Users' Guides to the Medical Literature

III. How to Use an Article About a Diagnostic Test

A. Are the Results of the Study Valid?

Roman Jaeschke, MD, MSc; Gordon Guyatt, MD, MSc; David L. Sackett, MD, MSc;
for the Evidence-Based Medicine Working Group

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 roentgenogram reveals a small right pleural effusion, but this is the first roentgenogram since the operation. Arterial blood gases show a P_{O_2} of 70 mm Hg, with a saturation of 92%. The electrocardiogram shows only nonspecific 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 telephone is "intermediate probability" for PE. Though still somewhat uncertain about the diagnosis, you order full anticoagulation. Although you have used the V/Q scan 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 local practice rather than on 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 the software program GRATEFUL MED and use this for your search. The program provides a listing of Medical Subject Headings (MeSH), 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 with some reference standard). This search yields 31 articles, of which you exclude 11 that evaluate new diagnostic techniques, nine that relate to the diagnosis and treatment of deep venous thrombosis, and one that 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 Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study[1] 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 articles. 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.

This article in the "Users' Guides to the Medical Literature" series and the one that follows will demonstrate an approach to making optimal use of the article.

INTRODUCTION

Clinicians regularly confront dilemmas when ordering and interpreting diagnostic tests. The continuing proliferation of medical technology renders the clinician's ability to assess articles about diagnostic tests 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 clini-

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

A complete list of the members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (Chair), MD, MSc; Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; K. S. Trout, FRCE; Peter Tugwell, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; and Mark Wilson, MD, MPH.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

cal scenario with the PIOPED article, 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 articles on therapy[2-4] (Table 1).

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—ie, 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 the V/Q scan 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 2 and 3. We'll get to the differences between these tables later; for now, let's apply the first of the three questions to this article.

Table 1.—Evaluating and Applying the Results of Studies of Diagnostic Tests

Are the results of 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?
What were the results?
Are likelihood ratios for the test results presented or data necessary for their calculation provided?
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?

Table 2.—The Relationship Between the Results of Pulmonary Angiograms and Ventilation-Perfusion Scan Results in Patients With Successful Angiograms

	Angiogram	
Scan Category	Pulmonary Embolus Present	Pulmonary Embolus Absent
High probability	102	14
Intermediate probability	105	217
Low probability	39	199
Near normal/normal	5	50
Total	**251**	**480**

Table 3.—The Relationship Between the Results of Pulmonary Angiograms and Ventilation-Perfusion Scan Results*

	Angiogram	
Scan Category	Pulmonary Embolus Present	Pulmonary Embolus Absent
High probability	102	14
Intermediate probability	105	217
Low probability	39	273
Near normal/normal	5	126
Total	**251**	**630**

*Includes 150 patients with low probability and near normal/normal ventilation-perfusion scans, no (136) or uninterpretable (14) angiograms, and no clinically important thromboembolism on follow-up.

ARE THE RESULTS OF THE STUDY VALID?

Primary Guides

Was There an Independent, Blind Comparison With a Reference Standard?—The accuracy of a diagnostic test is best determined by comparing it with 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.[5] In the PIOPED study, the pulmonary angiogram was used 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 computed tomographic scan, we see the previously undetected lesion on the chest roentgenogram; once we learn the results of the echocardiogram, we hear the previously inaudible cardiac murmur. The more likely it is 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 article. However, one could deduce from the effort they put into ensuring reproducible, independent readings that the interpreters were in fact blinded, 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.

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 carcinoembryonic antigen (CEA) in colorectal cancer. Carcinoembryonic antigen levels, when measured in 36 people with known advanced cancer of the co-

lon or rectum, were elevated in 35 of them. At the same time, much lower levels were found in normal people and in a variety of other conditions.[6] The results suggested that measurement of CEA levels 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 measurements plummeted, and the use of CEA levels for cancer diagnosis and screening was abandoned. Carcinoembryonic antigen is now recommended only as one element in the follow-up of patients with known colorectal cancer.[7]

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.

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.

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"[8,9] or "work-up bias",[10,11] 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 or 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. The results of the PIOPED study restricted to those patients with successful angiography are presented in Table 2.

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 angiographic 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 1 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 2, the result is Table 3. We hope you agree with us that the better estimate of the accuracy of V/Q scanning comes from Table 3, which includes the 150 patients who, from follow-up, did not have clinically important PE. Accordingly, we will use these data in subsequent calculations.

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 subpopulation. Therefore, we can proceed with relative confidence in the PIOPED results.

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.

References

1. 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.
2. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, I: how to get started. *JAMA*. 1993;270:2093-2095.
3. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA*. 1993;270:2598-2601.
4. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, Mass: Little Brown and Co; 1991:53-57.
6. Thomson DMP, 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.
7. Bates SE. Clinical applications of serum tumor markers. *Ann Intern Med*. 1991;115:623-638.
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 BCK. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45:581-586.