Methods Guide for Medical Test reviews

Paper 6

Assessing Applicability

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

DRAFT

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.

DRAFT

Paper 6. Assessing Applicability

Reviews of medical test studies, like most systematic reviews, are conducted for a practical purpose: to support clinicians, patients, and policymakers—decisionmakers—in making informed decisions. To make informed decisions, decisionmakers need not merely to understand whether a particular test strategy is worthwhile in some context, but whether it is worthwhile in the specific context—patient and disease characteristics, downstream management options, setting, and so on—relevant to their particular decisions. Is this test robust over a wide range of patient types and scenarios of use, or is it relevant only to a narrow set of circumstances?

As systematic reviewers, we approach this concern of decisonmakers about the applicability of *tests* to *contexts* by focusing our attention on the applicability of *evidence about tests* to one or more *key questions,* in which context is (or should be) stipulated (*see* Paper 2). To the extent that available evidence is applicable to a particular context of interest, and the evidence is supportive of the use of the test in that context, then the test is also applicable to that context.

AHRQ's *General Methods Guide*[1] describes four steps for assessing and reporting applicability of individual studies and of a body of evidence. These steps are relevant for evaluating medical tests, but with some important considerations, which we highlight here.

As with assessing interventions, it is useful to distinguish the assessment of applicability, or external validity, from that of quality, often considered synonymous with internal validity.[1] Consider, as is often the case, that the accuracy of a medical test is highly sensitive to the severity of disease—that is, there is a spectrum *effect*.[2] Imagine a study that purports to address the accuracy of a particular test for sepsis in a general population of individuals with abnormal white blood cell count, but in which only extreme subjects with white blood cell counts higher than twice normal are chosen for study. In this study, the selection of subjects introduces a spectrum *bias*, leading to estimates of operating characteristics that are likely to have low external validity, or low applicability to the "general" population of individuals with abnormal white blood cell counts;[3] in many reviews, such studies may be deemed to be of low "quality" (*see* Paper 5). However, in reality, the assessment of the test may be valid for a particular context, and the task at hand is to describe the context so that the utility of the test is clearly understandable for the end-users. As recommended when assessing applicability for systematic reviews on interventions,[1] applicability should be reported separately from quality or strength of evidence, and should not be assessed by a universal rating scheme.

Applicability is relevant to all elements in the causal chain in which test use is purportedly linked to health outcomes. We may ask, as for stress treadmill testing, "Is this test equally accurate for diagnosing coronary heart disease in women as in men?," or "How often does a thyrotropin-stimulating hormone test identify a thyroid condition that needs treatment among individuals with a complaint of fatigue across the age spectrum?" We may also ask about the applicability of studies that speak to more distal steps in the causal chain, such as, "Does fetal fibronectin testing of women with preterm labor symptoms (versus in routine clinical care) change the care they receive?" Or, for an appropriate clinical study, we may ask about its applicability to an overarching question, such as, "Does use of PET scanning for individuals with mild cognitive impairment lead to better survival or quality of life?" In this document, we focus on assessing

applicability of studies of test accuracy since these are the outcomes that are most unique to medical tests; however, the principles are relevant to all links of the causal chain, as well as to an overarching question about the value of testing on overall health outcomes.

Common Challenges

Evaluating the applicability of a body of evidence regarding medical tests presents several common challenges to reviewers.

Lack of Clarity in Key Questions

Early formulations of key questions may not provide clear context for determining the applicability of a study. For example, for test accuracy questions, not all potentially relevant contextual factors are stipulated in the key questions. This complicates further decisions in the review process. First, which studies should be included and which excluded? The issue typically arises either because the study evaluated a context different from that described in the key question, or because the study failed to provide sufficient information to assess applicability. 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 decisionmaking, and health outcomes in some significant way.

Another point in the review process where lack of clarity in key questions affects assessments of applicability is in the presentation, analysis, and summarization of data. What study features should be included in evidence tables? How should aggregate tables and meta-analyses approach studies that describe tests in somewhat different contexts? The challenge of when to "lump" and when to "split" is evident in the complexity of the evidence concerning lipid and triglyceride screening for prevention of cardiovascular disease. While it is possible to compile data relating cholesterol subtypes and total levels to health risks, we know clinically that they are interrelated and that effects may vary with age and gender. The reviewer must consider how to organize the data by important clinical or policy contexts so as to maximize the relevance of the findings to decisionmakers.

Review Covers a Wide Range of Contexts

A second common challenge faced by reviewers is that the key questions relate to the wide range of potential applications of one or more tests singly or in various combinations, and across multiple populations and settings. In addition to the problems noted above, when the review is intended to be broad, it is tempting to "let the evidence speak" by, for example, examining test accuracy as a function of the numerous possibly relevant contextual factors. The problem with this approach is that it introduces a potential bias in that some factors that appear to influence test accuracy or subsequent outcomes may actually be spurious. This is similar to the challenge of evaluating efficacy of treatments in subgroups and is addressed in the *General Methods Guide*.[1]

Tests are Rapidly Evolving

A third major challenge to assessing applicability especially relevant to medical tests is that, even more that treatments, tests are often changing rapidly, both in degree (enhancements in existing technologies) and type (incorporate substantively new technologies). 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 that it will produce a new version that provides only a continuous, quantitative measure. In this situation, reviewers must weigh how best to capture data relating the two versions of the test and decide whether there is merit 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.

Principles for Addressing the Challenges

The root cause of the challenges above 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, and analysis/summarization. In applying the principles described below, the PICOTS typology can be useful (*see* Papers 1 and 2).

Principle 1: Identify Important Contextual Factors

In an ideal review, all possible factors related to impact of a test use on health outcomes should be considered. However, this is usually not practical, and so some tractable list of factors must be considered before initiating a detailed review. First, consider factors that may affect the step in the causal chain of direct relvance to the key question (e.g., for assessing accuracy of cardiac MRI for atherosclerosis, slice thickness is a relevant factor in assessing applicability). However, even if the key question relates only to test accuracy, it is also important to consider factors that may 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 Principle 1, consider contextual issues that are especially relevant to tests. Factors that may be less obvious but of significant importance in the interpretation of medical tests are time effects and secular trends. Factors that may be especially relevant are listed below.

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, and 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, for example, would need to be examined by gestational dates and varied cut-offs for those stages in gestation. Awareness of these changes

guides review parameters such as date range selection and eligible test type for the included literature helps frame categorization and discussion of the review 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 the population characteristic changes 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 criteria to diagnose acquired immune deficiency syndrome (AIDS), the transition from the cystometrically defined condition detrusor instability or overactivity to the symptom complex "overactive bladder," and the continuous refinement of classifications of mental health conditions recorded in the *Diagnostic and Statistical Manual* updates. If the condition being tested for changes criteria, 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. Preventive care is an important component of understanding current practice to consider as a backdrop when contextualizing the applicability of a body of literature.

DRAFT

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 is 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 (*see* Papers 1 and 2) 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.

Core characteristics for assessing and describing the applicability of medical tests are organized by the elements of PICOTS in Table 6-1. This often requires generous use of the "NR" entry ("not reported") in evidence and summary tables.

Table 6-1. Using the PICOTS framework to assess and describe applicability of medical tests

PICOTS element	Core characteristics to document	Challenges	Example	Potential systematic approaches
Population	Method of identification/selection Inclusion & exclusion criteria Demographic characteristics of those included Prevalence of condition Spectrum of disease detected	Source of population not described Study population poorly specified Key characteristics not reported	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 How conducted By whom Cut-off/diagnostic thresholds applied Skill of assesors when interpretation of test required	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

DRAFT

PICOTS element	Core characteristics to document	Challenges	Example	Potential systematic approaches
				tables/text – Discuss implications – Model cut-offs and conduct sensitivity analyses
Comparator	Gold standard vs "alloy" standard Alternate or "usual" test 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	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	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., 2x2 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	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
Timing	Timing of availability of results to care team Place in the sequence of care Timing of assessment of disease status and outcomes	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

DRAFT

PICOTS element	Core characteristics to document	Challenges	Example	Potential systematic approaches
				– Discuss implications – Note need for challenge to be addressed in future research
Setting	Primary care vs. specialty care Hospital-based Routine processing vs. specialized lab or facility Specialized personnel	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 thromboembolism.

Principle 2: Be Prepared to Deal With Additional Factors Affecting Applicability

Despite best efforts, some contextual factors that appear to be 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 was 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 rises only to the level of a hypothesis.

Principle 3: Restriction of Scope may be Appropriate

An important decision is how to deal with studies that do not directly apply to the context described in the key questions, or when important details that allow the reviewers to assess applicability are missing. For instance, if the goal for a review is to synthesize the literature on the use of individual risk prediction tools for estimating risk of myocardial infarction or stroke among diabetics, the reviewer is faced with the decision whether to systematically exclude studies that were not limited to people with diabetes, or those that did not include some proportion of diabetics, or those that included diabetics but did not publish results stratified by the presence or absence of diabetes. This decision is particularly challenging when restriction leads to a scant body of evidence.

In general, if the review is intended to apply to 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), then excluding studies is appropriate if they (a) fail to address

that characteristic of interest, (b) do not include relevant individuals or settings, (c) do not stratify results by the feature of interest, or (d) do not include analysis of effect measure modification by that characteristic. 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 exclusion based on consensus, and inclusion but 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. 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.

Illustrations

Our work on the 2002 Cervical Cancer Screening Summary of the Evidence for the US Preventive Services Task Force[4] illustrates several of the challenges in determining applicability: the literature included many studies that did not use gold standards or testing of normals; many studies could not relate cytologic results to final histopathologic status; and few data were

available from the target "use case" populations. The evidence review also encountered significant examples of changes in secular trends and availability and format of medical tests.

When the update[4] was intitiated, much had changed since the prior 1996 review: liquid-based cervical cytology was making rapid inroads into practice; resources for reviewing conventional Pap smear tests were under strain from a relative shortage of cytotechnologists in the workforce and 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 the use of HPV testing to enhance accuracy or to triage which indiviudals needed evaluation with colposcopy and biopsies to evaluate for cervical dyplasia and cancer. No randomized controlled trials (RCTs) were available using, comparing, or adding new tests or technologies to prior conventional care. Some USPSTF members were interested in teens and younger women; however, in topic development, the spot searches and expert consensus was that the literature was insufficient to support a specific focus on the subgroup of women in their teens and twenties. Key questions related to stopping screening focused on the subgroups of older women and those after hysterectomy. We chose an exhaustive approach to describe the broadest possible picture of the state-of-the-science in cervical cancer screening, emphasizing the changes outlined above. We also sought to gather data for indirect comparisons among techniques, as well as examining outcomes of screening in two specific subgroups—older women and those who have had hysterectomy.

Because no data were available comparing the effects of new screening tools or strategies on outcomes, the report focused on medical test characteristics, 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. In fact, one of the companies involved had purchased the intellectual property rights of another with the plan to shelve the competitor's technology and reduce competition. This was publicly available knowledge that might have simplified the work and the findings. 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 confused users of the report. Only in retrospect after the decision to include all test was made were we able to see that this approach did not substantially add to understanding the findings because diagnostic characteristics of those tests that were not longer available or available only in restricted settings were not meaningfully superior..

We restricted *a priori* on these requirements: the study reported on tests obtained for screening; results were compared with a colposcopy and/or histology reference standard; the reference standard was assessed within 3 months of screening test, and data from the publication allowed completion of a 2-by-2 table, preferably for each level of dyplasia and presence of cancer. Overwhelmingly, failure to meet criteria resulted from failure to use a reference standard. Most excluded studies used split-sample correlations. Thus, there was little to report beyond the lack of high quality evidence to guide care.

DRAFT

Although clearly describing the dearth of information available to inform decisions, the review was not able to provide more information for care providers and women about expected usefulness of the tests. As a means of remediation, not planned in advance, we used prior USPSTF meta-analysis data on conventional Pap medical test performance,[5] along with the one included paper about liquid cytology,[6] to model the potential implications of its use; specifically, we explored the issue of overburdening care systems with detection of low-grade dysplasia while not substantively enhancing detection of severe disease or cancer.[7] The projections from the report have since been validated in prospective studies. In the time since the 2002 USPSTF Cervical Cancer Screening update,[4] liquid cytology has become the most widely used form of Pap testing, adopted by more than 80 percent of practices within 5 years of its introduction.

On the topic of HPV testing for the virus that causes dysplasia, the we identified 65 articles, of which 13 met the inclusion criteria. We compiled meta-estimates of the test performance characteristics including likelihood ratios and examples of sequential use of tests to demonstrate the potential influence of HPV testing in varied scenarios for underlying disease prevlance. Our intention was that these models would approximate expected outcomes across a wider range of settings than was represented in the literature, enhancing the usefulness of findings and improving applicability. In our desire to be exhaustive, however, we reviewed tests of single or combinations of virus types and methods of testing (polymerase chair reaction) that did not reflect tests likely to be used outside of specialized research settings. The data were confusing since some data understated performance compared with tools that were available and some overstated potential benefits by including highly sensitive tests not available for clinical use. Although the review discussed these nuances, we now recognize the report might have been clearer and would have been far more efficiently conducted had the team reviewed only those tests with current or likely near-term clinical availability, which could have been determined on the basis of patent and FDA regulatory decisions that indicated what products would be available and been designed for use in clinical care.

For two of the subgroup analyses (age and hysterectomy), it was necessary to include publications with information about underlying incidence and prevalence in order to provide context, as well as to drive modeling efforts. 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. Rather than restrict the literature, it was preferable not to use more restricted search terms in order to ensure that the most applicable evidence was found for the two target groups, even inside of larger studies.

As this illustration of challenges highlights, applicability of a report can be well served by restricting inclusion of marginally related or outdated studies and is rarely enhanced by uncritically extrapolating the results from one context to another. For example, we could not estimate 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 use of restrictions will benefit from seeking explicit guidance from clinical, medical testing, and statistical experts about applicability challenges. Review teams need to familiarize themselves with the availability and contemporary clinical use of the test, 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.

However, often the target of interest is 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—to be exhaustive or to narrow the eligible literature—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.

Summary

Key points are:

- Applicability is the ability of an individual study or a body of evidence to provide information to guide real-world decisions about the likely performance characteristics or outcomes of a particular medical test for a particular setting.
- Systematic reviews typically face one or more of the following challenges: (1) publications in the literature report insufficient detail to permit assessment of key elements of applicability; (2) the relevant literature is large and needs to be grouped or organized in ways that help convey applicability and that anticipate differences in test performance or outcomes across groups; and (3) medical tests are rapidly changing and the literature may reflect tests that are not yet available, or that are no longer in use.
- Using a PICOTS checklist approach to develop an inventory of applicability factors is key for designing and scoping reviews, summarizing large bodies of evidence and their relevance to smaller subgroups, and identifying potentially concerning trends over time that may influence interpretation of the literature.
- It is necessary to be prepared for the uncovering of relevant contextual factors after a substantial volume of literature has been reviewed. It is crucial to recognize that like any relationship discovered *a posteriori*, this may reflect a spurious association. 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 rises only to the level of a hypothesis.
- When the intended context and population for whom applicability is of interest are clear and focused in advance, EPCs should shape the review at the level of tightly focused inclusion and exclusion criteria. We also recommend that EPCs carefully weigh the implications of reviewing data for older and possibly outmoded test methods. The temptation to be exhaustive can dilute the focus.
- If a review is intended to be exhaustive, EPCs can use the factors that influence applicability to provide a structure within key question results and discussion for organizing findings. It may be useful to organize summary tables by key factors, for example, grouping data by sex when possible and indicating where this was not possible, or grouping by version of the test when there is more than one method in the literature.

DRAFT

- The importance of contextual factors that may affect applicability can be evaluated using evidence synthesis methods, such as meta-regression and decision modeling (*see* Papers 8 and 10).

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

2. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137(7):598-602.

3. 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.

4. Hartmann KE, Hall SA, Nanda K, et al. Cervical Cancer Screening: A Summary of the Evidence. (Prepared for the US Preventive Services Task Force under Contract No. 290-97-0011). AHRQ Publication No. 03-515A. Rockville (MD) Agency for Healthcare Research and Quality, January 2002. Available at: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=es25. Accessed on October 13, 2010.

5. McCrory DC, Matchar DB, Bastian L, et al. Evaluation of Cervical Cytology. Evidence Report/Technology Assessment No. 5. (Prepared by Duke University under Contract No. 290-97-0014.) AHCPR Publication No. 99-E010. Rockville, MD: Agency for Health Care Policy and Research. February 1999. Avaliable at: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=erta5. Accessed on October 13, 2010. 1999.

6. 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.

7. Hartmann KE, Nanda K, Hall S, et al. Technologic advances for evaluation of cervical cytology: is newer better? Obstet Gynecol Surv 2001;56(12):765-74.

DRAFT