Evaluation of diagnostic tests: from accuracy to outcome
Lijmer, J.G.

Citation for published version (APA):
Strategies for the evaluation of diagnostic technologies

Jeroen G. Lijmer, Ben W.J. Mol, Gouke J. Bonsel,
Martin H. Prins, Patrick M.M. Bossuyt

Submitted
Summary

We conducted a systematic review of the literature to collect existing hierarchical models for the evaluation of diagnostic tests. Such hierarchical evaluation models can prevent unnecessary research and inefficient use of available resources, by requiring documented evidence on one stage before one can move to the next one.

Eleven different models were found. These models showed a high similarity. In all models the evaluation of diagnostic accuracy, by comparing the results of the test with those of a reference test, preceded studies on patient outcome.

We investigated the practical usefulness of existing models by applying them to a series of diagnostic research proposals from a Dutch MTA program. The applicability of these models was found to be limited: in many diagnostic problems no diagnostic procedure could be identified as the reference test.

Hierarchical evaluation models for diagnostic procedures have their merits, but only if one is prepared to discard the universal use of the familiar concept of diagnostic accuracy, in terms of a blind comparison of an index test with a reference test in a clinical population. For many decision problems, such a comparison is not possible, not needed, or downright invalid. Instead, one should look at the decision problem that rises from the intended use of the test, and the information needed to support these decisions in an evidence-based approach.
Introduction

Over the last decades many new diagnostic methods have been developed and the number of available options is still increasing. Premature dissemination of such technologies can lead to erroneous diagnoses and unnecessary delays in starting appropriate therapy or, alternatively, to the initiation of unwarranted, sometimes even harmful therapy. Examples of premature dissemination and inappropriate use have been the dexamethasone suppression test for depression, the carcinoembryonic antigen for colon cancer, and $^{125}$I-fibrinogen leg scan for the diagnosis of deep venous thrombosis. In addition, the increasing costs of health care have put a pressure on available budgets, calling for the elimination of ineffective medical technologies. These are ample reasons why new diagnostic technologies should be thoroughly evaluated before they are introduced in daily practice.

The evaluation of medical technologies can in itself be a time-consuming and costly process. An efficient use of resources available for research calls for a well-planned evaluation strategy. In such a strategy, more elaborate and therefore expensive phases should only be performed if sufficient evidence has been obtained in previous steps of the evaluation process. Another advantage of such a hierarchical evaluation model is that it may facilitate the work of readers of medical journals and of decision-makers.

In the evaluation of new pharmaceutical compounds a four phase hierarchical model is well known. In phase one the toxicity and pharmacokinetics of the new drug are assessed. Phase two consists of small-scale clinical investigations, to estimate the effect of treatment and the adverse event rate in relation to dosage. If there is no treatment effect, or if the treatment effect is too small further evaluation will be discontinued. In phase three the effectiveness of the drug is assessed, by measuring patient outcome in randomised clinical trials. If the drug is effective, further surveillance after introduction to the market is necessary. In this fourth phase, the long-term effects and side effects are registered.

With the drug evaluation model as archetype, several comparable hierarchical models have been proposed for the evaluation of diagnostic tests. Analogous to the four-phase model for the evaluation of new drugs, these models require that in each phase certain conditions be fulfilled before the evaluation should proceed with the subsequent phase.

In this paper we review existing hierarchical models for the evaluation of diagnostic tests. We examine their feasibility by applying them to a series of diagnostic research proposals from a Dutch medical technology assessment
program. The paper closes with a discussion of the observed frictions between the existing hierarchical models and real life diagnostic questions, offering essentials for an alternative, problem-based evaluation model.

Hierarchical models for the evaluation of diagnostic tests
We conducted a systematic review of the literature to collect existing hierarchical models for the evaluation of diagnostic tests. Our goal was to find papers describing a phased approach to diagnostic test assessment. The data collection started with an electronic search of the Medline and EMBASE databases. In addition the references of included articles were screened for relevant articles. A Pubmed search for ‘related articles’ was also performed.

We could retrieve 14 papers on the subject.\textsuperscript{4-17} Three of these were based on the model proposed by Guyatt et al.\textsuperscript{5, 6, 13} Two papers referred to the model proposed by Fryback et al.\textsuperscript{9, 10}. In total, 11 different models were found. Table 1 shows the phases described by the different models. Phases that were identical with respect to content, but labelled differently by the original author are placed in footnotes. All models appear quite similar, despite some differences in terminology. Each model consists of 5 to 6 different phases. All start with a technical evaluation of the test, and an assessment of diagnostic accuracy always precedes patient outcome studies.

These models for diagnostic test assessment show an unmistakable similarity with the four-phased hierarchical model used for the evaluation of drug therapies. We can summarize the contents of these models for the evaluation of diagnostic tests into four phases, closely mimicking the drug evaluation models, printed in bold in table 1.

Phase one consists of test development. After this phase the test has to meet pre-specified technical requirements. Aspects that have to be documented in this phase include: required equipment and personnel, physical and/or biochemical parameters specific to the test (e.g. minimal detection level, circadian fluctuation, resolution, contrast level) and reproducibility. Guyatt et al. recommended that, in addition, the test should be applied to a large number of diverse conditions, in order to delineate the possible uses of the new test.\textsuperscript{5}
<table>
<thead>
<tr>
<th>Phase</th>
<th>Zweig et al(^a)</th>
<th>Guyatt et al(^b)</th>
<th>Freedman et al(^c)</th>
<th>Kobberling et al(^d)</th>
<th>Fryback et al(^e)</th>
<th>Kent et al(^f)</th>
<th>Taylor et al(^g)</th>
<th>Silverstein et al(^h)</th>
<th>Schouw et al(^i)</th>
<th>MacKenzie et al(^j)</th>
<th>Pearl et al(^k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Development</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>B</td>
<td>C</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>range of possible uses</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic accuracy</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>C</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>easy accessible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clinical population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contribution to arsenal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Effectiveness</td>
<td>diagnostic thinking</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therapeutic impact</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>patient outcome</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>D</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Post Marketing</td>
<td>societal efficacy</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A = subdivided in 4 types of easy accessible populations; B = subdivided in technical evaluation, standardization, tissue characterization, spectrum of appearance; C = diagnostic efficacy; D = therapeutic effectiveness.
In phase two the diagnostic accuracy of the test is assessed. In clinical investigations the results of the test under evaluation are compared to those from a reference test, in order to establish its discriminating capacity. The reference test is the best available test to detect the target condition of interest. Diagnostic accuracy can then be characterized in terms of sensitivity and specificity, predictive values, likelihood ratios, or receiver operating characteristic curves and derived measures. Some authors distinguish a series of sub-phases at this stage. They propose to evaluate the diagnostic accuracy first in a group of subjects with the disease of interest and a group of normal persons, for an easy comparison. Subsequently, the evaluation is extended to other parts of the disease spectrum. Finally, diagnostic accuracy is evaluated in a clinical population that closely resembles the spectrum of patients for which the test is intended. In addition two authors suggest to compare the diagnostic accuracy of the test with the performance of other tests intended to detect the same target condition before proceeding further.

The clinical effectiveness of the test, in terms of its effect on patient outcome, is evaluated in phase three. A diagnostic test can improve patient outcome in several ways. In some cases health gains are obtained by reducing test-associated morbidity, compared to an alternative test. Another effect can come from reassuring patients. The largest effect on patient outcome is to be expected from an improvement in decisions to select, start, modify, or withhold therapy. One way to estimate this effect is an assessment in randomized clinical trials, in which specific test-treatment combinations are compared. In some situations such an approach will not be feasible. In those cases, a decision analysis comparing different diagnostic strategies may provide an investigative alternative.

Several authors have subdivided the phase of the assessment of clinical effectiveness in three different levels: diagnostic impact, therapeutic impact, and patient outcome. The first two sublevels use proxy measures for assessing the impact of the test on a patient's health status. Diagnostic impact studies assess the degree of change in the subjective likelihood of a diagnosis induced by the test results. High diagnostic accuracy does not necessarily imply that the test is able to change the physician's thinking and subsequent management decisions. This will depend on the information available before the test is ordered and the diagnostic alternatives. The change in the physician's thinking can be measured by comparing the clinicians' pre-test and post-test probabilities for the same patients. Therapeutic impact concerns the way in which therapy decisions change through the results of the test. A test may have diagnostic impact yet fail to affect therapy decisions. Therapeutic impact can be studied by comparing treatment plans before and after test results become available. The validity of studies on "diagnostic impact/thinking" has been questioned as physicians, rather than the test itself, are to some extent the
subject of the study. A negative outcome in such a study tells us more on the included physicians than on the usefulness of the test itself.\textsuperscript{20, 21}

Seven authors have described a fourth phase beyond the assessment of clinical effectiveness. Freedman et al. have suggested studies to monitor changes in clinical practice after the introduction of a new test.\textsuperscript{7} In such studies, changes in diagnostic use and the frequency of tests results can be documented once the new procedure is introduced into routine clinical practice. Such an evaluation can be compared with the post-introduction surveillance in the fourth phase of the evaluation of new drugs. Others proposed the assessment of societal efficacy as a final phase.\textsuperscript{4, 9, 12, 14, 15, 17} This phase moves beyond the individual risks and benefits of a test to an appraisal of the use of resources and medical benefits on a societal level.

**Applying current models to diagnostic decision problems**

Although the multi-phased evaluation models have a strong intuitive appeal, none of the papers we found reported on an explicit, empirical evaluation of the applicability of these models to actual diagnostic problems. The question then remains whether these models are truly useful in structuring diagnostic research. To answer this question we conducted a study of research proposals, to examine the applicability of the multi-phased evaluation models on topical diagnostic questions.

To obtain a sample of diagnostic decision problems we obtained permission to study all grant proposals of a large Dutch MTA program. This program, run by the Netherlands' national health insurance board, started in 1988. Each year, researchers could send in proposals for research projects on the effectiveness of new or existing medical technologies to support decision-making on a national and/or clinical level. We examined the titles and summaries of all 104 grant proposals submitted in 1996 and 1997. Within this set of grant proposals we identified those that dealt with a decision on diagnostic tests.

For each study proposal a set of data were abstracted from the grant application and noted down on a standard form. The type of test under evaluation was classified as laboratory, imaging, microbiology, pathology, or functional. The existence of a reference test for the target condition of the test under evaluation was recorded. The intended use of the test was documented as diagnosis, therapy decisions, monitoring of therapy, or other. Two physicians experienced in medical evaluation research (GB and MP) extracted the data independently. In case of disagreement the judgment of a third physician (JL) was decisive.

Within the 104 proposals, a total of 26 could be identified that had a diagnostic test as topic. Approximately half of these tests (14/26) was used to obtain a diagnosis. In the majority of the other proposals (11/26) the intended use of the test
was to aid therapy decisions. In one proposal the test was used to monitor disease severity.

The hierarchical evaluation model could not be applied to 13 of the 26 (50%) decision problems, as no objective reference test was available for the tests in these proposals. In the subgroup of tests used for diagnosis this was the case in 4 out of 14 proposals. In the subgroup of tests used for therapy decisions, 8 out of 11 had no reference test. The objectives and availability of a reference test for the different test types are listed in Table 2.

<table>
<thead>
<tr>
<th>Type of test</th>
<th>N</th>
<th>Reference test available</th>
<th>Purpose</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Imaging</td>
<td>9</td>
<td>6</td>
<td>Diagnosis</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Therapy decision</td>
<td>5</td>
</tr>
<tr>
<td>Functional</td>
<td>4</td>
<td>0</td>
<td>Monitoring</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Therapy decision</td>
<td>3</td>
</tr>
<tr>
<td>Genetic</td>
<td>3</td>
<td>3</td>
<td>Diagnosis</td>
<td>3</td>
</tr>
<tr>
<td>Laboratory/Pathology</td>
<td>2</td>
<td>1</td>
<td>Diagnosis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Therapy decision</td>
<td>1</td>
</tr>
<tr>
<td>Microbiology</td>
<td>4</td>
<td>2</td>
<td>Diagnosis</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Therapy decision</td>
<td>2</td>
</tr>
<tr>
<td>Multiple</td>
<td>4</td>
<td>1</td>
<td>Diagnosis</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>13</td>
<td>Diagnosis</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Monitor</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Therapy decision</td>
<td>11</td>
</tr>
</tbody>
</table>

Problems with the current approach

All existing hierarchical evaluation models require proof of diagnostic accuracy as a prerequisite before assessing the affect of a diagnostic test on patient outcome. Our application to the proposals submitted to the Dutch MTA program demonstrates that a comparison of the results of a test with those of a reference test is not feasible for all decision problems on medical tests. For example, one of the proposals was an evaluation of human papillomavirus (HPV) type tests for women with cervical intracellular neoplastic lesions (CIN II-III). According to the authors, HPV type testing would distinguish high-risk lesions, with a large risk to develop into cancer, from low-risk lesions. Currently, all patients with CIN II-III lesions are treated by dissecting the lesion. The authors hypothesized that use of the test could reduce the number of treated lesions, if by only the high-risk patients are treated. They proposed a randomized controlled trial to compare the classic work-up, no testing and treating all patients, with a strategy in which therapy is guided by HPV type testing, which would be a phase 3 study. How would one assess the diagnostic accuracy of HPV type testing in these patients? What would be the appropriate reference test? These
Strategies for the evaluation of diagnostic technologies

questions are hard to answer as it considers a new test and a new classification for which currently no other tests are used.

The same problem can be illustrated with a proposal to evaluate the role of somatostatin receptor scintigraphy (Octreoscan) for the detection of metastases in the follow-up of patients treated for primary breast cancer. A pilot study had shown that in women with somatostatin positive tumors an Octreoscan could detect metastases before symptoms occurred or conventional follow-up imaging showed a lesion. The authors considered that the addition of the Octreoscan to the preoperative evaluation of women with primary breast cancer and subsequent follow-up with the scan of somatostatin receptor positive patients could improve survival. They proposed a study in which all patients would undergo a preoperative work-up with the Octreoscan whereupon somatostatin receptor positive patients would be randomized between the classic follow-up imaging strategy and a follow-up strategy including 6-monthly Octreoscans.

What would have been an appropriate phase 2 evaluation before one would want to fund this study? Some would argue that pathology could be used as a reference test in this situation. Besides the practical problem, that it will be difficult to obtain cytologic material in case other imaging procedures show no lesions, such a design will not show us the degree to which the scan can correctly identify metastases that will need treatment. In phase 2 evaluations, one should compare the findings from a test with those from a reference test currently used for the same target condition, guiding similar treatment decisions. A direct comparison with a reference test that detects patients who will benefit from Octreoscan follow-up is difficult, as no such tests are currently used to detect metastasis in such an early stage.

Another example in which the comparison with a reference standard was not feasible was a proposal to evaluate the use of a lipase breath test for the guidance of enzyme supplementation in patients with malabsorption due to cystic fibrosis. Currently the dosage of the enzyme supplementation is based on the clinical presentation of the patients. The authors considered the hypothesis that a dosage regime based on the activity of the lipase in the duodenum, as measured by the breath test, improves the weight gain of these patients. They proposed a randomized trial comparing the classic work-up and dosage regime with a lipase breath test guided regime. A phase 2 evaluation of the lipase test with as reference test clinical assessment, the currently used best available method, would be of limited value. One can expect these two tests to differ quite substantially and even suspect the lipase test to be a more objective, and hence, superior, method to guide treatment decisions.
Even if there is a reference test available there are situations in which establishing a high sensitivity and high specificity make less sense. Consider for example the use of D-dimer assays in the diagnosis of patients suspected of deep venous thrombosis, a topic of one of the other proposals. The classic work-up of these patients consists of serial compression ultrasonography. D-dimer assays measure the level of D-dimer, a fragment specific for the degradation of fibrin, in blood or plasma. In patients with normal D-dimer assays further diagnostic testing might be discontinued while in patients with abnormal D-dimer assays further objective testing is still needed. The goal of the D-dimer assays is therefore to refute deep venous thrombosis at referral and limit the amount of ultrasonographies. Optimal concordance with the reference test, serial compression ultrasonography or venography, should not be required. To be safe the D-dimer assay should have a very high negative predictive value; to be efficient, the negativity rate should be high enough to warrant its use in practice.

**Discussion**

In a hierarchical evaluation strategy, more elaborate and therefore more expensive types of studies should only be performed if sufficient evidence has been obtained in previous steps of the evaluation process. Multiple of such phased evaluation models have been proposed for diagnostic technologies. In most of them, evaluations of diagnostic accuracy - in terms of the correspondence of the findings with the new test with those obtained with the reference test - precede evaluations of patient outcome. The standard design to measure diagnostic accuracy consists of a blind comparison of an index test with the reference test in a clinical population.\(^{22}\)

\(^{23}\)  

In the "classical" situation, the new test is proposed as a substitute for the existing test. If applying the test is less costly (for example, because it is non-invasive) and the two tests are perfectly concordant, then patient outcome will remain equal if the old test is substituted for the new one. As subsequent decisions on the management of patients, including the decision to start, stop or withhold treatment, rely on test outcomes, substituting the source of these test results can never alter the decisions whenever the tests are perfectly concordant. In case of less than perfect agreement, the impact of the difference in management on patients with discordant test results – testing positive on one test but negative on another – will determine whether or not patient outcome is affected.

A quite different situation occurs if researchers feel that the new test is better than the current reference test, leading to an improvement in patient outcome. In those situations, discordance with the reference test is even a necessity, and less
than perfect accuracy is to be expected, as was the case with the lipase breath test example.

Substitution of the existing reference tests is not the only possible decision problem. Several other decision problems can also be distinguished. One of them is the addition of the new test to the diagnostic pathway. In that situation, at the end of the diagnostic pathway, all patients are treated similarly. With the new test, a distinction will be made, depending on the test’s results: a subset of patients, say test negative patients, will receive the same treatment, whereas the others will receive a different kind of treatment. Examples of this problem were the Octreoscan and HPV type testing proposals.

A second type of decision problem is triage. In this case, the test under evaluation is to be applied in patients before the reference test is ordered, where the latter can be omitted for certain categories of test results on the former. This makes sense if the new test is less invasive, more readily available, or otherwise less costly for one or more parties than the reference test, as was the case in the D-dimer assay example. Here also optimal concordance is not required. In these situations, one requires a perfect predictive value for the test categories that will lead to not ordering the second, existing test.

The different types of decision problems are outlined in figure 1. It goes without saying that we can also devise comparable decision problems where doubts have arisen as to the appropriateness of tests that already belong to the diagnostic pathway.

These examples show that correspondence of test results with particular categories of patient outcomes is the required property of tests, not so much the correspondence with “the truth”. For all practical decision problems, the improvement in patient outcome, or maintaining outcome while lowering costs, is the driving force.

Can a phased evaluation model still be used if we have to discard “diagnostic accuracy” in the classical sense? We think it can. One solution is to broaden the concept of “reference test” from best available test to detect a target condition to any method for categorizing patients that are similar with respect to management-outcome combinations. A better term for the latter might be “reference standard” rather than reference test. Phase 2 studies measure the level of agreement between the test under evaluation and the reference standard. Instead of measuring “diagnostic accuracy”, in terms of correspondence with a truth, these studies aim to measure “the possible effects on outcome” of the test studied. These studies will be either observational, documenting concordance at the same point in time (such as with substitution problems) or longitudinal, establishing the relation between test results and outcome after different forms of treatment (as in the addition problems).
Three of the phased evaluations models that we were able to retrieve from the literature already suggested the use of clinical follow-up as possible reference standard.\textsuperscript{5, 12, 15} However, they did so more in the sense as surrogate measure in the absence of a "gold standard" reference test. We argue that one should stop the hunt for the best available reference test and start considering a problem-oriented approach towards the evaluation of diagnostic tests.

**Figure 1.** Substitution, triage and addition
Strategies for the evaluation of diagnostic technologies

References


