Short Communication

Evidence Based Medicine in Nuclear Medicine Practice; Part II: Appraising and Applying the Evidence

Ramin Sadeghi, MD

Nuclear Medicine Research Center, Imam Reza Hospital,
Mashhad University of Medical Sciences Mashhad, Iran

(Received 22 April 2009, Revised 26 May 2009, Accepted 28 May 2009)

ABSTRACT

As described in the first part of this article, Evidence Based Medicine (EBM) is a growing part of medical practice which emphasizes on the best evidence. Finding this evidence by formulating an answerable question and searching strategies were described in the first part of this review. In this part, appraising the retrieved article (with the main focus on the diagnostic studies) and applying the appraised evidence in the medical practice are explained.

Key words: Evidence based medicine, Nuclear medicine, Critical appraisal, Sensitivity, Specificity

Iran J Nucl Med 2009; 17(1): 49-56

Corresponding author: Ramin Sadeghi, Assistant Professor, Nuclear Medicine Research Center, Imam Reza Hospital, Mashhad University of Medical Sciences, Ebn Sina Street. Mashhad, Iran.
E-mail: sadeghir@mums.ac.ir

INTRODUCTION

EBM is a new approach in health care, which has not been addressed fully in Nuclear Medicine (1, 2). In the first part of this article, the importance of Evidence Based Medicine (EBM) was described and the first two steps of this practice were explained (3). These two steps include asking answerable questions and searching for the best evidence. After completion of these steps, critical appraisal of the found evidence is of utmost importance. Subsequently, applying the best found evidence to the specific patient, whom is planned to be treated, should be chosen. In this review, these two steps are explained with focus on the diagnostic studies. It is worth mentioning that most of the issues described in this review are also relevant to radiology in general.

STEP III: CRITICAL APPRAISAL

As mentioned before, not all published studies meet the standards of high quality evidence. The process of evaluating studies to assure the high quality is called critical appraisal. An easy and efficient way to do this task is assigning a level of evidence to each article.

Levels of evidence

The Oxford Centre for Evidence Based Medicine provides a comprehensive Table of levels of evidence online which is freely available (4). Table 1 shows levels of evidence for diagnostic studies. These levels of evidence can be easily assigned to found articles and the highest quality studies can be picked and used for further evaluation and use. By this strategy, one does not need to read all article regarding a specific clinical scenario.

For assigning the level of evidence, several specific questions should be asked regarding an individual study. For this, there are some standard appraisal sheets which are available online (5, 6). These appraisal sheets usually have two main sections: 1) Evaluation of the study validity which can be found in the materials and methods section. 2) The effect size or strength of the study which can be found in the result section. Usually the first section is used to assign the level of evidence (7).

Question 1: What was the spectrum of patients who underwent the test in question?

A high quality test would cover a whole spectrum of target disease regarding severity and temporality (mild and severe disease, acute and chronic disease). For minimizing selection bias it is best to select patients randomly. The characteristics of the patients (such as gender, age, ethnicity, etc) should be considered to assure that the study is applicable to the patients whom the test is going to be used on (5, 8). The eligibility criteria should also be defined meticulously (8).

Question 2: Was the reference standard the best available test?

The reference standard is the test used in a survey to verify the diagnosis of the studied disease. It should be the best available test. For example, to validate the diagnosis of tuberculosis, the reference standard can be sputum culture and for pulmonary embolism it would be pulmonary angiography. Sometimes, follow up instead of reference standard is used to find out if the patient has the disease or not. In this case, the period of follow up should be long enough considering the nature of the disease (5, 8). In some situations, a single reference standard is not available and a combination of clinical findings and paraclinical tests should be used. For example for diagnosis of recurrent lymphoma lesions this strategy should be used (9).

Question 3: Were the reference standard and the index test applied to all patients in the study regardless of the indexed test results?

The reference standard should be applied to all patients blindly. Failure to meet this standard is the most common flaw in the diagnostic literature (7).

The period between reference standard and index test is also important and should be considered in critical appraisal. This period should be short enough to assure that the target disease did not change between the reference standard and the index test (6). For example, the time between a pulmonary embolism event and pulmonary angiography should not be too long, since the thrombolysis process can interfere in the diagnosis of pulmonary embolism (10, 11).

Table 1. Oxford centre for evidence-based medicine levels of evidence for diagnosis studies. (reproduced with permission)

Level	Diagnosis
1a	SR (with homogeneity*) of Level 1 diagnostic studies; CDR† with 1b studies from different clinical centres
1b	Validating** cohort study with good†† reference standards; or CDR† tested within one clinical centre; Independent blind comparison of an appropriate spectrum of consecutive patients, all of whom have undergone both the diagnostic test and the reference standard
1c	Absolute SpPins and SnNouts†††
2a	SR (with homogeneity*) of Level >2 diagnostic studies
2b	Exploratory** cohort study with good†††reference standards; CDR† after derivation, or validated only on split-sample§ or databases; Independent blind comparison but either in nonconsecutive patients or confined to a narrow spectrum of study patients (or both), all of whom have undergone both the diagnostic test and the reference standard; or a clinical decision rule not validated
3a	SR (with homogeneity*) of 3b and better studies
3b	Non-consecutive study; or without consistently applied reference standards
4	Case-control study, poor or non-independent reference standard
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"

* By homogeneity we mean a systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all worrisome heterogeneity need be statistically significant. As noted above, studies displaying worrisome heterogeneity should be tagged with a "-" at the end of their designated level.

** Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (e.g. using a regression analysis) to find which factors are 'significant'.

† Clinical Decision Rule. (These are algorithms or scoring systems which lead to a prognostic estimation or a diagnostic category).

†† Good reference standards are independent of the test, and applied blindly or objectively to applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test. Use of a non-independent reference standard (where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference') implies a level 4 study.

††† An "Absolute SpPin" is a diagnostic finding whose Specificity is so high that a Positive result rules-in the diagnosis. An "Absolute SnNout" is a diagnostic finding whose Sensitivity is so high that a Negative result rules-out the diagnosis.

§ Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into "derivation" and "validation" samples.

Question 4: Was the comparison between the index test and reference standard blind and independent?

The interpretation of the index and reference standard tests should be blind to the results of the other test. The failure to meet this standard is called expectation bias (8).

Question 5: Were the index test and reference standard explained fully in the article?

This is important since full replication of results is dependent on it. Every aspect of both tests and all commercial names should be explained fully. For example, the type of the collimator, the type of the gamma camera, etc should be mentioned in the study.

Question 6: What were the results of the study?

For each diagnostic test, sensitivity, specificity, positive & negative predictive values (PPV & NPV) are the main characteristics which should be presented in the results section of the study. Table 2 shows a 2×2 chart for a test with dichotomous results. The definition of the characteristics of the test is shown in Table 3.

Table 2. 2×2 chart for a test with dichotomous results

	Reference standard positive	**Reference standard negative**
Index test positive	True positive (a)	False positive (b)
Index test negative	False negative (c)	True negative (d)

Table 3. The definitions of the characteristics of the test shown in Table 2.

Characteristic of the test	**Definition**	**Formula**
Sensitivity	How good is this test at detecting people with the disease?	a/(a+c)
Specificity	How good is this test at correctly excluding people without the disease?	d/(b+d)
Positive predictive value	What is the probability that a person with positive test has the disease?	a/(a+b)
Negative predictive value	What is the probability that a person with negative test does not have the disease?	d/(c+d)
Accuracy	What proportion of all tests yielded the correct result?	(a+d)/(a+b+c+d)

For tests with multilevel or continuous results (such as quantitative gated SPECT studies), the cut-off of the index test is the main determinant of the sensitivity and specificity. In these conditions, the sensitivity and specificity have a negative correlation with each other. This is the cornerstone of Receiver Operating Characteristic (ROC) analysis. ROC curve shows the sensitivity and specificity for different cut-off points of the test results. Area under the curve in ROC analysis is of utmost importance. When this area equals to 0.5, the test would be useless. The closer the area is to unity, the better the performance of the test is (Figure 1).

Sensitivity and specificity are considered the internal properties of a test and are constant regardless of the prevalence of the disease in a population the test in used in. On the other hand, predictive values (both negative (NPV) and positive (PPV)) change according to the prevalence of the disease in the population. This is the main idea of the Bayes' Theorem which is more discussed in the end of this article alongside the concepts of likelihood ratio, pre-test and post-test probabilities (1-2, 12).

Question 7: Were the confidence intervals mentioned for the test results?

Providing the p-values, sensitivity, specificity, and other statistical test results, is not sufficient enough for interpretation. P-value is only a probability that an outcome has occurred by chance. It is by no means a surrogate of effect size. To realize the effect size of a study the confidence intervals should be provided. When the confidence intervals range is wide, usually the sample size is small (13). For example sensitivity of a test can be 60% in two different studies; however the confidence intervals for this sensitivity can be 25-80% and 50-70% respectively. The second study provides better evidence in this regard. Confidence intervals can be calculated using online calculators (14).

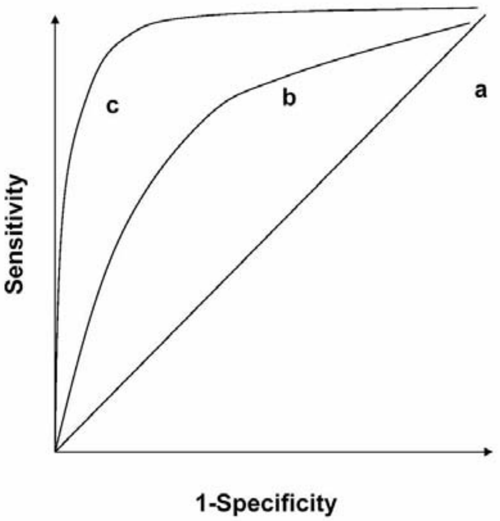

Figure 1. Receiver Operating Characteristic (ROC) curves of three different tests. The more the curve is to the left and top (or larger area under the curve), the better the performance of the test is. For example the test "c" performs the best and the test "a" the worst.

Full explanation of this issue is beyond the scope of this article and an excellent book written by Cohen addressed this issue in details (15).

Question 8: Was there any withdrawal from the study and if there was, what was the explanation for it?

If patients withdrew from the study and failed to be followed up, the results of the study may be biased. The full explanation for any withdrawal should be provided in the study.

By answering these questions we can attribute a level of evidence to any particular study. Articles with the highest level of evidence should be considered for EBM practice.

STEP IV: APPLYING THE EVIDENCE TO A PARTICULAR PATIENT.

As mentioned above, the positive and negative predictive values of a test change with the prevalence of a disease in a society. Usually the prevalence of a disease is called pre-test probability. This probability can be refined according to the test results to what is called post-test probability. For this purpose, likelihood ratios (LRs) are very useful. The definitions of LRs are shown in Table 4. For a given pre-test probability (prevalence), pre-test odds should be calculated.

PRE-TEST ODDS = PRE-TEST PROBABILITY/(1-PRE-TEST PROBABILITY)

The main property of odds and likelihood ratios is the ability to combine them by multiplication (7). No matter how many test are used to refine a probability, this method can be applied. For example if chest X-ray, D-dimer, and V/Q scans are all performed for a patient suspicious of pulmonary embolism, the LRs of all these tests can be used as follows:

POST-TEST ODDS = PRE-TEST ODDS × LR of test1 × LR of test2 × LR of test3 ...

Finally, the refined post-test probability of the disease can be calculated:

POST-TEST PROBABILITY = POST-TEST ODDS/(POST-TEST ODDS+1)

An alternative way for this task is using especial nomogram. This nomogram is shown in Figure 2. For using this nomogram, a line should be drawn which pass through the pre-test probability and LR of interest. The number where this line crosses the post-test probability line is the post-test probability.

A graphic way to express the concept of pre and post-test probability is shown in Figure 3.

Usually a threshold is set for each disease for treatment, which is called treatment threshold. This is the probability of the disease above which, the treatment of the disease is justified. For example, this threshold for osteosarcoma is very high (the diagnosis of osteosarcoma should be almost certain (near 100%) to start treatment). This is also true for pulmonary embolism: the probability of the presence of pulmonary embolism should be higher than 80% (high probability) to start treatment.

Table 4. The definitions of likelihood ratios (LRs) for the test shown in Table 2.

Characteristic of the test	Definition	Formula
LRs of a positive test	What is the likelihood of a positive test to be found in a person with the disease compared to a person without it?	Sensitivity / (l-specificity)
LRs of a negative test	What is the likelihood of a negative test to be found in a person without the disease compared to a person with it?	(l-sensitivity) /specificity

The main reason for requesting a paraclinical test is to refine the pre-test probability of a particular patient. If the post-test probability becomes higher than the treatment threshold, the treatment of the disease is justified and vice versa (Figure 3).

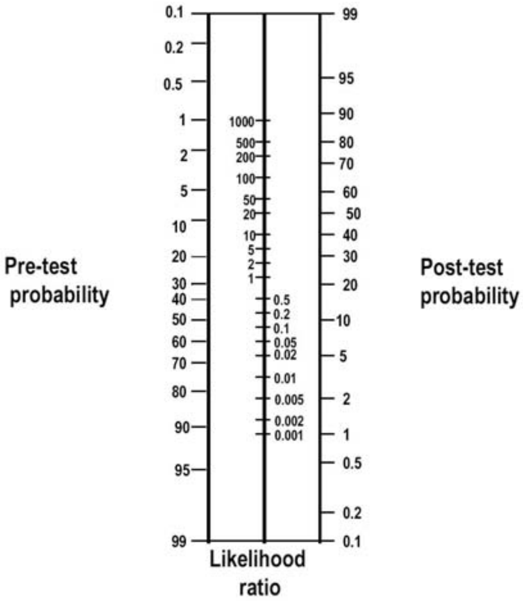

Figure 2. Nomogram for calculation of post-test probability from likelihood ratio and pre-test probability.

Other important issues to be addressed

When deciding to implement the best found evidence in real practice, cultural issues (such as religion, etc.) should be born in mind. Some tests and procedures may not be culturally acceptable and before requesting a test, this important fact should be addressed.

Economical issues are another aspect of health care practice. With limited resources for health care, too many expensive procedures and tests should not be requested for the patient management. This is beyond the scope of this review and you can find more detailed explanation elsewhere in the literature (16).

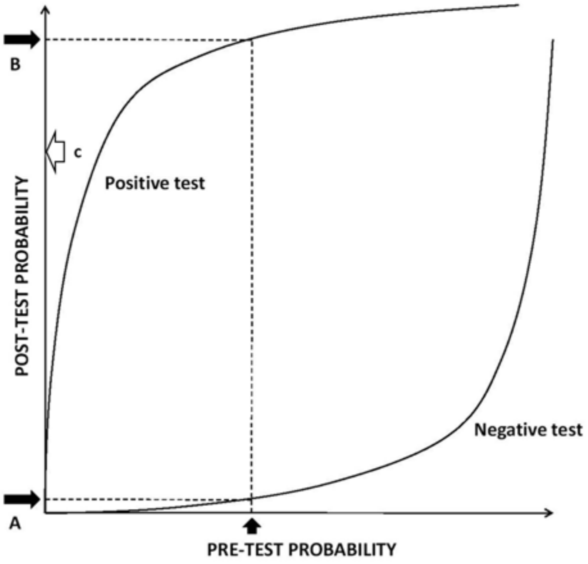

Figure 3. A graph which shows the correlation between pre-test probability and both negative and positive results of a particular test. If a particular patient has a pre-defined pre-test probability, which is shown by vertical arrow, the post-test probability after getting positive and negative results (A and B respectively) are shown by horizontal arrows. The treatment threshold is shown by white arrow (C). Since B>C if we get a positive result of this particular test, starting treatment in our patient is justified. On the other hand, A<C and this means that if we get a negative result of this particular test in our patient, the treatment is not justified.

FINAL WORD

In this article and the part I, the EBM practice has been briefly explained with major focus on diagnostic studies specially related to practice of nuclear medicine. By applying the techniques of EBM, we can provide the best service for our patients more efficiently and in a less time-consuming way.

REFERENCES

1. Wood Alvarez Ruiz S, Cortés Hernández J, Rodeño Ortiz De Zárate E, Alonso Colmenares JI, Alcorta Armentia P. Evidence based medicine. Generalizations on the application to nuclear medicine. Part I. Rev Esp Med Nucl 2001; 20(4): 313-328 [Spanish].

2. Alvarez Ruiz S, Canut Blasco A, Rodeño Ortiz de Zárate E, Barbero Martínez I, Alonso Colmenares JI, Cortés Hernández J et al. Evidence based medicine. Application to nuclear medicine. Diagnostic slope. Part II. 1: Rev Esp Med Nucl 2001; 20(5):393-412;
3. Sadeghi R. Evidence based medicine in nuclear medicine practice. Part I: Introduction, asking answerable questions and searching for the best evidence. Iran J Nucl Med 2009; 17(1): 41-48.
4. Levels of evidence. Oxford Centre for Evidence-Based Medicine Web site. http://www.cebm.net/index.aspx?o=1025.Accessed April 12, 2009.
5. Diagnostic Critical Appraisal Sheet. Oxford Centre for Evidence-Based Medicine Web site. http://www.cebm.net/index.aspx?o=1096.Accessed April 13, 2009.
6. Scottish Intercollegiate Guidelines Network (SIGN) Web site. http://www.sign.ac.uk/methodology/checklists.html. Accessed April 12, 2009.
7. Dodd JD. Evidence-based practice in radiology: Steps 3 and 4- Appraise and apply diagnostic radiology literature. Radiology 2007; 242(2): 342-354.
8. Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. BMJ 1997; 315(7107): 540-543.
9. Hoda S. Role of nuclear medicine in detection and management of hodgkin’s disease and non-Hodgkin’s lymphoma. Iran J Nucl Med 2002; 16-17: 17-25.
10. Worsley D, Silberstein EB, Alavi A, Elgazzar A. Very low probability lung scan findings: A need for change. Iran J Nucl Med 1997; 6-7: 27-29.
11. Dadparvar S, Woods K, Magno RM, Sabatino JC, Patil S, Dou Y. Diagnosis of thromboembolic disease: Combined ventilation perfusion lung scan and compression ultrasonography. Iran J Nucl Med 2002; 16-17: 26-28.
12. Sackett DL, Strauss SE, Richardson WS, Rosenberg W, Haynes RB. Evidence based medicine: how to practice and teach EBM. 2nd ed. Edinburgh, Scotland: Churchill Livingstone, 2000; 1–12.
13. Greenhalgh T. How to read a paper. Statistics for the non-statistician II: "Significant" relations and their pitfalls. BMJ 1997; 315(7105): 422-425.
14. CAT maker software. Oxford Centre for Evidence-Based Medicine Web site. http://www.cebm.net/index.aspx?o=1216.Accessed April 13, 2009.
15. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum Associates, 1988.
16. Greenhalgh T. How to read a paper. Papers that tell you what things cost (economic analyses). BMJ 1997; 315(7108): 596-599.