J Med Biochem 2011; 30 (3) DOI: 10.2478/v10011-011-0028-0

UDK 577.1 : 61 **ISSN 1452-8258**

J Med Biochem 30: 193–200, 2011

Review article
Pregledni članak

DEFINING BIOMARKER PERFORMANCE AND CLINICAL VALIDITY

DEFINISANJE PERFORMANSI I KLINIČKE VALIDNOSTI BIOMARKERA

Patrick M.M. Bossuyt

Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, the Netherlands

Summary: In the evaluation of biomarkers three questions can be answered: what is the analytical validity of the marker, what is its clinical validity, and does the marker have clinical utility? In most cases, clinical validity will be expressed in terms of the marker's accuracy: the degree to which it can be used to correctly identify diseased patients or, more generally, patients with the target condition. Diagnostic accuracy is evaluated in studies in which the biomarker values are compared to the outcome of the clinical reference standard in the same patients. There are several ways in which the results of diagnostic accuracy studies can be summarized, reported, and interpreted. In this paper we summarize and present the available measures. We classify these as error-based measures, information-based measures, and measures of the strength of the association. Clinical validity is linked to clinical utility. If the target condition is well defined and associated with unequivocal downstream management decisions, clinical validity, when defined in comparative terms, may sometimes act as a surrogate outcome measure for clinical utility.

Keywords: biomarker, accuracy, clinical validity, clinical utility

Kratak sadržaj: Evaluacija biomarkera može podrazumevati odgovore na tri pitanja: kakva je analitička validnost markera, kakva je njegova klinička validnost i da li je marker klinički koristan. U većini slučajeva, klinička validnost biće izražena kao tačnost markera: do koje se mere on može koristiti za tačno identifikovanje obolelih pacijenata ili, uopšteno, pacijenata sa ciljanim stanjem. Dijagnostička tačnost procenjuje se u studijama u kojima se vrednosti biomarkera porede sa ishodom kliničkog referentnog standarda kod istih pacijenata. Postoji nekoliko načina da se sumiraju, predstave i tumače rezultati studija dijagnostičke tačnosti. U ovom radu sumiraćemo i predstaviti dostupne mere. Podelili smo ih na mere zasnovane na greškama, mere zasnovane na informacijama i mere jačine asocijacije. Klinička validnost povezana je sa kliničkom korisnošću. Ukoliko je ciljano stanje tačno definisano i povezano sa ispravnim odlukama o terapiji, klinička validnost, kada se definiše u komparativnim terminima, ponekad može biti surogat mera ishoda za kliničku korisnost.

Ključne reči: biomarker, tačnost, klinička validnost, klinička korisnost

Introduction

In recent decades the number of medical tests and biomarkers has been rising at a rapid pace. New markers are proposed at an increasing rate and the technology of existing tests is continuously being improved. Like any other medical technology, new medical tests and biomarkers should be thoroughly evaluated prior to their introduction into daily practice. A rigorous evaluation process of diagnostic tests before introduction into clinical practice will not only improve patients' health but contribute to an efficient use of health care resources by preventing unnecessary testing.

Unfortunately, the evaluation of medical tests is less advanced than that of treatments. The methodology has been less well developed. There is uncertainty about what methods to use, and what the sources of bias are in biomarker studies.

Address for correspondence:

Patrick M.M. Bossuyt, Ph. D.
Dept. Clinical Epidemiology & Biostatistics
Academic Medical Center
University of Amsterdam
Room J1b-214
PO Box 22700; 1100 DE Amsterdam; the Netherlands
p.m.bossuyt@amc.uva.nl

The lack of progress in the methodology for biomarker evaluation can in part be attributed to the lower standards for the regulation of biomarkers. Unlike the evaluation of drugs, for which the threshold to marketing is relatively steep, entry to the market for developers of biomarkers has been less difficult for most products.

In this paper we present a triad of questions that can be asked about any new biomarker. The three questions deal with the analytical validity of the marker, its clinical validity and the clinical utility. We discuss in more detail how the clinical validity of diagnostic markers is expressed, in terms of diagnostic accuracy. We close by describing when and how the clinical validity of a biomarker can be regarded as a proxy measure for the clinical utility.

Three questions about biomarkers

The evaluation of medical technology can be a time-consuming and costly process. An efficient use of resources calls for a well-planned evaluation strategy, in which more elaborate and therefore more expensive forms of evaluation are only performed if satisfactory results have been obtained in the previous steps of the evaluation process. Such a phased approach, moving gradually from small to larger studies, may also protect the rights and integrity of human volunteers and patients.

Several comparable hierarchical models have been proposed for the evaluation of tests and biomarkers. Analogous to the 4-phase model for the evaluation of new drugs, these models require that in each phase certain conditions be fulfilled before the evaluation can continue with the subsequent phase.

In a systematic review, we identified 19 phased evaluation schemes (1). One of the best known schemes consists of the levels of efficacy proposed by Fryback and Thornbury (2). Their scheme, originally developed for imaging, has also been used for other forms of testing. In genetics, the ACCE and EGAPP frameworks have become more widely used (3). The identified schemes showed substantial similarity. *Table I* presents a simplified summary of these schemes, translated as a set of three questions.

The first question is »Can I trust the results of this marker?«. This is generally referred to as the analytical validity of the marker: the marker's ability to measure what it is supposed to measure. The analytical validity of a test refers to its ability to accurately and reliably measure the entity or analyte of interest.

Table I Three questions in the Evaluation of Biomarkers.

Question	Feature
Is it true?	Analytical Validity
Is it meaningful?	Clinical Validity
Is it useful?	Clinical Utility

There is no single statistic to express the level of analytical validity. Measures used include analytical sensitivity or limits of detection, precision, analytical specificity (cross-reactivity, interference), assay linearity, reliability and repeatability of test results, and assay robustness. The terms are not always used in an unambiguous way and there is little standardization in methods for this initial technical evaluation.

Analytical validity is usually evaluated in laboratory situations. For many methods, it is evaluated by using the test to detect and to measure the quantity of a known substance of a known concentration in a specimen. In this process, establishing reference methods can be problematic.

Other aspects that have to be documented in this initial evaluation phase include feasibility, required equipment and personnel, and physical and biochemical parameters specific to the test, such as the minimal detection level, circadian fluctuation, resolution, contrast level, and reproducibility.

The second question to be answered is »Are the results of this test meaningful?«. In this phase researchers show that the results of the test are meaningfully related to other clinical information. The type of information will depend on the purpose of testing. Markers can be used for diagnostic purposes. In that case, marker values will be correlated with the gold standard, or clinical reference standard outcome in the same patients, and the results will be expressed as the diagnostic accuracy of the marker. Diagnostic accuracy statistics will be discussed and compared later in this paper. For prognostic markers, the meaningful relations will include associations with future events, or future health or disease states. If the marker is proposed for monitoring therapy, the associations will be with target therapeutic levels of the drug, for example, or through the early identification of side-effects, or lack of effect.

The question that will be pivotal in decisions about markers is the one about its clinical utility: is using the marker helpful in improving or maintaining the health of patients? Or, alternatively: does using the marker lead to a more efficient use of health care resources, without compromising patient outcome? This is in principle an explicitly comparative question, with the current best strategy for managing patients as the comparator. It is also a strategy in which the effects of testing have to be expressed in terms of outcomes that matter to patients, such as survival, functional health, or health-related quality of life.

Below we first discuss the best known family of statistics for expressing the clinical validity of markers: expressions of diagnostic accuracy. The presentation

is based on a more general analysis of diagnostic accuracy statistics for medical tests (4).

Diagnostic accuracy

If the marker is to be used for diagnostic purposes, its clinical validity can be expressed in terms of its diagnostic accuracy. The diagnostic accuracy of a marker is the ability of a marker to distinguish between patients with and patients without disease or, more generally, between those with and without the target condition (4).

In studies of diagnostic accuracy, the outcomes from one or more tests are compared with outcomes of the reference standard, obtained in the same study participants. The clinical reference standard is the best available method for establishing the presence of the target condition in patients. The target condition can be a target disease, a disease stage, or some other condition that qualifies patients for a particular form of management. The reference standard can be a single test, a series of tests, a panel based decision, or some other procedure (5). For simplicity, we will assume that the results of the biomarker can be classified as positive, pointing to the presence of disease, or negative. We also assume that the target condition is either present or absent, and that the clinical reference standard is able to identify it in all patients.

Figure 1A shows the basic structure of a typical diagnostic accuracy study. *Figure 1B* offers an example of a diagnostic accuracy study of progastrin-releasing peptide (proGRP), to help identify small-cell lung cancer (SCLC) in patients with well-differentiated neuroendocrine tumors (WDNET) (6). ProGRP is a precursor form of gastrin-releasing peptide, a neuropeptide hormone. Gastrin-releasing peptide is produced by SCLC cells but it is extremely unstable and therefore not suitable as a tumor marker. In contrast, proGRP is a very stable peptide with a half-life of 19–28 days.

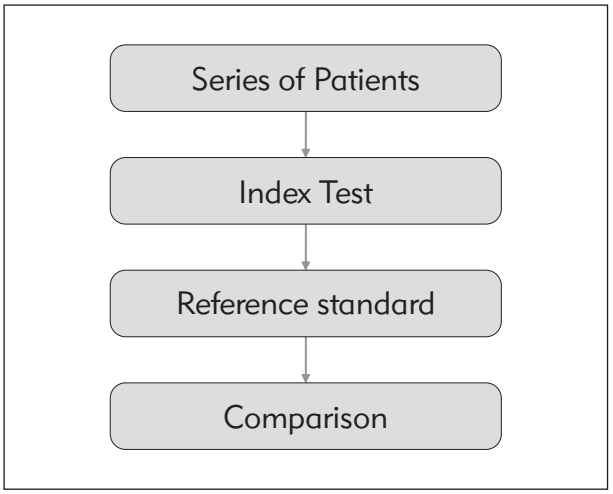

Figure 1 A schematic representation of a diagnostic accuracy study.

There are several potential threats to the internal and external validity of a study of diagnostic accuracy. Poor internal validity will produce bias, or systematic error, because the estimates do not correspond to what one would have obtained using optimal methods. Poor external validity limits the generalizability of the findings. In that case, the results of the study, even if unbiased, do not correspond to the data needs of the decision-maker.

The ideal diagnostic accuracy study examines a consecutive series of patients, enrolling all consenting patients suspected of the target condition within a specific period. All of these patients undergo the index test and then all undergo the reference test. Alternative designs are possible, some of which can be quite difficult to unravel. Some studies first select patients known to have the target condition, and then contrast the results of these patients with those from a control group. If the control group consists of healthy individuals only, diagnostic accuracy of the test will be overestimated. Studies of diagnostic tests with a suboptimal design are known to produce biased estimates.

Because of the biasing effect of design deficiencies, those who report of diagnostic accuracy studies should be well informed about how the study was designed and conducted. Unfortunately, studies of diagnostic accuracy suffer from poor reporting (7). The STARD initiative has been set up to improve the completeness and transparency of reporting (8).

Table II summarizes the results of the study of the accuracy of proGRP in the detection of SCLC in patients with WDNET. The numbers are based on the study performed and reported by Korse and colleagues in the Annals of Oncology. That study did not provide the raw numbers, so these numbers are reconstructed from the statistics reported in the paper. *Table II* reports four numbers, one in each cell, corresponding to the true and false positives, and the true and false negatives. *Table III* shows accuracy statistics that can be calculated from the data obtained (4).

Table II Diagnostic accuracy study results.

proGRP	Location	
	Lung	No Lung
Positive	TP = 18	FN = 2
Negative	FN = 2	TN = 238

TP: True positives; FP: False positives; TN: True negatives; FN: False negatives

Results of a diagnostic accuracy study to identify small-cell lung cancer in patients with well-differentiated neuroendocrine tumors. Cut-off value 90 ng/L.

Table III Measures for reporting diagnostic accuracy studies.

Measure	Definition	Estimate	95% CI
Error-based			
Sensitivity	TP / (TP+FN)	43%	34% to 46%
Specificity	TN / (TN+FP)	99%	98% to 100%
Overall Fraction Correct	(TP+TN) / N	91%	88% to 92%
Youden's J index	[TP/(TP+FN)] – [FP/(TN+FP)]	42%	32% to 46%
Information-based			
Positive Predictive Value	TP / (TP+FP)	90%	72% to 97%
Negative Predictive Value	TN / (TN+FN)	91%	89% to 91%
Positive Likelihood Ratio	[TP/(TP+FN)] / [FP/(TN+FP)]	51.4	14.5 to 196
Negative Likelihood Ratio	[FN/(TP+FN)] / [TN/(TN+FP)]	0.58	0.54 to 0.67
Association-based			
Diagnostic Odds Ratio	[TP × TN] / [FP × FN]	89	21 to 364
Kappa		0.54	0.41 to 0.59

One of the simplest measures of diagnostic accuracy is the overall fraction correct, sometimes also referred to as simple 'accuracy'. In the example in *Table II*, in 91% of the study patients the location of the tumor was correctly classified by the biomarker proGRP. However, the overall fraction correct is usually not a very helpful measure. With most conditions, there is a substantial difference between a false positive and a false negative test result. Two of the more frequently used measures of diagnostic accuracy take the differential nature of misclassification errors into account. The diagnostic or clinical sensitivity of the test is the proportion of the diseased correctly classified as such. In the example in *Table II*, the sensitivity of the test is estimated as 43%. Its counterpart is the diagnostic specificity: the proportion of the patients who do not have the target condition correctly classified as such. In the example in *Table II*, the specificity of the test is estimated as 99%. Sensitivity and specificity go hand in hand. If the positivity rate of the proGRP test were increased by selecting a lower positivity threshold than 90 ng/L, the number of true positives would go up, but so would the number of false positives.

Youden's J index is an alternative single measure of error-based accuracy (9). It can be defined in many ways, one of them being the true positive fraction minus the false positive fraction. If the positivity rate is the same in patients with lung cancer and patients without lung cancer, the test is useless and the Youden index is zero. With a perfect test, all lung cancer patients are positive and there are no false positives, so the Youden index is 1. Youden's index has the advantage of being a single measure, but it also loses the distinction between the false positives and the false negatives. So do other error-based measures, such as the area under the ROC curve. The paper on the proGRP has an ROC curve (6).

A number of alternative measures express the information value in specific test results. The positive predictive value of a test is the proportion of patients with a positive test result that actually have the target condition. Its counterpart, the negative predictive value, stands for the proportion of patients with a negative test result that do not have the target condition. For the study results summarized in *Table II*, the positive predictive value is estimated at 90% and the negative predictive value as 91%. If the proGRP test is used to rule out lung cancer, the negative predictive value tells us the proportion of patients with a negative proGRP who did not have SCLC. The positive predictive value tells us the prevalence of lung cancer in those with a positive proGRP result.

Like sensitivity and specificity, predictive values are essentially group based measures. Large differences may exist in the strength of pre-test suspicion of lung cancer in a clinical setting, based on the patient's presentation, his risk factors, and findings from history, and physical examination. The positive predictive value ignores all of that: it just tells us the proportion of lung cancer patients within those with a positive test result.

A different set of information-based measures has been proposed as an alternative: diagnostic likelihood ratios. The likelihood ratio (LR) of a particular test result is the proportion of subjects with the target condition who have that test result relative to the proportion without the target condition who have the same test result. Unlike sensitivity and specificity, LRs can also be obtained for tests that can have multiple, or even continuous, test results, without the need for dichotomization. If such a test is evaluated in an accuracy study, a set of LRs, one for each test result, or an LR function, will be reported.

Diagnostic LRs are supposed to be used with subjective pre-test probability expressions and Bayes' Theorem. The pre-test odds of the target condition being present – the pre-test probability relative to one minus that probability – must be multiplied with the LR of the obtained test result to produce the post-test odds: the posttest probability relative to one minus that probability.

LRs above unity increase the probability of disease, while LRs lower than one decrease that probability. In *Table II*, the LR of a positive result is estimated at 51 and the LR of a negative test result is estimated at 0.58. McGee has suggested a very rough simplification for the interpretation of LRs (10). For him, clinicians must remember only 3 LRs – 2, 5, and 10 – and the first 3 multiples of 15: 15, 30 and 45. For probabilities between 10 % and 90 %, an LR of 2 increases the probability with approximately 15 %, an LR of 5 with 30 %, and an LR of 10 with 45 %. For likelihood ratios less than 1, the rule works in the opposite direction.

Diagnostic accuracy studies can also be used to estimate the strength of the association between the index test results and the outcomes of the reference standard. Better tests will show stronger associations with the clinical reference standard. Some of the familiar epidemiology statistics are called upon to express the strength of associations. One such statistic is the odds ratio, in this context also known as the diagnostic odds ratio (11). The odds ratio expresses the odds of positivity in the diseased relative to the odds of positivity in the non-diseased. A property of the odds ratio is that this also equals the odds of disease in those with a positive test result relative to the odds of disease in those with a negative test result. Unlike the odds ratios in most other areas of epidemiology, diagnostic odds ratios tend to be quite high. In the example in *Table II*, the diagnostic odds ratio of proGRP is estimated at 89.

The odds ratio is a single measure, in contrast to many other measures, which come in pairs. Odds ratios can be used to make rapid comparisons of tests used for the same target condition. Yet the absence of the differential nature of the information and of the errors limits their use for decision-making.

Many more association-based measures have been used and proposed for diagnostic accuracy studies, such as the error odds ratio, relative risks, kappa statistics, and adjusted kappa statistics. Many of these have been criticized, and most are seldom used in practice.

In short, measures of the strength of the association between test and reference standard, borrowed from epidemiology, are rarely helpful for interpreting accuracy studies to support decision making. The diagnostic and test evaluation literature borrows probably too heavily from the general epidemiological literature, which may be seen as a sign of its immaturity.

Accuracy: a variable test property

Sensitivity and specificity tell us something about the conditional classification quality of the test, but they are not intrinsic test characteristics. The values that sensitivity and specificity take are conditional on the target condition, on how that target condition is defined, and on the clinical reference standard. Sensitivity and specificity do not only differ based on the target condition and the reference standard used. They are also likely to vary with the clinical setting. Test accuracy will also vary with the level of pretesting. Sensitivity and specificity should be looked upon as group averages, that vary across groups, but will also vary within subgroups (12, 13).

Most demonstrations of variability in accuracy have focused on sensitivity and specificity. Yet there is no evidence that other types of measures are unaffected, on the contrary. If the true positive and false positive fractions vary, so will the likelihood ratios, and this should lead to some caution in the unconditional application of Bayes' theorem to individual patients.

Researchers should spend more time on exploring and documenting this variability, because it could be useful for practice (14). Finding conditions that are more likely to generate false positives, for example, could be extremely helpful. The methodology for doing such analyses is well available (15). Unfortunately, we do not find such explorations on a regular basis in the literature. One reason could be the fact that many – if not most – diagnostic accuracy studies have limited sample size. In a survey of papers on tests used for reasons other than population screening, Bachman and colleagues showed that the median sample size

was 119, and the median number with the target condition was only 49 (16). These numbers are small for reliable estimates of diagnostic accuracy, and way too small for explorations of the sources of heterogeneity in accuracy. Meta-analysis may help in obtaining more precise estimates, but they cannot act as a substitute for exploring the sources of variability (17).

In principle, Bayes' theorem should only be used for a single patient on two conditions. The first is that the clinician's pretest probability is an adequate and substantiated expression of the strength of suspicion in the patient. The second condition is that the likelihood ratio expresses how much more likely a positive result, say, is in the patient, if diseased, compared to when that patient were not diseased. Applying a published, group-based likelihood ratio to a specific patient's test result to generate that patient's post-test probability is a leap of faith. Such a leap can only sensibly be made if it based on a solid understanding of that test, and the modifying effects of any condition, such as age, sex and co-morbidity, that may affect the rates of false positives or false negatives.

There seems to be some confusion about the statistics to report after a diagnostic accuracy study, and what to look for when interpreting the results. Some authors seem to be LR believers, while others believe in the logistic-regression based diagnostic function. Still others seem to be confused, and report every statistic mentioned in this paper. A paper on a study to evaluate ST-segment elevation in predicting acute occlusion in patients with acute coronary syndrome reported no less than a bewildering number of 11 accuracy measures (18).

How best to express diagnostic accuracy?

The concepts of sensitivity and specificity for test accuracy were proposed by Jacob Yerushalmy in the early 1940s, in his work on the consistency of chest X-ray reading in suspected tuberculosis (19). The notions became more prominent in medical science after the 1959 Science paper by Ledley and Lusted, although the terms themselves do not appear as such in that paper (20). Ledley and Lusted discussed the use of conditional probabilities and Bayes' theorem, and stated that these conditional probabilities can be grounded in medical knowledge, unlike the probabilities of disease in single patients.

A series of authors have challenged the prominence of these error-based notions and lamented their widespread use. We can arrange the criticism in two lines of thinking, each arriving at a plea for information-based accuracy measures.

One form of criticism stresses that, for practical reasons, predictive values are what matters most, expressed as probabilities, calculated with the help of diagnostic probability functions (21–23). These functions include not only the test result, but also all other available information of diagnostic value, such as data from history, from the physical examination, and all other test results. Using such functions, one can also evaluate the added value of new medical tests.

The second form of criticism discards error-based measures in favor of a more widespread use of diagnostic likelihood ratios. The proponents of this view can be found in many sectors of the evidence-based medicine movement (24, 25). The first users' guide to the medical literature, for example, published by the EBM group, used the availability of likelihood ratios as an essential element in the critical appraisal of reports of diagnostic accuracy studies (25).

Some of the arguments in favor of likelihood ratios can be regarded as a sign of proselyte EBM enthusiasm, and are not always grounded on the truth. An example: »While predictive values relate test characteristics to populations, likelihood ratios can be applied to a specific patient« (26). As discussed before, the likelihood ratios are just as well calculated on the basis of groups, they are subject to variability and bias. Applying them in individual patients is an act of judgment. Another quote from the same paper: »Moreover, likelihood ratios, unlike traditional indices of validity, incorporate all four cells of a 2×2 table, and thus are more informative than any of the other measures alone« (26). This is not a fair reflection of the truth; most statistics discussed so far, including the sensitivity-specificity pair (when considered together), rely on all four cells of the 2x2 table.

The argument that likelihood ratios are more intuitive can also be challenged (27). Most likely none of the measures presented here is very intuitive, as few of us humans are trained to think in terms of probabilities. Likelihood ratios are, in isolation, not very helpful for comparing tests or making decisions about tests.

Clinical validity and clinical utility

We would like to propose that measures have to be selected based on the type of clinical study question that is to be answered. If the study question regards the evaluation of the quality of one or more tests, the error-based measures are probably better suited to summarize the results. In the proGRP example, the quality of the assay to rule out lung cancer can be judged by an appraisal of its sensitivity, as this indicates the proportion of lung cancer patients that would be missed by the test.

Early evaluations of a new assay can determine the sensitivity for various subtypes of the target condition, or the specificity in different classes of non-diseased, such as those with specific co-morbidities or conditions that mimic the target condition, such as

other infections when evaluating a test for infectious diseases. Sensitivities and specificities can also come into play when considering the use of tests in guidelines and clinical flowcharts.

Many questions about tests are comparative in nature (28). Is proGRP better than other biomarkers? Is a qualitative point-of-care test as good as a quantitative test? For replacement questions, relative true and false positive fractions can be calculated, and hypotheses of superiority or equivalence can be statistically tested (29). Studies aimed at helping clinicians in interpreting test results, on the other hand, should consider reporting primarily in terms of information-based measures, possibly using likelihood ratios or diagnostic functions.

In the end, one important question has not yet been addressed in this paper: why accuracy? Some authors have argued that not only sensitivity and specificity are overrated, but that the whole accuracy paradigm is woefully inadequate for appropriately expressing the benefits and harms of testing (30, 31). It is based on a definition of true disease, on definitive evidence, that was formerly found only on postmortem examination, and it does not show the effects of testing on patient outcome.

Does that mean we then have to abandon the accuracy paradigm, and move to evaluations of clinical utility altogether? Probably not. To be useful, we only need a redefinition of what the clinical reference standard is supposed to detect. The pathological gold standard of disease has to be traded in for the notion of a target condition. When measuring test accuracy, the 'target condition' is the classification of disease one wishes to detect. Defining the target condition involves thinking about the clinical decisions the test will be used to guide and determining the most appropriate threshold or criteria to dichotomise the presence or absence of disease for these decisions.

This implies that the clinical reference standard should not be asked to distinguish the diseased from the non-diseased, but to identify those that are better off with a particular form of treatment or, more generally, management, versus those that are not. Such information could come from subgroup analysis in randomized clinical trials, or other evidence (32).

In terms of health outcomes and clinical utility, evaluating how well the target condition can be detected is an intermediate outcome measure. With claims of superior sensitivity for a new test, when a test identifies more cases, it is not unlikely that randomized clinical trials or other forms of research are needed to show that these additional cases benefit as much from treatment, or other interventions, as the cases identified by the older tests (33). After all, diagnosis is but a stepping stone to treatment, to changes in outcome, and to benefits in individual patients.

Conflict of interest statement

The author stated that there are no conflicts of interest regarding the publication of this article.

References

1. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009; 29 (5): E13–21.
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making 1991; 11 (2): 88–94.
3. Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genetics in medicine: official journal of the American College of Medical Genetics 2009; 11 (1): 3–14.
4. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008; 45 (3): 189–95.
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem Lab Med 2003; 41 (1): 68–73.
6. Korse CM, Taal BG, Bonfrer JM, Vincent A, Van Velthuysen ML, Baas P. An elevated progastrin-releasing peptide level in patients with well-differentiated neuroendocrine tumours indicates a primary tumour in the lung and predicts a shorter survival. Annals of oncology: official journal of the European Society for Medical Oncology / ESMO. 2011.
7. Smidt N, Rutjes AW, Van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005; 235 (2): 347–53.
8. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003; 49 (1): 1–6.
9. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's Index. Stat Med 1996; 15 (10): 969–86.
10. McGee S. Simplifying likelihood ratios. J Gen Intern Med 2002; 17 (8): 646–9.
11. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003; 56 (11): 1129–35.
12. Moons KG, Van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity,

likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997; 8 (1): 12–117.

13. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. Am J Cardiol 1986; 57 (13): 1175–80.
14. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002; 324 (7338): 669–71.
15. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction: Oxford University Press; 2003.
16. Bachmann LM, Puhan MA, Ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006; 332 (7550): 1127–9.
17. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149 (12): 889–97.
18. Rostoff P, Piwowarska W, Gackowski A, Konduracka E, El MN, Latacz P, et al. Electrocardiographic prediction of acute left main coronary artery occlusion. Am J Emerg Med 2007; 25 (7): 852–5.
19. Lilienfeld DE. Abe and Yak: the interactions of Abraham M. Lilienfeld and Jacob Yerushalmy in the development of modern epidemiology (1945–1973). Epidemiology 2007; 18 (4): 507–14.
20. Ledley R, Lusted L. Reasoning foundations of medical diagnosis science 1959; 130: 9–21.
21. Miettinen OS, Henschke CI, Yankelevitz DF. Evaluation of diagnostic imaging tests: diagnostic probability estimation. J Clin Epidemiol 1998; 51 (12): 1293–8.
22. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003; 10 (6): 670–2.
23. Guggenmoos-Holzmann I, Van Houwelingen HC. The (in)validity of sensitivity and specificity. Stat Med 2000; 19 (13): 1783–92.
24. Perera R, Heneghan C. Making sense of diagnostic tests likelihood ratios. Evid Based Med 2006; 11 (5): 130–1.
25. 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 1994; 271 (5): 389–91.
26. Chien PF, Khan KS. Evaluation of a clinical test. II: Assessment of validity. BJOG 2001; 108 (6): 568–72.
27. Puhan MA, Steurer J, Bachmann LM, Ter Riet G. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. Ann Intern Med 2005; 143 (3): 184–9.
28. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332 (7549): 1089–92.
29. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. Journal of Clinical Epidemiology 2010; 63 (8): 883–91.
30. Feinstein AR. Misguided efforts and future challenges for research on »diagnostic tests«. J Epidemiol Community Health 2002; 56 (5): 330–2.
31. Mrus JM. Getting beyond diagnostic accuracy: moving toward approaches that can be used in practice. Clin Infect Dis 2004; 38 (10): 1391–3.
32. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000; 356 (9244): 1844–7.
33. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006; 144 (11): 850–5.

Received: April 5, 2011

Accepted: April 20, 2011