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 crickles 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 P02 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 GRAETFUL 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 1986 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 81 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 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;270:2090-2096). The following members contributed to this article: Gordon Guyatt (Chair), MD, MSc; Eric Bass, MD, MPH; Patrick Blié-Edwards, MD; George Brown, MD; MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzig Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPh; Anne Holbrook, MD, Phd; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MSc, MSc; Hu-Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPh; Jin Nishikawa, MD; Andrew Oram, MD, MSc; FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephanie Sluiter, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD, K. S. Trout, FRCP; Peter Tugwiel, 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).

JAMA. February 2, 1994—Vol 271, No. 5
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 therapy2,4 (Table 1).

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 was 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 benefit 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

<table>
<thead>
<tr>
<th>Pollumary Embo Ms</th>
<th>Pulmonary Embo Ms</th>
<th>Pulmonary Embo Ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>High probability</td>
<td>102</td>
<td>14</td>
</tr>
<tr>
<td>Intermediate probability</td>
<td>105</td>
<td>217</td>
</tr>
<tr>
<td>Low probability</td>
<td>39</td>
<td>199</td>
</tr>
<tr>
<td>Near normal/normal</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>480</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pollumary Embo Ms</th>
<th>Pulmonary Embo Ms</th>
<th>Pulmonary Embo Ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>High probability</td>
<td>102</td>
<td>14</td>
</tr>
<tr>
<td>Intermediate probability</td>
<td>105</td>
<td>217</td>
</tr>
<tr>
<td>Low probability</td>
<td>39</td>
<td>199</td>
</tr>
<tr>
<td>Near normal/normal</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>480</td>
</tr>
</tbody>
</table>

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

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

390 JAMA, February 2, 1994—Vol 271, No. 5

Users' Guides to Medical Literature—Jaeschke et al
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.

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," 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 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