Medical Tests—White Paper Series

Proposals for a Phased Evaluation of Medical Tests

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of the Agency for Healthcare Research and Quality (AHRQ). Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

This report has been published in edited form: Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009 Sept-Oct;29(5): E13-E21. Epub 2009 Jul 15.

Suggested citation: Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a phased evaluation of medical tests. Medical Tests—White Paper Series. Agency for Healthcare Research and Quality. Rockville: MD. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=350

Proposals for a Phased Evaluation of Medical Tests

Authors:

Jeroen G. Lijmer, M.D., Ph.D.[a]
Mariska Leeflang, Ph.D.[a]
Patrick M.M. Bossuyt, Ph.D.[a]

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

Proposals for a Phased Evaluation of Medical Tests

Abstract

Background—In drug development, a four-phase hierarchical model for the clinical evaluation of new pharmaceuticals is well known. Several comparable phased evaluation schemes have been proposed for medical tests.

Purpose—To perform a systematic search of the literature, a synthesis, and a critical review of phased evaluation schemes for medical tests.

Data Sources—Literature databases: Medline, Web of Science, and Embase.

Study Selection and Data Extraction—Two authors separately evaluated potentially eligible papers and independently extracted data.

Data Synthesis—We identified 19 schemes published between 1978 and 2007. Despite their variability, these models show substantial similarity. Common phases are evaluations of technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal aspects.

Conclusions—The evaluation frameworks can be useful to distinguish between study types, but they cannot be seen as a necessary sequence of evaluations. The evaluation of tests is most likely not a linear but a cyclic and repetitive process.

Introduction

Over the last few decades many new medical tests have been developed, and the number of available options is still increasing. Premature dissemination of testing technologies can lead to erroneous diagnoses and preventable delays in starting appropriate treatment or, alternatively, to the initiation of unwarranted, sometimes dangerous therapy. Examples have been the dexamethason suppression test for depression, the carcinoembryonic antigen for colon cancer, and 125I-fibrinogen leg scan for the diagnosis of deep venous thrombosis.[1,2] In addition, the increasing costs of health care have put pressure on available budgets, calling for the elimination of ineffective medical technology. These are ample reasons why new medical tests should be thoroughly evaluated before they are introduced in clinical practice.

The ultimate benefit of any medical technology should be expressed in terms of its effects on health outcome, and tests are no exception.[3] Yet the evaluation of technology can be a time-consuming and costly process. An efficient use of resources calls for a well-planned evaluation strategy. In such a strategy, more elaborate and therefore more expensive forms of evaluation should only be performed if sufficient evidence has been obtained in 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.

In drug development, a four- or five-phase hierarchical model for the clinical evaluation of new products is well known. Phase 0 studies are exploratory first-in-human trials to evaluate whether the drug or agent behaves in human subjects as was expected from preclinical studies. In Phase I, the safety, tolerability and toxicity, pharmacodynamics, and pharmacokinetics of the new drug are assessed. Phase II usually consists of small-scale clinical investigations to obtain an initial estimate of the effect of treatment. If the treatment effect is too small, further evaluation will be discontinued. In Phase III, the effectiveness of the drug is assessed by measuring patient outcome in randomized clinical trials. If the drug is effective, further surveillance after introduction to the market is necessary. In Phase IV, the long-term effects and side effects can be registered.

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 certain conditions be fulfilled in each phase before the evaluation can continue with the subsequent phase. Several of these proposals are closely related to hierarchies of evidence. One of the best known are the levels of efficacy for imaging tests, proposed by Fryback and Thornbury in 1991.[4]

Several more proposals have appeared since then. It is unclear to what extent these models differ and in what elements they differ. We have performed a systematic search of the literature for phased or hierarchical models for the evaluation of medical tests. We present our findings, a synthesis of existing models, and a critical commentary.

Literature Search

Eligible for this review were papers that described a proposal for the phased evaluation of medical tests, from the first technological laboratory-based evaluation to the evaluation of the performance of the test in clinical practice. Studies that described only parts of this process and studies that advocated a less linear approach were also included in our review.

Papers describing hierarchical models for the evaluation of diagnostic tests use different words and descriptions for these models in their titles and abstracts. In general, these studies are

not indexed in a consistent way in electronic bibliographic databases. We first searched in Medline, Web of Science, and Embase for studies with the words "(phased approach[tiab] OR hierarchical model[tiab] OR phased evaluation[tiab] OR hierarchical approach[tiab] OR hierarchical evaluation[tiab]) AND (diagnosis[tw] OR diagnostic[tw] OR diagnosis[MeSH]) (239 hits – January 2009)."

The databases mentioned were then searched for similar or related articles and for articles that cited the included papers. We also manually checked the reference lists of identified papers. When a paper only made a reference to a previous proposal for a hierarchical evaluation without further modification, it was excluded.

Models for the Phased Evaluation of Tests

We identified 31 papers with a model for the phased evaluation of diagnostic tests. Two of these were based on a model previously proposed by Guyatt and colleagues. Two others referred to a Fineberg model, one was based on the Sackett and Haynes model, and seven papers referred to Fryback and Thornbury. In total, 19 different models were found. The first of these was published in 1978; the most recent paper appeared in 2007.

The oldest references we could identify appeared in a special issue of the *American Journal of Roentgenology* (*AJR*) on the evaluation of computed tomography. When computed tomography was widely adopted in the United States in the early 1970s, it became the focus of much debate on the evaluation of diagnostic imaging and new health technologies in general. In an editorial, Fineberg noted, "One of the difficulties in evaluating a diagnostic test is its remoteness from health outcome." Yet "the ultimate value of the diagnostic test is that difference in health outcome resulting from the test: In what ways, to what extent, with what frequency, in which patients is health outcome improved because of this test?"[5] Loop and Lusted reported how the American College of Radiology (ACR) had tried to deal with the problems of evaluating the health consequences of testing. The ACR had established an Efficacy Studies Committee in 1972, chaired by Lee B. Lusted. That committee decided: "The fullest and most long-range expression of efficacy ought to include some measure of the influence of the examination on the final outcome of the episode of ill health."[6] The committee distinguished between diagnostic efficacy (E-1), the change in the probability of diagnosis after radiographic results have become available; therapeutic efficacy (E-2), the change in therapy planning; and outcome efficacy (E-3): Was the patient better off as a result of the procedure having been performed?

Building on this model, Fryback and Thornbury developed their framework, which appeared in 1991 in a Lusted memorial issue of *Medical Decision Making*.[4] Both authors have described the framework in more detail in later publications.[7,8] Theirs is a six-tiered hierarchical model, which extends from the physics of imaging, through clinical use in decisions about diagnosis and treatment, to patient outcome and societal issues. Demonstration of efficacy at each lower level in this hierarchy, they wrote, is logically necessary but not sufficient to assure efficacy at higher levels. Kent and Larson used almost the same levels in discussing the efficacy of magnetic resonance imaging but added two other dimensions: the spectrum of diseases and the quality of research.[9] Another modification of the ACR framework was proposed by Mackenzie and Dixon.[10] Phelps and Mushlin combined medical decision theory and epidemiologic information in suggesting two hurdles for diagnostic technologies, linking the accuracy level with the societal level.[11]

Silverstein and colleagues translated the ACR approach to laboratory medicine, and Pearl applied it to tests in general.[12,13] The related ACCE framework for the evaluation of genetic tests

is a model process for evaluating data on emerging genetic tests. It derives its name from the four components: analytic validity; clinical validity; clinical utility; and ethical, legal, and social implications.[14]

Several others have translated the ACR levels of efficacy into phases of evaluation. In 1978, Freedman classified designs to evaluate and compare imaging techniques and observed a parallel with the standard classification of clinical trials.[15] Studies of diagnostic accuracy, he wrote, are analogous to Phase II trials, whereas studies evaluating the contribution to clinical management correspond to the Phase III category. The majority of studies he observed at the time were Phase II type accuracy studies, and more emphasis on Phase III studies was required. In a similar way, Taylor and colleagues classified 200 studies published in the *AJR* and in *Radiology* in 1988 and 1989 into one of five phases.[16] They found that the majority of studies focused on early technical assessment.

Guyatt and his colleagues from McMaster University also extended the ACR framework into a proposal for stepwise clinical evaluation of diagnostic technologies.[17] Diagnostic technology assessment should begin by establishing the capability of the technology under ideal or laboratory conditions, followed by an exploration of the range of possible uses and the accuracy of the test. Their proposal also contains a very strong plea for randomized clinical trials of test strategies and a critical discussion of some of the poorer study designs. Van der Schouw and colleagues and van den Bruel and colleagues similarly suggested stepwise evaluations of tests.[18,19]

Kobberling and colleagues proposed a four-phased model for test evaluation, explicitly emphasizing the similarity with the evaluation of therapeutic methods.[20] In 2000, Houn and her colleagues from the U.S. Food and Drug Administration (FDA) noticed a similarity in the evaluation of breast imaging technology and the FDA's phased approach in the clinical development of drugs and biologic products.[21] Phase I refers to the initial evaluation of a developing technology in human populations. Phase II refers to clinical studies involving limited numbers of human subjects to gather preliminary evidence regarding effectiveness and additional safety data. Phase III refers to controlled clinical studies intended to provide a reasonable assurance of safety and effectiveness in defined populations. Finally, Phase IV refers to studies performed once a technology has gained marketing approval; these studies address long-term safety and better characterize the performance of the technology within a larger population. In an accompanying editorial, Gatsonis introduced a paradigmatic matrix for the evaluation of imaging technology, with four phases and three possible endpoints for studies.[22] The four phases correspond to what he called the developmental age of the modality, starting from discovery and then moving to introduction, maturity, and dissemination. In the early phases the focus is on diagnostic performance, whereas later phases would focus on impact on the process of care and patient outcome.

While schemes inspired by the proposals by Lusted, Fineberg, and Guyatt made a distinction between accuracy, diagnostic impact, and therapeutic impact, other authors have proposed multiphase models for the evaluation of accuracy in itself. Zweig and Robertson suggested the label "Phase I Trial" for studies of the analytical precision, accuracy, sensitivity, and specificity of a laboratory test, while "Phase II Trials" would refer to studies determining the usual range of results encountered in healthy subjects or comparing the results obtained in various disease states with this usual range.[23] A prospective diagnostic trial of the actual clinical usefulness of a test in a realistic clinical setting would then be termed a "Phase III Trial." Multiple phases in the evaluation of accuracy have also been proposed by Sackett and Haynes,[24]

Pepe,[25] and Taube, Jacobson, and Lively.[26] Elsewhere, Obuchowski discussed how the questions and the number of readers should vary with a phased evaluation of imaging.[27]

Hierarchical Models: A Synthesis

In Table 1 we have summarized the levels and phases described by the 19 different models. Each model consists of four to seven different elements, with marked similarities between these proposals. Most models start with a Phase I, which consists of test development. During this phase the test has to meet prespecified technical requirements. Aspects that have to be documented in this phase include feasibility, required equipment and personnel, physical and biochemical parameters specific to the test, such as the minimal detection level, circadian fluctuation, resolution, contrast level, and reproducibility. Guyatt and colleagues recommended that, in addition, the test be applied to a large number of diverse conditions in order to delineate its possible uses.[17]

In most models, the diagnostic accuracy of the test is assessed in one or more subsequent phases. The results of the test under evaluation are compared to those from a reference standard in order to establish how well the test is able to identify patients with the target condition. 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 subphases at this phase. They propose to evaluate the diagnostic accuracy first in a group of subjects with the disease of interest and a group of healthy 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 study group that closely resembles the population of patients for which the test is intended. In addition, some authors suggest comparing the diagnostic accuracy of the test with the performance of other tests intended to detect the same target condition before proceeding further.

Most proposals continue with the evaluation of the clinical effectiveness of the test, assessed in terms of its effect on diagnostic thinking and patient management, therapeutic efficacy, and patient outcome. To investigate diagnostic-thinking efficacy, Fryback and Thornbury suggested studies to document the percentage of cases in which an image was judged "helpful" to making the diagnosis or to summarize the difference in clinicians' subjectively estimated diagnosis probabilities before and after receipt of test information.[4]

Studies of therapeutic efficacy should then establish the percentage of cases where images were judged helpful in planning the management of patients, the percentage of cases where a medical procedure could be avoided because of imaging findings, the number of times therapy planned before imaging changed after imaging information was obtained, or the percentage of cases in which clinicians' prospectively stated therapeutic choices changed after test information was obtained.

Evaluations in terms of patient outcome can be found in all of the retrieved models except the one by Taube et al.[26] This can be documented in randomized clinical trials, in which specific test-treatment combinations are compared. A decision analysis comparing different diagnostic strategies may provide an investigative alternative.

A subset of authors have described a last phase, beyond the assessment of clinical effectiveness, in which cost-effectiveness and other societal effects are studied. Freedman suggested studies to monitor changes in clinical practice after the introduction of a new test.[15] In such studies, changes in diagnostic use and the frequency of test 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.[4,13,16,18,23] 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.

Discussion

In a phased evaluation strategy, more elaborate and therefore more expensive types of studies are performed only if sufficient evidence has been obtained in previous steps of the evaluation process. In this review, we identified 31 proposals for a hierarchical model of evidence or a phased evaluation scheme for medical tests. We are aware that our review has its limitations, as we only searched papers in journals and did not look systematically for proposals described in books only. Because of poor indexing, we may not have been able to identify all existing schemes.

The variety in proposals may come as a surprise to those who are familiar with the four or five phases in drug development. Why have the phases in the clinical evaluation of drugs become so well engrained in our thinking, and why is there more variability in evaluations of tests? One of the reasons for this difference may be the absence of a strong regulatory framework. There are no clear international standards, and there is little agreement on what evidence is required in decisions about tests or by whom it is required.[28,29] Several authors have called for harmonization of regulatory standards internationally and for more transparency regarding the clinical evidence base for new tests. If this happens, a more standardized model may be developed in the process.

Most proposals are built on the chain of steps linking tests and outcome, and they can be traced back to the set of levels of efficacy identified for imaging in the 1970s. We present below a few critical thoughts on their use as phases in the assessment of tests.

Diagnostic Accuracy

Diagnostic accuracy plays a central role in most proposals. Unfortunately, the diagnostic accuracy literature suffers from poor study design, small study samples without power calculations, and suboptimal reporting.[30-33]

Design, conduct, and reporting can and should be improved.[34] Most accuracy studies focus on the test in isolation, although tests are never used in a vacuum. A number of prototypical roles of tests relative to existing ones can be distinguished: replacement, triage, or add-on.[35]

Several authors have questioned the central role of test accuracy in test evaluations.[36,37] Hunink and Krestin argued that results from accuracy studies are often too late to influence management and policy decisions, given the current rapid advances in technology.[38] Accuracy may be sufficient in providing evidence of improvement or equivalence in patient outcomes if there is a well-defined target condition linked to effective downstream management consequences, such as effective treatment.[39-41] Yet the pivotal position of the accuracy paradigm in the schemes identified in this review is somewhat problematic, especially whenever a new test leads to a classification in disease for which there is no clinical reference standard or when the new test is thought to be better than the current reference standard. Strategies exist to deal with cases in which the reference standard result is missing in some patients or when information can be used to build a substitute or proxy for the reference standard, but when there is no accepted reference standard, other approaches have to be used.[42]

There are other problems with a central position for diagnostic accuracy. A wide range of tests are not used for diagnosis but for other purposes, such as prognosis, prediction of treatment response, selecting therapy, or monitoring the course of disease or the effects of treatment. In these situations, there is not always a reference standard available, nor is it clear how the target condition should be defined.

Diagnostic-Thinking Efficacy

Because diagnostic tests are often remote from health outcome, in the short term, researchers rely on more proximate efficacy measures, such as the test's effect on clinical thinking. But studies of diagnostic-thinking efficacy or therapeutic efficacy are difficult to mount. At the University of Michigan in 1972 and 1973, a group of researchers tried to measure diagnostic thinking to support the work of the ACR Efficacy Committee mentioned previously. The team collected referring physicians' diagnosis prior to and after urography and their certainty in relation to receipt of the radiologic information. The change in these estimates was then transformed to log likelihood ratios.[43] The original intention was to measure the degree to which clinical management was influenced by the intravenous urogram. Unfortunately, clinicians balked at the prospect of formulating a treatment plan for a patient with, say, hematunia, who had not had a urographic contrast study.[5] Consequently, the American College of Radiology Efficacy Committee deferred all attempts to measure thinking efficacy.

Even if they could be done, are such studies also necessary? The ultimate question in decisions about testing is how much net gain from testing there will be for the patient in terms of improved treatment decisions and better health outcome.[44] Despite improvements in the methodology for measuring physician confidence, one can seriously question the validity of such studies as substitutes for improvement in patient outcome. In general, their object of study is clinician behavior, not patient outcome. A negative result in a judgment and decisionmaking study tells us something about the included physicians, and not necessarily a great deal about the qualities of the test itself or its potential for improving health outcome. When clinicians do not adjust pretest probabilities or change a management plan, we should not necessarily conclude that their failure to do so was correct. Alternatively, a confident adjustment of the probability of disease or the management plan after testing does not necessarily imply that patients are better off. Guyatt pointed out that clinicians differ systematically in their assessment of whether a given test result contributed to management, that it may be difficult to consistently be aware of clinicians' plans before the test results are available, and that clinicians' reports of what they would do before the test result is available may differ from what they actually would have done were the technology not available.[17]

This does not imply that there is no relevance at all in studying clinicians' judgment and decision-making, as patient outcome after testing will usually depend on the behavior and actions of one or more physicians. If one finds that a test does not improve patient outcome, it may be important to know that the ineffective link in the testing process is a modifiable behavior of the physician with regard to the test.

Randomized Trials

If the net gain from testing has to be expressed in terms of changes in patient outcome, one could consider jumping immediately to randomized clinical trials with patient-centered outcome measures, as Guyatt proposed.[3] Running randomized trials of tests and collecting

evidence of improved patient outcome after testing have almost become synonyms in many of the proposals. Is that justified?

Randomized trials of tests are more difficult to design than randomized studies of treatment. The benefits from testing may be limited to a subset of those tested, so sample size requirements can be substantial.[45] Trials of testing need a well-defined protocol that links testing, results, and downstream decisions. It is inevitable that such trials evaluate the effectiveness of testing as well as that of downstream management. These protocols may not always mimic the way the test will ultimately be used in practice, and physician compliance with such protocols may be difficult, limiting the external validity of the trial results. All of these practical problems are challenging but not insurmountable, and trials of testing can be found in the literature.

Evidence of an improvement in health is stronger than documented accuracy, but one may not always need to conduct a randomized trial to document the benefits of testing on patient outcome. Under specific circumstances, smaller scale studies of accuracy can suffice; or noncomparative studies of testing and test combinations or modeling may also suffice.[39] Elsewhere in this White Papers series, Lord and colleagues offer a more complete discussion of alternatives to randomized trials of testing.[41]

A Stepwise Approach?

The four phases in the development of drugs and devices have shown their merit. One proceeds to more costly or more risky evaluations only if there is enough evidence from previous phases. Trials in humans take place only after drugs have been tested thoroughly in the laboratory in animal studies. Trials in volunteers usually go before trials in patients. Can a similarly staged model be used for the evaluation of medical tests? In the early evaluation of new markers, a phased approach definitely makes sense. In the models proposed by Pepe, Sackett, and others, the first evaluations of a marker's accuracy are designed for selected subgroups, limited in size; only when enough evidence is gathered does one move to the more costly clinical evaluations. Should one also move cautiously through the other elements of the efficacy hierarchy, one level at a time? We do not think so. Accuracy studies are neither sufficient nor always necessary for showing improvement in patient outcomes from testing. Evaluations of physicians' judgments or their behavior are not necessary, nor can they be used as a satisfactory substitute for patient outcome.

In all fairness, the ACR committee distinguished between higher and lower levels of efficacy; they did not propose a phased evaluation of tests. Neither did Fryback and Thornbury, although their hierarchy has often been interpreted that way.[38] We do not think this is justified. The levels of efficacy should not be equated with a necessary succession of phases in the evaluation of tests, nor should they be connected with a hierarchy in study design.

More recent proposals for grading recommendations about testing, such as the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach, no longer refer to levels of evidence, but distinguish grading the quality of evidence, where study design obviously matters, from ranking levels of strength for recommendations.[3] The U.S. Preventive Services Task Force, for example, used to correlate its recommendations strongly with the research design of the most important studies, whereas currently it considers the evidence as a whole, using eight steps in an analytic framework, a causal pathway linking screening or other preventive services to health outcomes.[46]

Houn and colleagues recognized that the four-phased model for drug development is often thought of as a linear process from idea inception to product marketing, research, and

development.[21] They describe how it is actually a cyclic, repetitive process that begins with the recognition of a problem and continues through an expansive thinking phase to experimentation, assessment, and adoption. This process may be repeated as the technology is improved or modified for new uses. The process also cycles and moves "up the rungs" from laboratory to applied research and, ultimately, to clinical application, and it sometimes slips back to address unanticipated problems and then advances again as those problems are resolved. Similarly, Hunink and Krestin described the linear approach as a reflection of the philosophy prevalent in the industrial period. They felt that an interwoven circular approach for the evaluation of imaging, with concurrent development, assessment, and implementation of technology, would be more appropriate.[38] The same can be said for the evaluation of tests in general.

A classification of study types and outcomes has descriptive merit in understanding the published research and the gaps in knowledge. There is also value in thoughtful considerations of the quality of the available evidence when making decisions about large-scale evaluations of testing, requiring big budgets and large numbers of participants. Yet translating levels of efficacy into a linear series of phases in evaluating tests will ultimately prove to be too restrictive, and may fail to do justice to the myriad of tests and the wide range of testing purposes.

Table 1. Summary of proposals for the phased evaluation of medical tests

	Reference[a]																		
	Loo 1978	Zwe 1982	Guy 1986	Fre 1987	Kob 1990	Fry 1991	Ken 1992	Tay 1993	Sil 1994	Sch 1995	Mac 1995	Pea 1999	Hou 2000	Gat 2000	Sac 2002	Had 2003	Pep 2005	Tau 2005	Bru 2007
	6	23	17	15	20	4	9	16	12	18	10	13	21	22	24	14	25	26	19
Levels/Phases																			
Technical efficacy		1	1	1	1	1	1	1-3		1	1	1	1	1		1	1-3	1-2	1
Intended use			2															3	2
Diagnostic accuracy			3	2	2	2	2	4	1		2	2	2			2	4	4-6	3
Usual range/ subgroups		2			3					2				1	1-2				
Clinical population		3			4					3				2	3			7	
Diagnostic-thinking efficacy	1		4			3	3		2	4	3	3							
Therapeutic efficacy	2		5			4	4		3		4	4	3	3 -4					
Patient outcome efficacy	3		6	3	5	5	5	5	4		5	5	4	3-4	4	3	5		4
Societal efficacy				4		6				5		6	5			4			5

[a]For each reference, we show the first three letters of the first author's name, the publication date, and its number in the list of references at the end of this paper.

References

1. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. JAMA 1988;259:1699-1702.
2. Lensing AW, Hirsh J. 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. Thrombosis & Haemostasis 1993;69:2-7.
3. Schünemann H, Oxman A, Brozek J, et al. Rating quality of evidence and strength of recommendations: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies BMJ 2008;336:1106-1110.
4. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making 1991;11:88-94.
5. Fineberg HV. Evaluation of computed tomography: achievement and challenge. AJR Am J Roentgenol 1978;131:1-4.
6. Loop JW, Lusted LE. American College of Radiology Diagnostic Efficacy Studies. AJR Am J Roentgenol 1978;131:173-179.
7. Thornbury JR, Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol 1994;162:1-8.
8. Thornbury JR, Fryback DG. Technology assessment--an American view. . European Journal of Radiology 1992;14:147-156.
9. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Investigative Radiology 1992;27:245-254.
10. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clin Radiol 1995;50:513-518.
11. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Medical Decision Making 1988;8:279-89.
12. Silverstein MD, Boland BJ. Conceptual framework for evaluating laboratory tests: case-finding in ambulatory patients. Clin Chem 1994;40:1621-1627.
13. Pearl WS. A hierarchical outcomes approach to test assessment. Ann Emerg Med 1999;33:77-84.
14. Haddow JE, Palomaki GE. ACCE: A Model Process for Evaluating Data on Emerging Genetic Tests. In: Khoury M, Little J, Burke W, eds. Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Oxford: Oxford University Press; 2003:217-233.
15. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071-1081.
16. Taylor CR, Elmore JG, Sun K, et al. Technology assessment in diagnostic imaging. A proposal for a phased approach to evaluating radiology research. Invest Radiol 1993;28:155-161.
17. Guyatt GH, Tugwell PX, Feeny DH, et al. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-594.
18. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995;30:334-340.
19. Van den Bruel A, Cleemput I, Aertgeerts B, et al. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol 2007;60:1116-1122.
20. Working Group for Methods for Prognosis and Decision Making. Kobberling J, Trampisch HJ, Windeler J, eds. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990;28:873-879.
21. Houn F, Bright RA, Bushar HF, et al. Study design in the evaluation of breast cancer imaging technologies. Acad Radiol 2000;7:684-692.

22. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. Acad Radiol 2000;7:681-683.

23. Zweig MH, Robertson EA. Why we need better test evaluations. Clin Chem 1982;28:1272-1276.

24. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539-541.

25. Pepe MS. Evaluating technologies for classification and prediction in medicine. Stat Med 2005;24:3687-3696.

26. Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. Am J Pharmacogenomics 2005;5:357-364.

27. Obuchowski NA. How many observers are needed in clinical studies of medical imaging? AJR Am J Roentgenol 2004;182:867-869.

28. Walley T. Evaluating laboratory diagnostic tests. BMJ 2008;336:569-570.

29. Price CP, Christenson RH. Evaluating new diagnostic technologies: perspectives in the UK and US. Clin Chem 2008;54:1421-1423.

30. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-1066.

31. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

32. Smidt N, Rutjes AW, van der Windt DA, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347-353.

33. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006;332:1127-1129.

34. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-W12.

35. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-1092.

36. Feinstein AR. Misguided efforts and future challenges for research on “diagnostic tests.” J Epidemiol Community Health 2002;56:330-332.

37. Moons KG, van Es GA, Michel BC, et al. Redundancy of single diagnostic test evaluation. Epidemiology 1999;10:276-281.

38. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002;222:604-614.

39. 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;144:850-855.

40. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45:189-195.

41. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29:E1-E12.

42. Rutjes AW, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:iii, ix-51.

43. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology 1975;114:561-565.

44. Fineberg HV. Computerized tomography: dilemma of health care technology. Pediatrics 1977;59:147-9.

45. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-1847.

46. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20:21-35.