ELSEVIER

Seminars in
NUCLEAR
MEDICINE

How to Assess Quality of Primary Research Studies in the Medical Literature?

Grégoire Le Gal, MD, PhD,* and Pierre-Yves Le Roux, MD, PhD[†]

Accuracy studies are the cornerstone of the evaluation of new diagnostic tests. In an accuracy study, patients with a clinical suspicion of disease undergo both the new tests that is being evaluated, and the reference test or "gold-standard" test for the disease. Patients are stratified according to the test result and to whether or not the diagnosis was confirmed by the reference test. Main accuracy indices include the sensitivity (proportion of patients who test positive among those who have the disease), specificity (proportion of patients who test negative among those without the disease), as well as the positive and negative predictive values (proportion of patients with the disease among those with a positive test, and of patients without the disease among those with a negative test, respectively). In appraising an accuracy study, the reader should check that the study design and analysis follow methodological standards. The study population should be representative of the population in which the test will be used in practice. The new test should be interpreted in a blinded and independent fashion from the reference standard to avoid expectation bias, and should not be used to establish the diagnosis (incorporation bias). The reproducibility should be verified. Interpretation criteria and technical aspects should be described with enough details to allow replication. Provided that these conditions are met, the next step is to decide whether the test may be used for patient care. The clinical setting in which the test will be used, and the corresponding pretest probability of disease, will determine how best can the test be used in practice.

Semin Nucl Med 49:115-120 © 2018 Elsevier Inc. All rights reserved.

The question of how to assess quality of primary diagnostic research studies in the medical literature has wide applicability across all clinical domains and fields in medicine. As an illustrative example for this issue of Seminars in Nuclear Medicine, we will illustrate our concepts with a discussion of V/Q scintigraphy, organized around a clinical case scenario.

Clinical Case Scenario

An 18-year-old woman presents to the emergency room with a new onset of right-sided pleuritic chest pain and shortness of breath. She has no particularly relevant past medical history, no personal or family history of venous thromboembolism (VTE), and no known malignancy. She has no recent major risk factors for VTE, including surgery, trauma, or bed rest. She is a nonsmoker. On physical examination her heart rate is 94 beats per minute and her oxygen saturation is 96% on room air. She has no fever, no cough, or hemoptysis. She denies any leg symptoms. Her physical examination reveals no pain at calf palpation, no edema or redness on either leg. Her heart auscultation is unremarkable except for a regular tachycardia. Lung auscultation reveals normal vesicular breathing without any additional adventitious breath sounds. The electrocardiogram shows sinus tachycardia, normal axis, and no ST changes. The chest X-ray is within normal limits, without evidence of pneumonia, pleural effusion, or pneumothorax.

The attending physician suspects that she might have a pulmonary embolism (PE). As a first step, he estimates her pretest probability of PE, using a clinical decision rule. Both the Geneva score[1] and the Wells score[2] classify the patient as having a low pretest probability, corresponding to a PE

*Department of Medicine, University of Ottawa, Ottawa Hospital Research Institute, Thrombosis Research Group, Ottawa, Canada.
[†]Service de Médecine Nucléaire, Centre Hospitalier Régional et Universitaire, EA 3878 GETBO, Université de Brest, Brest, France.
Address reprint requests to Grégoire Le Gal, MD, PhD, The Ottawa Hospital, General Campus 501, Smyth Road, Box 201A, Ottawa, Ontario K1H 8L6, Canada. E-mail: glegal@ohri.ca

https://doi.org/10.1053/j.semnuclmed.2018.11.007
0001-2998/© 2018 Elsevier Inc. All rights reserved.

prevalence of less than 10%.[3] Her D-dimer test is positive. Given her young age and normal chest X-ray, the attending physician orders a ventilation-perfusion (V/Q) lung scan as suggested by the recent Appropriate Use Criteria for V/Q Imaging.[4,5] The nuclear medicine physician proposes performing imaging with SPECT rather than with conventional planar V/Q scintigraphy and explains the potential advantages of the technique: by using tri-dimensional imaging, SPECT V/Q has the ability to eliminate overlap or superimposition of normally perfused or ventilated tissues with defects, and hence allows better characterization of abnormalities and can detect mismatched defects unseen on planar imaging. The nuclear medicine physician claims that the test has a better diagnostic performance for PE and, as an additional advantage over planar V/Q, allows a binary response, that is, the test is either positive or negative for PE. The attending physician is concerned about this newer imaging modality. PE is a life-threatening disease, and wants to be sure that a negative test would safely rule out the disease. On the other hand, over-diagnosis should also be avoided: current guidelines suggest long-term anticoagulation in patients with unprovoked PE. A positive diagnosis could lead to lifelong anticoagulation, with a significant impact on the patient's lifestyle, contraception options, and future pregnancies. The two physicians decide to organize a Journal Club to critically appraise available evidence on the accuracy of V/Q SPECT for diagnosing PE. We will return to their conclusions following an analysis of the issues, as noted below.

Studies Evaluating the Performance of Diagnostic Tests

Following the development of a potential new diagnostic test (eg, new immunoassay or new imaging technique), it is critical to determine its accuracy for the disease that it intends to diagnose before it is implemented in clinical practice. The most commonly used study design for the evaluation of diagnostic tests is the *accuracy study*. Diagnostic randomized trials and cohort studies are discussed in a separate article in this special issue.[6] The objective of accuracy studies is to evaluate the diagnostic performance of the new test, that is, its ability to detect the disease in affected patients and to classify as disease-free the nonaffected patients. This is achieved by comparing, in a cohort of patients with a clinical suspicion of the disease, the new test results with that of the "reference" or "gold-standard" test. The latter is used to classify the patients as having the disease or not. The results are often summarized in a 2 × 2 table (Table 1). Affected patients are classified as "true positives" if they test positive with the new test, and "false negatives" otherwise. Disease-free patients are classified as "true negatives" if they test negative with the new test, and "false positives" otherwise.

Table 1 2 × 2 Contingency Table in an Accuracy Study

	Disease Positive*	Disease Negative*
Test positive	True positive (*a*)	False positive (*b*)
Test negative	False negative (*c*)	True negative (*d*)

*as determined by the gold-standard test.

Two main indices are then computed:

- Sensitivity—this reflects the ability of the test to detect that a patient is affected by the disease. It is defined as the likelihood of obtaining a positive test in an affected patient. Sensitivity is calculated by dividing the number of true positive patients by the total number of patients with the disease, that is a / (a + c).
- Specificity—this reflects the ability of the test to detect that a patient is free of disease. It is defined as the likelihood of obtaining a negative test in a disease-free patient. Specificity is calculated by dividing the number of true negative patients by the total number of patients without the disease, that is, d / (b + d).

These are the main accuracy indices, because they reflect the ability of the test to recognize affected and unaffected patients. However, clinicians are often interested in other indices that are more directly useful in practice:

- The positive predictive value is the likelihood of disease among patients who test positive that is, the proportion of patients with disease among patients with a positive test, a / (a + c).
- The negative predictive value is the likelihood of not having the disease in patients who test negative that is, the proportion of disease-free patients among those with a negative test, d / (b + d).

Importantly, these indices depend not only on the test sensitivity and specificity, but also on the disease prevalence in the tested population. It is often surprising for physicians that a positive or negative test result might have a different significance according the disease prevalence in the tested population. However, it is possible to use this property of diagnostic tests to improve their overall performance. In Figure 1, two contingency tables depict the use of the same diagnostic test, the D-dimer test, in an accuracy study of patients with suspected PE. On the left side, the test is applied in a population of 100 patients with PE and 100 patients without PE. With a sensitivity of 95% and a specificity of 50%, the negative predictive value of the test is only 91%. In other words, 9% of patients with a negative test actually have PE, a proportion that is too high for physicians to rely on a negative D-dimer test to rule out PE, given the potential adverse consequences of a missed diagnosis. On the right side of the figure, the same D-dimer test is applied in a population with a 10% prevalence of PE. With the same sensitivity and specificity, the negative predictive value is now 99%. Therefore, only 1% of patients with a negative test have PE, a proportion that is low enough to safely rule out the disease without further testing. This is what led to the widespread use of pretest probability assessment in patients with suspected PE. Clinical decision rules, such as the Wells or Geneva scores,[1,2] are able to

PE prevalence: 50%

	PE	No PE	
D-Dimer positive	95	50	145
D-Dimer negative	5	50	55
	100	100	200

Sn = 95 / 100 = 95%
Sp = 50 / 100 = 50%

NPV = 50 / 55 = 91%

PE prevalence: 10%

	PE	No PE	
D-Dimer positive	95	450	545
D-Dimer negative	5	450	455
	100	900	1000

Sn = 95 / 100 = 95%
Sp = 450 / 450 = 50%

NPV = 450 / 455 = 99%

Figure 1 Impact of the disease prevalence on the predictive values.

accurately classify patients with suspected PE in groups of low, intermediate, or high pretest probability, in which the prevalence of PE is known.[3] It is then possible to decide on how best to use diagnostic tests in combination with pretest probability to reach satisfactory positive or negative predictive values.

Positive and likelihood ratios express the impact of a positive or negative result on the probability of disease.

- Positive likelihood ratio—is the ratio of the likelihood of a positive test among patients with the disease divided by the likelihood of a positive test among disease-free patients, that is, the sensitivity divided by one minus the specificity, or (a / b) × (b + d) / (a + c).
- Negative likelihood ratio—is the ratio of the likelihood of a negative test among patients with the disease divided by the likelihood of a negative test among disease-free patients, that is, one minus the sensitivity divided by the specificity, or (c / d) × (b + d) / (a + c).

If the positive likelihood ratio is 5, it means that a positive test is five times more likely to be observed in a patient with, than in a patient without the disease. It is possible to calculate the post-test probability of disease based on the pretest probability and the likelihood ratios associated with a positive or negative test, using Fagan's nomogram. Nowadays, most smartphone medical calculator applications provide simple tools that can quickly compute the likelihood ratios and provide a post-test probability of disease based on the pretest probability and the likelihood ratios associated with a negative or positive test.

All these indices should be computed along with a confidence interval, giving the range of values between which the "true" value will probably lie. In Figure 1, the specificity is 50% in both tables. However, the 95% confidence varies based on the sample size. On the left side of the figure, the 95% confidence interval around the estimate of 50% (50/100) ranges from 40% to 60%. On the right side, the confidence interval around the same proportion of 50% (450/900) is much narrower, from 47% to 53%. A larger sample size enables a much more precise estimation.

How to Appraise an Accuracy Study?

Several user guides have been published to help readers identifying the main components and potential pitfalls of accuracy studies.[7-9] The use of reporting guidelines by authors helps ensure that readers will find the information they need to critically appraise a study.[10] These guides often recommend consulting a short list of questions that might help assess the key aspects of the validity of study results and their usefulness for clinical practice.

Was the Study Population Adequate?

A critical aspect is to ensure that the study population is representative of the population in which the test will be used in clinical practice, for example, consecutive patients presenting to the emergency room for a suspected PE. A frequent mistake is to use a sample of patients in whom the diagnosis has already been made, and a control group of "healthy" or "disease-free" patients. This is inappropriate for two reasons. The main reason is that it is important to evaluate the test in a population in which the whole spectrum of the disease may be observed. Including patients with a confirmed diagnosis could lead to selecting patients with a higher disease burden, or patients in whom the diagnosis was easy to confirm. The control group is often comprising of "healthy" patients, creating an artificially high contrast between the two groups, which could lead to an overestimation of the test accuracy. For example, comparing the D-dimer levels of patients with a confirmed PE to those of healthy blood donors creates a marked contrast, with significantly higher D-dimer levels in PE patients, as compared to those in the control group. The reality is that among patients with a clinical suspicion of PE, many of those without a PE may have an alternative diagnosis that shares symptoms and signs with PE and is also associated with higher D-dimer levels (eg, older age, pneumonia, chronic obstructive pulmonary disease exacerbation, or aortic dissection). Moreover, whenever the investigators decide on the number of patients to be included in each group, it becomes impossible to assess the positive and negative predictive values, which are dependent on the disease's prevalence in the tested population.

It is also important to look at the population characteristics to assess how similar or different the patients included in the study are to the ones in whom the test would be used in clinical practice. What were the inclusion and exclusion criteria? Which kind of symptoms did they present? In which setting were they approached for participation in the study (eg, primary care practice vs emergency room vs hospital ward vs specialized clinic vs imaging facility)? For example, if the recruitment in a study on patients with suspected PE is based in a department of nuclear medicine, it is likely that some patient selection will have already occurred. Patients with a low pretest probability and negative D-dimer are not usually referred for imaging tests. It is probable that referring physicians will have selected some specific patients for nuclear medicine imaging, such as young patients, patients with no abnormality on chest X-ray, or those with a contraindication to computed tomography pulmonary angiography. The potential impact of such a selection on the study results should be carefully discussed.

A study sample size calculation should also be provided. In accuracy studies, the number of patients to be included is usually selected based on the desired precision in the estimation of the accuracy indices, considering the anticipated disease prevalence and accuracy levels.

Was the New Test Evaluated in a Blinded and Independent Fashion From the Reference Standard?

It is of utmost importance that investigators in charge of interpreting the new test are blinded from patient status, including risk factors, signs and symptoms, and from the results of the reference test or strategy. Knowledge of this information could unconsciously bias their interpretation of the test (expectation bias), for example in case of dubious results. Similarly, the treating physicians should not be made aware of the results of the new test to ensure that their diagnostic and management decisions are not impacted by the results of the new test. The higher the chance of the new test interpretation being influenced by knowledge of the 'truth', or vice versa, the higher the importance of keeping interpreters and attending physicians blinded from each other's results. Many of the studies assessing the accuracy V/Q SPECT are limited by this important issue. A recent systematic review of V/Q SPECT for PE accuracy studies confirmed that all identified studies had a high risk of bias given the use of a composite standard as the reference standard that integrated V/Q SPECT results.[11] In other words, the test under consideration (the V/Q SPECT) was used to determine whether or not patients had PE. This is obviously a major limitation, leading to an overestimation of the test performances.

In an accuracy study, all patients in the study should undergo both the new test and the reference test. The result of the new test should not participate in the decision to perform the reference standard (and vice versa). That would be for example the case if in a V/Q SPECT study only patients with a positive test would undergo a confirmatory pulmonary angiography (work-up bias). The reference test should be the same regardless of the results of the new test.

The article should also clearly describe the methods employed to perform the new test including the criteria used for interpretation, the cut-off for positivity and enough details to enable replication. Although a consensus emerged around the criteria proposed by the European Association of Nuclear Medicine, the criteria for V/Q SPECT positivity significantly vary from one study to another. According to the European Association of Nuclear Medicine criteria, PE is diagnosed when a mismatch of at least one segment or two subsegments that conforms to the pulmonary vascular anatomy is present.[12,13] The reference test should also be clearly defined. The results of the study will only be valid if the reference test or strategy is one that is widely accepted for diagnosing the disease. Identifying a "gold-standard" test is not always easy. For PE, pulmonary angiography has long been considered the gold-standard test. However, this test is no longer performed in routine practice given its invasiveness, cost, risk of complication and higher radiation levels when compared to newer imaging techniques. Therefore, deciding which test to use as a reference is challenging. Most recent studies used a previously validated diagnostic algorithm rather than a single test.[14]

Validity assessment includes how missing values and inconclusive test results were handled. Not taking into account patients with inconclusive tests into the computation of accuracy indices leads to their overestimation.

Can I Use These Results for the Care of my Own Patients?

A first condition to be met before utilization in clinical practice is that the test should keep the same performance when implemented in a different setting. This implies that the test is described in enough detail to enable replication, that its accuracy is reproducible in a similar clinical situation, but also that interpretation is reproducible between different observers. This could be verified by an interobserver agreement assessment, conducted on part or the whole study sample.

For physicians to determine whether the study results are applicable to their clinical practice, it is useful to assess how the clinical setting and the patient population in the study differ from their patients. In fact, as already discussed, the disease prevalence in the study population will have a major impact on the negative and positive predictive values. The accuracy could also differ depending on the spectrum of disease severity (it is usually easier to diagnose most severe forms of a disease) or on the frequency of comorbid conditions: the performance of the V/Q scan for PE would be influenced by the presence of chronic obstructive pulmonary disease or other cardiopulmonary conditions if used in some settings such as elderly patients, or cardiac/respiratory intensive care units. The cost, resources, and skills required to implement the test should be weighed against the benefits vs current practice.

Will the test change management of my patient? This will mostly depend on the impact it will have on the pretest probability. Will a positive result be sufficient to instill confidence that the patient will benefit from treatment? Or will a negative result be reassuring enough to leave a patient untreated? In the case of the V/Q SPECT, given the limitations described above, it appears possible to use the V/Q SPECT result in combination with a pretest probability assessment tool to manage patients with suspected PE but further confirmatory studies are required. It is usually accepted that patients in whom the post-test probability of PE is above 90% will benefit from anticoagulant therapy, while patients in whom the post-test probability of PE is below 3% would clearly have more harm than benefit from anticoagulant therapy, and could be left untreated.

Implementation of a test is more likely to benefit patients when a missed diagnosis of the disease might be associated with poor outcome, the test has acceptable risks, and an effective treatment exists. It is important to assess the impact of using the test in daily practice in different type of studies (randomized controlled trial or prospective outcome study). For example, a randomized controlled trial could compare the outcome of patients managed using the new test as compared to those managed using the reference test. Alternatively, a management outcome study of patients with suspected PE would be another reasonable study design. Once the accuracy of the new test has been determined, its best use in combination with other available tests needs to be determined. The new diagnostic algorithm including the new test is then applied to consecutive patients seen with a clinical suspicion of PE. Those with a negative diagnostic work-up are left untreated and undergo clinical follow-up for 3 months. The 3-month risk of VTE is compared to that observed after a negative pulmonary angiography, the gold-standard test for PE.[15] In the case of the V/Q SPECT, such a formal management outcome study is yet to be performed. At least two studies provide reassuring data on patients followed after a negative V/Q SPECT.[16,17] However, in these studies, other diagnostic tests could have been performed at the discretion of the attending physician, preventing any definitive conclusion to be drawn.

Clinical Case Scenario Continued

After reading available studies, the two physicians agree that a V/Q scan is an appropriate test to perform in this patient. The V/Q SPECT appears as a promising modality but is less validated; many relevant research studies suffer from limitations that prevent a definitive conclusion to be drawn regarding the exact accuracy of V/Q SPECT for PE diagnosis. Moreover, a formal prospective management outcome study in which management would be decided on the basis of a standardized algorithm including a V/Q SPECT is lacking, limiting the wide adoption of the test by clinicians responsible for these patients. Based on the proven track record of planar scintigraphy, our concerned clinicians will fall back on the proven modality until further data are available.

Conclusion

Accuracy studies are the cornerstone of the assessment of diagnostic tests. When reading the results of such a study, assessment of its validity should mainly focus on the patient population that was included, and on whether or not the new test was compared in a blinded and independent fashion with a reference test. The test should be described in enough details to enable its implementation, and clear interpretation criteria should be provided. Before adopting a new test, it is important to understand the impact it will have on patients' management. Beyond accuracy indices, the clinical setting in which the test will be used, and the corresponding pretest probability of disease, will determine how best can the test be used in clinical practice.

Sn and SP refer to sensitivity and specificity, respectively. NPV refers to the negative predictive value of the test.

Acknowledgement

Dr. Le Gal holds a University of Ottawa Department of Medicine Chair in Diagnosis of Venous Thromboembolism, a "CP has heart" Cardiovascular Clinician-Scientist award from the Heart and Stroke Foundation of Ontario, and an Early Research Award from the Ontario Ministry of Research and Innovation.

References

1. Le Gal G, Righini M, Roy P-M, et al: Prediction of pulmonary embolism in the emergency department: The Revised Geneva Score. Ann Intern Med 144:165-171, 2006
2. Wells PS, Anderson DR, Rodger M, et al: Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: Increasing the models utility with the SimpliRED D-dimer. Thromb Haemost 83:416-420, 2000
3. Ceriani E, Combescure C, Le Gal G, et al: Clinical prediction rules for pulmonary embolism: A systematic review and meta-analysis. J Thromb Haemost 8:957-970, 2010
4. Waxman AD, Bajc M, Brown M, et al: Appropriate use criteria for ventilation-perfusion imaging in pulmonary embolism: Summary and excerpts. J Nucl Med 58, 2017. 13N-5N
5. Donohoe K, Ahuja S: Society of nuclear medicine and molecular imaging efforts toward standardization: From procedure standards to appropriate use criteria. Semin Nucl Med 2019. (in press)
6. Chassé M, Fergusson D: Diagnostic randomized trials vs. traditional diagnostic cohort studies. Semin Nucl Med 2019. (in press)
7. Jaeschke R, Guyatt G, Sackett DL: Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 271:389-391, 1994
8. Jaeschke R, Guyatt GH, Sackett DL: Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 271:703-707, 1994
9. Greenhalgh T: How to read a paper. Papers that report diagnostic or screening tests. BMJ 315:540-543, 1997

10. McInnes MDF, Lim CS, van der Pol CB, et al: Reporting guidelines for imaging research. Semin Nucl Med. 2019. (in press)
11. Phillips JJ, Straiton J, Staff RT: Planar and SPECT ventilation/perfusion imaging and computed tomography for the diagnosis of pulmonary embolism: A systematic review and meta-analysis of the literature, and cost and dose comparison. Eur J Radiol 84:1392-1400, 2015
12. Bajc M, Neilly JB, Miniati M, et al: EANM guidelines for ventilation/perfusion scintigraphy: Part 1. Pulmonary imaging with ventilation/perfusion single photon emission tomography. Eur J Nucl Med Mol Imaging. 36:1356-1370, 2009
13. Le Roux PY, Robin P, Delluc A, et al: V/Q SPECT interpretation for pulmonary embolism diagnosis: Which criteria to use? J Nucl Med 54: 1077-1081, 2013
14. Le Duc-Pennec A, Le Roux PY, Cornily JC, et al: Diagnostic accuracy of single-photon emission tomography ventilation/perfusion lung scan in the diagnosis of pulmonary embolism. Chest 141:381-387, 2012
15. Moores LK: Diagnosis and management of pulmonary embolism: are we moving toward an outcome standard? Arch Intern Med 166:147-148, 2006
16. Leblanc M, Leveillee F, Turcotte E: Prospective evaluation of the negative predictive value of V/Q SPECT using 99mTc-Technegas. Nucl Med Commun 28:667-672, 2007
17. Le Roux PY, Palard X, Robin P, et al: Safety of ventilation/perfusion single photon emission computed tomography for pulmonary embolism diagnosis. Eur J Nucl Med Mol Imaging 41:1957-1964, 2014