

**Developing the Topic and Structuring the Review:
Utility of PICOTS, Analytic Frameworks, Decision
Trees, and Other Frameworks**

Prepared for:

Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

**AHRQ Publication No. xx-EHCxxx-EF
<Month Year>**

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different medical tests, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care (EHC) Program, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (<http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318>), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and highlights particular areas where the inherently different qualities of medical tests necessitate a different or variation of the approach to systematic review compared to a review on treatments. We provide guidance in stepwise fashion for those conducting a systematic review.

The *Medical Test Methods Guide* is a living document, and will be updated as further empirical evidence develops and our understanding of better methods improves. Comments and suggestions on the *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 2. Developing the Topic and Structuring the Review: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks

“[We] have the ironic situation in which important and painstakingly developed knowledge often is applied haphazardly and anecdotally. Such a situation, which is not acceptable in the basic sciences or in drug therapy, also should not be acceptable in clinical applications of diagnostic technology.”

J. Sanford (Sandy) Schwartz, Institute of Medicine, 1985¹

Developing the topic creates the foundation and structure of an effective systematic review. Developing the topic includes understanding and clarifying how a test might be of value in practice and establishing the key questions to guide decisionmaking related to the claim. This typically involves specifying the clinical context in which the test might be used. Structuring the review refers to identifying the analytic strategy that will most directly achieve the goals of the review, accounting for idiosyncrasies of the data.

Topic development and structuring of the review are complementary processes. As EPCs develop and refine the topic, the structure the review should follow ideally becomes clearer. Moreover, success at this stage reduces the chance of major changes in the scope of the review and minimizes rework.

Common Challenges

The ultimate goal of a medical test review is to identify and synthesize evidence that will help evaluate the impacts on health outcomes of alternative testing strategies. Two common problems can impede achieving this goal. One is that the request for a review may state the claim for the test ambiguously. For example, a new medical test for Alzheimer’s disease might fail to specify the patients who may benefit from the test—from the “worried well” without evidence of deficit to those with frank impairment and loss of function in daily living. Similarly, the request for a review of tests for prostate cancer might neglect to consider the role of such tests in clinical decisionmaking, such as guiding the decision to biopsy.

Because of the indirect impact of medical tests on clinical outcomes, a second problem is how to identify the intermediate outcomes that link a medical test to improved clinical outcomes compared to an existing test. The scientific literature related to the claim rarely includes direct evidence, such as randomized controlled trial results, in which patients are allocated to the relevant test strategies and evaluated for downstream health outcomes. More commonly, evidence about outcomes in support of the claim relates to intermediate outcomes, such as test accuracy.

Principles for Addressing the Challenges

Principle 1: Engage Stakeholders Using the PICOTS Typology

In approaching topic development, EPCs should engage in a direct dialogue with the primary requestors or relevant users of the review (herein denoted “stakeholders”) to understand the objectives of the review in practical terms; in particular, EPC investigators should understand the sorts of decisions that the review is likely to affect. Such a discussion also serves to bring investigators and stakeholders to a shared understanding about the essential details of the tests and their relationship to existing test strategies (i.e., replacement, triage, or add-on), range of potential clinical utility, and potential adverse consequences of testing.

Operationally, the objective of the review is reflected in the key questions, which are normally presented in a preliminary form at the outset of a review. EPCs should examine the proposed key questions to ensure that they accurately reflect the needs of stakeholders and are likely to be answered given the available time and resources. Including a wide variety of stakeholders and experts (such as the U.S. Food and Drug Administration [FDA], manufacturers, technical and clinical experts, and patients) can help provide additional perspectives on the claim and use of the tests. A preliminary examination of the literature can identify existing systematic reviews and clinical practice guidelines that may summarize evidence on current strategies for using the test and its potential benefits and harms.

The PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting), defined in the Introduction to this *Medical Test Methods Guide* (Paper 1), is a typology for defining particular contextual issues, and this formalism can be useful in focusing discussions with stakeholders.

It is important to recognize that the process of topic refinement is iterative. Despite the best efforts of all participants, the topic may change even as the review is being conducted. EPCs should consider at the outset how such a situation will be addressed.²⁻⁴

Principle 2: Develop an Analytic Framework

The term “analytic framework” (sometimes called a causal pathway) is used here to denote a specific form of graphical representation that specifies a path from the intervention or test of interest to all important health outcomes, including intervening steps and intermediate outcomes.⁵ Each linkage relating test, intervention, or outcome represents a potential key question and, it is hoped, a coherent body of literature.

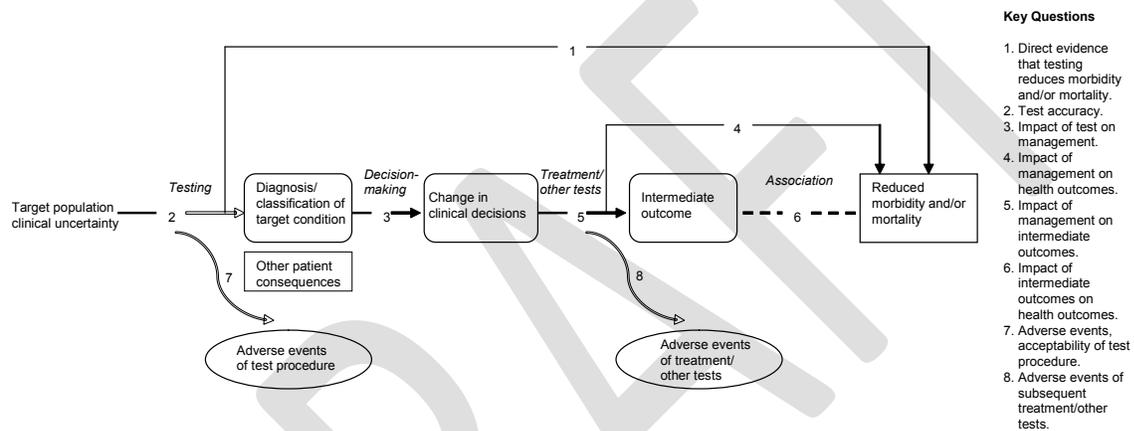
The AHRQ EPC program has described the development and use of analytic frameworks in systematic reviews of interventions. The analytic framework is developed iteratively in consultation with stakeholders to illustrate and define the important clinical decisional dilemmas and thus serves to clarify important key questions further.⁶

However, systematic reviews of medical tests present challenges not encountered in reviews of therapeutic interventions. The impact of medical tests on important outcomes is, by nature, more

indirect, and there are more potential pathways by which a medical test may affect important outcomes. Because of the often-convoluted linkage to clinical outcomes, research studies mostly focus on intermediate outcomes such as diagnostic accuracy. The analytic framework can help users to understand how these intermediate outcomes fit in the pathway to influencing clinical outcomes, and to consider whether these downstream issues may be relevant to the review.

Harris and colleagues have described the value of the analytic framework in assessing screening tests for the U.S. Preventive Services Task Force (USPSTF).⁷ A prototypical analytic framework for medical tests as used by the USPSTF is shown in Figure 2-1. Each number in Figure 2-1 can be viewed as a separate key question that might be included in the evidence review.

Figure 2-1. Application of USPSTF analytic framework to test evaluation



Adapted from Harris et al., 2001⁷

In summarizing evidence, reviewers should remember that studies for each linkage might vary in strength of design, limitations of conduct, and adequacy of reporting. The linkages leading from changes in patient management decisions to health outcomes are often of particular importance. The implication here is that the value of a test usually derives from its influence on some action taken in patient management. Although this is usually the case, sometimes the information alone from a test may have value independent of any action it may prompt.

Principle 3: Consider Using Decision Trees

An analytic framework is helpful when direct evidence is lacking, showing relevant key questions along indirect pathways between the test and important clinical outcomes. Analytic frameworks are, however, not well-suited to depicting multiple alternative uses of the particular test (or its comparators) and are limited in their ability to represent the impact of test results on clinical decisions, the specific potential outcome consequences of altered decisions. EPCs can use simple decision trees or flow diagrams alongside the analytic framework to illustrate details

of the potential impact of test results on management decisions and outcomes. Constructing decision trees may help clarify key questions by identifying which indices of diagnostic accuracy are relevant to the clinical problem and which range of possible pathways and outcomes (see Paper 3) practically and logically flow from a test strategy. Lord et al., describe how decision trees may be used for defining which steps and outcomes may differ with different test strategies, and thus what are the important questions to ask to compare tests according to whether the new test is a replacement, a triage, or an add-on to the existing test strategy.⁸

One example of how constructing decision trees can be useful comes from a review of noninvasive tests for carotid artery disease.⁹ This review found that common metrics of sensitivity and specificity that counted both high-grade stenosis and complete occlusion as “positive” studies would not be reliable guides to actual test performance because the two results would be treated quite differently. This insight was subsequently incorporated into calculations of noninvasive carotid test performance.⁹⁻¹⁰ Further examples are provided in the Illustrations, below.

Principle 4: Sometimes it is Sufficient to Focus Exclusively on Accuracy Studies

Once EPCs have diagrammed the decision tree by which diagnostic accuracy may affect intermediate and clinical outcomes, reviewers can determine whether it is necessary to include key questions regarding outcomes beyond diagnostic accuracy. For example, diagnostic accuracy may be sufficient when the new test is as sensitive as the old test *and* the new test’s value derives from avoiding the old test’s adverse effects (i.e., because the new test is safer or less invasive) or higher costs. Implicit in this example is the comparability of downstream management decisions and outcomes between the test under evaluation and the comparator test. Another instance when a review may be limited to evaluation of sensitivity and specificity is when the new test is as sensitive as, but more specific than, the comparator, allowing avoidance of harms of further tests or unnecessary treatment. This situation requires the assumptions that the same cases would be detected by both tests and that treatment efficacy would be unaffected by which test was used.¹¹

Particular questions that EPCs may consider in reviewing analytic frameworks and decision trees to determine if diagnostic accuracy studies alone are adequate include the following:

1. Are extra cases detected by the new, more sensitive test similarly responsive to treatment?
2. Are trials available that selected patients with the new test?
3. Do trials assess whether the new test results predict response?
4. If available trials selected only patients assessed with the old test, do extra cases represent the same spectrum or disease subtypes as trial participants?
5. Are tests’ cases subsequently confirmed by same reference standard?
6. Does the new test change the definition or spectrum of disease (e.g., earlier stage)?
7. Is there heterogeneity of test accuracy and treatment effect (i.e., do accuracy and treatment effects vary sufficiently according to levels of a patient characteristic to change the comparison of the old and new test)?

When the clinical utility of an older comparator test has been established, and the first five questions can all be answered in the affirmative, then diagnostic accuracy evidence alone may be sufficient to support conclusions about a new test.

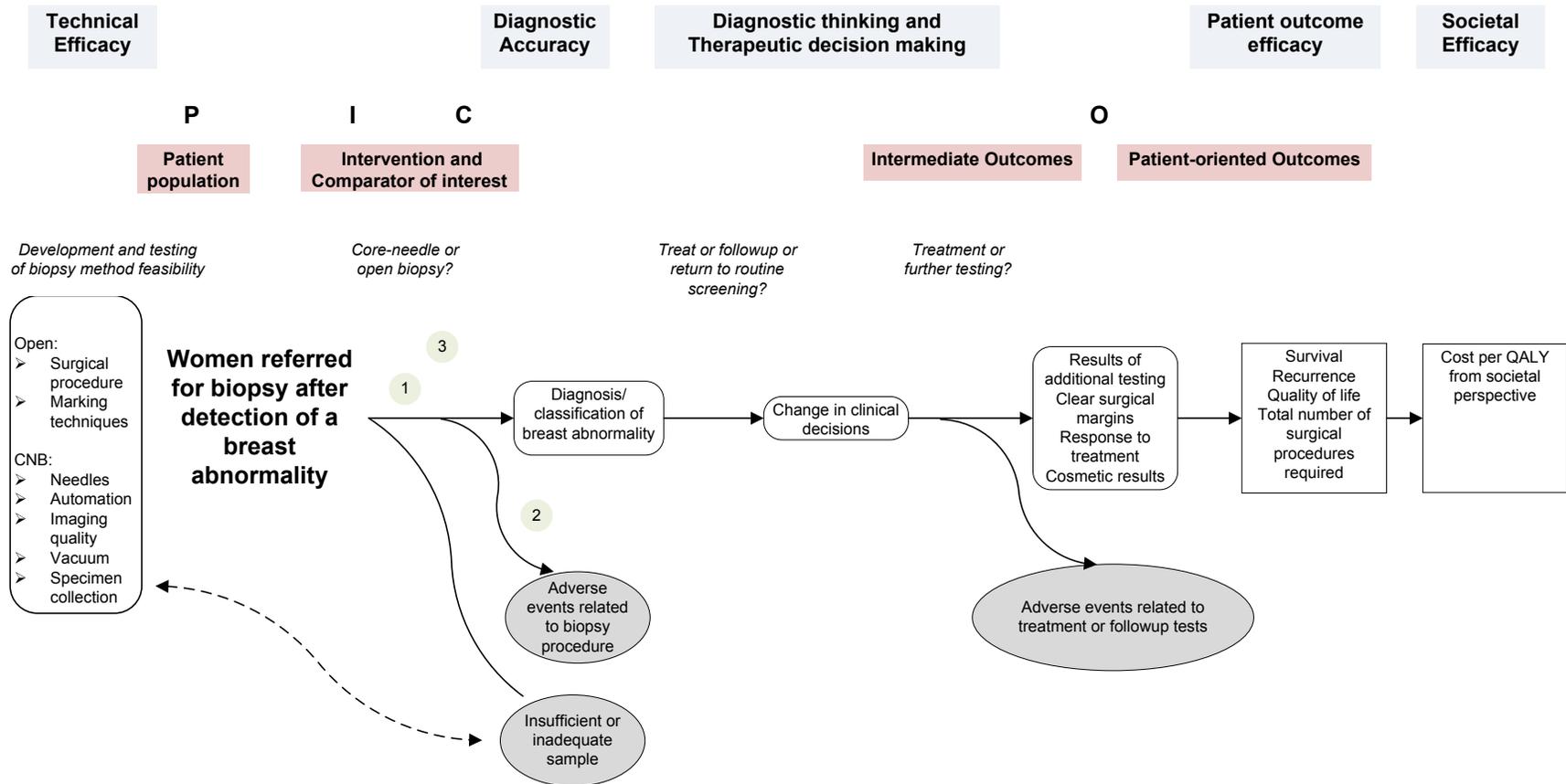
Principle 5: Other Frameworks May be Helpful

Various other frameworks (generally termed “organizing frameworks,” as described briefly in the Introduction to this *Medical Test Methods Guide* [Paper 1]) relate to categorical features of medical tests and medical test studies, and reviewers may find these frameworks useful. Lijmer and colleagues reviewed the different types of organizational frameworks and found 19 frameworks, which generally classify medical test research into 6 different domains or phases, including technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal aspects.¹²

These frameworks have been defined for a variety of purposes. Some researchers, such as Van Den Bruel and colleagues, proposed that these frameworks are a hierarchy and a model for how medical tests should be studied, with one level leading to the next (i.e., success at each level depends on success at the preceding level).¹³ Others, such as Lijmer and colleagues have argued that “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.”¹²

We suggest that rather than being a hierarchy of evidence, organizational frameworks are useful in categorizing key questions and which types of studies would be most useful for specific questions in the review. They may be useful in clustering studies to be reviewed together, and this may improve the readability of a review document. No specific framework is recommended, and indeed the categories of most organizational frameworks at least approximately line up with the analytic framework and the PICO(TS) elements as shown in Figure 2-2.

Figure 2-2. Example of an analytical framework within an overarching conceptual framework in the evaluation of breast biopsy techniques*



The numbers in the figure depict where the three key questions are located within the flow of the analytical framework.

Illustrations

To illustrate the principles above, we describe three examples. In each case the initial claim was at least somewhat ambiguous, and through the use of the PICOTS typology, the analytic framework, and simple decision trees, the systematic reviewers were able to work with stakeholders to clarify the objective and analytic approach to the evidence review (Table 2-1).

Table 2-1. Examples of initially ambiguous claims that were clarified through the process of topic development

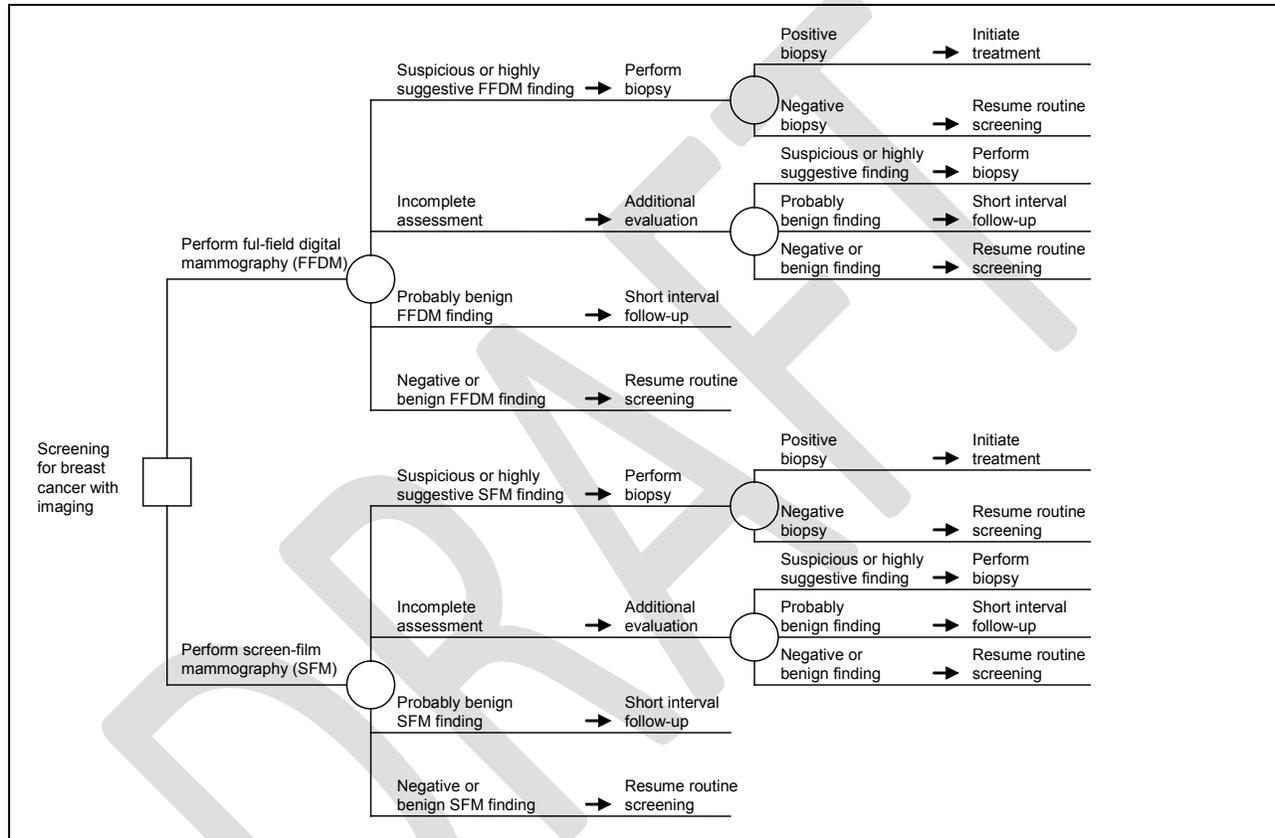
General topic	FFDM to replace SFM in breast cancer screening (Figure 2-3)	HER2 gene amplification assay as add-on to HER2 protein expression assay (Figure 2-4)	PET as triage for breast biopsy (Figure 2-5)
Initial ambiguous claim	FFDM may be a useful alternative to SFM in screening for breast cancer	HER2 gene amplification and protein expression assays may complement each other as means of selecting patients for targeted therapy	PET may play an adjunctive role to breast examination and mammography in detecting breast cancer and selecting patients for biopsy
Key concerns suggested by PICOTS, analytic framework, and decision tree	Key accuracy indices: sensitivity, diagnostic yield, recall rate; similar types of management decisions and outcomes for index and comparator test-and-treat strategies	Key accuracy indices: proportion of individuals with intermediate/equivocal HER2 protein expression results who have HER2 gene amplification; key outcomes are related to effectiveness of HER2-targeted therapy in this subgroup	Key accuracy indices: negative predictive value; key outcomes to be contrasted were benefits of avoiding biopsy versus harms of delaying initiation of treatment for undetected tumors
Refined claim	In screening for breast cancer, interpretation of FFDM and SFM would be similar, leading to similar management decisions and outcomes; FFDM may have a similar recall rate and diagnostic yield at least as high as SFM; FFDM images may be more expensive, but easier to manipulate and store	Among individuals with localized breast cancer, some may have equivocal results for HER2 protein overexpression but have positive HER2 gene amplification, identifying them as patients who may benefit from HER2-targeted therapy but otherwise would have been missed	Among patients with a palpable breast mass or suspicious mammogram, if FDG PET is performed before biopsy, those with negative scans may avoid the adverse events of biopsy with potentially negligible risk of delayed treatment for undetected tumor
Reference	Blue Cross and Blue Shield Association Technology Evaluation Center, 2002 ¹⁴	Seidenfeld et al., 2008 ¹⁵	Samson et al., 2002 ¹⁶

Abbreviations: FDG = fluorodeoxyglucose; FFDM = full-field digital mammography; HER2 = human epidermal growth factor receptor 2; PET = positron emission tomography; PICOTS = Patient population, Intervention, Comparator, Outcomes, Timing, Setting; SFM = screen-film mammography

The first example concerns full-field digital mammography (FFDM) as a replacement for screen-film mammography (SFM) in screening for breast cancer; the review was conducted by the Blue Cross and Blue Shield Association Technology Evaluation Center.¹⁴ Specifying PICOTS elements and constructing an analytic framework were straightforward, with the latter resembling Figure 2-2 in form. In addition, a simple decision tree was drawn (Figure 2-3) which revealed that the management decisions for both screening strategies were similar. The decision

tree also showed that the key indices of test performance were sensitivity, diagnostic yield, and recall rate, and given the symmetry of the tree and stakeholder input indicating that the outcomes of a breast cancer identified with one modality or the other was the same, downstream treatment outcomes were not a critical issue. These insights were useful as the project moved to abstracting and synthesizing the evidence, which focused on accuracy and recall rates. As a note, the reviewers concluded that FFDM and SFM had comparable accuracy and led to comparable outcomes; however, storing and manipulating images was much easier for FFDM than for SFM.

Figure 2-3. Replacement test example: full-field digital mammography versus screen-film mammography*

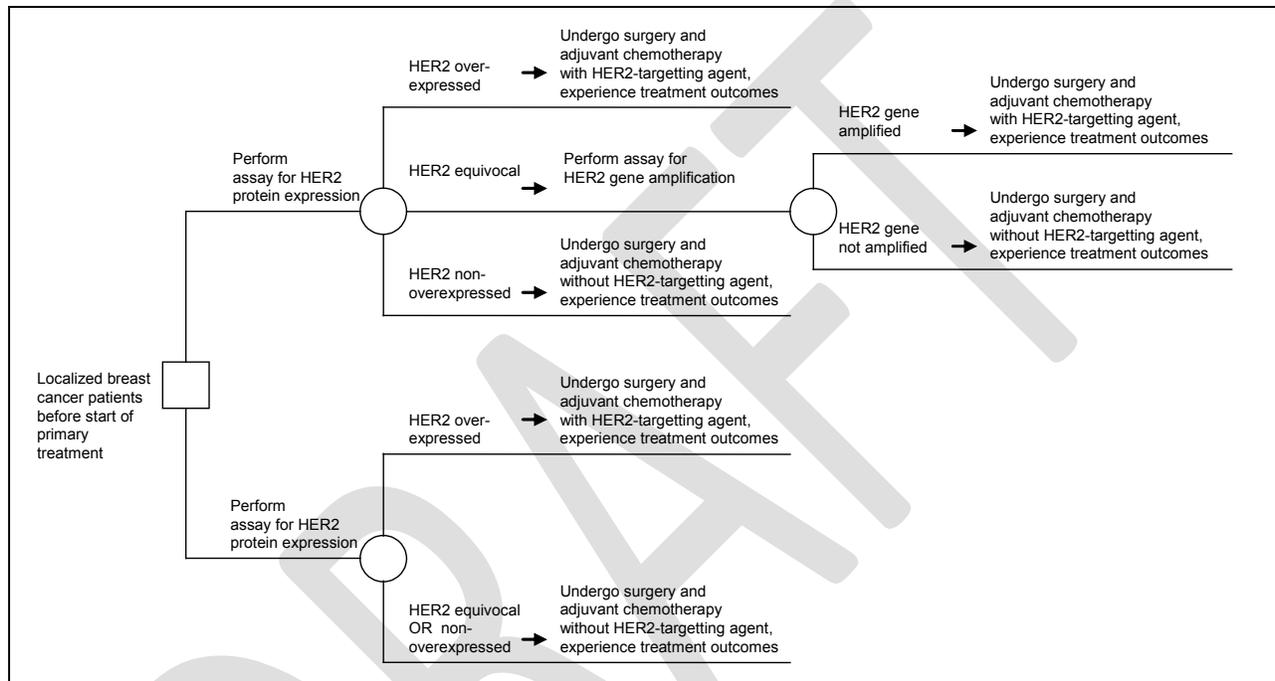


* Figure taken from Blue Cross and Blue Shield Association Technology Evaluation Center, 2002.¹⁴

The second example concerns use of the human epidermal growth factor receptor 2 (HER2) gene amplification assay after the HER2 protein expression assay to select patients for HER2-targeting agents as part of adjuvant therapy among patients with localized breast cancer.¹⁵ The HER2 gene amplification assay has been promoted as an add-on to HER2 protein expression assay. Specifically, individuals with equivocal HER2 protein expression would be followed up with a measure of amplified HER2 gene levels; in addition to those with increased HER2 protein expression, patients with elevated levels by amplification assay would also receive adjuvant chemotherapy that includes HER2-targeting agents. Again, PICOTS and an analytic framework were developed, establishing the basic key questions. In addition, a decision tree was constructed (Figure 2-4) that made it clear that the treatment outcomes affected by HER2 protein and gene assays were at least as important as the test accuracy. While in the first case, the reference

standard was actual diagnosis by biopsy, here the reference standard is the amplification assay itself. The decision tree identified the key accuracy index as the proportion of individuals with equivocal HER2 protein expression results who have positive amplified HER2 gene assay results. The tree exercise also indicated that one key question must be whether HER2-targeted therapy is effective for patients who had equivocal results on the protein assay but were subsequently found to have positive amplified HER2 gene assay results.

Figure 2-4. Add-on test example: HER2 protein expression assay followed by HER2 gene amplification assay to select patients for HER2-targetted therapy*



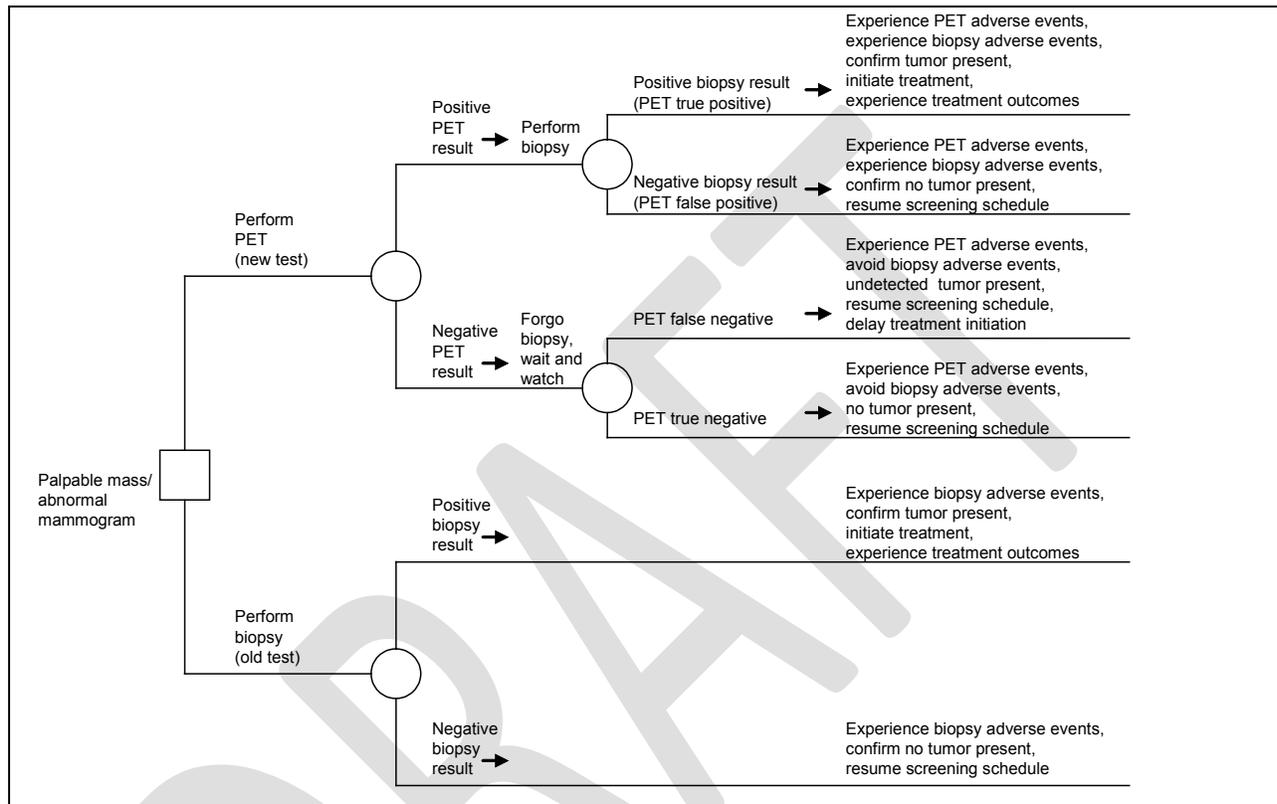
Abbreviation: HER2 = human epidermal growth factor receptor 2.

* Figure taken from Seidenfeld et al., 2008.¹⁵

The third example concerns use of fluorodeoxyglucose positron emission tomography (FDG PET) as a guide to the decision to perform a breast biopsy on a patient with either a palpable mass or an abnormal mammogram.¹⁶ Only patients with a positive PET scan would be referred for biopsy. Table 2-1 shows the initial ambiguous claim, lacking PICOTS specifications such as the way in which testing would be done. The utility of the analytic framework was limited as several possible testing strategies might be relevant and that is not represented explicitly in an analytic framework. A decision tree was then constructed (Figure 2-5). The testing strategy in the lower portion of the decision tree entails performing biopsy in all patients, while the triage strategy uses a positive PET finding to rule in a biopsy and a negative PET finding to rule out a biopsy. The decision tree helps us to see that the key accuracy index is negative predictive value: the proportion of negative PET results that are truly negative. The tree also reveals that the key contrast in outcomes involves any harms of delaying treatment for undetected cancer when PET is falsely negative versus the benefits of safely avoiding adverse effects of the biopsy when PET is truly negative. The review concluded that there is no net beneficial impact on outcomes when PET is used as a triage test to select patients for biopsy among those with a palpable breast mass

or suspicious mammogram. Thus, estimates of negative predictive values suggest that there is an unfavorable trade-off between avoiding the adverse effects of biopsy and delaying treatment of an undetected cancer.

Figure 2-5. Triage test example: positron emission tomography (PET) to decide whether to perform breast biopsy among patients with a palpable mass or abnormal mammogram*



* Figure taken from Samson et al., 2002.¹⁶

This case illustrates when a more formal decision analysis may be useful, specifically when new test has higher sensitivity but lower specificity than the old test, or vice versa. Such a situation entails tradeoffs in relative frequencies of true positives, false negatives, false positives, and true negatives, which decision analysis may help to quantify.

Summary

The immediate goal of a systematic review of a medical test is to evaluate efficiently the relative health impacts of use of the test in a particular context or set of contexts relative to one or more alternative strategies. The ultimate goal is to produce a review that promotes informed decisionmaking.

Key points are:

- Reaching the above-stated goals requires an interactive and iterative process of topic development and refinement aimed at understanding and clarifying the claim for a test. This work should be done in conjunction with the principal users of the review, experts, and other stakeholders.
- The PICOTS typology, analytic framework, simple decision trees, and other organizing frameworks are all tools that can minimize ambiguity, help identify where review resources should be focused, and guide the presentation of results.
- Sometimes it is sufficient to focus only on accuracy studies. For example, diagnostic accuracy may be sufficient when the new test is as sensitive as the old test *and* the new test's value derives from avoiding the old test's adverse effects (i.e., because the new test is safer or less invasive) or higher costs.

References

1. Institute of Medicine, Division of Health Sciences Policy, Division of Health Promotion and Disease Prevention, Committee for Evaluating Medical Technologies in Clinical Use. Assessing medical technologies. Washington, DC: National Academy Press; 1985. Chapter 3: Methods of technology assessment. p. 80-90.
2. Matchar DB, Patwardhan M, Sarria-Santamera A, et al. Developing a Methodology for Establishing a Statement of Work for a Policy-Relevant Technical Analysis. Technical Review 11. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025.) AHRQ Publication No. 06-0026. Rockville, MD: Agency for Healthcare Research and Quality. January 2006. Available at: <http://www.ahrq.gov/downloads/pub/evidence/pdf/statework/statework.pdf>. Accessed October 4, 2010.
3. Sarria-Santamera A, Matchar DB, Westermann-Clark EV, et al. Evidence-based practice center network and health technology assessment in the United States: bridging the cultural gap. *Int J Technol Assess Health Care* 2006;22(1):33-8.
4. Patwardhan MB, Sarria-Santamera A, Matchar DB, et al. Improving the process of developing technical reports for health care decision makers: using the theory of constraints in the evidence-based practice centers. *Int J Technol Assess Health Care* 2006;22(1):26-32.
5. Woolf SH. An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore SR, Siegel RA, editors. *Methodology perspectives: clinical practice guideline development*. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1994. p. 105-13.
6. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. *J Clin Epidemiol* 2010;63(5):484-90.

7. 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(3 Suppl):21-35.
8. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. *Med Decis Making* 2009;29(5):E1-E12. Epub 2009 Sep 22.
9. Feussner JR, Matchar DB. When and how to study the carotid arteries. *Ann Intern Med* 1988;109(10):805-18.
10. Blakeley DD, Oddone EZ, Hasselblad V, et al. Noninvasive carotid artery testing. A meta-analytic review. *Ann Intern Med* 1995;122(5):360-7.
11. Lord SJ, Irwig L, Simes J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? *Ann Intern Med* 2006;144(11):850-5.
12. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. *Med Decis Making* 2009;29(5):E13-21.
13. 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(11):1116-22.
14. Blue Cross and Blue Shield Association Technology Evaluation Center (BCBSA TEC). Full-field digital mammography. Volume 17, Number 7, July 2002.
15. Seidenfeld J, Samson DJ, Rothenberg BM, et al. HER2 Testing to Manage Patients With Breast Cancer or Other Solid Tumors. Evidence Report/Technology Assessment No. 172. (Prepared by Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center, under Contract No. 290-02-0026.) AHRQ Publication No. 09-E001. Rockville, MD: Agency for Healthcare Research and Quality. November 2008. Available at: www.ahrq.gov/downloads/pub/evidence/pdf/her2/her2.pdf. Accessed July 21, 2010.
16. Samson DJ, Flamm CR, Pisano ED, et al. Should FDG PET be used to decide whether a patient with an abnormal mammogram or breast finding at physical examination should undergo biopsy? *Acad Radiol* 2002;9(7):773-83.