

# Challenges in Systematic Reviews of Diagnostic Technologies

Athina Tatsioni, MD; Deborah A. Zarin, MD; Naomi Aronson, PhD; David J. Samson, BA; Carole R. Flamm, MD, MPH; Christopher Schmid, PhD; and Joseph Lau, MD

Diagnostic tests are critical components of effective health care. They help determine treatments that are most beneficial for a given patient. Their assessment is a complex process that includes such challenges as a dearth of studies that evaluate clinical outcomes and lack of data on use of the test in realistic clinical settings. The methodologic quality of studies of diagnostic tests also lags behind the quality of studies of therapeutic interventions. Statistical methods to combine diagnostic accuracy data are more complex and not as well developed, leading to difficulties in the interpretation of results. The Agency for Healthcare Research and Quality Technology Assessment Program has adopted a 6-level framework for evaluating diagnostic technologies. The

model emphasizes the need for systematic reviews of diagnostic test studies to go beyond the assessment of technical feasibility and accuracy to examine the impact of the test on health outcomes. In this paper, we use examples from 3 Evidence-based Practice Center reports to illustrate 3 challenges reviewers may face when reviewing diagnostic test literature: finding relevant studies, assessing methodologic quality of diagnostic accuracy studies, and synthesizing studies that evaluate tests in different patient populations or use different outcomes.

*Ann Intern Med.* 2005;142:1048-1055.  
For author affiliations, see end of text.

www.annals.org

Diagnostic tests, broadly construed, consist of any method of gathering information that may change a clinician's belief about the probability that a patient has a particular condition. Diagnosis is not an end in itself; rather, the purpose of a diagnostic test is to guide patient management decisions and thus improve patient outcomes. Because they are pivotal to health care decision making, diagnostic tests should be evaluated as rigorously as therapeutic interventions.

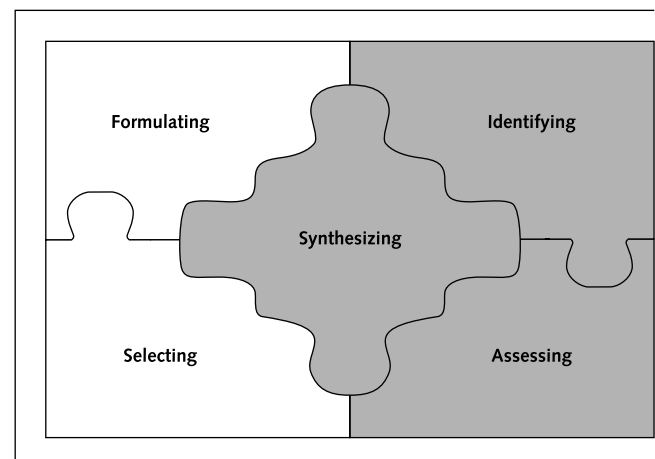
A cursory search of the literature for a diagnostic technology may reveal many articles dealing with various aspects of the test. But rarely do these include reports of trials to assess the outcomes of using the test to guide patient management. In the mid-1970s, several groups (1–4) developed a now widely adopted framework to evaluate diagnostic technologies by categorizing studies into 6 levels (5). This framework is hierarchal: Level 1 consists of studies that address technical feasibility, and level 6 consists of those that address societal impact. Table 1 summarizes the framework and the key questions addressed by studies in each level.

Evidence-based Practice Centers (EPCs) have produced several evidence reports and technology assessments of diagnostic technologies ([www.ahrq.gov/clinic/techix.htm](http://www.ahrq.gov/clinic/techix.htm)). This article uses 3 reports produced by the EPCs to illustrate the challenges involved in evaluating diagnostic technologies. The first assessed the use of magnetic resonance spectroscopy (MRS) to evaluate and manage brain mass. It exemplifies the challenges of identifying relevant studies and assessing the methodologic quality of diagnostic accuracy studies (6). The second, a report on technologies to diagnose acute cardiac ischemia, illustrates the problem of synthesizing studies that assess tests in different patient populations and report different outcomes (7). In particular, this report highlights the challenges in quantitatively combining data on test accuracy. The third report, on positron emission tomography (PET) for diagnosing and managing Alzheimer disease and dementia, exemplifies

the challenges of assessing the societal impact of a diagnostic test (8). Finally, we discuss the problem of publication bias, which may slant the conclusions of a systematic review and meta-analysis in a biased direction.

## CHALLENGE: IDENTIFYING RELEVANT PUBLISHED AND UNPUBLISHED STUDIES

A report that assessed the value of MRS to diagnose and manage patients with space-occupying brain tumors demonstrates that there are few higher-level diagnostic test studies (8). Table 1 shows the number of studies and patients available for systematic review at each of the 6 levels of evaluation. Among the 97 studies that met the inclusion criteria, 85 were level 1 studies that addressed technical feasibility and optimization. In contrast, only 8 level 2 studies evaluated the ability of MRS to differentiate tumors from nontumors, assign tumor grades, and detect intracranial cystic lesions or assessed the incremental value of MRS added to magnetic resonance imaging (MRI). These indications were sufficiently different that the studies could not be combined or compared. Three studies provided evi-



**Table 1. Hierarchy of Diagnostic Evaluation and the Number of Studies Available for Different Levels of Diagnostic Test in a Technology Assessment of Magnetic Resonance Spectroscopy for Brain Tumors\***

Level	Description	Examples of Study Purpose or Measures	Studies Available, <i>n</i>	Patients, <i>n</i>
1	Technical feasibility and optimization	Ability to produce consistent spectra	85	2434
2	Diagnostic accuracy	Sensitivity and specificity	8	461
3	Diagnostic thinking impact	Percentage of times clinicians' subjective assessment of diagnostic probabilities changed after the test	2	32
4	Therapeutic choice impact	Percentage of times therapy planned before MRS changed after the test	2	105
5	Patient outcome impact	Percentage of patients who improved with MRS diagnosis compared with those without MRS (e.g., survival, quality of life)	0	0
6	Societal impact	Cost-effectiveness analysis (e.g., use to detect tumor in asymptomatic population)	0	0

\* MRS = magnetic resonance spectroscopy.

dence that assessed impact on diagnostic thinking (level 3) or therapeutic choice (level 4). No studies assessed patient outcomes or societal impact (levels 5 and 6).

The case of MRS for use in diagnosis and management of brain tumors illustrates a threshold problem in systematic review of diagnostic technologies: the availability of studies providing at least level 2 evidence (since diagnostic accuracy studies are the minimum level relevant to assessing the outcomes of using the test to guide patient management). Although direct evidence is preferred, robust diagnostic accuracy studies can be used to create a causal chain for linking these studies with evidence on treatment effectiveness, thereby allowing an estimate of the effect on outcomes. The example of PET for Alzheimer disease, described later in this article, shows how decision analysis models can quantify outcomes to be expected from use of a diagnostic technology to manage treatment.

The reliability of a systematic review hinges on the completeness of information used in the assessment. Identifying all relevant data poses another challenge. The Hedges Team at McMaster University developed and tested special MEDLINE search strategies that retrieved up to 99% of scientifically strong studies of diagnostic tests (9). Although these search strategies are useful, they do not identify grey literature publications, which by their nature are not easily accessible. The Grey Literature Report is the first step in the initiative of New York Academy of Medicine ([www.nyam.org/library/grey.shtml](http://www.nyam.org/library/grey.shtml)) to collect grey literature items, which may include theses, conference proceedings, technical specifications and standards, noncommercial translations, bibliographies, technical and commercial documentation, and official documents not published commercially (10).

Diagnostic studies with poor test performance results that are not published may lead to exaggerated estimates of a test's true sensitivity and specificity in a systematic re-

view. Because there are typically few studies in the categories of clinical impact, unpublished studies showing no benefit by the use of a diagnostic test have even greater potential to cause bias during a review of evidence. Of note, the problem of publication bias in randomized, controlled trials has been extensively studied, and several visual and statistical methods have been proposed to detect and correct for unpublished studies (11). Funnel plots, which assume symmetrical scattering of studies around a common estimate, are popular for assessing publication bias in randomized, controlled trials. However, the appearance of the shape of the funnel plot has been shown to depend on the choices of weight and metric (12). Without adequate empirical assessments, funnel plots are being used in systematic reviews of diagnostic tests. However, their use and interpretation should be viewed with caution. The validity of using a funnel plot to detect publication bias remains uncertain. Statistical models to detect and correct for publication bias of randomized trials also have limitations (13). One solution to the problem of publication bias is the mandatory registration of all clinical trials before patient enrollment; for therapeutic trials, considerable progress has already been made in this area. Such a clinical trials registry could readily apply to studies of the clinical outcomes of diagnostic tests (14).

### CHALLENGE: ASSESSING METHODOLOGIC QUALITY

Diagnostic test evaluations often have methodologic weaknesses (15–17). Of the 8 diagnostic accuracy studies of MRS, half had small sample sizes. Of the larger studies, all had limitations related to patient selection or potential for observer bias. Methodologic quality of a study has been defined as “the extent to which all aspects of a study's design and conduct can be shown to protect against sys-

tematic bias, nonsystematic bias that may arise in poorly performed studies, and inferential error” (18). Test accuracy studies often have important biases, which may result in unreliable estimates of the accuracy of a diagnostic test (19–22). Several proposals have been advanced to assess the quality of a study evaluating diagnostic accuracy (23–25). Partly because of the lack of a true reference standard, there is no consensus for a single approach to assessing study quality (26). The lack of consistent relationships between specific quality elements and the magnitude of outcomes complicates the task of assessing quality (27, 28). In addition, quality is assessed on the basis of reported information that does not necessarily reflect how the study was actually performed and analyzed.

The Standards for Reporting of Diagnostic Accuracy (STARD) group recently published a 25-item checklist as a guide to improve the quality of reporting all aspects of a diagnostic study (29). The STARD checklist was not developed as a tool to assess the quality of diagnostic studies. However, many items in the checklist are included in a recently developed tool for quality assessment of diagnostic accuracy studies (the QUADAS tool). The QUADAS tool consists of 14 items that cover patient spectrum, reference standard, disease progression bias, verification and review bias, clinical review bias, incorporation bias, test execution, study withdrawals, and intermediate results (28, 30).

**CHALLENGE: ASSESSING APPLICABILITY OF STUDY POPULATIONS**

Studies beyond the level of technical feasibility must include both diseased and nondiseased individuals who reflect the use of the diagnostic technologies in actual clinical settings. Because of the need to understand the relationship between test sensitivity and specificity, a study that reports only sensitivity (that is, evaluation of the test only in a diseased population) or only specificity (that is, evaluation of the test only in a healthy population) results cannot be used for this evaluation.

In this section, we base our discussion on the evidence report on evaluating diagnostic technologies for acute cardiac ischemia in the emergency department (7). When the

spectrum of disease ranges widely within a diseased population, the interpretation of results in a diagnostic accuracy study may be affected if study participants possess only certain characteristics of the diseased population (15, 21). For example, patients in cardiac care units are more likely to have acute cardiac ischemia than patients in the emergency department. When patients with more severe illness are analyzed, the false-positive rate is reduced and sensitivity is overestimated. For example, biomarkers may have high sensitivity when used in patients with acute myocardial infarction in a cardiac care unit but may perform poorly in an emergency department because of their inability to detect unstable angina pectoris.

Table 2 shows the distribution of studies according to hospital settings and inclusion criteria for presenting symptoms used in each of the studies. Although there were a total of 108 studies, the number of studies available for analysis at each hospital setting and for each inclusion criteria definition is limited. If we apply a strict criterion of accepting only studies performed in the emergency department and on patients with the most inclusive definition of acute cardiac ischemia (category I in the table), only 13 studies are available. Furthermore, some studies used acute cardiac ischemia as the outcome of interest and some used acute myocardial infarction, further reducing the number of studies available for a specific assessment. As shown by the small numbers in some cells of the table, the information available from all studies using the 21 diagnostic technologies meeting the criteria is lacking for certain combinations of patients and settings.

The heterogeneity of study populations makes synthesizing study results difficult. Used in conjunction within the 6-level framework, this scheme of categorizing the population and settings into similar groups greatly facilitates study comparison and interpretation of results.

**CHALLENGE: SYNTHESIZING MEASURES OF TEST ACCURACY**

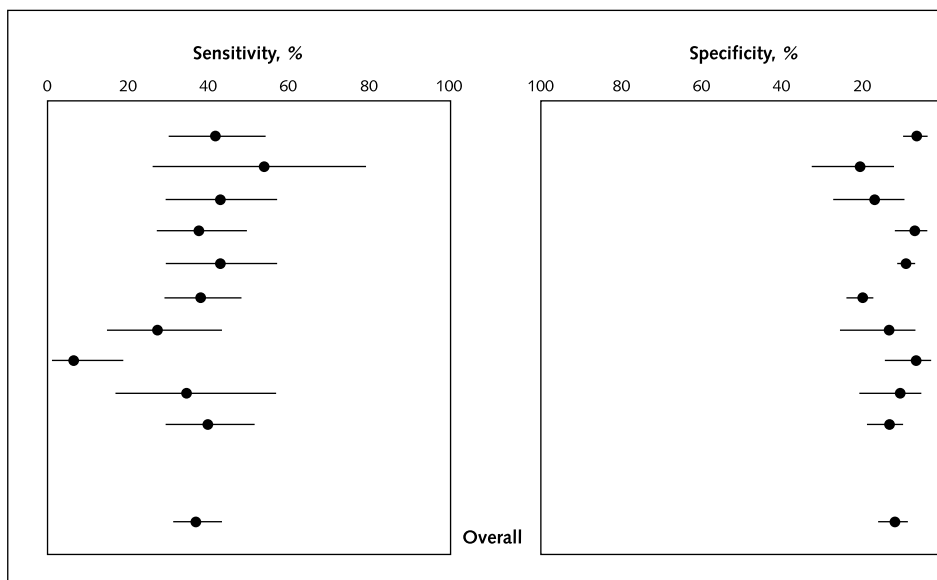
In the evidence report “Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Depart-

*Table 2. Categorization of Studies by Population Category and Setting: Diagnostic Technologies To Evaluate Acute Cardiac Ischemia in the Emergency Department*

Population Category*	Studies, n				Total
	Emergency Department	Hospital	Cardiac Care Unit	Prehospital	
I	13	0	2	4	19
II	27	16	3	7	53
III	20	7	3	1	31
IV	4	1	0	0	5
Total	64	24	8	12	108

\* Category I: studies that included all patients with signs and symptoms suggestive of acute cardiac ischemia, such as chest pain, shortness of breath, jaw pain, and acute pulmonary edema; category II: studies that used chest pain as the inclusion criteria; category III: studies that included patients with chest pain but excluded those with clinical or electrocardiographic findings diagnostic of acute myocardial infarction; category IV: studies in which all patients were hospitalized or that used additional criteria to enroll highly selected subpopulations or retrospective studies.

Figure 1. Meta-analysis of studies evaluating the use of a single creatine kinase measurement to diagnose acute myocardial infarction in the emergency department.



The sensitivity and specificity estimate of each study is plotted, along with the overall results obtained by combining the estimates independently using a random-effects model.

ments” (7), the authors used 3 methods to summarize test accuracy results: independent combining of sensitivity and specificity values, diagnostic odds ratios, and summary receiver-operating characteristics (ROC) analysis. Other publications have described the methods for meta-analysis of diagnostic test accuracy (31–36). We discuss the basic challenges of applying these methods and interpreting their results here. **Figures 1** and **2** illustrate examples of results obtained from 2 common methods.

Diagnostic test results are often reported as a numeric quantity on a continuous scale (for example, troponin level) but are then used as a binary decision tool by defining a threshold above which the test result is positive and below which it is negative. Results may then be summarized in a  $2 \times 2$  table reflecting the agreement between the test result and the disease state as the number of true-positive, false-positive, true-negative, and false-negative results. Changing this threshold changes both the sensitivity and specificity of the test. Test performance is described by the ROC curve, which displays the true-positive rate (sensitivity) on the vertical axis versus the false-positive rate ( $1 - \text{specificity}$ ) on the horizontal axis for all possible thresholds. This fundamental bivariate structure poses a challenge for constructing a single-number summary to describe test performance.

**Table 3** summarizes common single-number measures used to describe test performance. Only the last 4 combine information about both sensitivity and specificity. This can be illustrated by the different ROC curves that each measure implies. A constant sensitivity implies a horizontal line, a constant specificity implies a vertical line, and a constant likelihood ratio also implies a linear relationship

between sensitivity and specificity. The odds ratio, on the other hand, describes a curve symmetrical about the line where sensitivity equals specificity.

In combining data from several diagnostic tests, the first step should be to plot the sensitivity–specificity pairs in each study on one graph. Because the plot may suggest a curvilinear relationship, the use of summary sensitivity, specificity, or likelihood ratio measures may be inadequate. We should combine likelihood ratios only if sensitivity and specificity are linearly related, and we should combine sensitivity or specificity only if one is invariant to a change in the other. Combining both sensitivity and specificity independently is appropriate only if we believe the test always operates at one fixed combination of sensitivity and specificity. Nevertheless, because these summary measures are either proportions or ratios of proportions and thus are easily combined by using standard meta-analytic techniques, many meta-analyses of diagnostic tests have reported them. If curvilinearity is detected, it is better instead to summarize with the diagnostic odds ratio using standard methods for combining odds ratios.

The odds ratio should not be applied, however, if the ROC curve suggests asymmetry about the line of equal sensitivity and specificity. In this case, the decrease in sensitivity corresponding to increased specificity differs between high and low sensitivity settings. A summary ROC curve does allow for such asymmetry. It models the linear relationship between the odds ratio and the probability of a positive test result and then re-expresses this in terms of the ROC curve (33). It reduces to the odds ratio if the regression slope is zero so that the odds ratio is invariant to the test positivity threshold. In doing so, however, it treats the

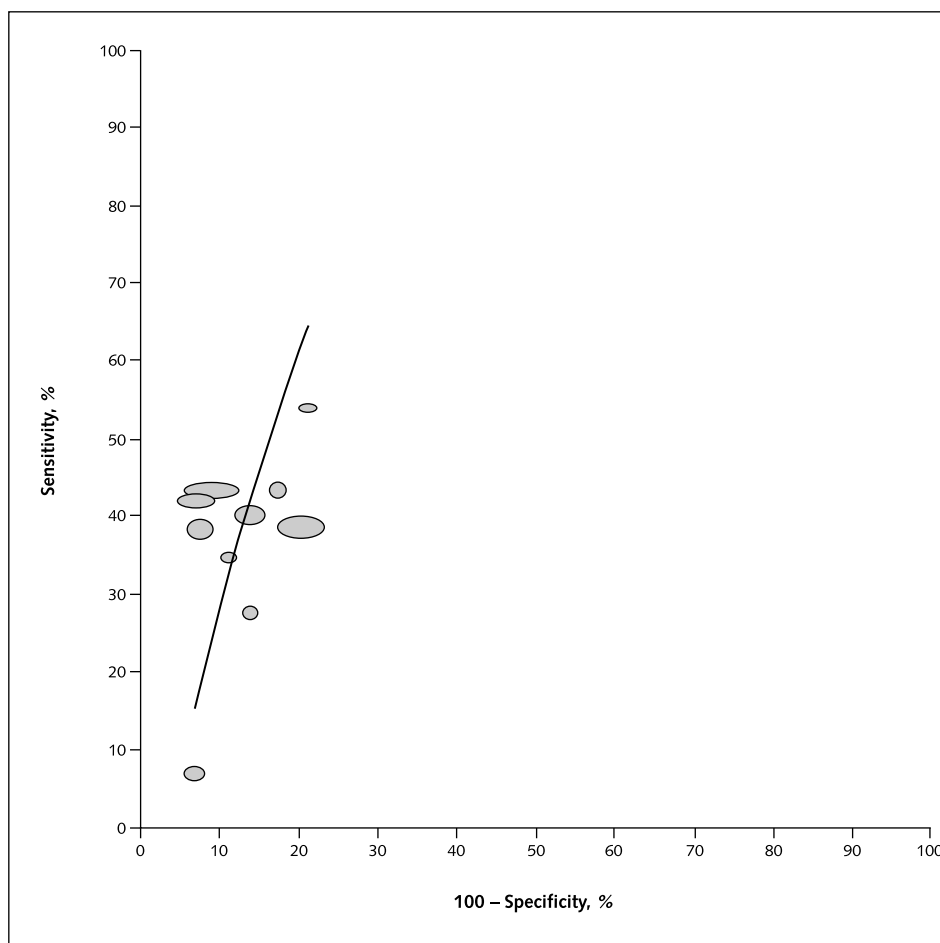
observed odds ratios as random effects but the probabilities of a positive result as fixed effects. Thus, the resulting summary measure fails to incorporate all the uncertainty in the observed test results. Recently, some analysts have used bivariate random-effects models to capture the randomness in both sensitivity and specificity simultaneously (37). Summary ROC curves may similarly be modeled by bivariate random-effects regressions. Further research is needed to compare and relate these methods to each other.

Meta-regression, the use of regression methods to incorporate the effect of covarying factors on summary measures of performance, has been used to explore between-study heterogeneity in therapeutic studies (38). In diagnostic studies, likewise, heterogeneity in sensitivity and specificity can result from many causes related to definitions of the test and reference standards, operating characteristics of the test, methods of data collection, and patient characteristics. Covariates may be introduced into a regression with any test performance measure as the dependent variable. As with any meta-regression, however, the sample

size will correspond to the number of studies in the analysis. A small number of studies will limit the power of regression to detect significant effects. As always, we should not assume that the lack of significance implies that no factors could influence the relationship between sensitivity and specificity. Although multivariate meta-regression has advantages, study characteristics are often strongly associated with each other; this leads to collinearity, which creates difficulty in interpreting meta-regression models. Warning signs of collinearity include large pairwise correlations between predictor variables, large changes in coefficients caused by the addition or deletion of other variables, and extremely large standard errors for coefficients.

To date, the most common approach to meta-regression has been to insert covariates into the summary ROC regression, allowing the odds ratio to vary by study characteristics other than test positivity (33, 39). This implies that performance is described by not one but multiple ROC curves. Although the literature does provide quite a few examples of the use of meta-regression with diagnostic

Figure 2. Summary receiver-operating characteristics curve analysis of studies evaluating the use of a single creatine kinase measurement to diagnose acute myocardial infarction in the emergency department.



The position of the ellipse corresponds to the sensitivity and specificity reported in the individual study. The size of the ellipse is weighted according to the study size. The summary receiver-operating characteristic curve is limited to the range where data are available.

Table 3. Metrics of Test Performance\*

Metric	Definition	Advantages	Disadvantages
Accuracy	$(TP + TN)/N$	Intuitive	Depends on prevalence
Sensitivity	$TP/N_D$	Does not depend on prevalence	Applies only to diseased persons
Specificity	$TN/N_W$	Does not depend on prevalence	Applies only to nondiseased persons
Positive predictive value	$TP/N_p$	Clinical relevance	Depends on prevalence
Negative predictive value	$TN/N_N$	Clinical relevance	Depends on prevalence
Positive likelihood ratio	$(TP/N_D)/(FP/N_W)$	Does not depend on prevalence	Applies only to positive tests
Negative likelihood ratio	$(FN/N_D)/(TN/N_W)$	Does not depend on prevalence	Applies only to negative tests
Odds ratio	$TP \times TN/FN \times FP$	Does not depend on prevalence; combines sensitivity and specificity	Values FP and FN errors equally; not intuitive
Area under curve	Area under ROC curve	Does not depend on prevalence; combines sensitivity and specificity	Lack of clinical interpretation

\* FN = false-negative; FP = false-positive; N = sample size;  $N_D$  = TP + FN;  $N_N$  = TN + FN;  $N_p$  = TP + FP;  $N_W$  = TN + FP; ROC = receiver-operating characteristic; TN = true-negative; TP = true-positive.

tests (40), the interpretation is mainly limited to determining whether study factors affect performance. Few studies have fully explicated these effects on the tradeoff between sensitivity and specificity.

Frequently, meta-analyses assess several diagnostic tests for the same condition. In such cases, we may wish not only to report the performance of each test but also to compare performance between tests. However, these comparisons are seldom reported. Among authors who do report them, many use paired or unpaired chi-square tests of single numeric summaries in the form of proportions such as sensitivity and specificity (41) or rank-sum tests comparing odds ratios (42). A few meta-analyses have incorporated the test as a covariate into the summary ROC model (43). Nevertheless, the major obstacle to comparing tests is that each research paper rarely reports on each test. Therefore, the cross-classification of papers and tests is incomplete and comparisons must deal with missing values. Siadat and colleagues have suggested a repeated-measures model fit by generalized estimating equations to overcome this problem, but more research is needed to test their model (44).

### CHALLENGE: ASSESSING SOCIETAL IMPACT OF DIAGNOSTIC TECHNOLOGY

Proponents of new diagnostic technologies often claim that their use would improve outcomes, avoid unnecessary procedures, and reduce costs. Although cost-effectiveness evaluations in actual clinical settings are desirable, such evaluations are rarely done on diagnostic technologies because of the time and resources required. Moreover, various diagnostic tests can be used singly, in sequence, or in combination with other tests to evaluate a condition. Performing clinical trials to compare many diagnostic strategies is simply not feasible. In face of the paucity of outcome studies, the use of an explicit model can allow a

decision maker to use available level 2, 3, and 4 data. Such models incorporate diagnostic test performance data obtained either from relevant individual studies or from systematic reviews and meta-analyses (45).

For example, in evaluating the use of PET scans to help manage patients with possible Alzheimer disease (8), the Duke EPC used a decision analysis model that incorporated test accuracy data to estimate health effects. Given the low efficacy and toxicity of current pharmacologic treatment for dementia, the strategy of empirically treating all patients is preferred over the alternative strategies of selectively treating patients according to PET scan results or expectant management in terms of quality-adjusted life-years. The PET scan strategy is preferred only in terms of the measure “percentage correct diagnosis.” When hypothetical treatments are considered, PET scan becomes the more attractive strategy as complications of the hypothetical treatment become more severe. But if efficacy is simultaneously increased with dangerous treatment—as it logically should be in order to be worth considering—the strategy of PET scan becomes less attractive (8).

In another example, a cost-effectiveness analysis was conducted as part of an evidence report to evaluate technologies for identifying acute cardiac ischemia in the emergency department (7). A decision analytic model with 21 diagnostic strategies was developed by using results from meta-analyses of the diagnostic technologies evaluated in the evidence report. A direct comparison of such a large number of diagnostic technologies would not be feasible in actual clinical settings.

In translating information on diagnostic test performance to life-years saved by the diagnostic test, cost-effectiveness analysis becomes inherently complex because treatments—which may vary greatly—must also be specified as a consequence of a specific diagnosis (46). Furthermore, estimating costs for consequences of diagnostic testing is

difficult since both correct and incorrect diagnoses must be considered. Where evidence on diagnostic performance is poor or sparse, cost-effectiveness analysis may mislead rather than elucidate. Finally, a cost-effectiveness analysis usually addresses the questions most relevant to the decision makers within the context of their health care system (46), which may limit the applicability of the results to other settings; however, this may be mitigated by relevant sensitivity analyses.

## CONCLUSION AND RECOMMENDATIONS

The advent of evidence-based medicine has propelled systematic reviews of interventions and diagnostic tests to the forefront of health care practice and policy. Diagnosis is not an end in itself; rather, the purpose of a test is to guide patient management decisions and thus improve patient outcome. Diagnostic tests should be evaluated as rigorously as therapeutic interventions in order to determine whether use of the technology improves patient outcomes. Table 4 lists recommendations for improving systematic reviews of diagnostic technologies.

Evidence-based Practice Centers have been guided by a 6-level framework that was introduced almost 30 years ago to evaluate diagnostic technologies. Ideally, we would like to have adequate studies for each level of the framework, but this is seldom the case. For most technologies, technical feasibility studies are not clinically relevant and are not considered in the assessment. Studies on diagnostic accuracy have received the most attention because these are the most available of the clinically relevant studies and may be used to model the outcomes of using the results of a diagnostic test for treatment management decisions.

While the 6-level framework provides a useful guide to evaluated diagnostic technologies, several methodologic challenges remain. First, publication bias may distort the findings of the systematic reviews of diagnostic technologies. The use of funnel plots of statistical models to correct for publication bias should be viewed with caution because validation of these methods is lacking. Second, diagnostic accuracy studies often have important biases that may result in overestimation of the performance of a diagnostic test. We hope that such initiatives as STARD and QUADAS will result in better-quality studies. Third, heterogeneity of study populations in diagnostic test evaluations makes synthesizing results difficult. Categorizing populations and settings into similar groups may facilitate study comparisons and interpretation. Fourth, choosing appropriate methods of quantitative synthesis of test accuracy results is difficult. In general, we believe that a bivariate method that simultaneously considers sensitivity and specificity in the analysis, such as the summary ROC method, is the preferred approach to summarizing test accuracy results. In clinical practice, various diagnostic tests can be used singly, in sequence, or in combination with other tests to evaluate a condition. Empirical assessment of myriad

**Table 4. Recommendations for Improving Systematic Reviews of Diagnostic Technologies**

View funnel plots or statistical models to detect or correct for publication bias with caution because validation of these methods is lacking.
Support coauthors' and editors' adherence to reporting standards for diagnostic tests.
Support categorizing and reporting of study populations and settings into similar groups to facilitate study comparisons and interpretation.
Use appropriate methods of quantitative synthesis of test accuracy results. In general, consider a bivariate method that simultaneously considers sensitivity and specificity.
Consider decision analytic and cost-effectiveness models when there is good-quality evidence regarding diagnostic performance and therapeutic effectiveness of the alternative technologies.

diagnostic strategies is not feasible. Decision analytic and cost-effectiveness models can be informative, but only where good-quality evidence on diagnostic performance and therapeutic effectiveness of the alternative technologies are available.

From Tufts-New England Medical Center Evidence-based Practice Center, Tufts-New England Medical Center, Boston, Massachusetts; Agency for Healthcare Research and Quality, Rockville, Maryland; and Blue Cross and Blue Shield Association Technology Evaluation Center, Chicago, Illinois.

**Grant Support:** In part by grant R01 HS13328 from the Agency for Healthcare Research and Quality (Christopher Schmid, Joseph Lau).

**Potential Financial Conflicts of Interest:** *Employment:* N. Aronson (Blue Cross and Blue Shield Association), D.J. Samson (Blue Cross and Blue Shield Association), C.R. Flamm (Blue Cross and Blue Shield Association). Authors of this paper have received funding for Evidence-based Practice Center reports.

**Corresponding Author:** Joseph Lau, MD, Tufts-New England Medical Center Evidence-based Practice Center, Tufts-New England Medical Center, 750 Washington Street, Boston, MA 02111.

Current author addresses are available at [www.annals.org](http://www.annals.org).

## References

1. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. *Radiology*. 1975; 114:561-5. [PMID: 1118556]
2. McNeil BJ, Adelstein SJ. Determining the value of diagnostic and screening tests. *J Nucl Med*. 1976;17:439-48. [PMID: 1262961]
3. Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography. Effect on diagnostic and therapeutic plans. *JAMA*. 1977;238:224-7. [PMID: 406427]
4. Fryback DG. A conceptual model for output measures in cost-effectiveness evaluation of diagnostic imaging. *J Neuroradiol*. 1983;10:94-6. [PMID: 6410018]
5. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11:88-94. [PMID: 1907710]
6. New England Medical Center EPC Boston, Massachusetts. Magnetic Resonance Spectroscopy for Brain Tumors. Accessed at [www.cms.hhs.gov/coverage/download/id52-2.zip](http://www.cms.hhs.gov/coverage/download/id52-2.zip) on 15 March 2005.

7. Lau J, Ioannidis JP, Balk EM, Milch C, Terrin N, Chew PW, Salem D. Evaluation of technologies for identifying acute cardiac ischemia in emergency departments. Evidence Report/Technology Assessment No. 26 (Prepared by the New England Medical Center Evidence-based Practice Center under contract 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality; May 2001. AHRQ publication no. 01-E006.
8. Duke Center for Clinical Health Policy Research and Evidence Practice Center. Positron emission tomography, single photon emission, computed tomography, functional magnetic resonance imaging, and magnetic resonance spectroscopy for the diagnosis and management of Alzheimer's dementia. Accessed at [www.cms.hhs.gov/coverage/download/id104b.pdf](http://www.cms.hhs.gov/coverage/download/id104b.pdf) on 15 March 2005.
9. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ*. 2004;328:1040. [PMID: 15073027]
10. Alberani V, De Castro Pietrangeli P, Mazza AM. The use of grey literature in health sciences: a preliminary survey. *Bull Med Libr Assoc*. 1990;78:358-63. [PMID: 2224298]
11. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol*. 2000;53:207-16. [PMID: 10729693]
12. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol*. 2000;53:477-84. [PMID: 10812319]
13. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Stat Med*. 2003;22:2113-26. [PMID: 12820277]
14. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors [Editorial]. *N Engl J Med*. 2004;351:1250-1. [PMID: 15356289]
15. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA*. 1988;259:1699-702. [PMID: 3278149]
16. Banta HD, Thacker SB. Electronic fetal monitoring. Lessons from a formative case of health technology assessment. *Int J Technol Assess Health Care*. 2002;18:762-70. [PMID: 12602077]
17. Blackmore CC, Black WC, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. *Acad Radiol*. 1999;6 Suppl 1:S8-18. [PMID: 9891161]
18. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. *Jt Comm J Qual Improv*. 1999;25:470-9. [PMID: 10481816]
19. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA*. 1984;252:2418-22. [PMID: 6481928]
20. Arroll B, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. *J Gen Intern Med*. 1988;3:443-7. [PMID: 3049967]
21. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA*. 1995;274:645-51. [PMID: 7637146]
22. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061-6. [PMID: 10493205]
23. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med*. 1989;4:288-95. [PMID: 2760697]
24. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120:667-76. [PMID: 8135452]
25. Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests: Recommended Methods. Updated 6 June 1996. Available at Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests: Recommended Methods, updated 6 June 1996. Accessed at [www.cochrane.org/cochrane/sadtdoc1.htm](http://www.cochrane.org/cochrane/sadtdoc1.htm) on 1 July 2004.
26. Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematic reviewing studies of diagnostic tests. *Clin Chem Lab Med*. 2000;38:577-88. [PMID: 11028761]
27. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002;287:2973-82. [PMID: 12052127]
28. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140:189-202. [PMID: 14757617]
29. 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. *Ann Intern Med*. 2003;138:40-4. [PMID: 12513043]
30. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25. [PMID: 14606960]
31. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ*. 2001;323:157-62. [PMID: 11463691]
32. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol*. 2002;2:9. [PMID: 12097142]
33. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12:1293-316. [PMID: 8210827]
34. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol*. 1995;48:119-30; discussion 131-2. [PMID: 7853038]
35. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests: a review of the concepts and methods. *Arch Pathol Lab Med*. 1998;122:675-86. [PMID: 9701328]
36. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med*. 2002;21:1237-56. [PMID: 12111876]
37. Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. *Radiology*. 2003;226:837-48. [PMID: 12601211]
38. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *J Clin Epidemiol*. 2004;57:683-97. [PMID: 15358396]
39. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20:2865-84. [PMID: 11568945]
40. Romagnuolo J, Bardou M, Rahme E, Joseph L, Reinhold C, Barkun AN. Magnetic resonance cholangiopancreatography: a meta-analysis of test performance in suspected biliary disease. *Ann Intern Med*. 2003;139:547-57. [PMID: 14530225]
41. Imran MB, Khan MA, Aslam MN, Irfanullah J. Diagnosis of coronary artery disease by stress echocardiography and perfusion scintigraphy. *J Coll Physicians Surg Pak*. 2003;13:465-70. [PMID: 12921688]
42. Bafounta ML, Beauchet A, Chagnon S, Saiag P. Ultrasonography or palpation for detection of melanoma nodal invasion: a meta-analysis. *Lancet Oncol*. 2004;5:673-80. [PMID: 15522655]
43. Vasbinder GB, Nelemans PJ, Kessels AG, Kroon AA, de Leeuw PW, van Engelsehoven JM. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension: a meta-analysis. *Ann Intern Med*. 2001;135:401-11. [PMID: 11560453]
44. Siadat MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *J Clin Epidemiol*. 2004;57:698-711. [PMID: 15358397]
45. Mushlin AI. Challenges and opportunities in economic evaluations of diagnostic tests and procedures. *Acad Radiol*. 1999;6 Suppl 1:S128-31. [PMID: 9891180]
46. Mushlin AI, Ruchlin HS, Callahan MA. Cost effectiveness of diagnostic tests. *Lancet*. 2001;358:1353-5. [PMID: 11684235]

---

**Current Author Addresses:** Drs. Tatsioni, Schmid, and Lau: Tufts-New England Medical Center Evidence-based Practice Center, Tufts-New England Medical Center, 750 Washington Street, Boston, MA 02111. Dr. Zarin: National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894.

Dr. Aronson, Mr. Samson, and Dr. Flamm: Blue Cross and Blue Shield Association, Technology Evaluation Center, 225 North Michigan Avenue, Chicago, IL 60601-7680.