REVIEW

Chapter 6: Assessing Applicability of Medical Test Studies in Systematic Reviews

K. E. Hartmann, MD, PhD[1,2,6], D. B. Matchar, MD[3,4], and S. Chang, MD, MPH[5]

[1]Vanderbilt AHRQ Evidence-based Practice Center, Vanderbilt University, Nashville, TN, USA; [2]Institute for Medicine and Public Health, Vanderbilt University, Nashville, TN, USA; [3]Center for Clinical Health Policy Research, Duke University, Durham, NC, USA; [4]Duke University Medical Center, Durham, NC, USA; [5]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD, USA; [6]Obstetrics & Gynecology and Medicine, Vanderbilt University School of Medicine, Vanderbilt University, Nashville, TN, USA.

Use of medical tests should be guided by research evidence about the accuracy and utility of those tests in clinical care settings. Systematic reviews of the literature about medical tests must address applicability to real-world decision-making. Challenges for reviews include: (1) lack of clarity in key questions about the intended applicability of the review, (2) numerous studies in many populations and settings, (3) publications that provide too little information to assess applicability, (4) secular trends in prevalence and the spectrum of the condition for which the test is done, and (5) changes in the technology of the test itself. We describe principles for crafting reviews that meet these challenges and capture the key elements from the literature necessary to understand applicability.

KEY WORDS: systematic evidence review; diagnostic test; screening test; prognostic test; applicability.
J Gen Intern Med 27(Suppl 1):S39–46
DOI: 10.1007/s11606-011-1961-9

INTRODUCTION

Most systematic reviews are conducted for a practical purpose: to support clinicians, patients, and policy makers—decision makers—in making informed decisions. To make informed decisions about medical tests, whether diagnostic, prognostic or those used to monitor the course of disease or treatment, decision makers need to understand whether a test is worthwhile in a specific context. For example, decision makers need to understand whether a medical test has been studied in patients and care settings similar to those in which they are practicing, and whether the test has been used as part of the same care management strategy that they plan to use. They may also want to know whether a test is robust over a wide range of scenarios for use or relevant only to a narrow set of circumstances.

Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers (EPC) review scientific literature on topics including clinical care and medical tests to produce evidence reports and technology assessments to inform coverage decisions, quality measures, educational materials and tools, guidelines, and research agendas. The EPCs use four principles for assessing and reporting applicability of individual studies and the overall applicability of a body of evidence. These principles may provide a useful framework for other investigators conducting systematic review of medical tests:[1]

- Determine the most important factors that affect applicability
- Systematically abstract and report key characteristics that may affect applicability
- Make and report judgements about major limitations to applicability of individual studies
- Consider and summarize the applicablity of the body of evidence

Comprehensive information about the general conduct of reviews is available in the AHRQ Evidence-based Practice Center Methods Guide for Comparative Effectiveness Reviews.[2] In this report we highlight common challenges in reviews of medical tests and suggest strategies that enhance interpretation of applicability.

COMMON CHALLENGES

Key Questions Lack Clarity. Key questions guide the presentation, analysis, and synthesis of data, and thus the ability to judge applicability. Key questions should provide a clear context for determining the applicability of a study. Lack of specificity in key questions can result in reviews of larger scope than necessary, failure to abstract relevant study features for evidence tables, less useful organization of summary tables, disorganized synthesis of results, and findings from meta-analysis that do not aggregate data in crucial groupings. In addition, key questions that do not distinguish the management context in which the test is being used can introduce misinterpretations of the literature. A common scenario for such confusion is when the research

compares the accuarcy of a new test to another test (i.e., as a replacement), but in reality, the test is proposed to be used as a triage test to guide further testing or as an add-on after another test.

If relevant contextual factors are not stipulated in the key questions, it also hinders decisions during the review process. Which studies should be included and which excluded? If the patient population and care setting are not explicitly described, the default can be to broadly lump all contexts and uses of the test together. However, decisions to "lump" or "split" must be carefully considered and justified. Inappropriate lumping without careful consideration of subgroups that should be analyzed separately may result in oversimplification. Decisions about meaningful subgroupings, for instance by age of participants, by setting (hopsital versus ambulatory), or version of the test, should be made in advance.

Conducting subgroup analyses after appraising the included studies may introduce type 1 error from *a posteriori* biases in interpretation, making it difficult to distinguish whether identified effects are spurious or real. Decisions in advance to split reporting of results for specific subgroups and contexts should be carefully considered and justified. Decisions should be based on whether there is evidence that a particular contextual factor is expected to influence the performance characteristics of the test or its effectiveness as a componenet of care.

Studies Are Not Specific to the Key Questions. When there is appropriate justification to "split" a review so that key questions or subquestions relate to a specific population, setting, or management strategy, the studies identified for inclusion may not reflect the same subgroups or comparisons identified in the key questions. The reviewer is faced with deciding when these deviations from ideal are minor, and when they are more crucial and are likely to affect test performance, clinical decision-making, and health outcomes in some significant way. The conduct and synthesis of the findings will require a method to track and describe how the reviewers dealt with two types of mismatches: (1) literature from other populations and contexts that does not directly address the intended context of the key question; and (2) studies that do not provide sufficient information about context to determine if they apply. Annotation througout the review, in tables and synthesis, can then note if these types of mismatch apply, how common they were, and what the expected impact is on interpreting applicability.

Tests Are Rapidly Evolving. A third challenge, especially relevant to medical tests, is that, even more than treatments, tests often change rapidly, in degree (enhancements in existing technologies), type (substantively new technologies), or target (new molecular targets). The literature often contains evidence about tests that are not yet broadly available or are no longer common in clinical use. Secular trends in use patterns and market forces may shape applicability in unanticipated ways. For instance, suppose that a test is represented in the literature by dozens of studies that report on a version that provides dichotomous, qualitative results (present versus absent), and that the company marketing the test subsequently announces production of a new version that provides only a continuous, quantitative measure. Or genetic tests for traits may evolve from testing for a single-nucleotide polymorphisms to determining the gene sequence. In these situations, reviewers must weigh how best to capture data relating the two versions of the test and decide whether there is value in reviewing the obsolete test to provide a point of reference for expectations about whether the replacement test has any merit or whether reviewing only the more limited, newer data better addresses the key question for contemporary practice.

PRINCIPLES FOR ADDRESSING CHALLENGES

The root cause of these challenges is that test accuracy, as well as more distal effects of test use, is often highly sensitive to context. Therefore, the principles noted here relate to clarifying context factors and, to the extent possible, using that clarity to guide study selection (inclusion/exclusion), description, analysis, and summarization. In applying the principles described below, the PICOTS typology can serve as a framework for assuring relevant factors have been systematically assessed (see Table 1).[3,4]

Principle 1: Identify Important Contextual Factors. In an ideal review, all possible factors related to the impact of a test use on health outcomes should be considered. However, this is usually not practical, and some tractable list of factors must be considered before initiating a detailed review. Consider factors that could affect the causal chain of direct relevance to the key question: for instance, in assessing the accuracy of cardiac MRI for detecting atherosclerosis, slice thickness is a relevant factor in assessing applicability. It is also important to consider applicability factors that could affect a later link in the causal chain (e.g., for lesions identified by cardiac MRI vs. angiogram, what factors may impact the effectiveness of treatment?).

In pursuing this principle, consider contextual issues that are especially relevant to tests, such as patient populations, management strategy, time effects, and secular trends:

Spectrum Effect. The severity or type of disease may effect the accuracy of the test. For example, cardiac MRI tests may be generally accurate at identifying cardiac anatomy and functionality, but certain factors may affect the test

Table 1. Using the PICOTS Framework to Assess and Describe Applicability of Medical Tests*

PICOTS element	Potential characteristics to describe and assess	Challenges when assessing studies	Example	Potential systematic approaches for decisions
Population	▪ Justification for lumping or splitting key questions ▪ Method of identification/selection ▪ Inclusion & exclusion criteria o for the review ▪ Demographic characteristics of those included in review ▪ Prevalence of condition in practice and in studies ▪ Spectrum of disease in practice and in studies	▪ Source of population not described ▪ Study population poorly specified ▪ Key characteristics not reported ▪ Unclear whether test performance varies by population	Education/literacy level not reported in study of pencil-and-paper functional status assessment	Exclude *a priori* if key element crucial to assessing intended use case is missing Or include but: – Flag missing elements in tables/text – Organize data within key questions by presence/absence of key elements – Include presence/absence as parameter in meta-regression or sensitivity analyses – Note need for challenge to be addressed in future research
Intervention	▪ Version of test used in practice and in studies ▪ How and by whom tests are conducted in practice and in studies ▪ Cutoff/diagnostic thresholds applied in practice and in studies ▪ Skill of assesors when interpretation of test required in studies	Version/ instrumentation not specified Training/quality control not described Screening and diagnostic uses mixed	Ultrasound machines and training of sonographers not described in study of fetal nuchal translucency assessment for detection of aneuploidy	Exclude a priori if version critical and not assessed Or include but: – Contact authors for clarification – Flag version of test or deficits in reporting in tables/text – Discuss implications – Model cutoffs and conduct sensitivity analyses
Comparator	▪ Use of gold standard vs. "alloy" standard in studies ▪ Alternate or "usual" test used in the studies ▪ How test is used as part of management strategy (e.g., triage, replacement, or add-on) in practice and in studies ▪ In trials is comparator no testing vs. usual care with ad hoc testing	Gold standard not applied Correlational data only	Cardiac CT compared with stress treadmill without use of angiography as a gold standard	Exclude a priori if no gold standard Or include but: – Restrict to specified comparators – Group by comparator in tables/ text
Outcome of use of the test	▪ How accuracy outcomes selected for review relate to use in practice: ▪ Accuracy of disease status classification • Sensitivity/specificity • Predictive values • Likelihood ratios • Diagnostic odds ratio • Area under curve • Discriminant capacity	Failure to test "normals," or subset, with gold standard Precision of estimates not provided Tests used as part of management strategy in which exact diagnosis is less important than "ruling out" a disease	P-value provided for mean of continuous test results by disease status but confidence bounds not provided for performance characteristics	Exclude a priori if test results cannot be mapped to disease status (i.e., 2 × 2 or other test performance data cannot be extracted) Exclude if subset of "normals" not tested Or include but: – Flag deficits in tables/text – Discuss implications – Assess heterogeneity in meta-analysis and comment of sources of heterogeneity in estimates
Clinical Outcomes from test results	▪ How studies addressed clinical outcomes selected for the review: • Earlier diagnosis • Earlier intervention • Change in treatment given • Change in sequence of other testing • Change in sequence/intensity of care • Improved outcomes, quality of life, costs, etc.	▪ Populations and study designs of included studies heterogeneous with varied findings ▪ Data not stratified or adjusted for key predictors	Bone density testing reported in relation to fracture risk reduction without consideration of prior fracture or adjustment for age	Exclude if no disease outcomes and outcomes key to understanding intended use case Or include and: – Document details of deficits in tables/text – Discuss implications – Note need for challenge to be addressed in future research

Table 1. (continued)

PICOTS element	Potential characteristics to describe and assess	Challenges when assessing studies	Example	Potential systematic approaches for decisions
Timing	▪ Timing of availability of results to care team in studies and how this might relate to practice ▪ Placement of test in the sequence of care (e.g., relationship of test to treatment or follow-on management strategies) of studies and how this might relate to practice ▪ Timing of assessment of disease status and outcomes in studies	▪ Sequence of use of other diagnostics unclear ▪ Time from results to treatment not reported ▪ Order of testing varies across subjects and was not randomly assigned	D-dimer studies in which it is unclear when results were available relative to DVT imaging studies	Exclude if timing/sequence is key to understanding intended use case Or include and: – Contact authors for information – Flag deficits in tables/text – Discuss implications – Note need for challenge to be addressed in future research
Setting	▪ How setting of test in studies relate to key questions and current practice: • Primary care vs. specialty care • Hospital-based • Routine processing vs. specialized lab or facility • Specialized personnel • Screening vs. diagnostic use	▪ Resources available to providers for diagnosis and treatment of condition vary widely ▪ Provider type/specialty vary across settings ▪ Comparability of care in international settings unclear	Diagnostic evaluation provided by geriatricians in some studies and unspecified primary care providers in others	Exclude if care setting known to influence test/outcomes or if setting is key to understanding intended use case Or include but: – Document details of setting – Discuss implications

**Abbreviations: CT= computed tomography; DVT= deep venous thrombosis*

performance, such as arrythmias, location of the lesion, or obesity. Reviews must identify these factors ahead of time and justify when to "split" questions or to conduct subgroup analyses.

Tests as Part of a Management Strategy. Studies on cardiac MRI often select patients with a relatively high pre-test probability of disease (i.e., presumably pre-screened with other non-invasive testing such as stess EKG) and evaluate the diagnostic accuracy when compared to a gold standard of x-ray coronary angiography. However, the test performance under these conditions does not necessarily apply when used in patients with lower pre-test probability of disease, such as when screening patients with no symptoms or when used as an initial triage test (i.e., compared to stress EKG) rather than an add-on test after initial screening. It is important for reviewers to clarify and distinguish the conditions in which the test is studied and in which it is likely to be used.

Methods of the Test Over Time. Diagnostics, like all technology, evolve rapidly. For example, MRI slice thickness has fallen steadily over time, allowing resolution of smaller lesions. Thus, excluding studies with older technologies and presenting results of included studies by slice thickness may both be appropriate. Similarly, antenatal medical tests are being applied earlier and earlier in gestation, and studies of test performance would need to be examined by varied cutoffs for stages of gestation, and genetic tests are evolving from detection of specific polymorphisms to full gene sequences. Awareness of these changes should guide review parameters such as date range selection and eligible test type for the included literature to help categorize findings and discussion of results.

Secular Trends in Population Risk and Disease Prevalence. Direct and indirect changes in the secular setting (or differences across cultures) can influence medical test performance and applicability of related literature. As an example, when examining the value of screening tests for gestational diabetes, test performance is likely to be affected by the average age of pregnant women, which has risen by more than a decade over the past 30 years, and by the proportion of the young female population that is obese, which has also risen steadily. Both conditions are associated with risk of type II diabetes. As a result, we would expect the underlying prevalence of undiagnosed type II diabetes in pregnancy to be increased, and the predictive values and cost-benefit ratios of testing, and even the sensitivity and specificity in general use, to change modestly over time.

Secular trends in population characteristics can have indirect effects on applicability when population characteristics change in ways that influence ability to conduct the test. For example, obesity diminishes image quality in tests, such as ultrasound for diagnosis of gallbladder disease or fetal anatomic survey, and MRI for detection

of spinal conditions or joint disease. Since studies of these tests often restrict enrollment to persons with normal body habitus, current population trends in obesity mean that such studies exclude an ever-increasing portion of the population. As a result, clinical imaging experts are concerned that these tests may not perform in practice as described in the literature because the actual patient population is significantly more likely to be obese than the study populations. Expert guidance can identify such factors to be considered.

Prevalence is inexorably tied to disease definitions that may also change over time. Examples include: (1) criteria to diagnose acquired immune deficiency syndrome (AIDS), (2) the transition from cystometrically defined detrusor instability or overactivity to the symptom complex "overactive bladder," and (3) the continuous refinement of classifications of mental health conditions recorded in the *Diagnostic and Statistical Manual* updates.[5] If the diagnostic criteria for the condition change, the literature may not always capture such information; thus, expert knowledge with a historical vantage point can be invaluable.

Routine Preventive Care over Time. Routine use of a medical test as a screening test might be considered an indirect factor that alters population prevalence. As lipid testing moved into preventive care, the proportion of individuals with cardiovascular disease available to be diagnosed for the first time with dyslipidemia and eligible to have the course of disease altered by that diagnosis has changed. New vaccines, such as the human papilloma virus (HPV) vaccine to prevent cervical cancer, are postulated to change the distribution of viral subtypes in the population and may influence the relative prevalence of subtypes circulating in the population. As preventive practices influence the natural history of disease, such as increasing proportions of a population receiving vaccine, they also change the utility of a medical test, like that for HPV detection. Knowledge of preventive care trends is an important component of understanding current practice to consider as a backdrop when contextualizing the applicability of a body of literature.

Treatment Trends. As therapeutics arise that change the course of disease and modify outcomes, literature about the impact of diagnostic tools on outcomes requires additional interpretation. For example, the implications of testing for carotid arterial stenosis are likely changing as treatment of hypertension and the use of lipid-lowering agents have improved.

We suggest two steps to ensure that data about populations and subgroups are uniformly collected and useful. First, refer to the PICOTS typology[3,4] (see Table 1) to identify the range of possible factors that might affect applicability and consider the hidden sources of limitations noted above. Second, review the list of applicability factors with stakeholders to ensure common vantage points and identify any hidden factors specific to the test or history of its development that may influence applicability. Features judged by stakeholders to be crucial to assessing applicability can then be captured, prioritized, and synthesized in the process of designing the process and abstracting data for an evidence review.

Principle 2: Be Prepared to Deal with Additional Factors Affecting Applicability. Despite best efforts, some contextual factors relevant to applicability may only be uncovered after a substantial volume of literature has been reviewed. For example, in a meta-analysis, it may appear that a test is particularly inaccurate for older patients, although age was never considered explicitly in the key questions or in preparatory discussions with an advisory committee. It is crucial to recognize that like any relationship discovered *a posteriori*, this may reflect a spurious association. In some cases, failing to consider a particular factor may have been an oversight; in retrospect, the importance of that factor on the applicability of test results may be physiologically sensible and supported in the published literature. Although it may be helpful to revisit the issue with an advisory committee, when in doubt, it is appropriate to comment on an apparent association and clearly state that it is a hypothesis, not a finding.

Principle 3: Justify Decisions to "Split" or Restrict the Scope of a Review. In general, it may be appropriate to restrict a review to specific versions of the test, selected study methods or types, or populations most likely to be applicable to the group(s) whose care is the target of the review such as a specific group (e.g., people with arthritis, women, obese patients) or setting (e.g., primary care practice, physical therapy clinics, tertiary care neonatal intensive care units). These restrictions may be appropriate (1) when all partners are clear that a top priority of a review is applicability to a particular target group or setting, (2) when there is evidence that test performance in a specific subgroup differs from the test performance in the broader population or setting or that a particular version of the test performs differently than the current commonly used version. Restriction of reviews is efficient when all partners are clear that a top priority of a review is applicability to a particular target group or setting. Restriction can be more difficult to accomplish when parties differ with respect to the value they place on less applicable but nonetheless available evidence. Finally, restriction is not appropriate when fully comprehensive summaries including robust review of limitations of extant literature are desired.

Depending on the intent of the review, restricting the review during the planning process to include only specific

versions of the test, selected study methods or types, or populations most likely to be applicable to the group(s) whose care is the target of the review may be warranted. For instance, if the goal of a review is to understand the risks and benefits of colposcopy and cervical biopsies in teenagers, the portion of the review that summarizes the accuracy of cervical biopsies for detecting dysplasia might be restricted to studies that are about teens; that present results stratified by age; or that include teens, test for interaction with age, and find no effect. Alternatively, the larger literature could be reviewed with careful attention to biologic and health systems factors that may influence applicability to young women.

In practice, we often use a combination of inclusion and exclusion criteria based on consensus along with careful efforts to highlight determinants of applicability in the synthesis and discussion. Decisions about the intended approach to the use of literature that is not directly applicable need to be tackled early to ensure uniformity in review methods and efficiency of the review process. Overall, the goal is to make consideration of applicability a prospective process that is attended to throughout the review and not a matter for *post hoc* evaluation.

Principle 4: Maintain a Transparent Process. As a general principle, reviewers should address applicability as they define their review methods and document their decisions in a protocol. For example, time-varying factors should prompt consideration of using timeframes as criteria for inclusion or careful descriptions and analyses as approprite of the possible impact of thes effects on applicability.

Transparency is essential, particularly when a review decision may be controversial. For example, after developing clear exclusion criteria based on applicability, a reviewer may find themselves "empty-handed." In retrospect, experts—even those accepting the original exclusion criteria—may decide that some excluded evidence may indeed be relevant by extension or analogy. In this event, it may be appropriate to include and comment on this material, clearly documenting how it may not be directly applicable to key questions, but represents the limited state of the science.

AN ILLUSTRATION

Our work on the 2002 Cervical Cancer Screening Summary of the Evidence for the US Preventive Services Task Force[6] illustrates several challenges and principles at work: the literature included many studies that did not use gold standards or testing of normals, and many did not relate cytologic results to final histopathologic status. We encountered significant examples of changes in secular trends and availability and format of medical tests: liquid-based cervical cytology was making rapid inroads into practice; resources for reviewing conventional Pap smear testing were under strain from a shortage of cytotechnologists in the workforce and from restrictions on the volume of slides they could read each day; several new technologies had entered the market designed to use computer systems to pre- or postscreen cervical cytology slides to enhance accuracy; and the literature was beginning to include prospective studies of adjunct use of HPV testing to enhance accuracy or to triage which indiviudals needed evaluation with colposcopy and biopsies to evaluate for cervical dysplasia and cancer. No randomized controlled trials (RCTs) were available using, comparing, or adding new tests or technologies to prior conventional care.

Because no data were available comparing the effects of new screening tools or strategies on cervical cancer outcomes, the report focused on medical test characteristics (sensitivity, specificity, predictive values, and likelihood ratios), reviewing three computer technologies, two liquid cytology approaches, and all methods of HPV testing. Restricting the review to techologies available in the United States, and therefore most applicable, would have reduced the scope substantially. Including all the technologies to determine if there were clear differences among techniques made clear whether potentially comparable or superior methods were being overlooked or no longer offered, but may have also unnecessarily complicated the findings. Only in retrospect, after the decision to include all tests was made and the review conducted, were we able to see that this approach did not substantially add to understanding the findings because the tests that were no longer available were not meaningfully superior.

Although clearly describing the dearth of information available to inform decisions, the review was not able to provide needed information. As a means of remediation, not planned in advance, we used prior USPSTF meta-analysis data on conventional Pap medical test performance[7], along with the one included paper about liquid cytology[8], to illustrate the potential risk of liquid cytology overburdening care systems with detection of low-grade dysplasia while not substantively enhancing detection of severe disease or cancer.[9] The projections from the report have since been validated in prospective studies.

For two specific areas of applicability interest (younger and older age, and hysterectomy status), we included information about underlying incidence and prevalence in order to provide context, as well as to inform modeling efforts to estimate the impact of testing. These data helped improve understanding the burden of disease in the subgroups compared with other groups, and improve understanding about the yield and costs of screening in the subgroups compared with others.

SUMMARY

Review teams need to familiarize themselves with the availability, technology, and contemporary clinical use of the test they are reviewing. They should consider current treatment modalities for the related disease condition, the potential interplay of the disease severity and performance characteristics of the test, and the implications of particular study designs and sampling strategies for bias in the findings about applicability.

As examples throughout this report highlight, applicability of a report can be well served by restricting inclusion of marginally related or outdated studies. Applicability is rarely enhanced by uncritically extrapolating results from one context to another. For example, we could not estimate the clinical usefulness of HPV testing among older women from trends among younger women. In the design and scoping phase for a review, consideration of the risks and advantages of restricting the scope or excluding publications with specific types of flaws, benefits from explicit guidance from clinical, medical testing, and statistical experts about applicability challenges.

Often the target of interest is intentionally large—for example, all patients within a health system, a payer group such as Medicare, or a care setting such as a primary care practice. Regardless of the path taken—exhaustive or narrow—the review team must take care to group findings in meaningful ways. For medical tests, this means gathering and synthesizing data in ways that enhance ability to readily understand applicability. Grouping summaries of the findings using familiar structures like PICOTS can enhance how clearly the applicability issues are framed, for instance grouping results by the demographics of the population included: all women, women and men, by the intervention, grouping together studies that used the same version of the test, or by outcomes, grouping together those studies that report an intermediate marker versus those that measured the actual outcome of interest. This may mean that studies are presented within the review more than once, grouping findings along different "applicability axes" to provide the clearest possible picture.

Since most systematic reviews are conducted for the practical purpose of supporting informed decisions and optimal care, keeping applicability in mind from start to finish is an investment bound to pay off in the form of a more useful review. The principles summarized in this review can assure valuable aspects of weighing applicability are not overlooked and that review efforts support evidence-based practice.

KEY POINTS

- Early in the review planning process, systematic reviewers should identify important contextual factors that may affect test performance (Table 1).
- Reviewers should carefully consider and document justification for how these factors will be addressed in the review—whether through restricting key questions or from careful assessment, grouping, and description of studies in a broader review.
  - A protocol should clearly document which populations or contexts will be excluded from the review and how the review will assess subgroups.
  - Reviewers should document how they will address challenges in including studies that may only partly fit with the key questions or inclusion/exclusion criteria, or that poorly specify the context.
- The final systematic review should include a description of the test's use in usual practice and care management and how the studies fit with the usual practice.

Acknowledgments: *This paper is based on the experiences of the EPC program in conducting systematic reviews of medical tests and on the proceedings of a working meeting held at AHRQ in 2008 (white papers published here and in MDM: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=350). We are grateful to our many peers across the AHRQ and EPC leadership who have sustained conversations about best practices and have continually advanced the methods for review of medical tests and their applicability.*

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Disclaimer: *This project was funded under contract no. 290-2007-10065-I and 290-2007-10066-I from the Agency for Healthcare Research and Quality, US Department of Health and Human Services. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services.*

Corresponding Author: *Katherine E. Hartmann, MD, PhD; Obstetrics Gynecology and Medicine, Vanderbilt University School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37203-8291, USA (e-mail: katherine.hartmann@vanderbilt.edu).*

REFERENCES

1. Agency for Healthcare Research and Quality (US). Methods guide for medical test reviews. Available at: http://www.effectivehealthcare.ahrq.gov/tasks/sites/ehc/assets/File/methods_guide_for_medical_tests.pdf. Accessed November 8, 2011.
2. Agency for Healthcare and Research Quality (US). Methods reference guide for effectiveness and comparative effectiveness reviews. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=318. Accessed November 8, 2011.
3. **Matcher DB.** Introduction to the Methods Guide for Medical Test Reviews. J Gen Intern Med. 2011; doi:10.1007/s11606-011-1798-2.

4. **Samson D.** Schoelles KM. Chapter 2: Medical Tests Guidance (2) Developing the Topic and Structuring Systematic Reviews of Medical Tests: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks. J Gen Intern Med. 2012; doi:10.1007/s11606-012-2007-7.
5. American Psychiatric Association. Task Force on DSM-IV. Diagnostic and statistical manual of mental disorders: DSM-IV-TR. 4th ed. Washington: American Psychiatric Association; 2000.
6. **Hartmann KE, Hall SA, Nanda K, Boggess JF, Zolnoun D.** Screening for cervical cancer. Available at: http://www.ahrq.gov/downloads/pub/prevent/pdfser/cervcanser.pdf. Accessed November 8, 2011.
7. **McCrory DC, Matchar DB, Bastian L, et al.** Evaluation of cervical cytology. Evid Rep Technol Assess (Summ). 1999;5:1–6.
8. **Hutchinson ML, Zahniser DJ, Sherman ME, et al.** Utility of liquid-based cytology for cervical carcinoma screening: results of a population-based study conducted in a region of Costa Rica with a high incidence of cervical carcinoma. Cancer. 1999;87(2):48–55.
9. **Hartmann KE, Nanda K, Hall S, Myers E.** Technologic advances for evaluation of cervical cytology: is newer better? Obstet Gynecol Surv. 2001;56(12):765–774.