to put the threshold of "positive" versus "negative" in a different place in Table 3. For example let's call only the normal/near normal V/Q scan result negative. This 2x2 table (Table 5C) shows the sensitivity is now 246/251, or 98% (among 251 people with PE 246 are diagnosed on V/Q scan), but what has happened to specificity? Among 630 people without PE only 126 have a negative test result (specificity of 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 also interpret intermediate and low probability results as if they increase the likelihood of pulmonary embolus, when in fact they decrease the likelihood. You can generate the third 2x2 table by setting the cut-point in the middle - if your sensitivity and specificity is 82 and 63% respectively, and associated likelihood ratios of a "positive" and "negative" test 2.25 and 0.28, you have it right [F2].

Table 5C. Comparison of the results of diagnostic test (V/Q scan) with the result of reference standard (pulmonary angiogram) assuming only normal/near normal scans are negative (truly normal).

Scan Category	Angiogram	
	Pulmonary Embolus Present	Pulmonary Embolus Absent
High, intermidiate, 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

You can see that in using sensitivity and specificity you have to either throw away important information, or recalculate sensitivity and specificity for every cut-point. We recommend the LR approach because it is much simpler and much more efficient.

Thus far, we have established that the results are likely true for the people who were included in the PIOPED study, and ascertained the LRs associated with different results of the test. How useful is the test likely to be in our clinical practice?

III. Will the results help me in caring for my patients?

1. Will the reproducibility of the test result and its interpretation be satisfactory in my 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 for 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 tell readers how reproducible the test results can be expected to be. This is especially important when expertise is required in performing or interpreting the test (and 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, it is likely that the test can be readily applied to your clinical setting. If reproducibility of a diangostic test is very high, and observer variation very low, either the test is simple and unambigious or those interpreting it are highly skilled. If the latter applies, less skilled interpreters in your own clinical setting may not do as well.

In the PIOPED study, the authors not only provided a detailed description of their diagnostic criteria for V/Q scan interpretation, they also reported on the agreement between their two independent readers. Clinical disagreements over intermediate and low probability scans were common (25% and 30% respectively), and they resorted to adjudication by a panel of experts.

2. Are the results applicable to my patient?

The issue here is whether the test will have the same accuracy among your patients as was reported in the paper. Test properties may change with a different mix of disease severity or 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 (sensitivity increases). If patients are all mildly affected, likelihood ratios move toward a value of 1 (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 and the test will appear less useful. In a different clinical setting in which fewer of the non-diseased have these competing conditions the likelihood ratios will move away from 1 and the test will appear more useful.

The phenomenon of differing test properties in different subpopulations has been most strikingly demonstrated 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 post-operative 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, low probability patients.
If you practice in a setting similar to that of the investigation and your patient 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 judgement is required. As with therapeutic interventions, you should ask whether there are compelling reasons why the results should *not* be applied to your patients, either because the severity of disease in your patients, or 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]. The patients in the PIOPED study were a representative sample of patients with suspected PE from a number of large general hospitals. The results are therefore readily applicable to most clinical practices in North America. There are groups to whom we might be reluctant to generalize the results, such as critically ill patients (who were excluded from the study, and who are likely to have a different spectrum of competing conditions than other patients).

3. Will the results change my management?

It is useful, in making, learning, teaching, and communicating mangement decisions, to link them explicitly to the probability of the target disorder. Thus, for any target disorder there are probabilities below which a clinician would dismiss a diagnosis and order no further tests (a "test" threshold). Similarly, there are probabilities above which a clinician would consider the diagnosis confirmed, and would stop testing and initiate treatment (a "treatment" threshold). When the probability of the target disorder lies between the test and treatment thresholds, further testing is mandated [2].

Once we decide what our test and treatment thresholds are, post-test probabilities have direct treatment implications. Let us suppose that we are willing to treat those with a probability of pulmonary embolus of 80% or more (knowing that we will be treating 20% of our patients unnecessarily). Furthermore, let's suppose we are willing to dismiss the diagnosis of PE in those with a post-test probability of 10% or less. You may wish to apply different numbers here; the treatment and test thresholds are a matter of judgement, and 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 post-test probability very low before abandoning the diagnostic search). In our young man, a high probability scan results in a post-test probability of 82% and may dictate treatment (or, at least, further investigation), an intermediate probability scan (23% post-test probability) will dictate further testing (perhaps bilateral leg venography, serial impedance plethysmography or ultrasound, or pulmonary angiography), while a low probability or normal scan (probabilities of less than 10%) will allow exclusion of the diagnosis of PE. In the elderly woman, a high probability scan dictates treatment (97% post-test probability of PE), an intermediate result (74% post-test probability) may be compatible with either treatment or further testing (likely a pulmonary angiogram), while any other result mandates further testing.

If most patients have test results with LRs near 1, 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 LRs so that the test result will move their probability of disease across a test or treatment threshold. as a result likely to be transported across a test or treatment threshold. In the patients suspected of having PE in our V/Q scan example, review of Table 3 allows us to determine the proportion of patients with extreme results (either high probability with a LR of over 10, or near normal/normal scans with an LR of 0.1). The proportion can be calculated as (102+14+5+126)/881 or 247/881 = 28%. Clinicians who have repeatedly been frustrated by frequent intermediate or low probability results in their patients with suspected PE will already know that this proportion (28%) is far from optimal. Thus, despite the high LR associated with a high probability scan, and the low LR associated with a normal/near normal result, V/Q scanning is of limited usefulness in patients with suspected PE.
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 pre-test 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 instance, 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 information [19].

4. Will patients be better off as a result of the test?

The ultimate criterion for the usefulness of a diagnostic test is whether it adds information beyond that otherwise available, and whether this information leads to a change in management that is ultimately beneficial to the patient [20]. The value of an accurate test will be undisputed when the target disorder, if left undiagnosed, is dangerous, the test has acceptable risks, and effective treatment exists. A high probability or near normal/normal results of a V/Q scan may well eliminate the need for further investigation and result in anticoagulants being appropriately given or appropriately withheld (either course of action having a substantial

influence on patient outcome).

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. Examples include right heart catheterization for many critically ill patients, or the incremental value of MRI scanning over CT for a wide variety of problems.

How you may use these guides for clinical practice and for reading.

By applying the principles described in this and the preceding paper you will be able to assess and use information from articles about diagnostic tests. You are now equipped to decide whether an article concerning a diagnostic test represents a believable estimate of the true value of a test, what the test properties are, and the circumstances under which the test should be applied to your patients. Because LRs now are being published for an increasing number of tests [11], the approach we have outlined has become directly applicable to the day-to-day practice of medicine.

Footnotes

F1. The equation to convert probabilities into odds is (probability / [1 - probability] -- equivalent to: probability of having the target disorder/probability of not having the target disorder). A probability of 0.5 represent odds of 0.50 / 0.50, or 1 to 1; a probability of 0.80 represents odds of 0.80 to 0.20, or 4 to 1; a probability of 0.25 represents odds of 0.25 to 0.75, or 1 to 3, or 0.33. Once you have the pre-test odds the post-test odds can be calculated by multiplying the pre-test odds by the LR. The post-tests odds can be converted back into probabilities using a formula of probability = odds / (odds + 1).
F2. If you were to create a graph where vertical axis will denote sensitivity (or true positive rate) for different cut-offs and the horizontal axis will display [1-specificity] (or false positive rate) for the same cut-offs, and you connect the points generated by using cut points, you would have what is called a Receiver Operating Characteristic (ROC curve). ROC curves can be used to formally compare the value of different tests by examining the area under each curve, but once one has abandoned the need for a single cut-point, have no other direct clinical application.

References

1. The PIOPED Investigators, Anonymous. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. J A M A 263. 2753-9 (1990).
2. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Second Edition. Clinical epidemiology, a basic science for clinical medicine.(1991) Second Edition.Boston/Toronto: Little, Brown and Company.
3. Thomson DMP, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proceedings of the National Academy of Sciences of the United States of America 64. 161-7 (1969).
4. Bates SE. Clinical applications of serum tumor markers. Ann Intern Med 115. 623-38 (1991).
5. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39. 207-15 (1983).
6. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 4. 151-64 (1984).
7. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299. 926-30 (1978).
8. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin

Epidemiol 45. 581-6 (1992).
9. Mayeski RJ.Panzer RJ, Black ER, Griner PF, Eds. Pulmonary Embolism.(1991) Diagnostic strategies for common medical problems. American College of Physicians.
10. Stein PD, Terrin NL, Hales CA, et al, Terrin ML, Palevsky HI, Saltzman HA, Thompson BT, Weg JG. Clinical, laboratory, roentgenographic, and electrocardiographic findings in patients with acute pulmonary embolism and no pre-existing cardiac or pulmonary disease. Chest 100. 598-603 (1991).
11. Panzer RJ, Black ER, Griner PF. Diagnostic strategies for common medical problems.(1991) 1.Philadelphia: American College of Physicians.
12. Sackett DL, Rennie D. The science and art of the clinical examination. J A M A 267. 2650-2 (1992).
13. Pozen MW, D'Agostino RB, Selker HP, et al, Sytkowski PA, Hood WB, Jr. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. A prospective multicenter clinical trial. N Engl J Med 310. 1273-8 (1984).
14. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 293. 257(1975).
15. Hlatky MA, Pryor DB, Harrell FE, Harrell FE, Jr., Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 77. 64-71 (1984).
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 181. 651-4 (1991).
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 117. 135-40 (1992).
18. Irwig L, Tosteson A, Gatsonis C, et.al. Guidelines for meta-analyses evaluating diagnostic tests. (1995). Unpublished.
19. Guyatt GH, Oxman A, Ali M. Diagnosis of iron deficiency. J Gen Intern Med 7. 145-53 (1992).
20. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 134. 587-94 (1986).